Analysing decisions in data analysis

H. Sherry Zhang Roger D. Peng

1 Introduction

2 Literature review

3 Construct decision databases

- give example from extracting decision from sentences of a paper
- adapt from the tidy data principle (**tidydata?**), each row is a decision

- 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:

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

PaperID	Model	variable	method	parameter	type	reason	decision
ostro 4	Poisson regression	model	NA	NA	spatial	to account for variation among cities	separate regression models fit in each city

4 Analysis pipeline

4.1 Austomatic reading of literature with LLM

- We use LLM to automatic read the paper through the `ellmer` package (Wickham, Cheng, and Jacobs 2025) and manually review the decision outputs. Both Antropic Claude and Google Gemini accept pdf inputs and we choose Claude. The prompt used to finetune the Claude LLM is available in the appendix.

4.2 Review the LLM output

(the shiny app)

- screenshot of the interface
- The current application includes three actions:

  1) modify a row (`dplyr::mutate(xxx = ifelse(CONDITION, "yyy" , xxx))`),
  2) delete unrelated decisions (`dplyr::filter(!(CONDITION))`), and
  3) 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

4.3 Calculate paper similarity

- define what does it mean by papers are similar: same reason? some decisions?
- lexical similarity through word embedding using the `text` package (Kjell, Giorgi, and Schwartz 2023)
- summarize paper similarity through item similarity

4.4 Visualize the decision database

- apply clustering algorithm and visualize the database through `sigma.js`

5 Examples

5.1 Air pollution mortality modelling

5.2 Another Poisson GAM

6 Conclusion

Reference

Kjell, Oscar, Salvatore Giorgi, and H. Andrew Schwartz. 2023. "The Text-Package: An r-Package for Analyzing and Visualizing Human Language Using Natural Language Processing and Deep Learning." *Psychological Methods*. https://doi.org/10.1037/met0000542.

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