International Journal of Population Data Science

Journal Website: www.ijpds.org

Zone design for statistical disclosure control in administrative and linked microdata

Robards, James^{1*}, Martin, David¹, and Gale, Chris¹

¹University of Southampton

Objectives

To explore the application of automated zone design tools to protect record-level datasets with attribute detail and a large data volume in a way that might be implemented by a data provider (e.g. National Statistical Organisation/Health Service Provider), initially using a synthetic microdataset. Successful implementation could facilitate the release of rich linked record datasets to researchers so as to preserve small area geographical associations, while not revealing actual locations which are currently lost due to the high level of geographical coding required by data providers prior to release to researchers. Data perturbation is undesirable because of the need for detailed information on certain spatial attributes (e.g. distance to a medical practitioner, exposure to local environment) which has driven demand for new linked administrative datasets, along with provision of suitable research environments. The outcome is a bespoke aggregation of the microdata that meets a set of design constraints but the exact configuration of which is never revealed. Researchers are provided with detailed data and suitable geographies, yet with appropriately reduced disclosure risk.

Approach

Using a synthetic flat file microdataset of individual records with locality-level (MSOA) geography codes for England and Wales (variables: age, gender, economic activity, marital status, occupation, number of hours worked and general health), we synthesize address-level locations within MSOAs using 2011 Census headcount data. These synthetic locations are then associated with a range of spatial measures and indicators such as distance to a medical practitioner. Implementation of the AZTool zone design software enables a bespoke, non-disclosive zone design solution, providing area codes that can be added to the research data without revealing their true locations to the researcher.

Results

Two sets of results will be presented. Firstly, we will explain the spatial characteristics of the new synthetic dataset which we propose may have broader utility. Secondly, we will present results showing changing risk of disclosure and utility when coding to spatial units from different scales and aggregations. Using the synthetic dataset will therefore demonstrate the utility of the approach for a variety of linked and administrative data without any actual disclosure risk.

Conclusion

This approach is applicable to a variety of datasets. The ability to quantify the zone design solution and security in relation to statistical disclosure control will be discussed. Provision of parameters from the zone design process to the data user and the implications of this for security and data users will be considered.

*Corresponding Author:

Email Address: james.robards@soton.ac.uk (J. Robards)