Westat®

*Improving lives
through research™*

Family Life, Activity, Sun, Health, and Eating (FLASHE)

GeoFLASHE Methods Report

March 2018

Prepared for:
National Cancer Institute
9609 Medical Center Drive
Bethesda, MD 20892-9760

Prepared by:
Westat
An Employee-Owned Research Corporation®
1600 Research Boulevard
Rockville, Maryland 20850-3129

Table of Contents

GeoFLASHE Methods Report

The Family Life, Activity, Sun, Health, and Eating (FLASHE) study collected survey data on psychosocial, generational (parent-adolescent), and environmental correlates of cancer-preventive behaviors from adolescents and their parents. The FLASHE survey included questions that asked about the locations of their homes and the schools that the adolescent attends. The answers to these questions are not available in the public use survey datasets but were used as a basis for developing GeoFLASHE, a dataset of neighborhood and contextual variables that can be linked with FLASHE variables to provide an additional analytic dimension based on home and school locations. For more information about FLASHE, please visit the study website at: https://cancercontrol.cancer.gov/brp/hbrb/flashe.html.

This guide provides detailed information about how these variables were created and how they might be used to augment FLASHE data analysis. Section 1 contains information about the geographic data of the home and school locations. Section 2 describes the measures that characterize neighborhood buffers and surrounding areas. Section 3 provides reference information for each of the GeoFLASHE measures including information about how to link the GeoFLASHE measures with FLASHE survey databases.

The GeoFLASHE codebook ("FLASHE GeoData Codebook.xlsx") contains additional detailed information about each GeoFLASHE variable including the variable name and label, description, default format, and valid values. For each variable derived from Census data tables, the codebook identifies the table source and provides the formula used to calculate the value.

A detailed listing of the GeoFLASHE variables are available in Section 3 of this document.

1. Geographic Information

The FLASHE demographic survey asked parents two sets of open-ended questions about the location of their home and their adolescent's school:

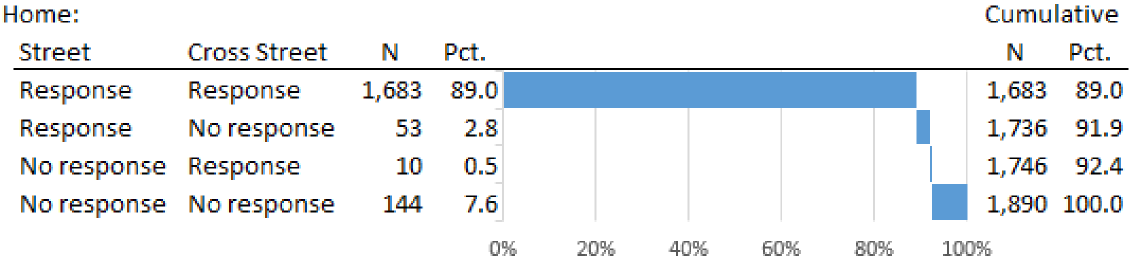
> "Can you tell me just the name of the street/road you live on?
> And what is the name of the nearest cross street/road?"
>
> "Can you tell me just the name of the street/road **teen's** school is on?
> And what is the name of the nearest cross street/road?"

In this section, we provide information about the completeness of these responses and how they were used to develop the geographic information that is included in the GeoFLASHE dataset. See Table 9 on page 22 for a detailed listing of the geographic variables.

1.1 Response Completeness

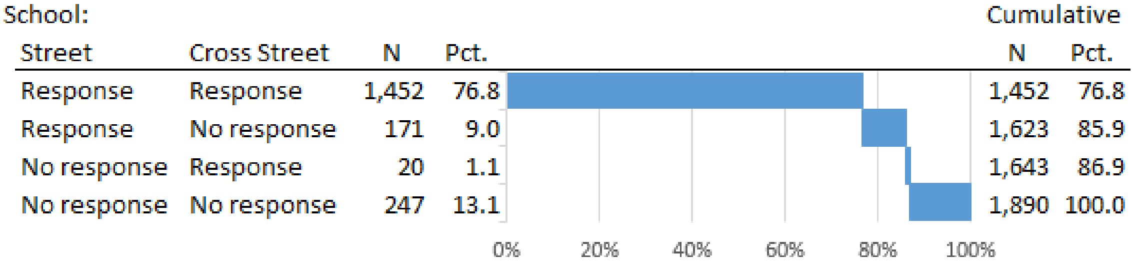
Over 85% (n=1683) of the respondents provided responses for both their home street and cross street. Over 75% (n=1452) of the respondents provided this information for their adolescents' school. Figures 1a and 1b provide a summary of the completeness of the responses to these questions.

Figure 1a – Completeness of responses to geographic location questions - home



Home:				Cumulative	
Street	Cross Street	N	Pct.	N	Pct.
Response	Response	1,683	89.0	1,683	89.0
Response	No response	53	2.8	1,736	91.9
No response	Response	10	0.5	1,746	92.4
No response	No response	144	7.6	1,890	100.0

Figure 1b – Completeness of responses to geographic location questions - school

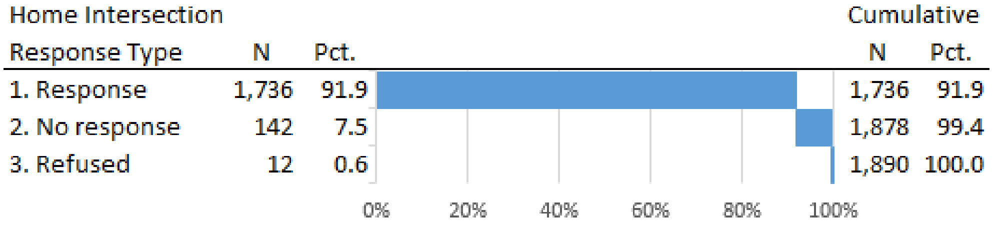


School:				Cumulative	
Street	Cross Street	N	Pct.	N	Pct.
Response	Response	1,452	76.8	1,452	76.8
Response	No response	171	9.0	1,623	85.9
No response	Response	20	1.1	1,643	86.9
No response	No response	247	13.1	1,890	100.0

The "No response" categories for both the school street and cross street includes "home-schooled" responses and "N/A" responses, as well as true non-responses to the question.

The survey responses about the home street and cross street provided insight into the attitudes of respondents toward geographic anonymity. The survey question asked, "Can you tell me just the name of the street/road you live on?" Responses such as "I'd prefer not to say," "No thank you," and "I could but won't" were coded to indicate respondents' preference for geographic anonymity. We classified these types of responses as refusals. We classified all other non-informative responses such as blanks, "None", and "N/A" as non-responses. Figure 2 gives a summary of the types of response values for home locations. The first row of Figure 2 corresponds to the first two rows of Figure 1a.

