MATHEMATICAL TRIPOS PART III 2023
Comments for Candidate

Essay: 39 - Measuring the Unmeasurable

Candidate: E567C - Gibbon, Jonah *Assessor:* Professor J. Aston

This essay looks at methods for multiple system estimation particularly for understanding the ExtremeBB dataset. This comprehensive dataset (curated by the Computer Lab in Cambridge) records posts on extremism forums and uses them for research into online extremist behaviour. This essay endeavours to use the data to get plausible estimates about the numbers of extremists using a multiple systems estimation approach.

The essay starts with a critique of the original classification method used in the ExtremeBB dataset, and then proposes an updated methods to generate the data table. This, in its own right, is a complicated piece of work, as it has to balance computational, statistical and pragmatic contraints. The new criteria is very well argued, and then used to create the dataset in the subsequent analysis.

The next section of the essay provides an overview of latent class modelling and then reviews the non-parametric latent class model approach. The approach is then implemented and tested rigorously through a Gibbs Sampling Algorithm. All the implementation was done by the candidate. The approach was then compared to other approaches available in the literature, but it was shown that most of these had computational bottlenecks which mean that these approaches are not applicable to the ExtremeBB dataset. It was very good to see the comparison with the Human Trafficking data set for all methodology though.

In the final section, the Gibbs Sampling Algorithm was applied to the ExtremeBB data set. This allowed a time varying estimate of the number of extremists associated with such boards (even those not posting) to be estimated. One of the really interesting findings (beyond the fact that an estimate can even be made, which is of significant interest in its own right), is that the time varying nature of the numbers changes whether you account for the missing population or not. This finding is quite strikingly shown in Figure 3.

A concluding section setting out some of the drawbacks of the work is given, as well as a few ideas for future further work.

Overall, this is an excellent example of an applied statistics essay. The candidate has really thought about all aspects of the dataset, from the data setup, through the modelling to the insights which can be derived from the analysis. They have done careful comparisons with other methodology and shown how they can overcome computational issues. The findings from this essay will likely be very useful in the analysis of extremism data. Overall, the candidate should be very proud of this essay.