**MESA Neighborhood Study (Ancillary Study AS023)**

**Racial Segregation Gi\* statistics**

Please acknowledge the following grant in manuscripts and abstracts: R01 HL071759 (Diez Roux)

Please include an acknowledgement:

We thank Paulina Kaiser for her contributions to creating the racial segregation measures.

# Overview:

Neighborhood racial segregation was calculated as the Getis-Ord Gi\* statistic. These measures are calculated for segregation for % non-Hispanic Black, % Hispanic, % non-Hispanic Asian, and % non-Hispanic White within census tracts. Tracts to include in each site were chosen to match the tracts that were used for calculating the built environment variables, which include the counties surrounding the MESA site and are intended to capture MESA participants at baseline and the majority of MESA participants who move within the site over time (see details section for further information on the counties selected).

The Gi\* statistics were calculated for various spatial definitions of neighbors: Rook, 1-mile, 2-mile, and 1-mile inverse weights.

Time-varying measures were created using data from the US Census 2000, American Community Survey 2005-2009, and American Community Survey 2007-2011. The measures were applied to the MESA Exams based on calendar years and linked to the census tract of residence.

Years 2000-2004 = Census 2000 data (Exams 1 and 2, part of Exam 3)

Years 2005-2007 = ACS 2005-2009 (Part of Exam 3, Exam 4)

Years 2008-2012 = ACS 2007-2011 (Exam 5)

# Recommendations:

For most analyses, interest has been in using racial segregation based on the participant’s own race group segregation. These measures have been created and are included in the dataset. It is recommended to use the measures based on either the rook or 1 mile definitions. The advantage of the rook definition is that the measures for all tracts will be based on at least one other neighboring tract. This may also standardize for differences in the relative sizes of tracts across the MESA study sites. The 1 mile definition ensures that the measure is always based on the same distance regardless of the size of the tract. The disadvantage is that if the tracts are large, then the Gi statistic is only based on the percent race within that tract, standardized to the total race distribution of the metro area.

Table 1: Recommended variables

| G\_rk | G statistic for own race, using rook definition of neighbors |
| --- | --- |
| G\_1mi | G statistic for own race, using 1-mile definition of neighbors |

Since the Gi statistic is a z-score, meaningful categories can be created based on statistical significance:

<-1.96 = Statistically significant low clustering

-1.96-1.96 = No clustering

>1.96 = Statistically significant high clustering

In many cases, for the low clustering (<-1.96), this yields very low sample size. A low cut-point of >0 has been used in manuscripts.

# Example Methodology Section for Manuscripts:

Neighborhood-level racial/ethnic residential segregation was measured separately for blacks, whites, and Hispanics by using the local *Gi*\* statistic(1), based on the geocoded addresses of MESA participants linked to US Census data. The *Gi*\* statistic returns a *Z* score for each neighborhood (census tract), indicating the extent to which the racial/ethnic composition in the focal tract and neighboring tracts deviates from the mean racial composition of some larger areal unit surrounding the tract (in our case, the set of counties represented in each MESA site). Higher positive *Gi*\* *Z* scores indicate higher racial/ethnic segregation or clustering (overrepresentation), scores near 0 indicate racial integration, and lower negative scores suggest lower racial/ethnic representation (underrepresentation), in comparison with the racial composition of the larger areal unit.

Most studies of neighborhood-level racial/ethnic residential segregation use racial/ethnic composition or the proportion of a race/ethnic group in a neighborhood, as a proxy for segregation(2). However, this measure is limited in that it does not incorporate any information on the racial composition of the larger area in which the neighborhood is embedded or on the distribution of groups in space(3). The *Gi*\* statistic, in contrast, better reflects both the contextual and spatial aspects of segregation. A given neighborhood will have a higher *Gi*\* statistic the larger the difference between its racial composition and the composition of the larger areal unit. In addition, a neighborhood surrounded by similarly segregated areas will have a higher *Gi*\* statistic than those surrounded by less segregated areas.

1 Getis A, Ord JK. The analysis of spatial association by use of distance statistics. *Geogr Anal*. 1992;24:189–206.

2 White K, Borrell LN. Racial/ethnic residential segregation: framing the context of health risk and health disparities. *Health Place*. 2011;17:438–448.

