An LLM-based pipeline for understanding decision choices in data analysis from published literature

ANONYMOUS AUTHOR(S)

Decision choices, such as those made when building regression models, and their rationale are essential for interpreting results and understanding uncertainty in an analysis. However, these decisions are rarely studied because tracing every alternatives considered by authors is often impractical, and reworking a completed analysis is generally of limited interest. Consequently, researchers must manually review large bodies of published analyses to identify common choices and understand how choices are made. In this work, we propose a workflow to automatically extract analytic decisions and their reasons from published literature using Large Language Models. Our method also introduces a paper similarity measure based on decision similarity and visualization methods using clustering algorithms. As an example, this workflow is applied to analyses studying the effect of particulate matter on mortality. This approach enables scalable and automated studies of decision choices in applied data analysis, providing an alternative to existing qualitative and interview-based studies.

CCS Concepts: • **Applied computing** → *Document analysis*; • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: Large language models

ACM Reference Format:

Anonymous Author(s). 2025. An LLM-based pipeline for understanding decision choices in data analysis from published literature. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI'26).* ACM, New York, NY, USA, 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Decisions are made at every stage of data analysis: from initial data collection and pre-processing to modelling choices. Different decision choices can have a direct impact to the final results, which can lead to different interpretation and policy recommendations that follow. When independent analysts analyzing the same dataset even to answer the same research questions, through many-analysts experiments, they often arrive at markedly different conclusions [8, 19, 46]. This variability in results can be attributed to the flexibility analysts have in making decisions throughout the data analysis process, which Gelman and Loken [18] describe as the "garden of forking paths". When such flexibility is misused, data analysis can lead to p-hacking, selective reporting, inflated effect sizes, and other issues, undermining the quality and credibility of the findings.

[this is not okay — Multiple recommendations have been proposed to improve data analysis practices, such as pre-registration and multiverse analysis. Bayesian methods also offer a different paradigm to p-value driven inference for interpreting statistical evidence. Most empirical studies of data analysis practices focus on specially designed and simplified analysis scenarios. While informative, these setups may not adequately capture the complexity of the data analysis with significant policy implications. [In practice, studying the data analysis decisions with actual applications is challenging.] Analysts may no longer be available for interviews due to job changes, and even when they are, recalling

Manuscript submitted to ACM

the full set of decisions and thinking process made during the analysis is often infeasible. Moreover, only until the last decades, analysis scripts and reproducible materials were not commonly required by journals for publishing. — up till here]

In this work, we develop a tabular format to record analytical decisions in data analysis and automate the extraction of these decisions from published papers using large language models (Gemini and Claude). The workflow also include a component to calculate paper similarity based on both the decisions and the semantic similarity of their rationales, and use clustering methods to visualize papers according to distance based on decision similarity. We apply this workflow to a set of 56 air pollution modelling studies estimating the effect size of particulate matter (PM2.5 or PM10) on mortality and hospital admissions, typically modeled using Poisson generalised linear models (GLMs) or generalized additive models (GAMs). Analysis of the extracted decisions reveals common choices in this type of analysis (number of knots or degree of freedom for smoothing methods for time, temperature and humidity) and find three distinct clusters corresponding to different smoothing methods (LOESS, natural spline, and smoothing spline) used in European and U.S. studies, consistent with findings from the APHENA project [28].

In summary, the contribution of this work includes:

- A new approach to study data analysis decision choices through automatic extraction of decisions from scientific literature using LLMs,
- A method to construct paper similarities based on the decisions and the semantic similarity of their rationale,
- A shiny GUI tool for validation LLM outputs in this context, and
- A dataset of decisions and rationale, along with metadata, compiled from 62 studies in air pollution mortality modelling.

2 Related work

2.1 Decision-making in data analysis

Data analysis involves making choices at every step, from initial data collection, data pre-processing to model specification, and post-processing. Each decision represents a branching point where analysts choose a specific path to follow, and the vast number of possible choices analysts can take forms what Gelman and Loken [18] describe as the "garden of forking paths". While researchers may hope their inferential results are robust to the specific path taken through the garden, in practice, different choices can lead to substantially different conclusions. This has been empirically demonstrated through "many analyst experiments", where independent research groups analyze the same dataset to address the same research questions with their own chosen analytic approach. A classic example is Silberzahn et al. [46], where researchers reported an odds ratio from 0.89 to 2.93 for the effect of soccer players' skin tone on the number of red cards awarded by referees. Similar variability has been observed in structural equation modeling [44], applied microeconomics [22], neuroimaging [8], and ecology and evolutionary biology [19].

Examples like above have rendered decision-making in data analysis as a subject to study in human computer interaction. To understand how analysts making decisions during data analysis and navigating the garden of forking path, researchers have conducted qualitative interviews with analysts on data analysis practices [2, 24, 31]. Visualization tools have also been explored to communicate the decision process through analytic decision graphics (ADG) [32]. In fairness machine learning literature, Simson et al. [47] contributed a reusable workflow that supports participatory input to democratize decisions in machine learning algorithms related to fairness, privacy, interpretability and performance. Conducting qualitative studies through interviews to study how assumptions and decisions are made in data analysis practices takes a significant amount of time and effort, and the findings may not generalize to other contexts. While published research papers may not provide a complete picture of the decision-making process, they do contain valuable information about the choices made by analysts and the rationale behind them. With recent advances in Large Language Models (LLMs), it has become possible to automatically extract structured information from unstructured text. This could provide a scalable way to study decision-making practices in data analysis.

On top of qualitative studies, software tools have also developed to incorporate potential alternatives in the analysis workflow. The `DeclareDesign` package [7] introduces the MIDA framework for researchers to declare, diagnose, and redesign their analyses to produce a distribution of the statistic of interest, which has been applied in the randomized controlled trial study [6]. The `multiverse` package [33, 43] provides a framework for researchers to conduct multiverse analysis to systematically explore how different choices affect results and to report the range of plausible outcomes that arise from alternative analytic paths.

2.2 Visualization on scientific literature

With the growing volume of scientific publications and the difficulty of navigating the literature to stay informed, there is increasing interest in developing tools to visualize and recommend scientific papers. These systems link papers based on their similarity and relevance, typically determined by keywords [23], citation information (e.g. citation list, co-citation) [13], or combinations with other relevant paper metadata (e.g. author, title) [5, 14, 17, 20]. Recent approaches incorporate text-based information using topic modelling [1], argumentation-based information retrieval [48], and text embedding [39]. While metadata and high-level text-based information are useful for finding relevant papers, researchers also need tools that help them *make sense* of the literature rather than simply *locating* it. In applied data analysis, one interest is to understand how studies differ or align in their analytical approaches. Capturing the decisions and reasoning expressed in analyses on a shared theme enables the calculation of similarity metrics based on these choice and their underlying rationale, which supports clustering and visualizing paper to identify common practices in the field.

3 Methods

TODO: a generic summary of the workflow, maybe an illustration

3.1 Record decisions in data analysis

Consider the following excerpt from Ostro et al. [40] that describes the modelling approach to provide evidence of an association between daily counts of mortality and ambient particulate matter (PM10):

> Based on previous findings reported in the literature (e.g., Samet et al. 2000), the basic model included a smoothing spline for time with 7 degrees of freedom (df) per year of data. This number of degrees of freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.

This sentence encode the following components of a decision:

- **variable**: time
- **method**: smoothing spline
- **parameter**: degree of freedom (df)
- **reason**: Based on previous findings reported in the literature (e.g., Samet et al. 2000); This number of degrees of freedom controls well for seasonal patterns in mortality and reduces and often eliminates autocorrelation.
- **decision**: 7 degrees of freedom (df) per year of data

To record these decisions in a tabular format, we follow the tidy data principle [51], which states each variable should be in a column and each observation in a row. For our purpose, each row represents a decision made by the authors in a paper and an analysis often include multiple decisions. To retain the original context of the decision, we extract the original text in the paper, without paraphrase or summarization. The decision choice above is a parameter choice of a statistical method applied to the variable. Analyses also include other types of decisions, such as temporal and spatial treatments, for example, the choice of lagged exposure for certain variables or whether the model is estimated collectively or separated for individual locations. These decisions don't have a specific method or parameter, but should still be recorded with the variable, type (spatial or temporal), reason, and decision fields.

Given the writing style and the quality of the analysis itself, multiple decisions may be combined in one sentence and certain fields, e.g. decision and reason, may be omitted. Consider the following excerpt from Ostro et al. [40]:

> Other covariates, such as day of the week and smoothing splines of 1-day lags of average temperature and humidity (each with 3 df), were also included in the model because they may be associated with daily mortality and are likely to vary over time in concert with air pollution levels.

This sentence contains four decisions: two for temperature (the temporal lag and the smoothing spline parameter) and two for humidity and should be structured as separate entries:

Paper	ID	variable	method	parameter	type	reason	decision
ostro	1	temperature	smoothing spline	degree of freedom	parameter	3 degree of freedom	NA
ostro	2	relative humidity	smoothing spline	degree of freedom	parameter	3 degree of freedom	NA
ostro	3	temperature	NA	NA	temporal	1-day lags	NA
ostro	4	relative humidity	NA	NA	temporal	1-day lags	NA

Notice in the example above, the reason field are recorded as NA. This is because the stated rationale ("and are likely to vary over time in concert with air pollution levels") only supports the general inclusion of temporal lags but does not justify the specific choice of 1-day lag over other alternatives, for example, 2-day average of lags 0 and 1 and single-day lag of 2 days. Similar scenario can happen when a direct decision is missing while a reason is provided ("done by minimizing Akaike's information criterion"), as in Katsouyanni et al. [27]:

> The inclusion of lagged weather variables and the choice of smoothing parameters for all of the weather variables were done by minimizing Akaike's information criterion.

