Quality Control Overview for the ONEdata Spatially Derived Variables

Author: Brian Frizzelle, Manager of the Spatial Analysis Unit, Carolina Population Center

The Obesity and Neighborhood Environment study generated spatially-derived datasets for Add Health waves I and III. Such a massive undertaking required significant attention to issues of quality control (QC), and the spatial analysts on the project devised a basic standard protocol that was used throughout.

Following the creation of each derived dataset, the output data file was passed on to the QC team, which consisted of one or more analysts, depending on the size and complexity of the dataset itself. Also delivered were the input source data, dataset documentation, the programmatic code used to generate the variables, and the log files created by the code. In some cases, to facilitate the verification of critical processing steps, the analyst who created the derived dataset also provided the QC team with intermediate datasets created at key checkpoints in the development sequence.

The following sections describe the elements of the quality control that were performed. They have been roughly organized in the order in which they were performed, although the QC teams had the latitude to perform the checks in whatever order they deemed best and to perform additional checks whenever appropriate. Please note that even though this document may not explicitly state that these checks resulted in corrections to source data, code, or other errors, be assured that all identified errors were corrected.

Code and Log Files

All datasets were created via the use of programmatic code, which avoided the risk of errors associated with manual processing and provided a thorough record for verification. Programs were created in a variety of languages – primarily Python, VBA, and SAS - and therefore one of the first items to be validated by the QC analyst was the log files generated by the code. The QC team checked the logs twice, once prior to looking at the data and once following the data checks. The first log check was performed to understand the processing and to identify any error codes that may have been generated. The second check was performed to verify the code itself. Input and output data paths and file names were validated, code syntax was inspected for errors, and formulas were reviewed. If any issues were discovered or the log file was incomplete or unclear, the QC analyst reviewed the originating script thoroughly and flagged items of concern for consideration by the programmer. For derived datasets containing missing values, the QC team verified that the correct replacement codes were used.

Formulas

For variables created through the use of mathematical or statistical formulas, the values in the output file were verified by replicating the formulas within other software packages, such as Microsoft Excel. Input dataset values for a sampling of respondents were plugged into the formula within Excel, and the resultant values were compared against the programmatic output to ensure that they matched.

Summary Statistical Checks of Output Distribution

One of the first checks of the output values was a review of summary statistics through procedures such as PROC MEANS in SAS. These checks were run on datasets that had not yet had missing values changed to replacement codes, and were done to verify that the range of output values fit range of input values,

and that the distribution of values was legitimate. The insertion of replacement codes was then performed and the procedures were rerun to verify the changes.

Manual Checks of Output Values

This part of the QC process had the most variability, given that the output values of datasets were generated in different ways. In some datasets, the values in the variables are the same as some element of the input data, and the values were copied over from the source data based on some geographic linkage. In other datasets, the values in the variables were calculated using some formula or spatial methodology. In both cases, a random sample of respondent records was checked using ArcMap GIS software. The respondent locations and the source data were loaded into ArcMap and the QC team replicated the procedures in the code that produced the output. The resultant QC values were used to validate the values in the output dataset. In situations where the QC team felt that certain geographic areas or sets of values necessitated further validation (e.g. minimum or maximum values in a variable), additional checks were performed on a targeted set of records meeting the criteria of concern.

Missing Values and Zeroes

The QC team checked all variables for missing values and zeroes. Any missing value was investigated to determine if it was legitimate (e.g. no source data for the respondent) or a result of coding errors or mistakes in the source data. These checks were always performed prior to the execution of the code that replaced missing values with replacement codes. A similar check was performed on any record with a zero for some or all variables, again to determine if the zero was legitimate or erroneous.

Anomalies in Source Data

This was often performed as part of the data development phase prior to the writing and execution of the processing code. However, it was always possible to discover previously missed anomalies in the source data that only showed up during the QC stage. In those situations, a thorough review of the source data was performed, and any anomalies in the source data were corrected if possible or noted in the documentation or through flag variables.

Counts, Variable Names, and Labels

Following the checks of the output values, the QC team used the SAS procedure PROC CONTENTS to verify that the output file contained the correct number of records and variables, and that the variable names and labels were correct, and followed the project's naming convention throughout the entire dataset.

Geographic Identifiers

The QC team performed a preliminary deductive disclosure check to make sure that the output dataset contained no overt geographic identifiers in the variables. An overt geographic identifier would be the inclusion of a value in any variable that is unique to a particular place and which could be used to identify a respondent's area of residence.

Documentation

The final step of QC was to check the documentation. The QC team verified that the dataset was named properly, the variable count was correct, the procedural description was accurate, the variable names in the documentation matched those in the output file and adhered to naming conventions, the formulas listed in the documentation were correct, the text was legible, and there were no typographic errors.