3 Kramer MR, Hogue CR. Is segregation bad for your health? *Epidemiol Rev*. 2009;31:178–194.

# Published MESA Manuscripts Using the Data:

Jones, M.R., et al., *Race/Ethnicity, Residential Segregation, and Exposure to Ambient Air Pollution: The Multi-Ethnic Study of Atherosclerosis (MESA).* Am J Public Health, 2014: p. e1-e8

Kershaw, K.N., et al., *Neighborhood-level racial/ethnic residential segregation and incident cardiovascular disease: the multi-ethnic study of atherosclerosis.* Circulation, 2015. **131**(2): p. 141-8.

# Dataset Description:

Participant inclusion: MESA participants who agreed to participate in the MESA Neighborhood study (N=6191)

Data set-up: Panel (stacked) dataset with 1 row per participant per exam; an exam indicator is included

Notes: Any addresses that were unable to geocode OR those addresses that are outside the boundaries for the study area will have missing data for all Gi\* variables. Indicators are included to distinguish the reason for missingness.

**Data Set Name: MESANBH\_RACIALSEG07072016**

| **Variable Order** | **Variable name** | **Description** | **Coding** |
| --- | --- | --- | --- |
| 1 | idno | MESA ID number |  |
| 2 | accuracy | Geocoding accuracy indicator. It is recommended that that only those with accuracy to at least zip code + 4 centroid are used in analyses, at least for sensitivity analysis. | 1 = Street level2 = Zip+4 centroid level3 = Zip+2 centroid level4 = Zip code centroid5 = Unable to geocode |
| 3 | cenid | Fake census tract id for clustering analysis. This should be used when using census tract level data in models. This is from 2000 census since census 2000 and ACS use this geography. |  |
| 4 | EXAM | MESA Exam number |  |
| 5 | race\_hisp | % Hispanic in census tract |  |
| 6 | race\_whiteNH | % non-Hispanic White in census tract |  |
| 7 | race\_blackNH | % non-Hispanic Black in census tract |  |
| 8 | race\_asianNH | % non-Hispanic Asian in census tract |  |
| 9 | edge | Indicator if the census tract is on the edge of the county definitions. May want to exclude those near the edge at least for sensitivity analysis. | 0 = Not on edge1 = On edge |
| 10 | G\_asi\_rk | G statistic for % Asian, using rook definition of neighbors |  |
| 11 | p\_asi\_rk | P value for G statistic for % Asian, using rook definition of neighbors |  |
| 12 | G\_bla\_rk | G statistic for % Black, using rook definition of neighbors |  |
| 13 | p\_bla\_rk | P value for G statistic for % Black, using rook definition of neighbors |  |
| 14 | G\_his\_rk | G statistic for % Hispanic, using rook definition of neighbors |  |
| 15 | p\_his\_rk | P value for G statistic for % Hispanic, using rook definition of neighbors |  |
| 16 | G\_whi\_rk | G statistic for % White, using rook definition of neighbors |  |
| 17 | p\_whi\_rk | P value for G statistic for % White, using rook definition of neighbors |  |
| 18 | G\_asi\_1mi | G statistic for % Asian, using 1-mile definition of neighbors |  |
| 19 | p\_asi\_1mi | P value for G statistic for % Asian, using 1-mile definition of neighbors |  |
| 20 | G\_bla\_1mi | G statistic for % Black, using 1-mile definition of neighbors |  |
| 21 | p\_bla\_1mi | P value for G statistic for % Black, using 1-mile definition of neighbors |  |
| 22 | G\_his\_1mi | G statistic for % Hispanic, using 1-mile definition of neighbors |  |
| 23 | p\_his\_1mi | P value for G statistic for % Hispanic, using 1-mile definition of neighbors |  |
| 24 | G\_whi\_1mi | G statistic for % White, using 1-mile definition of neighbors |  |
| 25 | p\_whi\_1mi | P value for G statistic for % White, using 1-mile definition of neighbors |  |
| 26 | G\_asi\_2mi | G statistic for % Asian, using 2-mile definition of neighbors |  |
| 27 | p\_asi\_2mi | P value for G statistic for % Asian, using 2-mile definition of neighbors |  |
| 28 | G\_bla\_2mi | G statistic for % Black, using 2-mile definition of neighbors |  |
| 29 | p\_bla\_2mi | P value for G statistic for % Black, using 2-mile definition of neighbors |  |
| 30 | G\_his\_2mi | G statistic for % Hispanic, using 2-mile definition of neighbors |  |
| 31 | p\_his\_2mi | P value for G statistic for % Hispanic, using 2-mile definition of neighbors |  |
| 32 | G\_whi\_2mi | G statistic for % White, using 2-mile definition of neighbors |  |
| 33 | p\_whi\_2mi | P value for G statistic for % White, using 2-mile definition of neighbors |  |
| 34 | G\_asi\_1in | G statistic for % Asian, using 1-mile inverse weighted definition of neighbors |  |
| 35 | p\_asi\_1in | P value for G statistic for % Asian, using 1-mile inverse weighted definition of neighbors |  |
| 36 | G\_bla\_1in | G statistic for % Black, using 1-mile inverse weighted definition of neighbors |  |
| 37 | p\_bla\_1in | P value for G statistic for % Black, using 1-mile inverse weighted definition of neighbors |  |
| 38 | G\_his\_1in | G statistic for % Hispanic, using 1-mile inverse weighted definition of neighbors |  |
| 39 | p\_his\_1in | P value for G statistic for % Hispanic, using 1-mile inverse weighted definition of neighbors |  |
| 40 | G\_whi\_1in | G statistic for % White, using 1-mile inverse weighted definition of neighbors |  |
| 41 | p\_whi\_1in | P value for G statistic for % White, using 1-mile inverse weighted definition of neighbors |  |
| 42 | n\_nbr\_rk | Number of neighbors identified by the rook definition of neighbors. If this is 0, then the corresponding Gi\* statistics are missing. |  |
| 43 | n\_nbr\_1mi | Number of neighbors identified by the 1-mile definition of neighbors. If this is 0, then the corresponding Gi\* statistics are missing. |  |
| 44 | n\_nbr\_2mi | Number of neighbors identified by the 2-mile definition of neighbors. If this is 0, then the corresponding Gi\* statistics are missing. |  |
| 45 | n\_nbr\_1mi\_inv | Number of neighbors identified by the 1-mile inverse weighted definition of neighbors. (Identical to the number of neighbors identified by the 1-mile fixed distance definition, though the weight of the neighbors will be different.) If this is 0, then the corresponding Gi\* statistics are missing. |  |
| 46 | RACE | Race/ethnicity | 1=White, Caucasian2=Chinese American3=Black, African-American4=Hispanic |
| 47 | site\_gstat | MESA site corresponding to the Census Tract | 3=NC4=NY5=MD6=MN7=IL8=CA |
| 48 | G\_rk | G statistic for own race, using rook definition of neighbors |  |
| 49 | p\_rk | P value for G statistic for own race, using rook definition of neighbors |  |
| 50 | G\_1mi | G statistic for own race, using 1-mile definition of neighbors |  |
| 51 | p\_1mi | P value for G statistic for own race, using 1-mile definition of neighbors |  |
| 52 | G\_1in | G statistic for own race, using 1-mile inverse weighted definition of neighbors |  |
| 53 | p\_1in | P value for G statistic for own race, using 1-mile inverse weighted definition of neighbors |  |
| 54 | G\_2mi | G statistic for own race, using 2-mile definition of neighbors |  |
| 55 | p\_2mi | P value for G statistic for own race, using 2-mile definition of neighbors |  |