3.2 Extract decisions automatically from literature with LLMs

Manually extracting decisions from published papers is labor-intensive and time-consuming. With Large Language Models (LLMs), it has become possible to automatically extract structured information from unstructured text by supplying a set of PDF documents and a prompt for instruction. Text recognition from PDF document relies on Optical Character Recognition (OCR) to convert scanned images into machine-readable text – capability currently offered by Antropic Claude and Google Gemini. In the prompt, we assign the LLM a role as an applied statistician and instruct it to generate a markdown file containing a JSON block that extract decisions from the PDF in the format described in

Section 3.1. We also provide a set of instructions and examples on the potential missing of reason and decision fields. Prompt engineering techniques [12, 54] are used to optimize the prompt script. The full prompt feed to the LLM is provided in the Appendix. We use the `chat_PROVIDER()` functions from the `ellmer` package [53] in R to obtain the output with Gemini and Claude API.

3.3 Validate and standardize LLM outputs

The LLM outputs need to be validated and standardized before further analysis. Validation focuses on ensuring the correctness of the extracted decisions by LLMs, while standardization aims to ensure consistency in variable and model names across papers, given authors may express the same concept in different ways. For example, "mean temperature", "average temperature", and "temperature" all refer to the same variable, which can be all standardized to "temperature" for consistency. To help with the validation and standardization process, we developed a Shiny application that provides an interactive interface for users to review and edit the LLM outputs. A Shiny application takes a CSV of extracted decisions as input and allows three types of edits: 1) *overwrite* – modify the content of a particular cell, 2) *delete* – remove a particular irrelevant decision, and 3) *add* – manually enter a missing decision. Figure 1 illustrates the *overwrite* action for standardizing the variable NCtot (The number concentration of urban background particles <100 nm in diameter) to "pollution": the user enters a predicate function in the filter condition box on the left panel, and the filtered data will appear interactively in the right panel. The user can then specify the variable to overwrite and the new value and the corresponding cells in the right panel will be updated. This change need to be confirmed by pressing the "Apply changes" button to update the full dataset. The corresponding `tidyverse` [52] code will then be generated in the left panel to be included in an R script, and the edited table can be downloaded for future analysis.

3.4 Calculate paper similarity and visualization

Once the output has been extracted and validated, the decisions can be treated as data for further analysis. In this section, we construct a distance metric between pairs of papers based on the similarity of their decision choices. This metric can then be used as a distance matrix among papers for clustering, dimension reduction, and visualization.

For each paper pair, a decision is considered comparable if the papers share the same variable and decision type, for example, a parameter decision on temperature or the temporal decision on humidity. For two decisions to be considered similar, both the decision choice and the rationale are taken into account. A similar choice indicates a similar final decisions are made in the analysis, whereas a similar reason reflects a shared rationale or justification for the choice, even when the choices themselves differ, potentially due to differences in the underlying data. To assign numerical value for measuring the similarity, we use the semantic similarity from text model, using the `text` package [29]. We first obtain the text embedding for all the reason and decisions and calculate the cosine similarity between the matched reason and decisions. For parameter type decisions, the statistical method used also contributes to the similarity of the decision. Since semantic similarity cannot fully capture the difference betweenit statistical methods (the difference between smoothing spline and natural spline is not well represented by the textual difference of "smoothing" and "natural"), method similarity is encoded as binary: 1 if the two papers used the same method, and 0 otherwise. The paper similarity is then computed as the average similarity across all the matched methods, decisions, and reasons. The resulting paper similarity metric can be interpreted as a distance measure to cluster and visualize papers based on their decision choices.

Because analyses vary in the decisions they report, the number of matched decisions differs across paper pairs. In practice, some studies may not fully report the decision and reason for every choice made, leading to missing data for

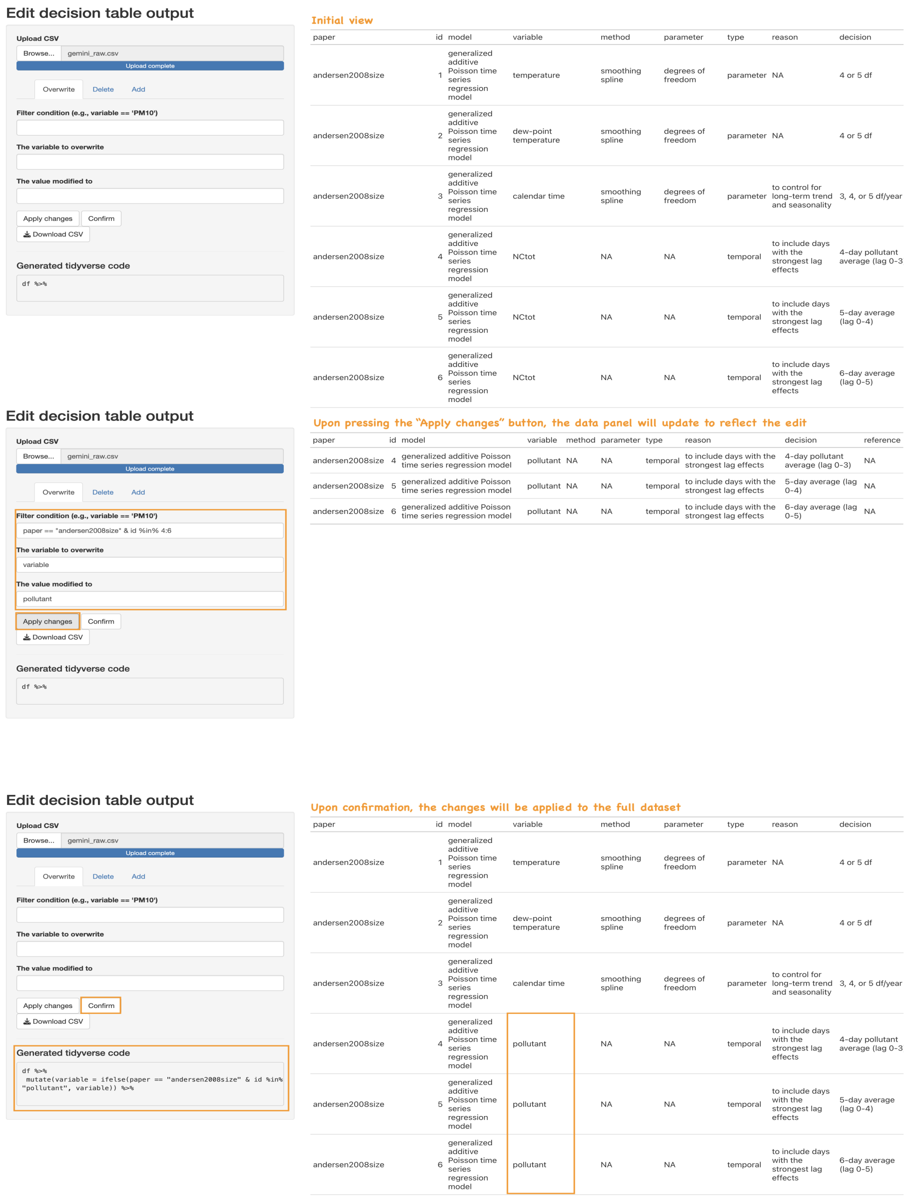

Fig. 1. The Shiny application interface to validate and standardize Large Language Model (LLM)-generated output. (1) the default interface after loading the input CSV file. (2) The table view will update interactively to reflect the edit: for paper with handle "andersen2008size" and id in 4, 5, 6, replace the variable NCtot with "pollutant". (3) After clicking the Confirm button, the corresponding `tidyverse` code is generated, and the table view returns to its original unfiltered view with the edits applied. The edited data can be downloaded by clicking the Download CSV button.

the matched decisions. Although paper similarity can be calculated based on all available matched decisions, cares should be taken for pairs with only a small number of matches, as the paper similarity may be overly influenced by one or two decisions. To address this, users may focus on a set of decisions shared across papers and on papers that report a minimal number of these decisions when calculating paper similarity.

4 Results

In the study of the health effects of outdoor air pollution, one area of interest is the association between short-term, day-to-day changes in particulate matter air pollution and daily mortality counts. This question has been studied extensively by researchers across the globe and in the US, it serves to provide scientific evidence for to guide public policy on setting the National Ambient Air Quality Standards (NAAQS) for air pollutants. While individual modelling choices vary, these studies often share a common structure: they adjust for meteorological covariates such as temperature and humidity, apply temporal or spatial treatments, like including lagged variables and may estimate the effect by city or region before combining results. This naturally forms a "many-analyst" experiment setting where different researchers analyze similar data to address the same scientific question and the analyses are documented in published papers.

From the 56 studies examining the effect of particulate matters (PM_{10} and $PM_{2.5}$) on mortality and hospital admission, we focus on the baseline model reported in each paper, excluding secondary models (e.g. lag-distributed models) and sensitivity analysis. We also exclude decisions on other pollutants, such as nitrogen dioxide (NO_2). This yields 242 decisions extracted using Gemini, averaging approximately 4 decisions per paper.

4.1 Validation and standardization of LLM outputs

Table 2. Summary of validation and standardization edits made during the review process.

Reason	Count
Remove decisions out of scope: other pollutants and sensitivity analysis	50
Edit made to recode smoothing parametser unit to per year	45
Duplicates	9
Fix incorrect capture	9
Edit made due to decisions are too general, e.g. minimum of 1 df per year was required	6
Remove decisions related to definition of variables, e.g. season	5
Total	124

Table 2 summarizes the number of edits made during the review process using the Shiny application. These edits fall into two main categories: 1) correcting LLM outputs and 2) standardizing extracted decision. The first category includes fixing incorrect captures, removing non-decision (e.g. definition of variables), removing duplication, excluding irrelevant decisions (e.g. sensitivity analyses), and excluding decisions whose stated reasons reflect general guidelines rather than actual choices (e.g. "minimum of 1 degree of freedom per year is required").

Standardization addresses variation in how authors express variable names and decisions. For example, variable names such as "mean temperature" and "average temperature" refer to the same variable and should be aligned for comparison for later decision similarity calculation. Variable names are manually standardized into four main categories:

Table 3. Missingness of decision and reason fields in the Gemini-extracted decisions. Most decisions report the choice (35.5 + 57.1 = 92%), but 57.1% lacks a stated reason.

