UNIVERSITY OF TORONTO
DEPARTMENT OF SOCIOLOGY
PH.D. COMPREHENSIVE EXAMINATION IN COMPUTATIONAL AND QUANTITATIVE METHODS
AUGUST 19-24, 2024
MAJOR OPTION

You are required to answer THREE (3) QUESTIONS (ONE QUESTION FROM EACH OF PARTS A, B & C). Each answer should be 10-12 pages (12-point Times New Roman font, standard margins, and double-spaced) in length. The complete exam should not total more than 36 pages in length (12-point Times New Roman font, standard margins, and double-spaced), not including references. THE COMMITTEE WILL NOT READ PAST THE 36TH PAGE.
You must copy and paste the questions you choose at the top of your written answers

Part A

1. The increasing availability of online data offers enormous potential for social research, but also introduces several challenges. Outline some of the ways modern online digital technologies can be used to generate data and describe the advantages these offer over methods that do not make use of these technologies. Then, describe the challenges that these new approaches face and critically evaluate the effectiveness of the main solutions that have been offered. Finally, pose a research question that you might answer using data collected online. Describe how you would design your data collection, including the steps you would take to maximize data quality. Be sure to ground your justifications in the principles you described earlier.
2. Statistical analyses in the social sciences are fundamentally concerned with detecting real patterns that link together social phenomena. Stated more forcefully, we want to discover what is true about the social world (i.e., make proper inferences from our samples to the population). Outline and justify criteria for identifying true findings. In other words, when should we trust research results? Next, discuss some of the challenges inherent in searching for truth using quantitative methods. Pay attention to issues that arise at all stages of the research process, such as sampling, measurement, analysis, and interpretation. What problems does a researcher face, and what solutions exist to address them? How effective are these solutions? Finally, imagine that you were given a generous budget to conduct a study. Describe how you would design your study (from start to finish) to maximize the chance of obtaining a real/true result.

Part B

1. Directed acyclic graphs (DAGs) have become an increasingly important tool for causal inference in sociology. This question has four parts. First, explain the key components and principles of DAGs and how they represent causal relationships. In your answer, make sure to explain the concept of d-separation and the backdoor criterion. Second,

discuss how DAGs can be used to identify potential sources of bias in causal inference, such as confounding, selection bias, and overcontrol. Third, provide an example of how a DAG might reveal a non-obvious source of bias in a sociological research context. For this part, visually present the DAG and clearly label the nodes, and walk through the potential sources of biases. Finally, critically evaluate the utility of DAGs for sociological research. In answering this question, consider to what extent DAGs can help address long-standing challenges in making causal inferences from observational data. Also, consider some of the potential pitfalls or limitations of using DAGs to understand complex social phenomena.

2. Clustering techniques are an important part of unsupervised learning, used to identify "natural" groupings within data without predefined labels. This question has three parts. First, explain the basic concepts and principles of clustering methods, such as K-means clustering and hierarchical clustering, and their underlying assumptions. Include in your explanation the key practical considerations when clustering sociological data, such as the choice of dissimilarity measure to be used. Second, discuss the potential applications of clustering techniques in sociology, such as identifying subgroups within a population or uncovering hidden patterns in survey responses. Provide examples to illustrate these applications. Finally, critically evaluate the advantages and limitations of using clustering techniques in sociological research. In crafting your answer, consider issues such as interpretability, the influence of scaling and normalizing data, determining the number of clusters, and the potential for overfitting.

Part C

1. Supervised learning approaches are increasingly used by sociologists for various prediction tasks. Discuss some of the commonly used supervised learning methods for (a) classification and (b) regression and be sure to note the relative advantages/disadvantages of each modeling approach. Finally, identify and describe a specific research task where a sociologist could apply supervised learning. Using this example, explain how you would select and train an appropriate model for your prediction task. In your answer, be sure to discuss the following concepts: training and test samples, bias-variance trade-off, and cross-validation.
2. Sociological studies often use data structures that are nested. For example, students are nested in classrooms and/or schools, households are nested in census tracts, employees are nested within firms, and monthly observations of individual behaviour are nested within the observed individuals. At different times, this nested nature of data is seen as either an obstacle to causal inference or an important feature of the social world that demands sociological attention. Discuss these two alternative ways of understanding nested data structures. Identify and explain the common approaches to modeling nested data with these alternative goals in mind.