Figure 2 – Home intersection non-responses and refusals



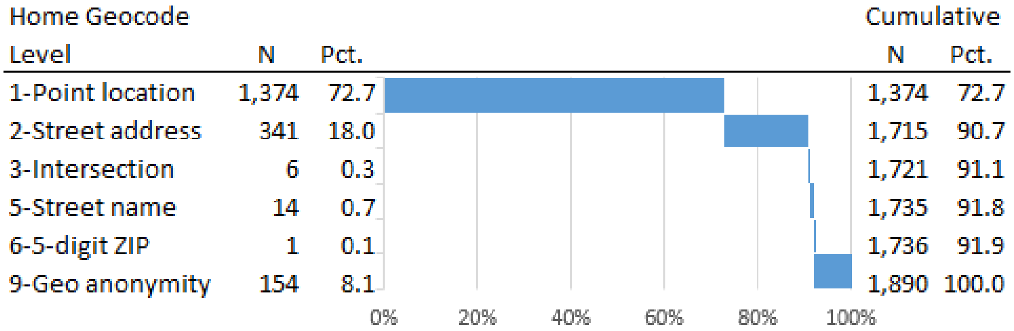
Home Intersection			Cumulative	
Response Type	N	Pct.	N	Pct.
1. Response	1,736	91.9	1,736	91.9
2. No response	142	7.5	1,878	99.4
3. Refused	12	0.6	1,890	100.0

For the GeoFLASHE data, we treated both the "no response" and "refused" categories as requests for geographic anonymity. Thus, 1,736 (about 92%) of the completed FLASHE responses have linkable supplemental neighborhood and contextual information in the GeoFLASHE data. The variable Home_Geo_Status identifies these 1,736 dyads. The geographic anonymity of the remaining 154 dyads was applied to both the home and school locations.

1.2 Geocoding Accuracy

In addition to the home intersection information derived from the survey questions, a home street address was available for each FLASHE respondent from the address that was used to mail the survey invitations. The street address of the home, supplemented with home intersection information, was geocoded to obtain the geographic latitude and longitude values for each respondent. Geocoding was done using Esri's ArcGIS Desktop 10.4.1 software with data from Esri's StreetMap Premium (Navteq) 2015 version 1. Figure 3 provides a summary of the accuracy levels of the home address geocoding (the first digit in the geocoding level reflects the relative accuracy).

Figure 3 – Geocoding accuracy of home locations

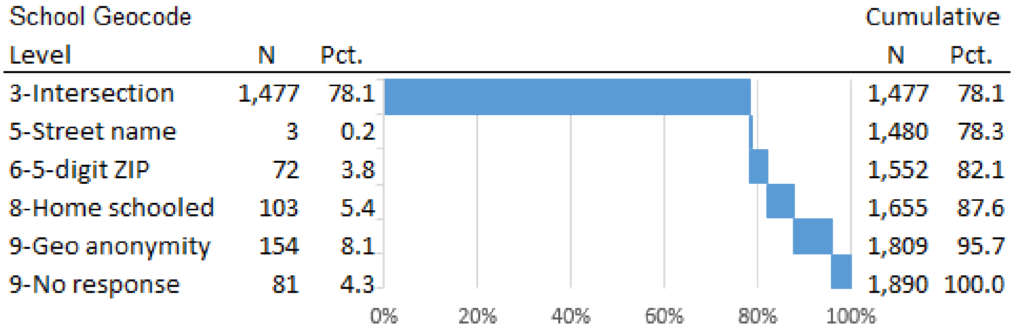


Home Geocode Level	N	Pct.	Cumulative N	Cumulative Pct.
1-Point location	1,374	72.7	1,374	72.7
2-Street address	341	18.0	1,715	90.7
3-Intersection	6	0.3	1,721	91.1
5-Street name	14	0.7	1,735	91.8
6-5-digit ZIP	1	0.1	1,736	91.9
9-Geo anonymity	154	8.1	1,890	100.0

The first two geocoding levels based on a point location or street address are the most accurate. These were available for about 90% of the FLASHE respondents and over 98% (1,715/1,736) of the responded who did not indicate a preference for geographic anonymity. For six of the respondents, the intersection information from the survey provided the most accurate geographic information. The variable Home_Geocode_Level provides the geocoding accuracy of the home location for each dyad.

Geographic information about the school location was limited to information provided in the survey responses. Figure 4 provides a summary of the accuracy levels of the school intersection geocoding.

Figure 4 – Geocoding accuracy of school locations



School Geocode			Cumulative	
Level	N	Pct.	N	Pct.
3-Intersection	1,477	78.1	1,477	78.1
5-Street name	3	0.2	1,480	78.3
6-5-digit ZIP	72	3.8	1,552	82.1
8-Home schooled	103	5.4	1,655	87.6
9-Geo anonymity	154	8.1	1,809	95.7
9-No response	81	4.3	1,890	100.0

For the school geocoding process, we used the ZIP code from the home address information to help locate the intersection. However, for the cases where the intersection could not be found (3 at the street name level and 72 at the ZIP code level), we set the school geocoding status to "missing" since we did not really know the ZIP code or even the city that the school was located in. This resulted in valid school geocoding information for the 1,477 cases where we found the school intersection as well as the 103 home schooled cases for a total of 1,580 dyads with geographic information about the school. The variable School_Geocode_Level provides the geocoding accuracy of the school location for each dyad.

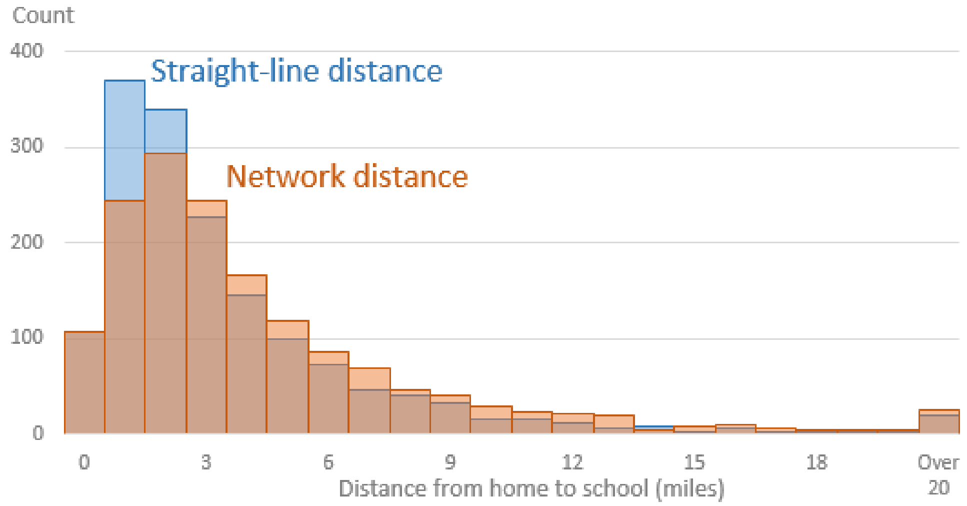
The "home schooled" geocode level in Figure 4 includes respondents who indicated that the adolescent was home schooled in the original FLASHE survey results (TSCHLTYPE value of 3, n=101) as well as respondents who indicated the adolescent was home schooled in the school intersection information (n=2). In all of these cases, the latitude and longitude of the school was set to the latitude and longitude values of the home location. For more information about identifying home-schooled adolescents, see the discussion of the Geo_Notes variable in Section 3.1.

1.3 Home to School Distance Variables

For cases where we have both a home and school location, we calculated three additional variables: the straight-line distance between them ("as the crow flies" - SLDistMi), the distance along the street network (NETDistMi), and the ratio of the network distance to the straight-line distance (DistRatio). Note that the network distance is always at least as long as the straight-line distance and is generally longer. Thus, the distance ratio is always greater than or equal to one. The ratio of the network distance to the straight-line distance can be used as a measure of connectivity. A network distance that is only marginally longer than the straight-line distance indicates that an efficient route exists between the home and school [Thornton et al., 2011].