# Details:

## Theory: Getis-Ord Gi\* Statistic

Arthur Getis and Keith Ord developed G statistics[[1]](#footnote-1) as a way to quantify spatial clustering that could distinguish between clusters of higher-than-expected values and clusters of lower-than-expected values. G statistics determine whether the local average is higher or lower than the global average (in contrast to Moran’s I, which determine whether an index unit is similar to its neighbors). Figure 1 shows the formula for the Getis-Ord Gi\* statistic, which is the statistic that has been calculated for use in MESA. Gi\* is calculated as the sum of X values for a unit and its neighbors is compared to the expected value if there is no clustering (e.g. the global average of X), and then standardized. The Gi\* statistic is a z-score, meaning that Gi\* values greater than 1.96 or less than -1.96 are considered statistically significant at $α$=0.05. A statistically significant positive Gi\* value (Gi\*>1.96) represents a ‘hot spot,’ indicating that there is clustering of high values around unit *i*. A statistically significant negative Gi\* value (Gi\*<-1.96) is a ‘cold spot,’ indicating that clustering of low values is present around *i*. (Note that Gi\* is a local statistic, while the similar ‘general’ G statistic is a global measure of clustering.[[2]](#footnote-2))



Figure . (From ESRI, “How Hot Spot Analysis (Getis-Ord Gi\*) works”) <http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//005p00000011000000>

The numerator of the formula is the weighted sum of X for the neighbors of unit *i* minus the global average of X. So the choice of area for which Gi\* statistics are being calculated affects the results; for example, being a statistically significant ‘hot spot’ of percent black residents in relation to the five-county area around Los Angeles is different than being a statistically significant ‘hot spot’ of percent black residents in relation to the urban Los Angeles area.