	Decision	
Reason	Non-missing	Missing
Non-missing	90 (37.2%)	14 (5.8%)
Missing	134 (55.4%)	4 (1.7%)

- **temperature**: "mean temperature", "average temperature", "temperature", "air temperature", "ambient temperature"
- **humidity**: "dewpoint temperature" and its hyphenated variants, relative humidity", "humidity"
- **PM**: "pollutant", "pollution", "particulate matter", "particulate", "PM10", "PM2.5"
- **time**: "date", "time", "trends", "trend"

Notice that "dewpoint temperature" is standardized under humidity because it serves as a proxy for temperature in achieving a 100% relative humidity.

Decisions themselves also require standardization. For example, the smoothing parameter (number of knots and degree of freedom) may be expressed *per year* or *in total*, and temporal lag decision may be expressed in different formats (e.g. "6-day average", "mean of lags 0+1", "lagged exposure up to 6 days"). Smoothing parameter units are manually recoded to a *per year* basis for consistency, asreflected in Table 2. Temporal decision show a wider variety, generally falling into two categories:

- **multi-day average lags**, such as "6-day average", "3-d moving average", "mean of lags 0+1", "cumulative lags, mean 0+1+2" and
- **single-day lags**, such as "lagged exposure up to 6 days", "lag days from 0 to 5".

This variability makes manual standardization impractical, hence we apply a secondary LLM process (claude-3-7-sonnet-latest) using the `ellmer` package to convert temporal decisions into a consistent format: `multi-day: lag [start]-[end]` and `single-day: lag [start], . . . ,lag [end]`. For instance, "6-day average" is converted to "multi-day: lag 0-5" and "lagged exposure up to 6 days" is converted to "single-day: lag 0, lag 1, lag 2, lag 3, lag 4, lag 5".

4.2 Exploratory analysis of decision choices

As raised in Section 3.1, not all decisions reported in the literature include both the decision choice and the rationale. Some decisions may only report the choice without a stated reason, while others may provide a reason without specifying the exact choice made. Table 3 summarizes the missingness of the decisions and reason for the extracted decisions. While 2% of decisions are complete for both decision and reasons, 55% of decisions lack a stated rationale for the choice. This reflects a common reporting practice in the field, where authors often present the decision itself without providing a justification, e.g. "We decide to use x degree of freedom for variable y_1 and y_2". This also includes cases where authors provide general guidelines for selecting the parameter, but the rationale is too broad to justify the specific choice made (hence validated as `NA` in Section 4.1).

Table 4. Count of variable-type decisions in the Gemini-extracted decisions. The most commonly reported decision are the parameter choices and temporal lags for for time, PM, temperature, and humidity.

Variable	Type	Count
time	parameter	44
PM	temporal	39
temperature	parameter	35
humidity	parameter	25
temperature	temporal	23
humidity	temporal	19
PM	parameter	9
time	temporal	3

Table 4 lists the eight most frequently reported decision: parameter and temporal choice for time, PM, temperature, and humidity. While a wider list of variables have been used in the analysis, these four variables are most commonly included in baseline models. Parameter choices for time, temperature, and humidity are typically made on the use of smoothing parameter for the smoothing method (natural spline and smoothing spline), whereas temporal choices are commonly reported for PM, temperature, and humidity for the number of lag to consider in the model.

Table 5. Options captured for parameter choices for time, humidity, and temperature variables in the Gemini-extracted decisions. The choices for natural spline knots are generally less varied than the degree of freedom choices for smoothing spline. Choices for temperature and humidity tend to be close, given they are both weather related variables, while the choices for time are more varied inherently.

Method	Variable	Decision
natural spline	humidity	3, 4
natural spline	temperature	3, 4, 6
natural spline	time	1, 1.5, 3, 4, 6, 7, 8, 12, 15, 30
smoothing spline	humidity	2, 3, 4, 6, 8, 50% of the data
smoothing spline	temperature	2, 3, 4, 6, 8, 50% of the data
smoothing spline	time	1, 3, 4, 5, 6, 7, 7.7, 8, 9, 10, 12, 30, 100, 5% of the data

Table 5 presents the parameter-related decisions extracted for spline methods (natural and smoothing spline) applied to variable time, humidity and temperature. These decisions concern the number of knots or degree of freedom, with all values standardized to a *per year* scale for consistency. The selection of knot for natural spline has less variation than the degree of freedom choices for smoothing spline. Choices for temperature and humidity are generally similar, given they are both weather related variables, whereas choices for time are more varied. This tabulation provides a reference set for common parameter choices for future studies and help to identify anomalies and special treatment in practice. For example, the choice of 7.7 degree of freedom reported in Castillejos et al. [11] may prompt analysts to seek further justification. By cross comparing with other reporting, some decisions appear ambiguous. For example, in Moolgavkar [37] and Moolgavkar [38], the reported value of 30 and 100 degrees of freedom for time may be understandable for experienced domain researcher, it could be unclear for junior analysts as to whether they apply to the full 9 year period or on a per-year basis. We also observe a different report style from Schwartz [45], where smoothing spline parameters are expressed as a proportion of the data ("5% of the data" and "5% of the data") rather than fixed numerical value.

Table 6. Options captured for temporal lag choices for PM, temperature, and humidity variables in the Gemini-extracted decisions. Both single-day lags and multi-day average lags are commonly used, generally considering up to five days prior (lag 5).

Lag type	Variable	Decision
multi-day average	PM	lag 0-1, 0-2, 0-3, 0-4, 0-5, 0-6
multi-day average	humidity	lag 0-1, 0-2, 0-3, 0-5, 1-5, 2-4
multi-day average	temperature	lag 0-1, 0-2, 0-3, 0-5, 2-4
single-day lag	PM	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
single-day lag	humidity	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13
single-day lag	temperature	lag 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

Similarly, Table 6 summarizes the temporal lag choices for PM, temperature, and humidity. For single-day lags, the lags are considered up to 13 days (approximately two weeks). For multi-day averages, 3-day and 5-day averages are most common, although other choices such as 2-4 day average are also observed as in López-Villarrubia et al. [35]:

> In particular, lags 0 to 1 and lags 2 to 4 averages of temperature, relative humidity, and barometric pressure were considered as meteorological variables.

4.3 Paper similarity and clustering

Given the number of decisions reported in Table 4, we focus on the six most common variable-type decisions for calculating paper similarity: parameter choices for time, temperature, and humidity, and temporal lag choices for PM, temperature, and humidity. We also restrict our analysis to papers that report at least three of these six decisions, resulting in 48 papers for the similarity analysis. This ensures that the paper similarity metric is based on a sufficient number of comparable decisions. We use the default text embedding model (BERT) in the `text` package and cosine similarity to compute the similarity score. Sensitivity analysis on different text embedding model is checked in Section 4.4.3. Paper similarity is then calculated as the average of decision similarity for each paper pair. The resulting distance matrix is then used for hierarchical clustering in Figure 2 and multi-dimensional scaling (MDS) in Figure 3. The two MDS dimension reveals three clusters correspond to the three smoothing methods used in these analyses: LOESS, natural spline, and smoothing spline. This grouping aligns with the modelling strategies seen in large-scale analysis, such as the U.S. NMMAPS study [42] and the European APHEA [26] and APHEA2 [27] project.

To reconcile these differences, the APHENA project [28] was launched with the aim to "assess the consistency across Europe and North America when estimated using a common analytic protocol and to explore possible explanations for any remaining variation". While multi-dimensional scaling in Figure 3 shows the match of three clusters with three smoothing methods, this is not inconsistent with the APHENA project [28] that the amount of smoothing to have a more important role than the method of smoothing for estimating the effect of PM on public health variables. The similarity metric we proposed focuses on the variation of choices across analyses, without directly assessing how those choices influence results. By pooling decision choices from multiple studies with LLMs, it becomes much easier to reveal common practices and difference in research practices, highlighting decisions that require further sensitivity

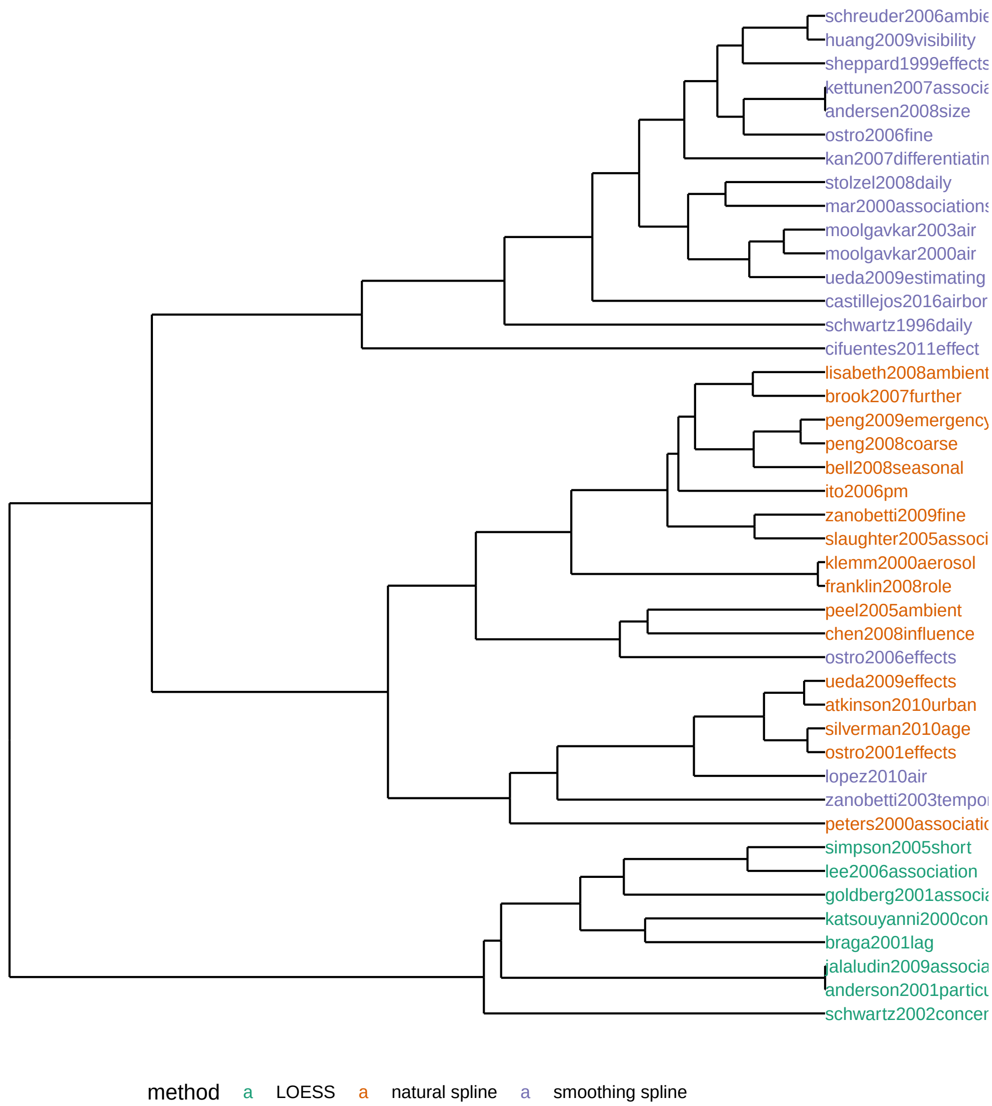

Fig. 2. Maybe we don't need teh hierarchical clustering plot because it has the same information as the next MDS one???