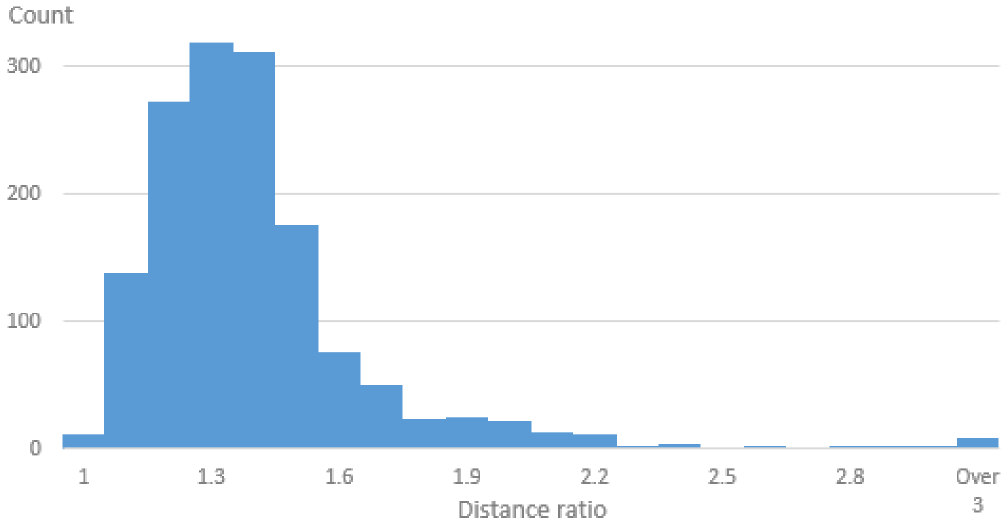
Two overlapping histograms comparing the distribution of the values for the straight-line and network distance variables are provided in Figure 5. A histogram showing the distribution of the distance ratio is provided in Figure 6. In Figure 5, notice how the network distances are generally longer than the straight-line distances.

Figure 5 – Distribution of the straight-line and network distance variables



	N	Minimum	Median	Mean	Maximum	Standard Deviation
Straight-line distance	1,580	0	1.86	4.71	2,266.5	57.27
Network distance	1,580	0	2.58	5.83	2,602.0	65.74

Figure 6 – Distribution of the distance ratio (network to straight-line)



	N	Minimum	Median	Mean	Maximum	Standard Deviation
Distance ratio	1,472	1	1.30	1.35	4.16	0.30

Note that the maximum distance in Figure 5 (both straight-line and network) is over two thousand miles. There is one dyad where the school appears to be an out-of-state boarding school. This dyad is identified in the Geo_Notes variable so that researchers can exclude it if desired.

The responses for 108 dyads specified the same intersection for both the home and the school and, thus, the distance measures are zero. This can occur because the residence is in the same block as the school or because the student is home-schooled. Of these, 103 indicated that the adolescent is home schooled. We assume the other five live in the same block as the school.

1.4 Missing Data

The GeoFLASHE dataset includes variables to indicate the reasons for any missing numeric data. Oftentimes, values are missing because the location could not be geocoded or the respondent indicated a preference for geographic anonymity. In both of these cases, a value of -9 is coded indicating that a geocode is not available. The variables Home_Geo_Status or School_Geo_Status are coded with a value of "2-Missing" to identify these cases. Note that the two distance variables require both a home and school location and will have a value of -9 if either is missing.

For the GeoFLASHE variables describing neighborhood characteristics (see next section), values can also be missing if the original source data is missing. For example, variables describing the median rent will have missing values when there are no rental units in the neighborhood. In these cases, the values are coded as -99 to indicate that source data is not available.

When a ratio is calculated, values are missing if the denominator is zero. In these cases, the values are coded as -999. Currently, this only occurs during the calculation of the ratio of the street-network distance to the straight-line distance between the home and school (DistRatio).

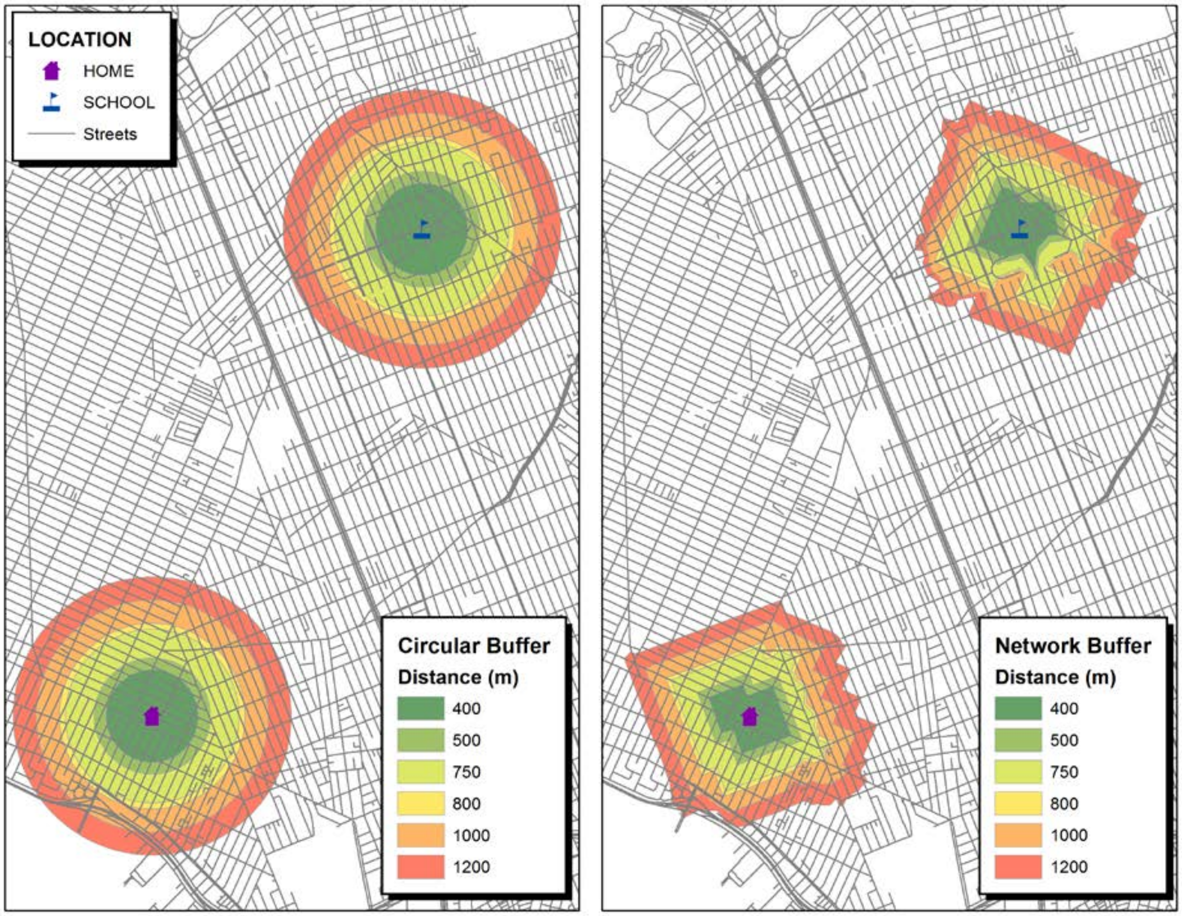
2. Neighborhood Buffers and Measures

To capture information about the neighborhood context, sets of buffers were constructed with differing characteristics. These buffers provide researchers a range of options to define the appropriate spatial extent of a neighborhood for particular research questions. Refer to Table 10 on page 25 for a detailed listing of neighborhood variables.

