Mikko Hyyryläinen

Type: *Writing an response to data science article of your choice discussing its methodological choices: 5 points per each response*

Link to the article:
https://www.sciencedirect.com/science/article/abs/pii/S0304422X13000570

This paper is a response to Emily A. Marshall's article *Defining population problems: Using topic models for cross-national comparison of disciplinary development*. It was published in *Poetics* 41:6 in 2013.

Emily A. Marshall's article reports a study which used correlated topic modeling "for a cross-national comparison of the development of research agendas in the discipline of demography" (701). The results were that "demographic research agendas reflected both cultural and institutional differences that shaped different understandings of fertility decline" (701). The article was a contribution to the discussion on how the trajectories of academic disciplines vary between national contexts.

My response starts from how the author frames the data that is used. The data consists of "journal articles and newspaper items" (704) and the author defines the role of journals for a discipline as that they are a forum (which consists of scholars who review each other's work) where the discipline and its core concerns are defined. However, I think that the way that the author views the whole publishing process is somehow too idealistic. At least if the discipline is fragmented in a way that social sciences usually tend to be, journal articles (especially in leading journals) do not necessarily tell us about the discipline and its development in general. Instead it can tell us about, for example, what its editors think is a good example of the work that is done in that discipline – i.e. it does not represent the who discipline. I think that it would perhaps have been more beneficial to take data also from those journals that are not leading journals so there would have been a much wider perspective on the discipline and its development. Of course, this would have meant that there would have been even more data and that would probably have been quite impractical.

The combination of methods that is used are correlated topic models which are then combined with interpretive analysis (706). The article explains quite well why correlated topic modeling was chosen instead of the more traditional Latent Dirichlet Allocation (707). The process of choosing the right number of topics was described in a clear and convincing way and at least I do not have much questions or critiques towards it since it seemed to be quite logical. However, the author discussed only in a very short passage of how "some evidence suggests that the structure of newspaper articles is different from that of journal articles". I think that this is a quite relevant thing for the analysis and should have been discussed in a much more detail. In addition, thinking about the differences between journal and newspaper articles could have made the topic modeling task easier.

The role of interpretation in topic modeling is highlighted. Author says that the interpretation of the topics it is "guided by the substantive knowledge, research question, and the interpretation of the investigator" (711). The way the author verified the results of topic modeling was that they used top 25 words generated for each topic. In addition, they searched for articles that was a good representation of the topic and looked at them closer (in some cases read the article). I think that in relatively many articles (that I've read..) which use machine learning methods the analysis is verified with the help of some other person who, for example, attempts to interpret at least a few topics. I think that this is a good practise and something that should have been used also in this study. Or at least, the analysis could have been even more convincing "if this practice was used in this study.

The analysis, or results section is clearly written and presented. The interpretations are clearly supported by the data and the analysis is really interesting since it compares two nations and in the case of the UK it compares "normal newspaper" to the academic journal. However, my critique is about the data and what the data is about. I am not sure if by looking at only one (leading) journal we can say about the "differences in the kinds of research questions addressed by academic demographers in Great Britain and France during the analytic period [1946–2005]" (713). For example, British journal *Population Studies* may have been a good window to the field demographics in the 1950's. However, I would suggest that since the number of academic journals is probably much higher in the 1990's than in the 1950's, after some point analyzing only one journal does not necessarily tell us much about the development of the discipline in general. In addition, I am not sure if comparing British and French fields of demographics tell us that much about the national differences in the development of disciplines. Since it could be claimed that the academic world of France is quite often characterised an intellectual universe of its own which tends to develop on its own since it uses its own language and outsiders are not that eager to influence it in any way (or at least this is the expression I have obtained – I may be wrong). Of course, the goal of the article was to discuss the differences between national fields, but by choosing French demographics, it was quite certain that there would be differences and the result that there would not be differences was quite improbable.

Although I find some things to critique (most of which are about what the data is actually about and the choice of the field that are compared), I found this article quite interesting and it was clearly written. I think that it is a very good article that illustrates what we can do with topic modeling. Topic modeling seemed to be a good choice of a method for this task.