analyses to assess their impact. The different smoothing methods revealed in Figure 3 are consistent with the analysis by Peng et al. [41] and Touloumi et al. [49] that compares different smoothing methods and rationale for selecting smoothing parameters.

4.4 Sensitivity analysis

A series of sensitivity analysis has been conducted to explore the reproducibility for using LLMs for text extraction tasks (Section 4.4.1), discrepancies in decision extraction between different LLM models: Gemini (`gemini-2.0-flash`)

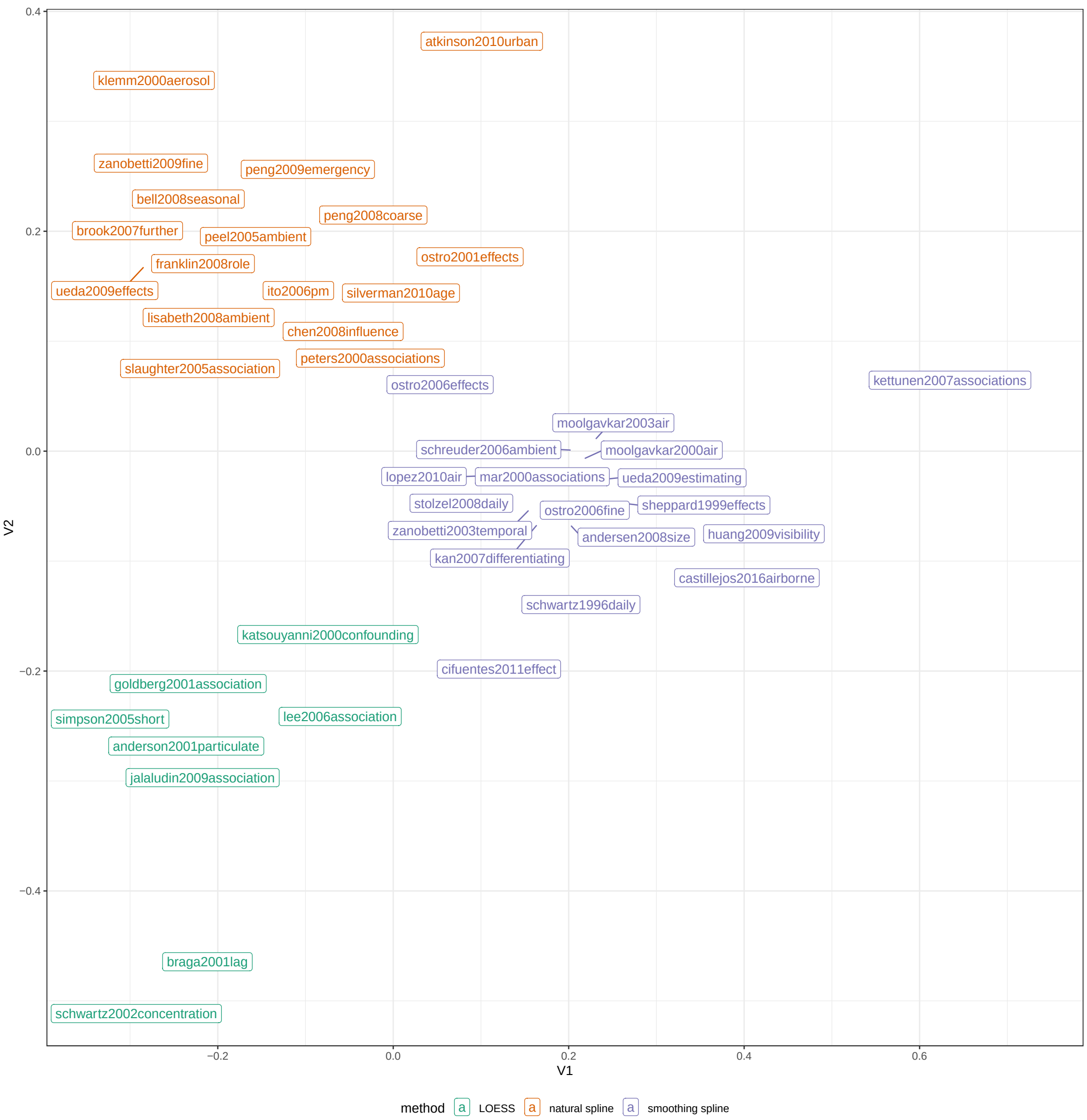

Fig. 3. The multi-dimensional scaling (MDS) based on paper similarity distance for `length(good_pp)` air pollution mortality modelling papers, colored by the smoothing method used. The MDS reveals the three distinct groups of papers, corresponds to LOESS, natural spline, and smoothing spline. These groups corresponds to the different modelling strategies debated in the European and U.S. studies, as documented in the APHENA project [28].

and Claude (`claude-3-7-sonnet-latest`) (Section 4.4.2), and the sensitivity of text model for computing the semantic decision similarity (Section 4.4.3).

4.4.1 LLM reproducibility. We assess the reproducibility of Gemini's text extraction (`gemini-2.0-flash`) by repeating the task five times for each of the 62 papers and perform pairwise comparison between runs. This generates $5 \times 4/2 \times 62 =$

620 possible comparisons for both "reason" and "decisions" fields. Comparisons where the runs produced a different number of decisions were excluded, as this would require manual alignment. After filtering, 449 out of 620 (72%) remained. Table 7 prints the decisions in Andersen et al. [3] across two runs and all the four decisions are identical with no difference.

Table 7. Example comparing Gemini's text extraction for Andersen et al. [3] across two runs. The extracted decisions are identical in both runs.

Variable	Run1	Run2
NCtot	6day average (lag 05)	6day average (lag 05)
calendar time	3 4 or 5 dfyear	3 4 or 5 dfyear
dew-point temperature	4 or 5 df	4 or 5 df
temperature	4 or 5 df	4 or 5 df

Table 8. Number of differences in the reason and decision fields across Gemini runs for papers with consistent number of decisions across runs.

Num. of difference	Count	Proportion (%)
0	358	79.73
1	12	2.67
2	8	1.78
3	0	0.00
4	24	5.35
5	12	2.67
6	3	0.67
7	0	0.00
8	10	2.23
9	6	1.34
10	10	2.23
11	6	1.34
Total	449	100.00

Table 8 summarizes the number of differences observed in each pairwise comparison. Among all comparisons, 80% produce the identical text in reason and decision. The discrepancies come from the following two reasons: 1) Gemini extracted different length for the same decision, e.g. in Kan et al. [25], some runs may extract "singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving average **of current and previous day concentrations** (lag=01)", while others extract "singleday lag models underestimate the cumulative effect of pollutants on mortality 2day moving average (lag=01)". Similarity, for decisions, some runs yield "10 df for total mortality", while other runs yield "10 df". 2) Gemini fails to extract reasons in some runs but not others, e.g. in Burnett et al. [9], the first run generates NAs in the reasons, but the remaining four runs are identical. In Ueda et al. [50] and Castillejos et al. [11]

, runs 1 and 5 fail to extract the reasons and produce the same incomplete version, whereas runs 2, 3, and 4 produce accurate versions with reasons populated.

