Disclosure Risk Assessment for Official Statistics

Chris Skinner*
London School of Economics and Political Science, London, United Kingdom - c.j.skinner@lse.ac.uk

Abstract

A key challenge of statistical disclosure control is the assessment of disclosure risk. This talk will review alternative approaches, with an emphasis on more recent developments. Risk assessment varies according to the mode of access. In official statistics in the UK, approaches of growing importance are (i) open data, including public use files, occasionally containing synthetic data, and (ii) secure settings, which may include an element of remote access and may also include forms of linked data. In addition, there continues to be the release or more traditional tabular outputs as well as microdata under end-user licences. The assessment of the risk of disclosure for alternative modes of access generally requires assumptions about the nature of scenarios of attack, including possible key variables which might be used for matching. Some evidence to inform such assumptions may be derived from 'intruder testing', where some 'friendly intruders' are invited to test the safety of possible output releases. Such testing may also provide direct measures of risk via proportions of claimed disclosures which are correct. Such experiments may be extended to other empirical approaches to risk assessment, where external databases are constructed, possibly using the original data source under study, and systematic ways of linking the external database to the proposed output for release are studied. An alternative to such directly empirical methods is to formulate theoretical measures of risk based upon assumptions about the behaviour of a hypothetical intruder. One line of development has been to formulate ways of estimating probabilities of identification or predictive disclosure based upon statistical models. This includes models which seek to take account of the riskiness of data elements which are unique in the population. A rather different recent line of development has introduced the notion of differential privacy. A final set of methods are more 'rule-based', identifying a proposed output as disclosive or not, according to whether it obeys a specified set of rules. Some justification for such rules may come from empirical testing or from more theoretical arguments.