## Theory: Spatial Weights Matrices

Spatial analysis explores the impact of neighboring locations on an association of interest. Spatial weights matrices (SWMs) specify which tracts are ‘neighbors’ to any index tract. There are many ways to define ‘neighbors’; we used four distinct conceptualizations of the spatial relationships between tracts:



*Credit: Paul Voss & Katherine Curtis, ICPSR “Spatial Regression” 2012*

Figure 2. Simple illustration of queen vs. rook definition of neighbors.

1. First order rook (contiguity edges only)
	1. By this definition, a tract that shares a border with the index tract is considered to be a neighbor. ArcGIS calls this definition “contiguity edges only” (in contrast to a “contiguity edges corners,” also called a ‘queen’ definition, in which any tract that touches the index tract – even at just a point – is considered a neighbor; see Figure 2 for an illustration).
	2. Because of the way that we selected tracts to include in the spatial weights matrices, every tract has at least one neighbor using the first order rook definition.
2. 1 mile fixed distance
	1. By this definition, a tract whose centroid is within a 1-mile (1609.3 meters) radius around the centroid of the index tract is considered to be a neighbor.
	2. Where census tracts are small, each tract will tend to have more neighbors; if a tract is large enough so that its centroid is more than one mile from any edge, it will have no neighbors. Some MESA tracts have no neighbors by this definition (see Table 3 for more information).
3. 2 mile fixed distance
	1. By this definition, a neighbor is a tract whose centroid is within a 2-mile (3218.7 meters) radius around the centroid of the index tract.
	2. Similar to the 1 mile fixed distance definition, it’s possible for tracts to have no neighbors; the range in the number of neighbors is largest with this definition (see Table 3).
4. 1 mile inverse distance
	1. Inverse distance is similar to fixed distance, except that the actual weight assigned to each neighboring tract depends on the distance from the index tract. Specifically, tracts whose centroid is closer to the centroid of the index tract have a higher weight than tracts whose centroid is further away. We used an exponent of one, e.g. a linear relationship between weight and distance.

## Selecting Census Tracts

Tracts to include in each site were chosen to match the tracts that were used for calculating the built environment variables, which include the counties surrounding the MESA site and are intended to capture MESA participants at baseline and the majority of MESA participants who move within the site over time. The counties included are listed in Table 2. All Census tract boundaries were defined by the 2000 Census. These files have the race variables (% Hispanic, % non-Hispanic white, % non-Hispanic black, and % non-Hispanic Asian) linked to the census tract IDs*.*

Table 2. Counties included in MESA spatial weights matrices from each site.

|  |  |  |
| --- | --- | --- |
| **Site** | **County Names** | **UTM Projection** |
| Los Angeles, CA | Venture, Los Angeles, Orange, Riverside, and San Bernardino | Zone 11 |
| St Paul, MN | Anoka, Hennepin, Ramsey, Washington, Carver, Scott, and Dakota | Zone 15 |
| Chicago, IL | Kane, DuPage, Cook, and Will | Zone 16 |
| Winston-Salem, NC | Forsyth | Zone 17 |
| Baltimore, MD | Baltimore City and Baltimore County | Zone 18 |
| New York, NY | Queens, Kings, New York, and Bronx | Zone 18 |

Spatial weights matrices were calculated for each site separately. UTM projections were used to ensure that Euclidean distances are measured accurately. All spatial weights matrices were calculated with row-standardized weights.

The Spatial Weight Matrices (SWMs) created for use with MESA data are limited to include only census tracts in the counties surrounding the MESA site. Thus, for tracts on the ‘edge’ of the input counties, some neighboring tracts (e.g. the tracts outside the defined input area) are not considered in the calculation of numbers of neighbors or for spatially weighted statistics such as Gi\* statistics. This means that the number of neighbors and the Gi\* statistics provided are not fully reflective of reality, and may introduce bias in analyses. These tracts have been identified in order to allow analysts to explore the impact that these boundary tracts may have on analyses. Overall, 2.43% of all input census tracts are edge tracts. The highest percentage is in North Carolina (29.33% of tracts) and the lowest in California (0.86%).

Figure 3. North Carolina input tracts, with ‘edge’ tracts in darker blue.