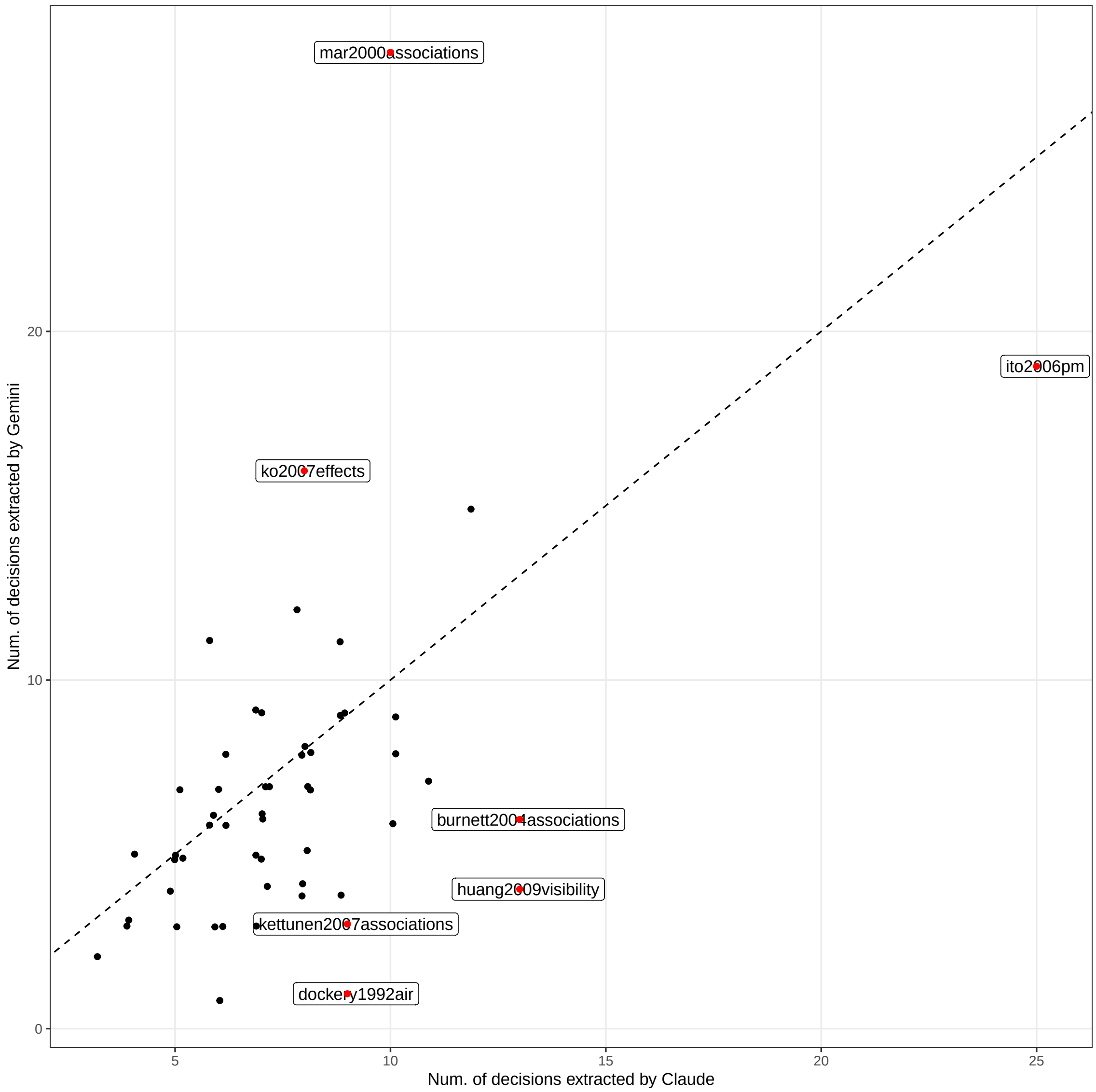

Fig. 4. Comparison of decisions extracted by Claude and Gemini. Each point represents a paper, with the x- and y-axes showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. More points fall below this line, suggesting Claude extracts more decisions – often including noise from data pre-processing or secondary data analysis steps – which requires additional manual validation.

4.4.2 LLM models. Reading text from PDF document requires Optical Character Recognition (OCR) to convert images into machine-readable text, which currently is only supported by Antropic Claude (`claude-3-7-sonnet-latest`) and

Google Gemini (`gemini-2.0-flash`). We compare the number of decisions extracted by Claude and Gemini across all 62 papers in Figure 4. Each point represents a paper, with the x- and y-axes showing the number of decisions extracted by Claude and Gemini, respectively. The dashed 1:1 line marks where both models extract the same number of decisions. While both models extract decisions irrelevant to our analysis, such as sensitivity analyses and secondary analyses, Claude's extractions tend to include more of these irrelevant decisions, examples of these include 1) the definition of "cold day" and "hot day" indicators in Dockery et al. [16] ("defined at the 5th/ 95th percentile"), 2) decisions relate to other pollutants: NO_2, O_3, and SO_2 using a "24 hr average on variable" in Huang et al. [21], and 3) the definition of black smoke and in Katsouyanni et al. [27] for secondary analysis ("restrict to days with BS concentrations below 150 $\mu g/m^2$"). While Gemini also capture these irrelevant decisions, such as "0-4 lag days" for air pollution exposure variables (CO, EC, K_S, NO_2, O_3, OC, Pb, S, SO_2, TC, Zn) in Mar et al. [36]. However, these cases are less frequent than Claude's extraction and has been validated and standardized in Section 4.1.

For both Claude and Gemini, we find they fail to link the general term "weather variables" to the specific weather variables (e.g. Dockery et al. [16] and Burnett et al. [10] for Gemini and Dockery et al. [16] and Katsouyanni et al. [27] for Claude). Although our prompt specified that some decisions may require linking information across sentences and paragraphs to identify the correct variable, this instruction doesn't appear to be applied consistently.

4.4.3 Text model. We have conducted sensitivity analysis on the text model for obtaining the decision similarity score from the Gemini outputs. The tested language models tested include 1) BERT by Google [15], 2) RoBERTa by Facebook AI [34], trained on a larger dataset (160GB v.s. BERT's 15GB), 3) XLNnet by Google Brain [55], and two domain-trained BERT models: 4) sciBERT [4], trained on scientific literature, and 5) bioBERT [30], trained on PubMed and PMC data.

Figure 5 shows the distribution of the decision similarity and the corresponding multi-dimensional scaling visualization, where distance are calcualted from the paper similarity for each text model. At decision level, the BERT model produces the widest variation across all five models, while the similarity scores from XLNet are all close to 1. While the raw scores are not directly comparable across models due to the difference in the underlying transformer architecture, the multi-dimensional scaling (MDS) based on paper similarity scores shows a similar clustering pattern corresponding to the three main smoothing methods (LOESS, natural spline, and smoothing spline).

5 Discussion

While the extraction of decisions from literature could be largely automated with LLMs, manual validations remains essential to ensure the quality of the extracted decisions for further analysis. The quality from the LLM ouput directly affects the amount of manual effort needed for validation and standardization. Using a default temperature of 1 and instructing the model to extract original text rather than paraphrase, we find hallucination is not a major issue with Claude and Gemini for this application. While prompt engineering is used in this work to optimize the prompt for decision extraction, an alternative is to fine-tune a local model to improve LLM performance. Such approach could be beneficial for a systematic literature review, although it would require a labelled decision dataset for training and significantly more training efforts.

As a demonstration, we focus on the modelling decision for the baseline model in the air pollution epidemiology literature. Analyses in this fields often fit multiple models for different health outcomes. Other models, such as distributed lag models and multi-pollutant models are also commonly used to estimate relative risks and the interaction among pollutants. These factors increase the complexity of the decision extraction for LLM, as for additional models,authors often describe only the differences from the baseline model specification, assuming other decisions remain unchanged.

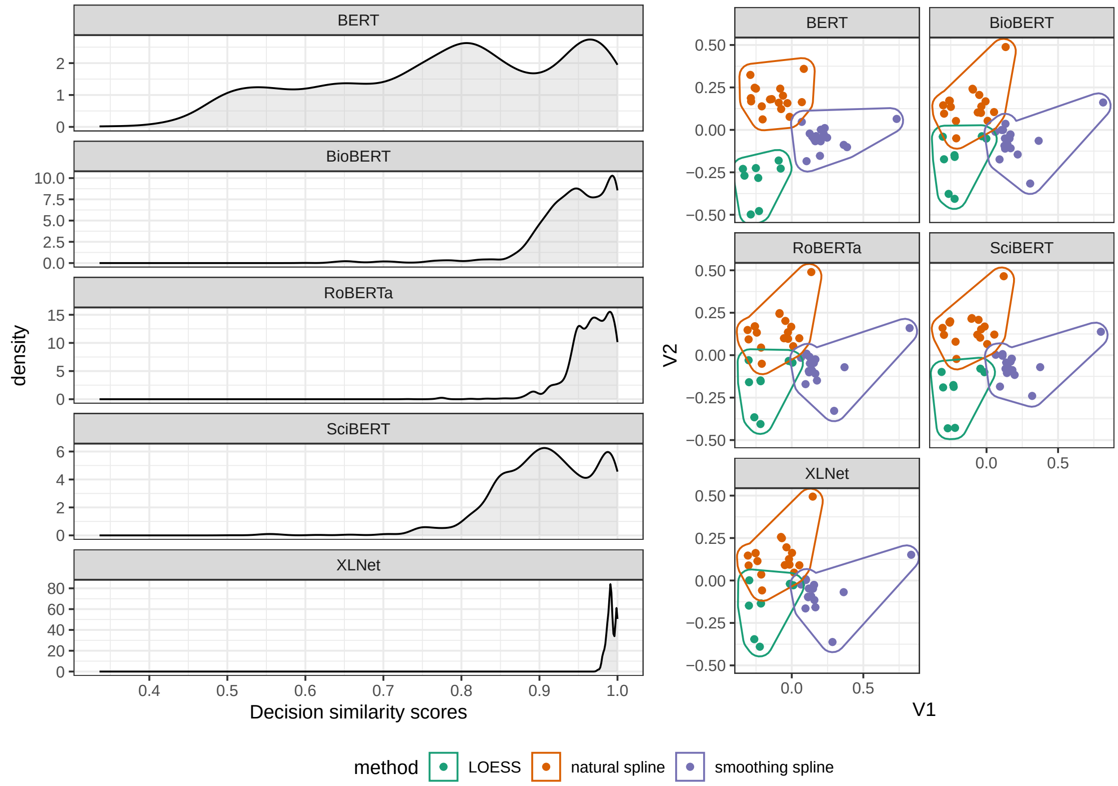

Fig. 5. Distribution of decision similarity (left) and multi-dimensional scaling (MDS) of the paper similarity scores (right) computed for five different text models (BERT, BioBERT, RoBERTa, SciBERT, and XLNet). The default language model, BERT, produces the widest variation across the five models, while the similarity scores form XLNet are all close to 1. The model BioBERT, RoBERTa, and SciBERT yield decision similar scores mostly between 0.7 to 1. All the text models shows a similar clustering structure based on the three main smoothing methods (LOESS, natural spline and smoothing spline).

The LLMs will need to be able to link the decisions across different models and identify the full set of decision for each model for cross-comparison among papers. Apart from modelling choices, other decisions in data pre-processing are also worth comparing. This would include how variables are defined and computed from the raw data.

With the advocacy for reproducibility in science, it is expected that more papers will share their code and data. Code availability can serve as a supplementary source for understanding the choices made in the analysis and cross-check against the description in the manuscript. However, decision choices could be extracted from the scripts, but the rationale behind these choices may not be easily discernible given the lack of comments in the current practice.

