Models are not exploratory artifacts but rather a continuum between bottom-up and top-down approaches in updating prior beliefs

Edwin Pos, Quantitative Biodiversity Dynamics and Botanic Gardens, Utrecht University

Introduction: In the classic paradigm, statistical models can be used to either test explicit hypotheses or to explore potential underlying patterns within data without a priori (specific) expectations (referred to in Hullman and Gelman, 2021 by Confirmatory and Exploratory Data Analysis respectively). In simpler terms, this paradigm constitutes either a top-down or bottom-up approach to data analysis. Either we have preconceived formulations or objectives we want to validate, or we have no explicit a priori hypothesis but rather seek to update our (prior) beliefs by building these from the data at hand. Whether to use one or the other depends on the type of question and stage of analysis.

What it brings: Model building is not always about *exploring* structure and/or outcomes. There are times in which there is already a clear distinct question that asks for a concrete test, i.e. whether two populations are expected to be drawn for the same underlying distributions or not. For this, however, one already needs to have a predefined idea of what the processes are that might account for these similarities or differences (e.g. bottom-up, we have a preconceived notion of what might be the case and wish to test this). It is when we do not know what to expect or what the underlying processes are that a more explorative (top-down) approach can be beneficial. The latter, however, are sensitive to our prior beliefs (talking in a Bayesian frame of reference): stating simply we are prone to see what we are familiar wit

Consequences: Focusing on either CDA or EDA restricts our potential of exploring data. Too much emphasis on EDA might lead to superficial insights if not validated further and can lead to subjective conclusions whereas too much emphasis on CDA might restrict us from noticing patterns that might support alternative hypotheses. Either way, focusing on one or the other leads to a narrow view of data exploration and testing. In addition, where for CDA there are usually straightforward, mathematical, means to validate hypotheses (i.e. confidence intervals, p-values and R_2 interpretations of data fits), such approaches either do not exist or are not commonplace when it comes to EDA.

Future: Formulating a more quantitative way of EDA using a Bayesian approach as proposed by Hullman and Gelman (2021) but in a more rigorous manner could create a conceptual bridge between CDA and EDA. A more Bayesian way of thinking about data exploration by incorporating statistical inference rather than explicit testing as a start of EDA, which can then logically be followed by the more traditional CDA. One potential interesting starting point of exploration could be to integrate principles from information theory to objectively update beliefs about patterns and underlying relationships with user-informed prior information as a starting point. Introducing such approaches as a bridge in the existing realm of CDA and EDA can lead to data-driven exploration of data that simultaneously still allows for quantitative validation.

References:

Hullman, J., & Gelman, A. (2021). Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference. Harvard Data Science Review, 3(3). https-//doi.org/10.1162/99608f92.3ab8a587