## Calculating Gi\* Statistics: Hot-Spot Analysis Tool

Gi\* statistics were calculated based on racial composition measures for Census 2000, American Community Survey 2005-2009 (ACS0509), and American Community Survey 2007-2011 (ACS0711). All of these measures are calculated based on Census2000 boundaries. Census2000 and ACS0509 use Census2000 boundaries in sampling. ACS0711 uses Census2010 boundaries in sampling so a crosswalk[[3]](#footnote-3) was used to convert these measures to Census2000 boundaries.

The Hot Spot Analysis tool in ArcGIS 10.1 was used to calculate the Gi\* statistics. For each site, an input shapefile was created with the selected census tracts (see documentation on creating SWMs for more details on the selection of tracts for each site) joined with a data table of racial composition from Census 2000, ACS0509, or ACS0711. Four sets of Gi\* statistics were calculated for each Census Tract per year, one each for the four methods of “Conceptualization of Spatial Relationships” (e.g. neighbor definitions, quantified by spatial weights matrices [SWM files]) calculated for MESA (rook, 1-mile, 2-mile, and 1-mile inverse). For each conceptualization of spatial relationships, Gi\* statistics were calculated for % Black, % Asian, % Hispanic, and % White. Thus, 16 Gi\* statistics have been calculated for each Census Tract per year. An indicator for census tracts that fall on the edge of the counties included in the SWM was merged with the final datasets.

The default conceptualization of spatial relationships for Hot Spot Analysis in ArcGIS is fixed distance (e.g. the 1-mile or 2-mile SWMs calculated for MESA). A contiguity (rook) conceptualization is also appropriate. Using an inverse distance relationship can be problematic with distances less than one between units; this is unlikely with the MESA data (the shapefiles are projected to UTM zones with units in feet), but offers no clear benefit over another SWM.[[4]](#footnote-4)

Tracts with no neighbors have a Gi\* value, but this value is just the index value divided by the global average – there is no spatial adjustment or influence from neighboring tracts. Therefore it may be appropriate to remove tracts with no neighbors from the analysis. Getis & Ord recommended that each unit should have at least 8 neighbors to ensure the stability of the G statistics; ArcGIS adopts this recommendation as well.[[5]](#footnote-5) We did not limit the calculation of the Gi\* statistics by the number of neighbors; however, the number of neighbors (using each SWM) is available in the dataset if you would like to constrain the tracts used in your analysis.

## Which SWM Should I Use?

It depends! A good place to start is with the rook (contiguity edges only) definition, because all tracts will have at least one neighbor, and because the number of neighbors is most consistent across sites (compared to the other SWMs). However, where census tracts are small, the area covered by an index tract and its first order rook neighbors will be relatively small; where census tracts are large, the area covered by an index tract and its first order rook neighbors will be much larger.

With the threshold distance SWMs (either fixed or inverse), the variability between MESA sites – a one mile radius in New York may be very meaningful for social interactions or food availability, while one mile radius in North Carolina may not be very meaningful.

For additional discussion on using spatial weights matrices, see Getis and Aldstadt.[[6]](#footnote-6)

Spatial variograms (or semi-variograms) are a way to plot the similarity between observations by the distance between those observations; it can help to define the range of the spatial influence in the data. It may also be helpful to do sensitivity analyses using different SWMs and comparing the results.

1. Getis, A., & Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. *Geographical analysis*, *24*(3), 189-206. [↑](#footnote-ref-1)
2. <http://www.sce.lsu.edu/cego/documents/reviews/geospatial/spatial_autocorrelation.pdf> [↑](#footnote-ref-2)
3. John R. Logan, Zengwang Xu, and Brian Stults. 2012. “Interpolating US Decennial Census Tract Data from as Early as 1970 to 2010: A Longitudinal Tract Database” Professional Geographer, forthcoming [↑](#footnote-ref-3)
4. See <http://www.sce.lsu.edu/cego/documents/reviews/geospatial/spatial_autocorrelation.pdf> and the ArcGIS Help 10.1 page about Hot Spot Analysis. [↑](#footnote-ref-4)
5. <http://resources.esri.com/help/9.3/arcgisengine/java/gp_toolref/spatial_statistics_tools/hot_spot_analysis_getis_ord_gi_star_spatial_statistics_.htm> [↑](#footnote-ref-5)
6. Getis, Arthur, and Jared Aldstadt. "Constructing the spatial weights matrix using a local statistic." *Perspectives on Spatial Data Analysis*. Springer Berlin Heidelberg, 2010. 147-163. [↑](#footnote-ref-6)