6 Conclusion

In this paper, we aim to study how analysts make decisions in their data analysis practice. While classic interviews are often conducted in small scale with toy examples, we developed a pipeline for automatically extracting decisions using LLMs (Claude and Gemini) from scientific literature. We also introduced a method for calculating paper similarity through comparing the similarities among decisions and the similarity metric can be used as a distance to cluster papers by their decision choices and visualization with dimension reduction algorithms, such as multidimensional scaling. We applied this pipeline to a set of air pollution modelling literature that associates daily particulate matter and daily mortality and hospital admission. From the extracted modelling decisions, we identify the most common decision choices in this type of analysis and the paper similarity score calculation revealed the three clusters of paper corresponding to different modelling strategies. These findings are all consistent with the general understanding of the field, as documented in the APHENA project [28] and other methodological comparison studies [41, 49].

While sensitivity analyses are commonly used to assess the robustness of findings to different analytical choices, the set of choices tested is often limited and selected subjectively by the authors. Our approach offers a new perspective by pooling decisions made in analyses across studies in the fields. This allows for a holistic account on the alternatives in the field and identification of both consensus and divergence within the field, providing insights for future research and methodological development.

References

[1] Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014 ieee conference on visual analytics science and technology (vast). pages 173–182, 10 2014. doi: 10.1109/VAST.2014.7042493. URL https://ieeexplore.ieee.org/document/7042493.

[2] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):22–31, 01 2019. doi: 10.1109/TVCG.2018.2865040. URL https://ieeexplore.ieee.org/document/8440815.

[3] Z. J. Andersen, P. Wahlin, O. Raaschou-Nielsen, M. Ketzel, T. Scheike, and S. Loft. Size distribution and total number concentration of ultrafine and accumulation mode particles and hospital admissions in children and the elderly in copenhagen, denmark. *Occupational and Environmental Medicine*, 65(7):458–466, 07 2008. doi: 10.1136/oem.2007.033290. URL https://oem.bmj.com/content/65/7/458. Publisher: BMJ Publishing Group Ltd Section: Original article PMID: 17989204.

[4] Iz Beltagy, Kyle Lo, and Arman Cohan. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp). pages 3613–3618, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL https://www.aclweb.org/anthology/D19-1371.

[5] Steven Bethard and Dan Jurafsky. Cikm '10: International conference on information and knowledge management. pages 609–618, Toronto ON Canada, 10 2010. ACM. doi: 10.1145/1871437.1871517. URL https://dl.acm.org/doi/10.1145/1871437.1871517.

[6] Dorothy V. M. Bishop and Charles Hulme. When alternative analyses of the same data come to different conclusions: A tutorial using declaredesign with a worked real-world example. *Advances in Methods and Practices in Psychological Science*, 7(3):25152459241267904, 07 2024. doi: 10.1177/25152459241267904. URL https://doi.org/10.1177/25152459241267904. Publisher: SAGE Publications Inc.

[7] Graeme Blair, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. Declaring and diagnosing research designs. *American Political Science Review*, 113(3):838–859, 08 2019. doi: 10.1017/S0003055419000194. URL https://www.cambridge.org/core/product/identifier/S0003055419000194/type/journal_article.

[8] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A. Mumford, R. Alison Adcock, Paolo Avesani, Blazej M. Baczkowski, Aahana Bajracharya, Leah Bakst, Sheryl Ball, Marco Barilari, Nadège Bault, Derek Beaton, Julia Beitner, Roland G. Benoit, Ruud M. W. J. Berkers, Jamil P. Bhanji, Bharat B. Biswal, Sebastian Bobadilla-Suarez, Tiago Bortolini, Katherine L. Bottenhorn, Alexander Bowring, Senne Braem, Hayley R. Brooks, Emily G. Brudner, Cristian B. Calderon, Julia A. Camilleri, Jaime J. Castrellon, Luca Cecchetti, Edna C. Cieslik, Zachary J. Cole, Olivier Collignon, Robert W. Cox, William A. Cunningham, Stefan Czoschke, Kamalaker Dadi, Charles P. Davis, Alberto De Luca, Mauricio R. Delgado, Lysia Demetriou, Jeffrey B. Dennison, Xin Di, Erin W. Dickie, Ekaterina Dobryakova, Claire L. Donnat, Juergen Dukart, Niall W. Duncan, Joke Durnez, Amr Eed, Simon B. Eickhoff, Andrew Erhart, Laura Fontanesi, G. Matthew Fricke, Shiguang Fu, Adriana Galván, Remi Gau, Sarah Genon, Tristan Glatard, Enrico Glerean, Jelle J. Goeman, Sergej A. E. Golowin, Carlos González-García, Krzysztof J. Gorgolewski, Cheryl L. Grady, Mikella A. Green, João F. Guassi Moreira, Olivia Guest, Shabnam Hakimi, J. Paul Hamilton, Roeland Hancock, Giacomo Handjaras, Bronson B. Harry, Colin Hawco, Peer Herholz, Gabrielle Herman, Stephan Heunis, Felix Hoffstaedter, Jeremy Hogeveen, Susan Holmes, Chuan-Peng Hu, Scott A. Huettel, Matthew E. Hughes, Vittorio Iacovella, Alexandru D. Iordan, Peder M. Isager, Ayse I. Isik, Andrew Jahn, Matthew R. Johnson, Tom Johnstone, Michael J. E. Joseph, Anthony C. Juliano, Joseph W. Kable, Michalis Kassinopoulos, Cemal Koba, Xiang-Zhen Kong, Timothy R. Koscik, Nuri Erkut Kucukboyaci, Brice A. Kuhl, Sebastian Kupek, Angela R. Laird, Claus Lamm, Robert Langner, Nina Lauharatanahirun, Hongmi Lee, Sangil Lee, Alexander Leemans, Andrea Leo, Elise Lesage, Flora Li, Monica Y. C. Li, Phui Cheng Lim, Evan N. Lintz, Schuyler W. Liphardt, Annabel B. Losecaat Vermeer, Bradley C. Love, Michael L. Mack, Norberto Malpica, Theo Marins, Camille Maumet, Kelsey McDonald, Joseph T. McGuire, Helena Melero, Adriana S. Méndez Leal, Benjamin Meyer, Kristin N. Meyer, Glad Mihai, Georgios D. Mitsis, Jorge Moll, Dylan M. Nielson, Gustav Nilsonne, Michael P. Notter, Emanuele Olivetti, Adrian I. Onicas, Paolo

Papale, Kaustubh R. Patil, Jonathan E. Peelle, Alexandre Pérez, Doris Pischedda, Jean-Baptiste Poline, Yanina Prystauka, Shruti Ray, Patricia A. Reuter-Lorenz, Richard C. Reynolds, Emiliano Ricciardi, Jenny R. Rieck, Anais M. Rodriguez-Thompson, Anthony Romyn, Taylor Salo, Gregory R. Samanez-Larkin, Emilio Sanz-Morales, Margaret L. Schlichting, Douglas H. Schultz, Qiang Shen, Margaret A. Sheridan, Jennifer A. Silvers, Kenny Skagerlund, Alec Smith, David V. Smith, Peter Sokol-Hessner, Simon R. Steinkamp, Sarah M. Tashjian, Bertrand Thirion, John N. Thorp, Gustav Tinghög, Loreen Tisdall, Steven H. Tompson, Claudio Toro-Serey, Juan Jesus Torre Tresols, Leonardo Tozzi, Vuong Truong, Luca Turella, Anna E. van 't Veer, Tom Verguts, Jean M. Vettel, Sagana Vijayarajah, Khoi Vo, Matthew B. Wall, Wouter D. Weeda, Susanne Weis, David J. White, David Wisniewski, Alba Xifra-Porxas, Emily A. Yearling, Sangsuk Yoon, Rui Yuan, Kenneth S. L. Yuen, Lei Zhang, Xu Zhang, Joshua E. Zosky, Thomas E. Nichols, Russell A. Poldrack, and Tom Schonberg. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582(7810): 84–88, 06 2020. doi: 10.1038/s41586-020-2314-9. URL https://www.nature.com/articles/s41586-020-2314-9. Publisher: Nature Publishing Group.

[9] Richard T. Burnett, Sabit Cakmak, Mark E. Raizenne, David Stieb, Renaud Vincent, Daniel Krewski, Jeffrey R. Brook, Owen Philips, and Haluk Ozkaynak. The association between ambient carbon monoxide levels and daily mortality in toronto, canada. *Journal of the Air & Waste Management Association*, 48(8):689–700, 08 1998. doi: 10.1080/10473289.1998.10463718. URL https://www.tandfonline.com/doi/full/10.1080/10473289.1998.10463718.

[10] Richard T. Burnett, Stieb ,Dave , Brook ,Jeffrey R. , Cakmak ,Sabit , Dales ,Robert , Raizenne ,Mark , Vincent ,Renaud , , and Tom Dann. Associations between short-term changes in nitrogen dioxide and mortality in canadian cities. *Archives of Environmental Health: An International Journal*, 59(5):228–236, 05 2004. doi: 10.3200/AEOH.59.5.228-236. URL https://doi.org/10.3200/AEOH.59.5.228-236. Publisher: Routledge _eprint: https://doi.org/10.3200/AEOH.59.5.228-236 PMID: 16201668.

[11] Margarita Castillejos, Borja-Aburto ,Victor H. , Dockery ,Douglas W. , Gold ,Diane R. , , and Dana. Loomis. Airborne coarse particles and mortality. *Inhalation Toxicology*, 12(sup1):61–72, 01 2000. doi: 10.1080/0895-8378.1987.11463182. URL https://doi.org/10.1080/0895-8378.1987.11463182. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0895-8378.1987.11463182.

[12] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. Unleashing the potential of prompt engineering for large language models. *Patterns*, 6(6):101260, 06 2025. doi: 10.1016/j.patter.2025.101260. URL https://www.sciencedirect.com/science/article/pii/S2666389925001084.

