Appendix to "Analysing decisions in data analysis"

H. Sherry Zhang Roger D. Peng

LLM prompt

Task

Consider yourself as an applied statistician and you will read a PDF file to extract the decisions made in the data analysis. Your output should contain a JSON code block under the top-level key "decisions". Each item in "decisions" should be a dictionary with the following fields:

- `model`: the main model used to model mortality vs. air pollution (e.g., "generalized additive model", "distributed lag model"). Should be one model per paper.
- `variable`: the variable the statistical method is applied to.
- `method`: standard statistical method applied, for example, "LOESS", "smoothing spline", and "natural spline". If the `type` is "spatial" or "temporal", write "NA". There may be multiple methods discussed in the paper. Only extract the method that is actually used on the variable.
- `parameter`: the parameter name of the statistical method discussed, for example, "degrees of freedom", and "number of knots". If the `type` is "spatial" or "temporal", write "NA". Do not use the value of the parameter, for example, "3 df".
- `type`: one of "parameter", "spatial", or "temporal", indicating the nature of the decision.
- `reason`: the reason for the decision made. If none is given, write "NA".
- `decision`: the decision made regarding the parameter, spatial, or temporal aspect. If not specified, write "NA".
- `reference`: if the reason for the decision has any reference, find the corresponding reference in the reference section and write the reference in the following format: abc, where a is the last name of first author (without accent and all lower letters), b is the year in four digit format, and c is the first word of the title, excluding a/an/the, e.g. braga2001lag. If multiple references are given for a decision, include all of them in the reference field, separated by commas and a space. Do not stop at the first reference. Do not split into multiple rows

An example looks like the following:

```json
{
  "decisions": [
    {
      "model": "generalized additive model",
      "variable": "PM2.5",
      "method": "smoothing spline",
      "parameter": "smoothing parameter",
      "type": "parameter",
      "reason": "based on published litearture",
      "decision": "5 df",
      "reference": "smith2005trend"
    }
  ]
}
```

Rules

- Some decisions may require linking information across sentences and/or paragraphs to identify the correct variable. Example:
  - Text: if a paragraph states "… model adjusted for overdispersion, to estimate associations between day-to-day variations in pollutant concentrations (lag 0–5) and day-to-day variations in hospital admission counts for three health outcomes separately. A basic model without pollutants was built first, ….", Followed by the next paragraph: "After assessing the effects of single day concentrations (lags 0–5), …"
  - Guidance: The temporal decision ("lag 0-5" ") refers to pollutant concentrations, not the model.
- If a sentence includes multiple decisions that vary by outcome or variable, each decision should be treated and recorded separately.
  - Example 1:
    - Text: "resulting in a 4-day pollutant average (lag 0–3) for CVD, a 5-day average (lag 0–4) for RD, and a 6-day average (lag 0–5) for asthma"
    - Output: three separate decisions of type "temporal" " (variable/ decision): 1) CVD/ 4-day average (lag 0–3), 2) RD/ 5-day average (lag 0–4), 3) Asthma/ 6-day average (lag 0–5)
  - Example 2:
    - Text: "smoothing splines of one-day lags of humidity [each with 3 degrees of freedom (df)]"

        - Output:
            - Decision 1: variable: “humidity”, type: “temporal”, decision: “one-day lag”
            - Decision 2: variable: “humidity”, type: “parameter”, decision: “3 df”
    - Example 3:
        - Text: “three- day averaged temperature and dew point temperature with a natural cubic spline with three d.f..”
        - Output:
            - Decision 1: variable: “temperature”, method: “NA”, type: “temporal”, decision: “3-day averaged”
            - Decision 2: variable: “dew point temperature”, method: “NA”, type: “temporal”, decision: “3-day averaged”
            - Decision 3: variable: “temperature”, method: “natural cubic spline” , type: “parameter”, decision: “3 df”
            - Decision 4: variable: “dew point temperature”, method: “natural cubic spline”, type: “parameter”, decision: “3 df”

- Temporal terms may indicate parameter choices: words like “monthly,” “weekly,” or “daily” may appear in sentences, but they do not always imply temporal decisions. Sometimes they describe how a smoothing function is applied, which is a parameter decision. Example:

    - Text: “We set time-based knots at monthly midpoints.”
    - Interpretation: This describes a spline function applied to time, using 12 knots (one per month)
    - Output: This is a *parameter* decision, not a temporal one.

- Please use cognitive reasoning to determine whether the extracted reason is *specific* to the decision made. If the reason is general and doesn’t justify the particular choice, write “NA”. Example:

    - Text: “To control for weather variables, the 14-day lagged moving average (including concurrent day and previous 13 days) of temperature, dew point temperature, and barometric pressure (as both linear and quadratic terms) were included in the conditional logistic regression models.”,
    - Decision: “14-day lagged values”
    - Extracted reason: “to control for weather variables”
    - Output: “NA” because the reason is general to why the weather variables are included and do not justify why this temporal decision of 14-day lag is chosen

- Only include rows where a clear decision is made on a parameter, spatial, or temporal aspect of the modeling. For ambiguous or vague decisions (e.g., “not below 2 months”), do not record a row.

- Do NOT include decisions about inclusion of variables, for example, reasons as “to filter out cyclical patterns” and decision as “include day-of-the-week variable.
- A paper may contain multiple models; label them precisely (e.g., “generalized additive Poisson regression”, “lag distributed model”).
- Convert numbers written in words to numerals. For example, rewrite “one-day lag” as “1-day lag”. Apply this consistently across all entries.