Analysing decisions in data analysis

H. Sherry Zhang Roger D. Peng

this is the abstract

1 Introduction

In this work, we design a tabular format to record the choices made by analysts during data analysis. Using large language models, we automatically extract these choices from a set of research papers focused on specific topics, e.g. air pollution modelling. This allows us to analyze these choices as data – tracking how they've changed over time or query the possible methodologies used in similar studies. We also introduce a workflow to cluster paper based on decision similarity, using both the decisions themselves and the justifications authors provide for their choices.

2 Background

Data analysis as an complicated, iterative process to make sense [ref] of the data collected. The iterative process of formulating hypothesis Jun et al. (2022).

Choices are made at nearly every stage of data analysis, ranging from variable pre-processing variables, variable and lag selection in model formulation, to the specification of smoothing parameter during model construction. These possible choices contribute to what Gelman and Loken (2014) describe as the "garden of forking paths". These choices can introduce substantial variability in results, which has been demonstrated in many-analyst experiments, where independent teams analyzing the same dataset to answer a pre-defined research question often arrive at markedly different conclusions. A prominent example is Silberzahn et al. (2018) where researchers reported a wide range of point estimates and 95% confidence intervals for the effect of soccer players' skin tone on the number of red cards awarded by referees (odds ratio from 0.89 to 2.93). Similar findings have emerged in other domains, including structural equation modeling (Sarstedt et al. 2024), applied microeconomics (Huntington-Klein et al. 2021), neuroimaging (Botvinik-Nezer et al. 2020), and ecology and evolutionary biology (Gould et al. 2025).

Another line of work focuses on developing software tools to support analysts in making more informed decisions. For example, the `Tisane` package (Jun et al. 2022) integrates conceptual ideas, such as DAGs, and modelling structure (group/ cluster/ hierarchical structure), to assist junior researchers in specifying GLM and GLMM model. The `DeclareDesign` package (Blair et al. 2019) introduces the MIDA framework for researchers to declare, diagnose, and redesign their analyses to produce a distribution of the statistic of interest. This approach has been applied in randomized controlled trial (Bishop and Hulme 2024) .

The `multiverse` package

- facilitates the specification and execution of multiple parallel choices for sensitivity analysis, allowing researchers to systematically explore how different choices affect results and to report the range of plausible outcomes that arise from alternative analytic paths.

Study decisions in data analysis:

- interview analysts and researchers to provide recommendation for data analysis practices (Kale, Kay, and Hullman 2019; Alspaugh et al. 2019; Yang Liu, Althoff, and Heer 2020).
- Yang Liu, Althoff, and Heer (2020) provides visualization to communicate the decision processes through the Analytic Decision Graphs (ADG)
- Simson et al. (2025) conducts a participatory AI study to demonstrate the "garden of forking paths" of decisions in data analysis and how it affects ML fairness

2.1 Recording decisions in data analysis

- give example from extracting decision from sentences of a paper
- adapt from the tidy data principle - each row is a decision Wickham (2014)

- some decisions are related to how the variable is estimated spatially and temporally
- model level decisions on how the model is estimated spatially (for multi-site analyses) and/or temporally (different treatments for years or seasons)
- sometimes the decisions are not explicitly stated in the paper (use AIC to choose the degree of freedom in a smoothing spline)
- sometimes the reason is not explicitly stated (e.g., why 3 degree of freedom)

A hypothetical database of decisions may look as follows:

Paper	ID	Model	variable	method	parameter	type	reason	decision
ostro	1	Poisson regression	temperature	smoothing spline	degree of freedom	parameter	NA	3 degree of freedom

Paper	ID	Model	variable	method	parameter	type	reason	decision
ostro	2	Poisson regression	temperature	smoothing spline	degree of freedom	temporal	NA	1-day lag
ostro	3	Poisson regression	relative humidity	LOESS	smoothing parameter	parameter	to minimize Akaike's Information Criterion	NA
ostro	4	Poisson regression	model	NA	NA	spatial	to account for variation among cities	separate regression models fit in each city

2.2 Austomatic reading of literature with LLM

- We use LLM to automatic process the literature to output analysis decisions. Currently, two LLMs, Antropic Claude and Google Gemini, are able to take input of pdf documents and the results reported in this paper is based on Gemini's output. See the section sensitivity analysis for the comparison of the two models.
- Claude is decoder only, Gemini is an encoder–decoder model
    - these models may paraphrase or hallucinate unless explicitly told not to since it is generative in nature based on the predicted probability of the next word from the text and the instruction
- prompt: instruct the LLM to produce a markdown file with decisions included in a JSON block with the fields described in Section xxx
- use the `ellmer` package (Wickham, Cheng, and Jacobs 2025) to connect to Gemini API to process the pdf documents in R.
- experiment with seed and temperature
- Our task differs from conventional text extraction using large language models (LLM) in that ~~it requires higher-order reasoning to identify the decisions made by the authors~~.