[13] Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006. doi: 10.1002/asi.20317. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20317. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/asi.20317.

[14] J. K. Chou and C. K. Yang. Papervis: Literature review made easy. *Computer Graphics Forum*, 30(3):721–730, 2011. doi: 10.1111/j.1467-8659.2011.01921.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2011.01921.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2011.01921.x.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Naacl-hlt 2019. page 4171–4186, Minneapolis, Minnesota, 06 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

[16] Douglas W. Dockery, Joel Schwartz, and John D. Spengler. Air pollution and daily mortality: Associations with particulates and acid aerosols. *Environmental Research*, 59(2):362–373, 12 1992. doi: 10.1016/S0013-9351(05)80042-8. URL https://www.sciencedirect.com/science/article/pii/S0013935105800428.

[17] Marian Dörk, Nathalie Henry Riche, Gonzalo Ramos, and Susan Dumais. Pivotpaths: Strolling through faceted information spaces. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2709–2718, 12 2012. doi: 10.1109/TVCG.2012.252. URL https://ieeexplore.ieee.org/document/6327277.

[18] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 12 2014. URL https://www.proquest.com/docview/1616141998/abstract/5E050DCE82414037PQ/1. Num Pages: 6 Place: Research Triangle Park, United States Publisher: Sigma XI-The Scientific Research Society.

[19] Elliot Gould, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, Daniel G. Hamilton, Robin N. Abbey-Lee, Jessica K. Abbott, Luis A. Aguirre, Carles Alcaraz, Irith Aloni, Drew Altschul, Kunal Arekar, Jeff W. Atkins, Joe Atkinson, Christopher M. Baker, Meghan Barrett, Kristian Bell, Suleiman Kehinde Bello, Iván Beltrán, Bernd J. Berauer, Michael Grant Bertram, Peter D. Billman, Charlie K. Blake, Shannon Blake, Louis Bliard, Andrea Bonisoli-Alquati, Timothée Bonnet, Camille Nina Marion Bordes, Aneesh P. H. Bose, Thomas Botterill-James, Melissa Anna Boyd, Sarah A. Boyle, Tom Bradfer-Lawrence, Jennifer Bradham, Jack A. Brand, Martin I. Brengdahl, Martin Bulla, Luc Bussière, Ettore Camerlenghi, Sara E. Campbell, Leonardo L. F. Campos, Anthony Caravaggi, Pedro Cardoso, Charles J. W. Carroll, Therese A. Catanach, Xuan Chen, Heung Ying Janet Chik, Emily Sarah Choy, Alec Philip Christie, Angela Chuang, Amanda J. Chunco, Bethany L. Clark, Andrea Contina, Garth A. Covernton, Murray P. Cox, Kimberly A. Cressman, Marco Crotti, Connor Davidson Crouch, Pietro B. D'Amelio, Alexandra Allison de Sousa, Timm Fabian Döbert, Ralph Dobler, Adam J. Dobson, Tim S. Doherty, Szymon Marian Drobniak, Alexandra Grace Duffy, Alison B. Duncan, Robert P. Dunn, Jamie Dunning, Trishna Dutta, Luke Eberhart-Hertel, Jared Alan Elmore, Mahmoud Medhat Elsherif, Holly M. English, David C. Ensminger, Ulrich Rainer Ernst, Stephen M. Ferguson, Esteban Fernandez-Juricic, Thalita Ferreira-Arruda, John Fieberg, Elizabeth A. Finch, Evan A. Fiorenza, David N. Fisher, Amélie Fontaine, Wolfgang Forstmeier, Yoan Fourcade, Graham S. Frank, Cathryn A. Freund, Eduardo Fuentes-Lillo, Sara L. Gandy, Dustin G. Gannon, Ana I. García-Cervigón, Alexis C. Garretson, Xuezhen Ge, William L. Geary, Charly Géron, Marc Gilles, Antje Girndt, Daniel Gliksman, Harrison B. Goldspiel, Dylan G. E. Gomes, Megan Kate Good, Sarah C. Goslee, J. Stephen Gosnell, Eliza M. Grames, Paolo Gratton, Nicholas M. Grebe, Skye M. Greenler, Maaike Griffioen, Daniel M. Griffith, Frances J. Griffith, Jake J. Grossman, Ali Güncan, Stef Haesen, James G. Hagan, Heather A. Hager, Jonathan Philo Harris, Natasha Dean Harrison, Sarah Syedia Hasnain, Justin Chase Havird, Andrew J. Heaton, María Laura Herrera-Chaustre, Tanner J. Howard, Bin-Yan Hsu, Fabiola Iannarilli, Esperanza C. Iranzo, Erik N. K. Iverson, Saheed Olaide Jimoh, Douglas H. Johnson, Martin Johnsson, Jesse Jorna, Tommaso Jucker, Martin Jung, Ineta Kačergytė, Oliver Kaltz, Alison Ke, Clint D. Kelly, Katharine

Keogan, Friedrich Wolfgang Keppeler, Alexander K. Killion, Dongmin Kim, David P. Kochan, Peter Korsten, Shan Kothari, Jonas Kuppler, Jillian M. Kusch, Malgorzata Lagisz, Kristen Marianne Lalla, Daniel J. Larkin, Courtney L. Larson, Katherine S. Lauck, M. Elise Lauterbur, Alan Law, Don-Jean Léandri-Breton, Jonas J. Lembrechts, Kiara L'Herpiniere, Eva J. P. Lievens, Daniela Oliveira de Lima, Shane Lindsay, Martin Luquet, Ross MacLeod, Kirsty H. Macphie, Kit Magellan, Magdalena M. Mair, Lisa E. Malm, Stefano Mammola, Caitlin P. Mandeville, Michael Manhart, Laura Milena Manrique-Garzon, Elina Mäntylä, Philippe Marchand, Benjamin Michael Marshall, Charles A. Martin, Dominic Andreas Martin, Jake Mitchell Martin, April Robin Martinig, Erin S. McCallum, Mark McCauley, Sabrina M. McNew, Scott J. Meiners, Thomas Merkling, Marcus Michelangeli, Maria Moiron, Bruno Moreira, Jennifer Mortensen, Benjamin Mos, Taofeek Olatunbosun Muraina, Penelope Wrenn Murphy, Luca Nelli, Petri Niemelä, Josh Nightingale, Gustav Nilsonne, Sergio Nolazco, Sabine S. Nooten, Jessie Lanterman Novotny, Agnes Birgitta Olin, Chris L. Organ, Kate L. Ostevik, Facundo Xavier Palacio, Matthieu Paquet, Darren James Parker, David J. Pascall, Valerie J. Pasquarella, John Harold Paterson, Ana Payo-Payo, Karen Marie Pedersen, Grégoire Perez, Kayla I. Perry, Patrice Pottier, Michael J. Proulx, Raphaël Proulx, Jessica L. Pruett, Veronarindra Ramananjato, Finaritra Tolotra Randimbiarison, Onja H. Razafindratsima, Diana J. Rennison, Federico Riva, Sepand Riyahi, Michael James Roast, Felipe Pereira Rocha, Dominique G. Roche, Cristian Román-Palacios, Michael S. Rosenberg, Jessica Ross, Freya E. Rowland, Deusdedith Rugemalila, Avery L. Russell, Suvi Ruuskanen, Patrick Saccone, Asaf Sadeh, Stephen M. Salazar, Kris Sales, Pablo Salmón, Alfredo Sánchez-Tójar, Leticia Pereira Santos, Francesca Santostefano, Hayden T. Schilling, Marcus Schmidt, Tim Schmoll, Adam C. Schneider, Allie E. Schrock, Julia Schroeder, Nicolas Schtickzelle, Nick L. Schultz, Drew A. Scott, Michael Peter Scroggie, Julie Teresa Shapiro, Nitika Sharma, Caroline L. Shearer, Diego Simón, Michael I. Sitvarin, Fabrício Luiz Skupien, Heather Lea Slinn, Grania Polly Smith, Jeremy A. Smith, Rahel Sollmann, Kaitlin Stack Whitney, Shannon Michael Still, Erica F. Stuber, Guy F. Sutton, Ben Swallow, Conor Claverie Taff, Elina Takola, Andrew J. Tanentzap, Rocío Tarjuelo, Richard J. Telford, Christopher J. Thawley, Hugo Thierry, Jacqueline Thomson, Svenja Tidau, Emily M. Tompkins, Claire Marie Tortorelli, Andrew Trlica, Biz R. Turnell, Lara Urban, Stijn Van de Vondel, Jessica Eva Megan van der Wal, Jens Van Eeckhoven, Francis van Oordt, K. Michelle Vanderwel, Mark C. Vanderwel, Karen J. Vanderwolf, Juliana Vélez, Diana Carolina Vergara-Florez, Brian C. Verrelli, Marcus Vinícius Vieira, Nora Villamil, Valerio Vitali, Julien Vollering, Jeffrey Walker, Xanthe J. Walker, Jonathan A. Walter, Pawel Waryszak, Ryan J. Weaver, Ronja E. M. Wedegärtner, Daniel L. Weller, and Shannon Whelan. Same data, different analysts: variation in effect sizes due to analytical decisions in ecology and evolutionary biology. *BMC Biology*, 23(1):35, 02 2025. doi: 10.1186/s12915-024-02101-x. URL https://doi.org/10.1186/s12915-024-02101-x.

[20] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. Citerivers: Visual analytics of citation patterns. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):190–199, 01 2016. doi: 10.1109/TVCG.2015.2467621. URL https://ieeexplore.ieee.org/document/7192685/authors.

[21] Wei Huang, Jianguo Tan, Haidong Kan, Ni Zhao, Weimin Song, Guixiang Song, Guohai Chen, Lili Jiang, Cheng Jiang, Renjie Chen, and Bingheng Chen. Visibility, air quality and daily mortality in shanghai, china. *Science of The Total Environment*, 407(10):3295–3300, 05 2009. doi: 10.1016/j.scitotenv.2009.02.019. URL https://linkinghub.elsevier.com/retrieve/pii/S004896970900165X.

