CUSP Data Product Export Policy

Table of Contents

Background

The CUSP Data Facility's Secure Computing Environment (SCE) provides researchers with a collaborative workspace in which they can analyze non-public datasets for urban policy research projects. The facility provides to researchers analytical tools and computing power, including access to a Hadoop cluster. All CUSP datasets that are non-public can only be accessed within the SCE.

Researchers can request access to non-public datasets using the Research Proposal Request form. Researchers will sign the non-public dataset use agreement upon submitting this request.

Why Disclosure Review?

Data products that are generated through the analysis of non-public datasets can only be exported from the SCE after a disclosure review by the CDF team for compliance with data privacy guidelines, **to ensure that individuals and companies cannot be reidentified from the aggregated datasets**. This disclosure review policy is in place because micro data (data at the level of an individual person or business) that has been aggregated are not automatically considered de-identified. The CDF disclosure review process is modeled after the process implemented by the Research Data Centers at the German Federal Employment Agency[1] and is informed by the framework laid out in Lane et al, 2008[2].

Note on Nontabular Data Products

Traditional methods of statistical disclosure review have been developed around tabular data outputs and associated charts and figures. CUSP does encourage researchers to produce tabular datasets and charts and figures for export review. In the case where the CDF researcher is producing nontabular data products for publication, the results will be reviewed on a case-by-case basis.

Preparing Data Products for CDF Export Review

Code

Researchers in the CDF should be recording all code in a project-level CUSP GIT repository. Include in the data export request a link to portion of the GIT repo that was used to create the data product.

---

[1] Hochfellner et al, Privacy in Confidential Administrative Micro Data: Implementing Statistical Disclosure Control in a Secure Computing Environment, J Empirical Res on Human Res Ethics, 2014

[2] Lane et al, Data Access in a Cyber World: Making Use of Cyberinfrastructure, Transactions on Data Privacy, 2008.

Code should be appropriately commented for review by CDF staff. CDF staff should be able to run the code on the specified dataset, within the secure data enclave.

The export review will be most efficient if researchers describe in their code how person or business data were aggregated in generating the derived dataset.

Tabular Data

All tabular data for export review should be saved in the export folder and labeled with the ProjectID and date of export request. After files are saved, fill out a Data Export Request form to initiate review.

For each cell in the table (or the source data underlying any graph representation), provide the counts of each entity (n), the percentage (p) of any value accounted for by the top 4 (k) entities. If the aggregate is a ratio, the number of valid cases has to be computed for each subgroup of the aggregate (e.g. number of men in state X and number of women in state X in addition to the ratio of women in state X).

- All result files generated after an aggregation must contain the following information at the beginning:
  - at which level the aggregation took place,
  - during which step in the program the aggregation took place (also specify in code)
  - the minimum number of individuals and establishments per data line (e.g.: cells containing <20 establishments were deleted), and
  - name of the variable that contains the observation count per data line.

File Naming Convention

Files names should be in the following format: *projectname_date_description_version*

Example: *citymanagement_11022015_landusestats_1b*

Charts, Graphs, and Maps

The data underlying all charts, graphs and maps should be saved in the export folder and labeled with the ProjectID and date ofexport request. After files are saved, fill out a Data Export Request form to initiate review.

For each underlying data point, provide the counts (n), the percentage (p) of any value accounted for by the top 4 (k) entities.

File Naming Convention

Files names should be in the following format:

For charts, graphs, maps: *projectname_date_description_version*

Example: *citymanagement_11022015_landusestatsgraph_1b*

For underlying data: *projectname_date_description_data_version*

Example: *citymanagement_11022015_landusestatsgraph_data_1b*

Export Review Turnaround Time

The CDF disclosure review team will generate an approval or a request for more information within 2 business days after a request and all code and data products are submitted. The disclosure review process will be expedited by good data hygiene practices – clean and well-commented code and the inclusion of metadata in tabular datasets, as described below.

External Resources on Best Practices for Reducing Disclosure Risks

US Census Bureau Statistical Disclosure Control Guidelines

The "checklist" provides a set of questions that help to guide new researchers through the data preparation for disclosure process.

American Statistical Association Methods for Reducing Disclosure Risks When Sharing Data

Data Without Boundaries - Review of Statistical Disclosure Control Software Tools