2.1 Buffer Creation

Buffers were created surrounding the geocoded home and school locations. Circular buffers were created with radii of 400, 500, 750, 800, 1,000 and 1,200 meters (Figure 7a). Street-network buffers were generated using the same distances, however these buffers were defined by the distance traveled from the home and school locations along the street network (Figure 7b). For the street-network calculation, we used all roads other than limited access highways. Radii were selected based on previous studies prior to linking geographic data with FLASHE domains (e.g. James et al. 2014).

Figure 7 - Example of circular and street-network buffers created for GeoFLASHE
(the home and school locations were randomly selected for this example)



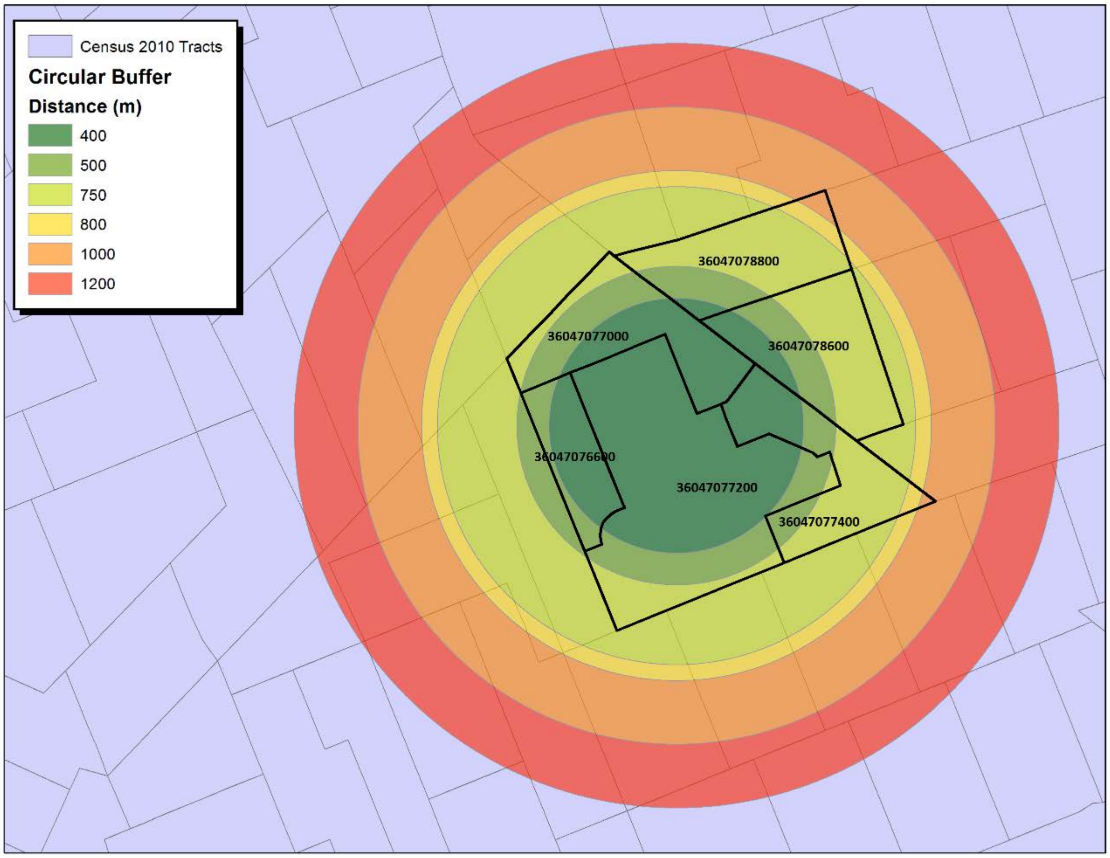
a. circular

b. network

2.2 Calculation of Buffer Percentages

Buffer percentages were calculated by performing a spatial intersection of the buffers with boundaries of census tracts. Figure 8 gives an example showing six census tracts that intersect the 400 meter circular buffer (the dark green circle) using a randomly selected location (not associated with any of the FLASHE respondents).

Figure 8 - Example of the six tracts that intersect a 400 meter circular buffer
(the location was randomly selected for this example)



For each intersecting tract, we calculated the percentage of the buffer area that is contained in the tract. As an example, Table 1 gives the results for the 400-meter buffer shown in Figure 8 for the school location of dyad N.

Table 1 – Example of buffer percentage calculations

DYADID	Location	Buffer Type	Buffer Size	Tract	Buffer Percentage
N	SCHOOL	CIRCULAR	400	36047076600	11.61%
N	SCHOOL	CIRCULAR	400	36047077000	17.36%
N	SCHOOL	CIRCULAR	400	36047077200	55.68%
N	SCHOOL	CIRCULAR	400	36047077400	8.71%
N	SCHOOL	CIRCULAR	400	36047078600	5.41%
N	SCHOOL	CIRCULAR	400	36047078800	1.24%

In this table, DYADID is the identifier for the dyad, Location is the location type (home or school), Buffer Type indicates whether a circular or network buffer was used, Buffer size indicates the radius of the buffer in meters, Tract is the census tract identifier, and Buffer Percentage is the percentage of the buffer area that is contained in the tract. Notice that the total of the buffer percentage is 100%.

Finally, these calculated buffer percentages were used as weights to calculate weighted averages for each of the census-based measures using the following formula:

$$\frac{\sum_{i=1}^{n} M_i w_i}{n}$$

Where M_i is the census-based measure for tract *i* and w_i is the buffer percentage for tract *i*.

For example, the calculation of the percent Hispanic population for this same 400-meter buffer is given in Table 2.

Table 2 – Example of a weighted neighborhood variable calculation

Tract	Tract % Hispanic	Buffer Percentage	Weighted % Hispanic
36047076600	11.55%	11.61%	1.34%
36047077000	15.69%	17.36%	2.72%
36047077200	12.20%	55.68%	6.79%
36047077400	12.46%	8.71%	1.09%
36047078600	8.25%	5.41%	0.45%
36047078800	4.40%	1.24%	0.05%
		Total % Hispanic:	12.44%

If a census-based measure is missing for one of the tracts, the weighted average ignores that tract (we assume a missing value is similar to the weighted average of the non-missing values). The resulting buffer measure is missing only if values for all of the component tracts are missing.

Using these methods, the set of neighborhood variables describing demographic, socio-economic, and built environment characteristics for FLASHE home and school locations were generated.

In addition, values for the geographic area of the home and school census tracts have been included in GeoFLASHE datasets for the categorical SES index and the walkability factor variables (see Sections 2.5 and 2.6). We have not included census tract measures for the continuous variables to protect confidentiality.

2.3 Variable Suffixes for Neighborhood Measures

There is a different instance of each GeoFLASHE neighborhood measure for each neighborhood buffer size (400, 500, 750, 800, 1,000 and 1,200 meters), buffer type (circular or distance along the street network) and location (home or school) for a total or 24 instances for each measure. The name of each GeoFLASHE neighborhood variable consists of a common root followed by a suffix that describes the location and buffer configuration.

The following convention is used for coding variable name suffixes:

- The first character is the location: H=home, S=school
- The second character is the buffer type: C=circular, N=network

- The third through sixth characters give the buffer size in meters

For the categorical measures, instances for the census tract areas have "Tract" as the second through sixth characters. Table 3 provides a list of the variable name suffixes.

