Provocation

Pre-registration locks researchers into facile statistical tests

Without extensive exploratory analysis and visualization, data-driven analysis will lead to weak or straw-men hypotheses

Roger Beecham, June 2024

Introducing provocation

Pre-registration practices are an appealing pathway out of the replication crisis. However, the implied workflow – exploratory analysis > pre-register hypotheses > collect new data to test pre-registered hypotheses – is restrictive. Pre-registration works in experimental research designs, where the collected data map cleanly to the processes under investigation, and in mature research settings where prior studies and data exist that can be used to judge and evaluate observed effects. In (Transport) Geography, my 'home' discipline, the space for doing this form of pre-registration seems quite small. Target populations, and proxy samples, need to be defined with precision; important conditioning context needs to be known, described and represented in a meaningful way with data; hypothesis tests need to be constructed with anticipated effect sizes (differences from expectation) that are the right amount of "detectable" (or else what's the point in testing?).

An example from my research: our attempts to quantify the impact of contraflow cycling infrastructure on observed cycle volumes and crashes, looking at 20+years of data in London. Although we did spec up a research design, much analytic work – exploratory analysis – was needed to derive outcome variables that made sense. For example, non-straightforward decisions around infrastructure data; linking estimates of recorded crashes and modelled cycle counts (exposure) to sections of infrastructure; allocating linked crashes to pre- and post-infrastructure periods. Each of these analysis steps were implemented quite differently to what was laid out in our initial research design. Separate from this was the muddiness of the 'global' statistics used to claim findings: those estimating change in cycling rates and crash risk pre- and post- infrastructure. When disaggregating by different context variables, we found that these estimated rates varied enormously, including directionally. Some of the factors contributing to variation could be captured and incorporated into our analysis; other factors, suggested via more individual analysis of edge cases, could be described but not so obviously captured. The assumption baked into the statistics presented in our analysis was that random differences in context 'balance out' in our global-level statistic. It didn't take much thinking-time for us to loose faith in this assumption. So key findings were presented with corresponding levels of humility; our analysis of cycling volumes and contraflow emphasised more descriptive elements of the analysis.

Consequences: What it brings

The contraflow cycling case is typical in geographical analysis and applied data analysis more generally (Beecham and Lovelace 2023). If this provocation emphasises flexibility, removing the straightjacket or preoccupation with pe-registered statistical testing, then there are useful consequences for applied analysis. The provocation gives license to doing and publishing exploratory work that seeks to **describe** processes, interactions, relationships in rich detail. So this is the main benefit of the provocation: *describing* the space under which some events, processes or observed changes happened. In the example of contraflow cycling infrastructure, for those road sections that experienced large changes in cycling volumes and crash rates, what is it about contraflow in those cases that might have contributed to change, and what does this tell us about the design and provision of cycling facilities?

Future

This emphasis on rich description might seem overly “qualitative” and at odds with the scientific method. In fact, it invites us to think *causally* about the phenomena under observation. By characterising in detail the context under which some events happen, and do not happen, we naturally begin to think about how events were generated, the regularity with which they appear. The challenge becomes how to represent (tools) and evaluate (statistical models) this detailed /rich space of context-dependent outcomes in a reasonably formal, standardised and accessible way.

Relevant reading (to update)

- Beecham, R. and Lovelace, R. (2023), “A Framework for Inserting Visually Supported Inferences into Geographical Analysis Workflow: Application to Road Safety Research”, *Geographical Analysis*, 55: 345-366. https://doi.org/10.1111/gean.12338
- Meyer, M. and Dykes, J. (2020) "Criteria for Rigor in Visualization Design Study", *IEEE Transactions on Visualization and Computer Graphics*, 26(1):87-97, doi: 10.1109/TVCG.2019.2934539. *> Describes interpretivist approach to presenting Design Studies. May be useful ideas here for presenting and structuring descriptive work in applied data analysis.*