Task Force on
"Spatial Anonymization in Public-Use Household Survey Datasets"

1. Background

Household survey data collection frequently involves the use of handheld GPS units and GPS-enabled tablets to geo-reference dwellings, agricultural parcels and plots, facilities and other locations directly associated with surveyed communities and respondents. These precise location data are considered confidential, as they could be used to identify individual respondents. In order to maximize the analytical use of survey data, efforts have been made to disseminate anonymized spatial data that meets a spectrum of research needs while maintaining confidentiality. These include a method for generating modified coordinates – first developed and used in the dissemination of data from the USAID-funded Demographic and Health Surveys (DHS) (Burgert et al., 2013), and later adopted by the World Bank Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) program, which additionally publicly disseminates a set of spatial variables that are computed based on actual GPS coordinates (that are otherwise not made available) and that capture the geography, infrastructure and natural resource characteristics of survey sites.

While these approaches have thus far provided a satisfactory solution to most research applications, several issues are driving a renewed interest in the topic of spatial anonymization, fueled by the potentially huge analytical gains of greater access. First, the concept of data privacy and obligation of data providers to safeguard personal data has come into greater focus. The EU General Data Protection Regulation (GDPR), which came into effect in May 2018, includes location data in its definition of "personal data", holding collectors of such data to high standards in data protection and security. At the same time there are numerous examples of unintentional but consequential data exposure, such as the inadvertent public mapping of US military bases through the use of fitness tracking apps. Such examples are made increasingly likely because of the dramatic expansion of publicly available data and imagery, as well as open source tools and cloud computing that facilitate integration of data from different sources and data mining. Lastly, this rich data landscape and analytical applications in turn spur greater demand for access to more precise location information for valuable cutting-edge research applications.

In response to this changing context, a review of existing protocols is warranted, with an overarching objective of maximizing the analytical usefulness of confidential location data collected with surveys within the constraints of maintaining confidentiality of respondents. The ultimate objective would be to create guidelines in spatial anonymization of public use microdata.

2. Activities

At least two distinct and complementary lines of research will be sought in addressing the existing constraints and converging on widely-accepted global guidelines based on sound empirical evidence.

The first line of methodological research should evaluate the risk of disclosure associated with different coordinate masking strategies, including the DHS methods, and the dissemination of spatial variables, as promoted by the World Bank LSMS-ISA. As part of this work program, there is a need to first agree on the alternative coordinate masking strategies that should be evaluated. The on-going research that is being conducted by the World Bank LSMS in this domain is currently considering the following alternatives: (1) the existing DHS method, (2) a variant of the DHS method, using an interior exclusion zone, (3) adaptive masking, (4) coordinate truncation, and (5) the use of lowest administrative unit centroid/polygon. This work program could be expanded to include additional coordinate masking strategies, to be discussed and agreed with other members of the Task Force.

The choice of a particular masking technique will ultimately depend on the specific use. In order to guide users in choosing the most appropriate strategy, a complementary line of work will assess the relative empirical utility associated with the use of anonymized data obtained under different masking strategies. The work will initially focus on two areas of research currently being pursued by the LSMS team at the World Bank, namely: (i) small-area estimation of poverty through the combination of household survey data, high-resolution satellite imagery and machine learning models, and (ii) estimation of agricultural productivity effects of climate shocks, depending on whether the weather variables are computed based on unmodified household/plot coordinates versus a range of modified alternatives. The Task Force could consider supporting additional empirical applications for understanding the analytical impact of spatial anonymization in a more comprehensive manner so as to provide users with a broader set of options, conditional on the specific use.

3. Outputs and Timeline

The analytical outputs that will be created in each research track should be distilled and feed into the guidelines on spatial anonymization of public use microdata. The guidelines should provide operational recommendations on spatial anonymization for survey practitioners at-large and the National Statistical Offices in particular. This will include tools that may be used in creating

spatially-anonymized public use datasets and that will yield themselves to customization to accommodate different degrees of sensitivity and specific uses.

Draft guidelines will be prepared by the end of 2019.

4. Membership of the Task Force

The Task Force will be led by the World Bank. Other members will include …