Table 3 – Suffixes for neighborhood variables

Variable Name Suffix	Variable Label Suffix	Variable Description
*_HC0400	* - home circular 400 m	Uses a circular 400 meter buffer from the home location
*_HC0500	* - home circular 500 m	Uses a circular 500 meter buffer from the home location
*_HC0750	* - home circular 750 m	Uses a circular 750 meter buffer from the home location
*_HC0800	* - home circular 800 m	Uses a circular 800 meter buffer from the home location
*_HC1000	* - home circular 1000 m	Uses a circular 1000 meter buffer from the home location
*_HC1200	* - home circular 1200 m	Uses a circular 1200 meter buffer from the home location
*_HN0400	* - home network 400 m	Uses a network 400 meter buffer from the home location
*_HN0500	* - home network 500 m	Uses a network 500 meter buffer from the home location
*_HN0750	* - home network 750 m	Uses a network 750 meter buffer from the home location
*_HN0800	* - home network 800 m	Uses a network 800 meter buffer from the home location
*_HN1000	* - home network 1000 m	Uses a network 1000 meter buffer from the home location
*_HN1200	* - home network 1200 m	Uses a network 1200 meter buffer from the home location
*_SC0400	* - school circular 400 m	Uses a circular 400 meter buffer from the school location
*_SC0500	* - school circular 500 m	Uses a circular 500 meter buffer from the school location
*_SC0750	* - school circular 750 m	Uses a circular 750 meter buffer from the school location
*_SC0800	* - school circular 800 m	Uses a circular 800 meter buffer from the school location
*_SC1000	* - school circular 1000 m	Uses a circular 1000 meter buffer from the school location
*_SC1200	* - school circular 1200 m	Uses a circular 1200 meter buffer from the school location
*_SN0400	* - school network 400 m	Uses a network 400 meter buffer from the school location
*_SN0500	* - school network 500 m	Uses a network 500 meter buffer from the school location
*_SN0750	* - school network 750 m	Uses a network 750 meter buffer from the school location
*_SN0800	* - school network 800 m	Uses a network 800 meter buffer from the school location
*_SN1000	* - school network 1000 m	Uses a network 1000 meter buffer from the school location
*_SN1200	* - school network 1200 m	Uses a network 1200 meter buffer from the school location
*_HTract	* - home census tract	Uses the area of the census tract containing the home location
*_STract	* - school census tract	Uses the area of the census tract containing the school location

The appropriate buffer size and type to use will depend on the nature of the neighborhood context that the researcher is trying to capture. For constructs involving travel from one place to another such as access to parks and recreation facilities or access to healthy food sources, network buffers might be more appropriate. For constructs that capture the nature of the surrounding area such as poverty levels or unemployment, circular buffers might be more appropriate. The appropriate buffer size might be selected based on the degree of spatial proximity needed. Smaller buffers sizes might be used to represent activities within a short walk from the home or school whereas activities within a longer walk or a short drive could use larger buffer sizes. Of course, researchers may select a buffer type and size to match what was used in a prior study so that results are comparable.

2.4 Urban/Rural Categorizations

For each of the GeoFLASHE buffer configurations, we developed measures of the urban/rural environment of the buffer from Census 2010 urban and rural area determinations [Census 2010]. A continuous measure of urbanicity is available in the PctUrban_* variable which gives the percent of the population living in an urban area. A categorical measure of urbanicity is available in the UrbRurlCat_* variable which uses the National Center for Education Statistics urban-centric categories [NCES 2010]. A related variable is PopDen_* which gives the population density of the buffer. Note that some types of urban development have low population densities such as industrial areas and urban green spaces.

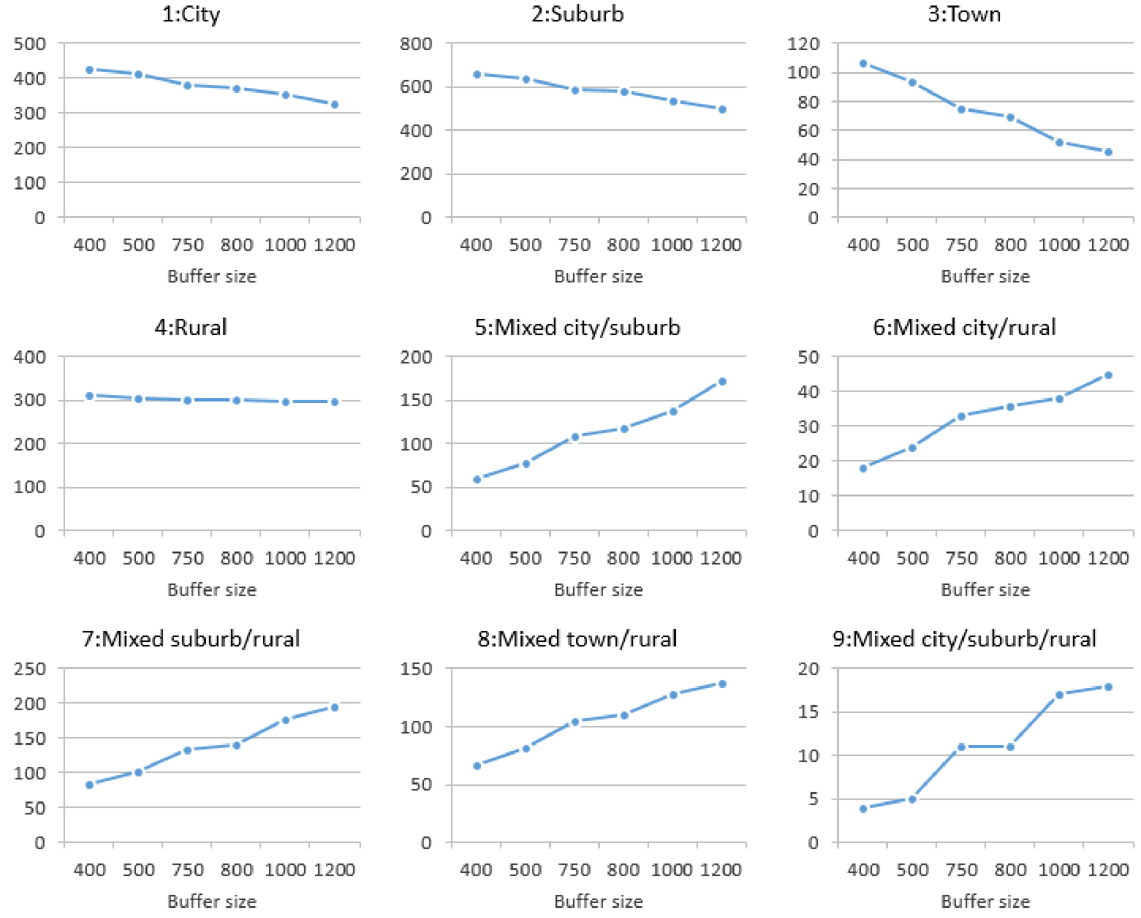
In creating the urban/rural categories, a 90% threshold was used to determine when a buffer area had more than one type of development. For instance, if 90% or more of the area was in a particular category, that category was assigned to the buffer; otherwise, a "mixed" category was assigned consisting of the categories with 10% or more of the area. The nine resulting categories are shown in Table 4.

Table 4 – Base GeoFLASHE urban/rural categories

Category	Definition
1	City (in an Urban Area and a Principal City)
2	Suburb (in an Urban Area but not a Principal City)
3	Town (in an Urban Cluster)
4	Rural (not in an Urban Area or Urban Cluster)
5	Mixed city/suburb
6	Mixed city/rural
7	Mixed suburb/rural
8	Mixed town/rural
9	Mixed city/suburb/rural

The number of FLASHE dyads in each category varies. In general, there are more dyads in the four main categories than in the mixed categories. In addition, as the buffer size increases, the number of dyads in the mixed categories increases because the likelihood of having a mix of urban types is higher for larger areas. Figure 9 shows the relationship between the buffer size and the number of dyads in each urban/rural category for the circular home buffers.