[22] Nick Huntington-Klein, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, Paul Grieco, Godwin Ekpe, Todd Pugatch, Martin Saavedra, and Yaniv Stopnitzky. The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3):944–960, 2021. doi: 10.1111/ecin.12992. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/ecin.12992. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecin.12992.

[23] Petra Isenberg, Tobias Isenberg, Michael Sedlmair, Jian Chen, and Torsten Möller. Visualization as seen through its research paper keywords. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):771–780, 01 2017. doi: 10.1109/TVCG.2016.2598827. URL https://ieeexplore.ieee.org/document/7539364.

[24] Alex Kale, Matthew Kay, and Jessica Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. CHI '19, page 1–14, New York, NY, USA, 05 2019. Association for Computing Machinery. doi: 10.1145/3290605.3300432. URL https://dl.acm.org/doi/10.1145/3290605.3300432.

[25] Haidong Kan, Stephanie J. London, Guohai Chen, Yunhui Zhang, Guixiang Song, Naiqing Zhao, Lili Jiang, and Bingheng Chen. Differentiating the effects of fine and coarse particles on daily mortality in shanghai, china. *Environment International*, 33(3):376–384, 04 2007. doi: 10.1016/j.envint.2006.12.001. URL https://www.sciencedirect.com/science/article/pii/S0160412006002108.

[26] K Katsouyanni, J Schwartz, C Spix, G Touloumi, D Zmirou, A Zanobetti, B Wojtyniak, J M Vonk, A Tobias, A Pönkä, S Medina, L Bachárová, and H R Anderson. Short term effects of air pollution on health: a european approach using epidemiologic time series data: the aphea protocol. *Journal of Epidemiology and Community Health*, 50(Suppl 1):S12–S18, 04 1996. doi: 10.1136/jech.50.suppl_1.s12. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1060882/. PMID: 8758218 PMCID: PMC1060882.

[27] Klea Katsouyanni, Giota Touloumi, Evangelia Samoli, Alexandros Gryparis, Alain Le Tertre, Yannis Monopolis, Giuseppe Rossi, Denis Zmirou, Ferran Ballester, Azedine Boumghar, Hugh Ross Anderson, Bogdan Wojtyniak, Anna Paldy, Rony Braunstein, Juha Pekkanen, Christian Schindler, and Joel Schwartz. Confounding and effect modification in the short-term effects of ambient particles on total mortality: Results from 29 european cities within the aphea2 project. *Epidemiology*, 12(5):521, 09 2001. URL https://journals.lww.com/epidem/fulltext/2001/09000/confounding_and_effect_modification_in_the.11.aspx.

[28] Klea Katsouyanni, Jonathan M. Samet, H. Ross Anderson, Richard Atkinson, Alain Le Tertre, Sylvia Medina, Evangelia Samoli, Giota Touloumi, Richard T. Burnett, Daniel Krewski, Tim Ramsay, Francesca Dominici, Roger D. Peng, Joel Schwartz, and Antonella Zanobetti. Air pollution and health: A european and north american approach (aphena). Research Report 142, Health Effects Institute, Boston, MA, 2009.

[29] Oscar Kjell, Salvatore Giorgi, and H. Andrew Schwartz. The text-package: An r-package for analyzing and visualizing human language using natural language processing and deep learning. *Psychological Methods*, 2023. doi: 10.1037/met0000542. URL https://pubmed.ncbi.nlm.nih.gov/37126041/.

[30] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 02 2020. doi: 10.1093/bioinformatics/btz682. URL https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[31] Jiali Liu, Nadia Boukhelifa, and James R. Eagan. Understanding the Role of Alternatives in Data Analysis Practices. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):66–76, January 2020. ISSN 1941-0506. doi: 10.1109/TVCG.2019.2934593. URL https://ieeexplore.ieee.org/document/8805460/.

[32] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths explored, paths omitted, paths obscured: Decision points & selective reporting in end-to-end data analysis. CHI '20, page 1–14, New York, NY, USA, 04 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376533. URL https://dl.acm.org/doi/10.1145/3313831.3376533.

[33] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. Boba: Authoring and visualizing multiverse analyses. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1753–1763, 02 2021. doi: 10.1109/TVCG.2020.3028985. URL https://ieeexplore.ieee.org/document/9216579/.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. doi: 10.48550/arXiv.1907.11692.

[35] Elena López-Villarrubia, Ferran Ballester, Carmen Iñiguez, and Nieves Peral. Air pollution and mortality in the canary islands: a time-series analysis. *Environmental Health*, 9:8, 02 2010. doi: 10.1186/1476-069X-9-8. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2843667/. PMID: 20152037 PMCID: PMC2843667.

[36] T F Mar, G A Norris, J Q Koenig, and T V Larson. Associations between air pollution and mortality in phoenix, 1995-1997. *Environmental Health Perspectives*, 108(4):347–353, 04 2000. doi: 10.1289/ehp.00108347. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/ehp.00108347. Publisher: Environmental Health Perspectives.

[37] Suresh H. Moolgavkar. Air pollution and hospital admissions for diseases of the circulatory system in three u.s. metropolitan areas. *Journal of the Air & Waste Management Association*, 50(7):1199–1206, 07 2000. doi: 10.1080/10473289.2000.10464162. URL https://doi.org/10.1080/10473289.2000.10464162. Publisher: Taylor & Francis.

[38] Suresh H. Moolgavkar. Air pollution and daily mortality in two u.s. counties: Season-specific analyses and exposure-response relationships. *Inhalation Toxicology*, 15(9):877–907, 01 2003. doi: 10.1080/08958370390215767. URL https://doi.org/10.1080/08958370390215767. Publisher: Taylor & Francis.

[39] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. Vitality: Promoting serendipitous discovery of academic literature with transformers & visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):486–496, 01 2022. doi: 10.1109/TVCG.2021.3114820. URL https://ieeexplore.ieee.org/document/9552447/.

[40] Bart Ostro, Rachel Broadwin, Shelley Green, Wen-Ying Feng, and Michael Lipsett. Fine particulate air pollution and mortality in nine california counties: Results from calfine. *Environmental Health Perspectives*, 114(1):29–33, 01 2006. doi: 10.1289/ehp.8335. URL https://ehp.niehs.nih.gov/doi/10.1289/ehp.8335. Publisher: Environmental Health Perspectives.

[41] Roger D. Peng, Francesca Dominici, and Thomas A. Louis. Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(2):179–203, 03 2006. doi: 10.1111/j.1467-985X.2006.00410.x. URL https://doi.org/10.1111/j.1467-985X.2006.00410.x.

[42] Jonathan M. Samet, Francesca Dominici, Frank C. Curriero, Ivan Coursac, and Scott L. Zeger. Fine particulate air pollution and mortality in 20 u.s. cities, 1987–1994. *New England Journal of Medicine*, 343(24):1742–1749, 12 2000. doi: 10.1056/NEJM200012143432401. URL https://www.nejm.org/doi/full/10.1056/NEJM200012143432401. Publisher: Massachusetts Medical Society _eprint: https://www.nejm.org/doi/pdf/10.1056/NEJM200012143432401.

[43] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks (version 0.6.2). *OSF Preprints*, 2021. URL https://github.com/MUCollective/multiverse.

[44] Marko Sarstedt, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D. Liengaard. Same model, same data, but different outcomes: Evaluating the impact of method choices in structural equation modeling. *Journal of Product Innovation Management*, 41(6):1100–1117, 2024. doi: 10.1111/jpim.12738. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12738. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jpim.12738.

[45] Joel Schwartz. The distributed lag between air pollution and daily deaths. *Epidemiology*, 11(3):320–326, 2000. URL https://www.jstor.org/stable/3703220. Publisher: Lippincott Williams & Wilkins.

[46] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356, 09 2018. doi: 10.1177/2515245917747646. URL https://doi.org/10.1177/2515245917747646. Publisher: SAGE Publications Inc.

[47] Jan Simson, Fiona Draxler, Samuel Mehr, and Christoph Kern. Preventing harmful data practices by using participatory input to navigate the machine learning multiverse. CHI '25, page 1–30, New York, NY, USA, 04 2025. Association for Computing Machinery. doi: 10.1145/3706598.3713482. URL https://dl.acm.org/doi/10.1145/3706598.3713482.

[48] Imad Tbahriti, Christine Chichester, Frédérique Lisacek, and Patrick Ruch. Using argumentation to retrieve articles with similar citations: An inquiry into improving related articles search in the medline digital library. *International Journal of Medical Informatics*, 75(6):488–495, 06 2006. doi: 10.1016/j.ijmedinf.2005.06.007. URL https://www.sciencedirect.com/science/article/pii/S1386505605000894.

[49] G. Touloumi, E. Samoli, M. Pipikou, A. Le Tertre, R. Atkinson, and K. Katsouyanni. Seasonal confounding in air pollution and health time-series studies: effect on air pollution effect estimates. *Statistics in Medicine*, 25(24):4164–4178, 2006. doi: 10.1002/sim.2681. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.2681. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.2681.

[50] Kayo Ueda, Nitta ,Hiroshi , Ono ,Masaji , , and Ayano Takeuchi. Estimating mortality effects of fine particulate matter in japan: A comparison of time-series and case-crossover analyses. *Journal of the Air & Waste Management Association*, 59(10):1212–1218, 10 2009. doi: 10.3155/1047-3289.59.10.1212. URL https://doi.org/10.3155/1047-3289.59.10.1212. Publisher: Taylor & Francis _eprint: https://doi.org/10.3155/1047-3289.59.10.1212.

[51] Hadley Wickham. Tidy data. *Journal of Statistical Software*, 59:1–23, 09 2014. doi: 10.18637/jss.v059.i10. URL https://doi.org/10.18637/jss.v059.i10.

[52] Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

[53] Hadley Wickham, Joe Cheng, and Aaron Jacobs. *ellmer: Chat with Large Language Models*, 2025. URL https://CRAN.R-project.org/package=ellmer. R package version 0.1.1.

[54] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. doi: 10.48550/arXiv.2312.17617.

[55] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. doi: 10.48550/arXiv.1906.08237.