PhenX Toolkit

Data Collection Worksheet

Please Note: The Data Collection Worksheet (DCW) is a tool to aid integration of a PhenX protocol into a study. The PhenX DCW is not designed to be a data collection instrument. Investigators will need to decide the best way to collect data for the PhenX protocol in their study. Variables captured in the DCW, along with variable names and unique PhenX variable identifiers, are included in the PhenX Data Dictionary (DD) files.

The Dissimilarity Index is based on U.S. Census Bureau data. This protocol describes how to make calculations using the decennial census Summary File 1, referred to as the SF1. The SF1 is the short form of the U.S. Census collected from everyone; it is also referred to as 100% data. These calculations can be made using the 1990, 2000, and 2010 decennial Census.

The 2000 and 2010 SF1 data can be downloaded at https://census.gov.

2010 SF1 MS Access data:
https://www.census.gov/data/datasets/2010/dec/summary-file-1.html

2000 SF1 MS Access data:
https://www.census.gov/data/datasets/2000/dec/summary-file-1.html

Web version for 2010 data:
https://data.census.gov/cedsci/table?q=p5&tid=DECENNIALSF12010.P5

A repository of resources for decennial Census data can be found at U.S. Census https://www.census.gov/programs-surveys/decennial-census/data.html. Users not familiar with Census data should consult the technical materials. The technical documentation for the 2010 Census is available at https://www2.census.gov/programs-surveys/decennial/2010/technical-documentation/complete-tech-docs/summary-file/sf1.pdf. Technical documentation for the 1990 and 2000 Census SF1 data is provided in the references section.

This protocol focuses on the race and ethnicity data in the 2010 SF1 file. The key race/ethnicity data in the 2010 Census are found in "Table P5: Hispanic or Latino Origin by Race." This table is preferred over other possible race and race/ethnic tables available, as it provides data on the main race/ethnic groups in the United States and explicitly incorporates data on Hispanic or Latino populations, otherwise not available in the race-only tables.

https://www.census.gov/programs-surveys/decennial-

census/about/rdo/summary-files.html#P1

Variable Code	Variable Name
P0050001	Total:
P0050002	Not Hispanic or Latino:
P0050003	White alone
P0050004	Black or African American alone
P0050005	American Indian and Alaska Native alone
P0050006	Asian alone
P0050007	Native Hawaiian and Other Pacific Islander alone
P0050008	Some other race alone
P0050009	Two or more races:
P0050010	Hispanic or Latino:
P0050011	White alone
P0050012	Black or African American alone
P0050013	American Indian and Alaska Native alone
P0050014	Asian alone

P0050015	Native Hawaiian and Other Pacific Islander alone
P0050016	Some other race alone
P0050017	Two or more races:

The race/ethnic data are available for all small census geographies-such as census block, census block group, and census tract-and can be easily extracted for almost any geographic level.

Researchers can use the data in this table to easily calculate basic variables (e.g., the percentage of any race and/or ethnicity group) or to combine groups (e.g., all minorities).

The Dissimilarity Index provides data on larger areas (e.g., metropolitan statistical areas) using smaller level data.

The most common conceptualization of residential segregation is based on the dimension of evenness. Evenness refers to the differential distribution of the subject population across neighborhoods in a large area (e.g., metropolitan area). It ranges from 0 (complete integration) to 1 (complete segregation) and indicates the percentage of a group's population that would have to change residence for each neighborhood to have the same percentage of that group as the metropolitan area overall. It is computed as:

$$D = .5 * \sum_{i=1}^{n} \left| x_i / X - y_i / Y \right|$$

where

n is the number of tracts in the larger area (e.g., a metropolitan area),

x_i is the population size of the minority group of interest in tract *i*,

X is the population of the minority group in the larger area (e.g., metropolitan area) as a whole,

y_i is the population of the reference group (usually non-Hispanic Whites) in tract *i*, and,

Y is the population of the reference group in the larger area (e.g., metropolitan) area as a whole.

The calculation requires the computation of the totals for each group across all subareas within a larger region (e.g., all census tracts within a county), the proportion of each group within each subarea, the absolute difference between the proportions, and the sum of the absolute differences. The latter number is multiplied by 0.5 to generate a result between 0.0 and 1.0. A value of 0.0 would indicate there were the same proportions of majority and minority group populations in each subarea, as in the larger regions' population. If all subareas within the region contain members of just one group (i.e., there is no co-residence) then *D* equals 1.0, indicating complete segregation.

Extending the Dissimilarity Index: The Multigroup Analog

While much early research on segregation looked at two groups (e.g., Black and White, or majority and minority), today's society is multiethnic. Two-group measures are useful but limited for describing complex patterns of segregation. The choice to use a two-group or multigroup *D* depends on the specific question of interest. In a region where the population is composed of three groups (e.g., White non-Hispanic, Black non-Hispanic, and Hispanic), we may be interested in

a) segregation between two specific groups (e.g., How segregated are White from Black residents?); or

b) segregation among all three groups (e.g., How segregated are White non-Hispanic, Black non-Hispanic, and Hispanic residents from each other?).

The two-group measure can still be used by comparing all possible pairs of population groups (Morrill, 1995), but these are not comprehensive, and multiple groups are not treated simultaneously. To address segregation among multiple groups requires a multigroup analog to *D* (Morgan et al., 1975; Sakoda, 1981). The multigroup analog describes the extent to which two or more population groups are similarly distributed among subareas. The formula for multigroup dissimilarity (from Reardon & Firebaugh, 2002) is:

$$D = \sum_{m=1}^{M} \sum_{j=1}^{J} \frac{t_j}{2TI} \left| \pi_{jm} - \pi_m \right|$$

where

T is total population,

M is the number of groups *m*,

J is the number of subareas or units *j*,

t_j is number of individuals in subarea *j*,

π_m is the proportion in group *m*,

π_{jm} is the proportion in group m, of those in unit *j*, and

I is the Simpson's Interaction Index, given by

The interpretation of multigroup *D* (sometimes labeled as *D*(*m*)) is the same as *D* (see Wong, 1993).

In the Stata statistical software package, the command *seg* (installed by typing "ssc install seg" from within Stata) will compute *D* (Reardon, 2002).

Researchers have extended segregation measures by incorporating the spatial dimension (see White, 1983; Wong, 1993; Reardon & O'Sullivan, 2004). There are spatially modified versions of the *D* index (see Wong, 1993).