Figure 9 – Relationship between buffer size and the number of dyads in each urban/rural category (circular home buffers)



GeoFLASHE includes all nine categories for flexibility but we expect the categories will be collapsed into a smaller number of categories for analysis. Researchers should consider how they define and conceptualize urbanicity in deciding how to collapse the categories. Options for collapsing the urban/rural categories are summarized in Table 5.

Table 5 – Alternative approaches to collapsing the urbanicity categories

Original categories	Collapse the mixed categories:	Put mixed with the most urban category	Urban/ Suburban/ Rural (access)	Urban/ Suburban/ Rural (context)	Urban/ Rural (access)	Urban/ Rural (context)
1:City	City	City	Urban	Urban	Urban	Urban
2:Suburb	Suburb	Suburb	Suburban	Suburban	Urban	Urban
3:Town	Town	Town	Rural	Urban	Rural	Urban
4:Rural	Rural	Rural	Rural	Rural	Rural	Rural
5:Mixed city/suburb	Mixed	City	Urban	Urban	Urban	Urban
6:Mixed city/rural	Mixed	City	Urban	Urban	Urban	Urban
7:Mixed suburb/rural	Mixed	Suburb	Suburban	Suburban	Urban	Urban
8:Mixed town/rural	Mixed	Town	Rural	Urban	Rural	Urban
9:Mixed city/suburb/rural	Mixed	City	Urban	Urban	Urban	Urban
Number of categories	5	4	3	3	2	2

The five mixed categories could be collapsed into a single "mixed" category. Alternatively, they could be assigned to one of the main categories based on the most "urban" category in the mix. For example, "5:Mixed city/suburb" and "6:Mixed city/rural" could be assigned to "1:City", "7:Mixed suburb/rural" could be assigned to "2:Suburb", etc. Most analyses that include an urbanicity variable use just two or three categories. To get down to two categories, suburban is usually included in an urban category. With just two or three categories, a choice needs to be made for how to treat the "Town" areas. If the purpose of the urbanicity variable is to describe expected *access* to state-of-the-art medical facilities or other services found only in large cities, we recommend including "Town" with "Rural". If the purpose of the urbanicity variable is to describe the residential *context* (e.g., the likely presence of sidewalks and local parks), we recommend including "Town" with "Urban".

2.5 SES Index

To add a summary measure of neighborhood inequalities to the FLASHE survey results, we include an index of socio-economic status (SES) based on a number of SES-related measures including income, education, employment, occupation, and housing. We calculated the Yost SES index [Yost et al. 2001] using a method similar to that of Yu and colleagues [Yu et al. 2014]. Yu and colleagues calculated the Yost SES index at the census tract level for all census tracts in the geographic areas covered by the SEER program's cancer registries using data from the American Community Survey (ACS) for 2005-2009. We expanded this to calculate the index for all census tracts in the U.S. using ACS data for 2010-2014 and then applied the tract-level results to the GeoFLASHE buffers using the same method of weighted averages across component tracts used with the other neighborhood context variables (Section 2.2).

Table 6 compares the GeoFLASHE principal component analysis results for the Yost SES index with those of Yu et al. (2014).

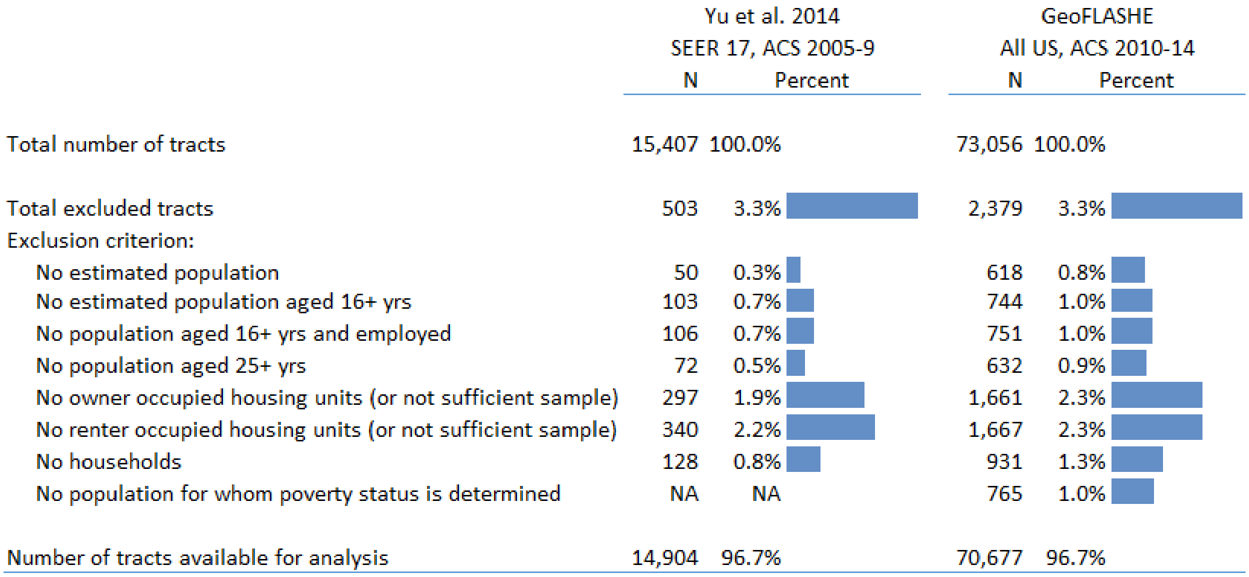
Table 6 – Comparison of principal component analysis results

		Yu et al. 2014	GeoFLASHE
Geographic area		SEER Registries	All US
Source data		ACS 2005-9	ACS 2010-14
% of common variance explained		91.82%	93.54%
SES Domain	Variable	Factor loadings	
Occupation	% working class	-0.109	-0.087
Unemployed	% aged ≥ 16 years who are unemployed	-0.037	-0.038
Poverty	% of persons below 150 % of poverty line	-0.233	-0.256
Income	Median HH income	0.490	0.494
Education	Education index (weighted school years)	0.118	0.096
Housing	Median house value	0.047	0.066
Housing	Median rent	0.059	0.049

Both the Yu et al. (2014) and GeoFLASHE results explain a large percentage of the variance. The individual factor loadings for GeoFLASHE had the same direction and a similar order of magnitude.

The Yost SES Index can only be calculated for tracts that have valid values for all of the input variables. The supplemental material for the Yu and colleagues' report provided data on the number of tracts that were excluded due to missing data. Figure 10 compares these values.

