

Statistical Disclosure Risk Assessment in Remote Analysis Systems

Bronwyn Loong^a and Christine M. O'Keefe^b

Remote analysis systems enable users to analyse confidential microdata collected by statistical agencies, without accessing any microdata. Users submit queries to a secure environment where the original data is held, the system runs the query, and results are returned to the analyst. The statistical agency must assess the disclosure risk of any output delivered to the user from the remote access server. This paper reviews the literature on disclosure risk assessment methods and discusses their applicability to disclosure risk assessment for remote analysis system output. Key elements of the assessment are a clear definition of disclosure risk, identification of the features of the data that pose high disclosure risk, description of the potential intruder attack to exploit the high disclosure risk features, and a measure to quantify the disclosure risk impact of the intruder attack. The literature prescribes several statistical models which may be used by agencies to quantify disclosure risk. However, it is also necessary for the agency to make some ad-hoc assumptions which are specific to the data set concerned. We conclude that the disclosure risk from remote analysis system output is driven by the explanatory power of the queries submitted to the remote analysis system.

^aResearch School of Finance, Actuarial Studies and Applied Statistics, The Australian National University, Canberra, ACT 0200, Australia

^bCSIRO Mathematics, Informatics and Statistics GPO Box 664, Canberra ACT 2601 Australia