2.3 Review the LLM output

(the shiny app)

- screenshot of the interface
- The current application includes three actions:
  - modify a row (`dplyr::mutate(xxx = ifelse(CONDITION, "yyy" , xxx))`),
  - delete unrelated decisions (`dplyr::filter(!(CONDITION))`), and
  - manually add a decision (`dplyr::bind_rows()`)
- All the actions will generate the corresponding codes.
- The download button will download the modified decision database as a csv file

2.4 Decision quality summary

3 Calculate paper similarity

- pre-processing
  - standardize statistical methods its corresponding parameters (LOESS, smoothing spline, etc)
  - group variables into broader categories: time, temperature, humidity, PM
- identify the most frequent analysis decisions across papers
- retain only papers that report more than x such decisions
- measure similarity between decisions and their justificaiton using NLP
  - word embedding with attention mechanism, instead of bag of word,
  - specific NLP models (default to `bert-base-uncased`), aggregation methods from word to text
- compute paper similarity score for each paper pair by aggregating decision-level compoarisons
  - check/ report on the number of decisions compared in each paper pair
- similarity score can serve as the distance matrix to cluster papers by their similarity on decision choices

3.1 Sensitivity analysis

- standard BERT (Devlin et al. 2019), Roberta (Yinhan Liu et al., n.d.): trained on a much larger dataset (160GB v.s. BERT's 15GB), transformer-xl (Dai et al., n.d.), xlnet by Google Brain(Yang et al., n.d.), and two domain-trained BERT models: sciBert (Beltagy, Lo, and Cohan 2019) and bioBert(Lee et al. 2020), trained on PubMed and PMC data.

4 Applications

4.1 Air pollution mortality modelling

- look at for one type of decision (time) - what are the choices made by different papers
- look at whether decisions changes across time
- Visualize the decision database: apply clustering algorithm and visualize the database through `sigma.js`

4.2 Species distribution modelling

5 Discussion

- Only prompting engineering is used to extract decisions from the literature. We expect that fine-tuning the model on statistical or domain-specific literature to yield more robust performance on the same document, though it would require substantially more training effort.

Reference

Alspaugh, Sara, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A. Hearst. 2019. "Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices." *IEEE Transactions on Visualization and Computer Graphics* 25 (1): 22–31. https://doi.org/10.1109/TVCG.2018.2865040.

Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)." In, 3613–18. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371.

Bishop, Dorothy V. M., and Charles Hulme. 2024. "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. https://doi.org/10.1177/25152459241267904.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–59. https://doi.org/10.1017/S0003055419000194.

Botvinik-Nezer, Rotem, Felix Holzmeister, Colin F. Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, et al. 2020. "Variability in the Analysis of a Single Neuroimaging Dataset by Many Teams." *Nature* 582 (7810): 84–88. https://doi.org/10.1038/s41586-020-2314-9.

Dai, Zihang, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. n.d. "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context." https://doi.org/10.48550/arXiv.1901.02860.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "NAACL-HLT 2019." In, edited by Jill Burstein, Christy Doran, and Thamar Solorio, 41714186. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102 (6): 460–65. https://www.proquest.com/docview/1616141998/abstract/5E050DCE82414037PQ/1.

Gould, Elliot, Hannah S. Fraser, Timothy H. Parker, Shinichi Nakagawa, Simon C. Griffith, Peter A. Vesk, Fiona Fidler, et al. 2025. "Same Data, Different Analysts: Variation in Effect Sizes Due to Analytical Decisions in Ecology and Evolutionary Biology." *BMC Biology* 23 (1): 35. https://doi.org/10.1186/s12915-024-02101-x.

Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, Naibin Chen, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59 (3): 944–60. https://doi.org/10.1111/ecin.12992.

Jun, Eunice, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and René Just. 2022. "Hypothesis Formalization: Empirical Findings, Software Limitations, and Design Implications." *ACM Transactions on Computer-Human Interaction* 29 (1): 1–28. https://doi.org/10.1145/3476980.

Kale, Alex, Matthew Kay, and Jessica Hullman. 2019. "Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths." In, 114. CHI '19. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300432.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. "BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining." Edited by Jonathan Wren. *Bioinformatics* 36 (4): 1234–40. https://doi.org/10.1093/bioinformatics/btz682.

Liu, Yang, Tim Althoff, and Jeffrey Heer. 2020. "Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis." In, 114.

CHI '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3313831.3376533.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. n.d. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." https://doi.org/10.48550/arXiv.1907.11692.

Sarstedt, Marko, Susanne J. Adler, Christian M. Ringle, Gyeongcheol Cho, Adamantios Diamantopoulos, Heungsun Hwang, and Benjamin D. Liengaard. 2024. "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. https://doi.org/10.1111/jpim.12738.

Silberzahn, R., E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, et al. 2018. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1 (3): 337–56. https://doi.org/10.1177/2515245917747646.

Simson, Jan, Fiona Draxler, Samuel Mehr, and Christoph Kern. 2025. "Preventing Harmful Data Practices by Using Participatory Input to Navigate the Machine Learning Multiverse." In, 130. CHI '25. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3706598.3713482.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59 (September): 1–23. https://doi.org/10.18637/jss.v059.i10.

Wickham, Hadley, Joe Cheng, and Aaron Jacobs. 2025. *Ellmer: Chat with Large Language Models*. https://CRAN.R-project.org/package=ellmer.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. n.d. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." https://doi.org/10.48550/arXiv.1906.08237.