Figure 10 – Comparison of number of excluded tracts



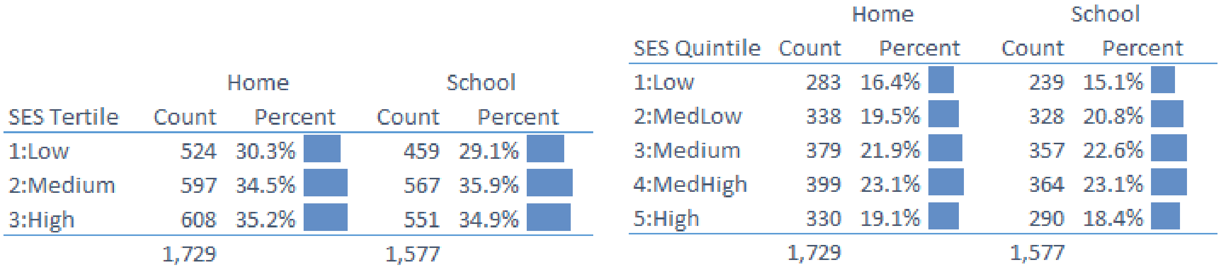
	Yu et al. 2014 SEER 17, ACS 2005-9		GeoFLASHE All US, ACS 2010-14	
	N	Percent	N	Percent
Total number of tracts	15,407	100.0%	73,056	100.0%
Total excluded tracts	503	3.3%	2,379	3.3%
Exclusion criterion:				
No estimated population	50	0.3%	618	0.8%
No estimated population aged 16+ yrs	103	0.7%	744	1.0%
No population aged 16+ yrs and employed	106	0.7%	751	1.0%
No population aged 25+ yrs	72	0.5%	632	0.9%
No owner occupied housing units (or not sufficient sample)	297	1.9%	1,661	2.3%
No renter occupied housing units (or not sufficient sample)	340	2.2%	1,667	2.3%
No households	128	0.8%	931	1.3%
No population for whom poverty status is determined	NA	NA	765	1.0%
Number of tracts available for analysis	14,904	96.7%	70,677	96.7%

Both Yu et al. (2014) and GeoFLASHE results excluded the same percentage of tracts due to missing values (3.3%). The proportions of excluded cases due to other criteria were also similar in the two analyses.

Following the procedures used by Yu and colleagues, we calculated cut points to divide the tract-level SES index values into tertiles and quintiles with roughly equal populations in each class.

Finally, we calculated SES index values for each of the FLASHE buffer configurations using the same method of weighted averages across component tracts used with the other neighborhood context variables (Section 2.2). The tract-level cut points were then applied to the SES index values for each of the FLASHE buffer configurations to generate tertile and quintile SES classes. Figure 11 gives the distribution of SES classes averaged across all buffer configurations for FLASHE home and school locations.

Figure 11 – Distribution of SES classes for GeoFLASHE home and school locations



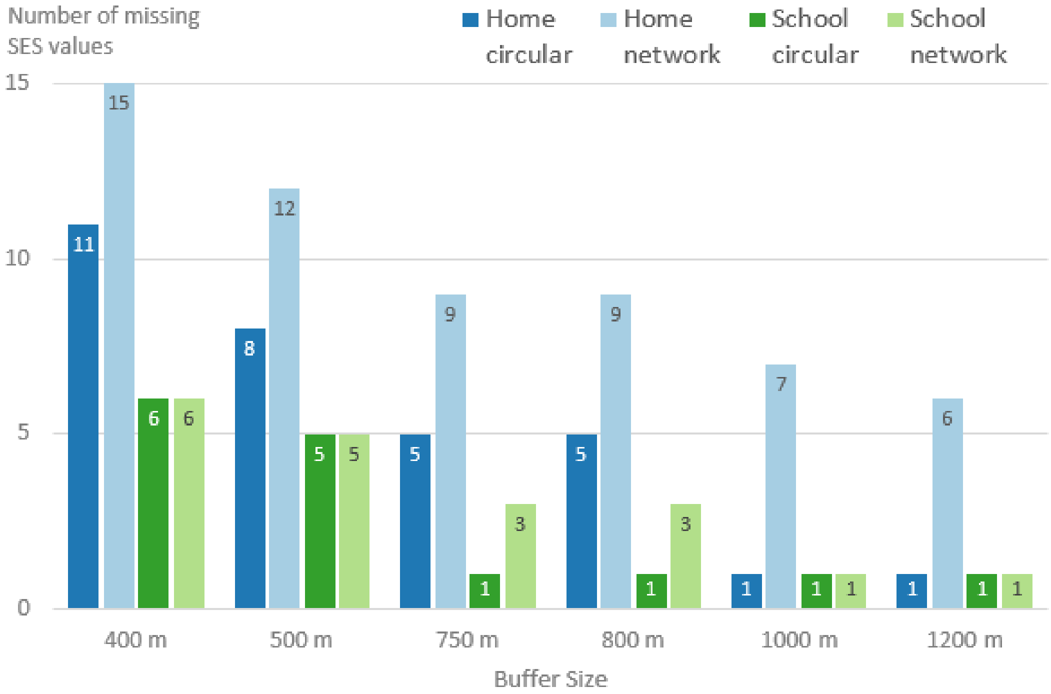
	Home		School	
SES Tertile	Count	Percent	Count	Percent
1:Low	524	30.3%	459	29.1%
2:Medium	597	34.5%	567	35.9%
3:High	608	35.2%	551	34.9%
	1,729		1,577	

	Home		School	
SES Quintile	Count	Percent	Count	Percent
1:Low	283	16.4%	239	15.1%
2:MedLow	338	19.5%	328	20.8%
3:Medium	379	21.9%	357	22.6%
4:MedHigh	399	23.1%	364	23.1%
5:High	330	19.1%	290	18.4%
	1,729		1,577	

Slightly fewer homes and schools are in the lowest SES tertile (29-30%); the rest of the homes and schools are fairly evenly distributed between the medium and high SES tertiles (35-36%). Similarly, slightly fewer homes and schools are in the lowest SES quintile (15-16%) and the highest SES quintile (18-19%); the rest of the homes and schools are fairly evenly distributed between the middle SES quintiles (20-23%).

Some census tracts do not have an SES index because one of the component measures is missing (for example, some tracts have no renter-occupied housing units so the median rent variable is missing). Hence, the number of non-missing SES variables can vary across neighborhood buffer configurations. The smaller buffer sizes are sometimes entirely in a single tract and, if the SES index for this tract is missing, the SES index for the buffer will also be missing. As the buffer size increases, neighboring tracts are included and, if these tracts have an SES index, an SES index for the buffer will be available. In general, the number of non-missing SES variables increases with buffer size. Figure 12 provides a summary of the number of missing SES values for home (n=1,736) and school (n=1,580) locations.

Figure 12 – Number of missing SES values by buffer size

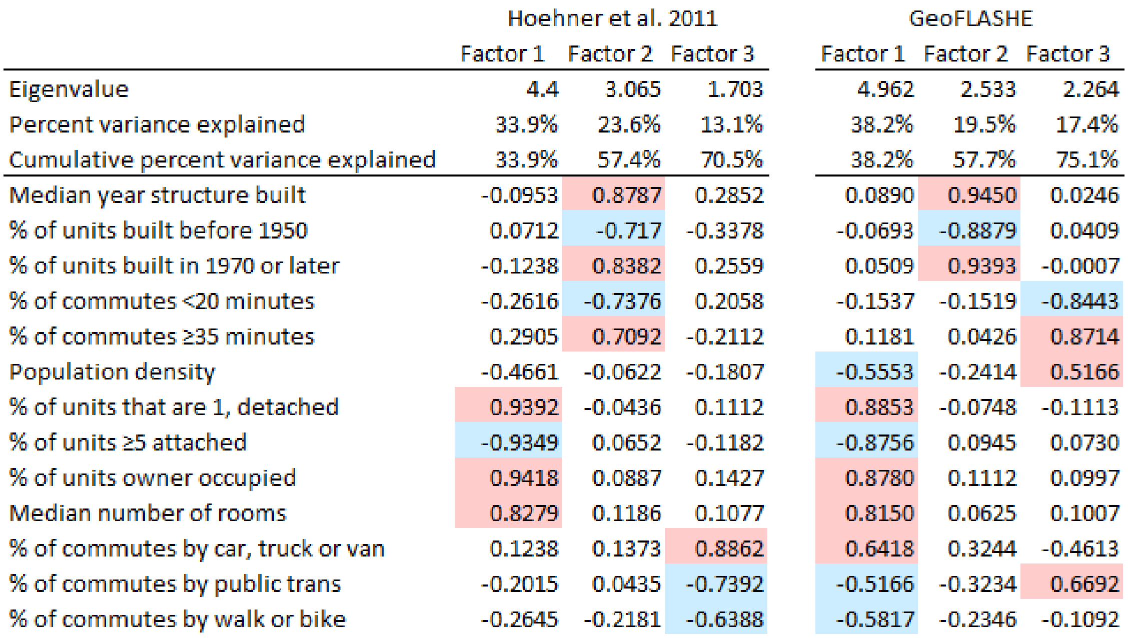


2.6 Neighborhood Characteristics Associated with Walkability

To create a set of summary measures of neighborhood walkability for the FLASHE survey results, a principal component analysis (PCA) was conducted using a set of built-environment measures from the U.S. Census Bureau. We followed the method described in a paper by Hoehner and colleagues [Hoehner et al. 2011] in a similar analysis of Census block groups in the state of Texas using data from the 1990 and 2000 decennial censuses. We performed a PCA for all census tracts in the U.S. using 13 measures: population density from the 2010 decennial census and 12 measures of the built environment from the 2010-2014 American Community Survey (ACS). Like Hoehner and colleagues, we used orthogonal varimax rotation and extracted three factors with eigenvalues greater than one.

Table 7 compares the Hoehner et al. (2011) PCA factor loading results with those of GeoFLASHE. To highlight the main contributors to each factor, the larger positive loadings (higher than 0.5) are shaded in red and larger negative loadings (lower than -0.5) are shaded in blue. Note that all of the variables were used to compute each of the factors, not just the highlighted ones.

Table 7 – Comparison of factor analysis results



	Hoehner et al. 2011			GeoFLASHE		
	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Eigenvalue	4.4	3.065	1.703	4.962	2.533	2.264
Percent variance explained	33.9%	23.6%	13.1%	38.2%	19.5%	17.4%
Cumulative percent variance explained	33.9%	57.4%	70.5%	38.2%	57.7%	75.1%
Median year structure built	-0.0953	0.8787	0.2852	0.0890	0.9450	0.0246
% of units built before 1950	0.0712	-0.717	-0.3378	-0.0693	-0.8879	0.0409
% of units built in 1970 or later	-0.1238	0.8382	0.2559	0.0509	0.9393	-0.0007
% of commutes <20 minutes	-0.2616	-0.7376	0.2058	-0.1537	-0.1519	-0.8443
% of commutes ≥35 minutes	0.2905	0.7092	-0.2112	0.1181	0.0426	0.8714
Population density	-0.4661	-0.0622	-0.1807	-0.5553	-0.2414	0.5166
% of units that are 1, detached	0.9392	-0.0436	0.1112	0.8853	-0.0748	-0.1113
% of units ≥5 attached	-0.9349	0.0652	-0.1182	-0.8756	0.0945	0.0730
% of units owner occupied	0.9418	0.0887	0.1427	0.8780	0.1112	0.0997
Median number of rooms	0.8279	0.1186	0.1077	0.8150	0.0625	0.1007
% of commutes by car, truck or van	0.1238	0.1373	0.8862	0.6418	0.3244	-0.4613
% of commutes by public trans	-0.2015	0.0435	-0.7392	-0.5166	-0.3234	0.6692
% of commutes by walk or bike	-0.2645	-0.2181	-0.6388	-0.5817	-0.2346	-0.1092

The three primary factors accounted for 70.5% of the total variance in the Hoehner results and 75.1% of the variance in our results. The main contributors to each factor were fairly similar, although the results differ in some ways. Commuting times move from factor 2 to factor 3 in our results and the modes of transportation move from factor 3 to factor 1 with public transit also appearing in our factor 3 but with an opposite sign.

Table 8 compares the interpretation of the three PCA factors between the Hoehner study and the GeoFLASHE results.

Table 8- Description of Factor Interpretations

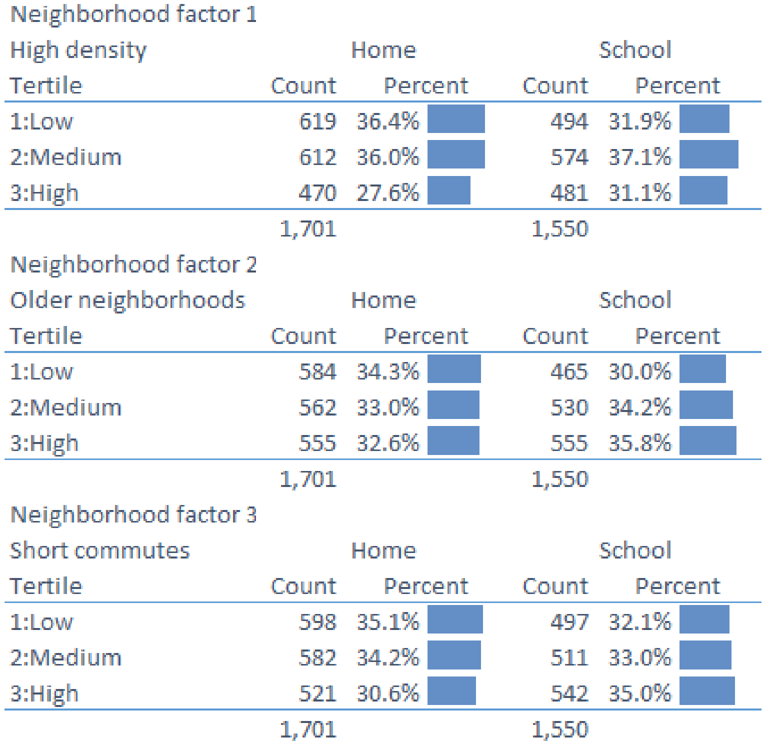
	Factor 1	Factor 2	Factor 3
Hoehner	High density	Traditional core	Non-auto commutes
GeoFLASHE	High density	Older neighborhoods	Short commutes

We use a slightly modified interpretation of the three factors. Hoehner and colleagues interpreted factor 1 as "high density" and we use this same term although our factor 1 includes high use of public transit and walk or bike commuting. Factor 2 was interpreted by Hoehner and colleagues as "traditional core" to include both older homes and shorter commutes. Since our factor 2 does not include shorter commutes, we interpret factor 2 as "older neighborhoods." Hoehner and colleagues interpreted factor 3 as "non-auto commutes" because it was based primarily on mode of transit. Our factor 3 is based primarily on the length of the commute so we interpret factor 3 as "short commutes." We interpret the large positive loading for the use of public transit in factor 3 as the use of commuter rail for long commutes. These interpretations of the factors are included in the labels of the GeoFLASHE variables: "neighborhood factor 1 high density," "neighborhood factor 2 older neighborhoods," and "neighborhood factor 3 short commutes."

Following the Hoehner analysis, we have coded the neighborhood factors so that high positive factor values are associated with high density, older neighborhoods, and short commutes. To do this, we have reversed the sign of the factor values so that negative loadings in Table 7 indicate variables that contribute to higher neighborhood factors and positive loadings in Table 7 indicate variables that contribute to lower neighborhood factors. For example, for neighborhood factor 2 (older neighborhoods), the loading for the percent of units built before 1950 is negative so areas where this percentage is high will have a higher value of neighborhood factor 2.

We calculated cut points to divide the tract-level neighborhood factors into tertiles and quintiles with roughly equal populations in each class for all U.S. census tracts. We then calculated factor values for each of the FLASHE buffer configurations using the same method of weighted averages across component tracts that we used with the other neighbor context variables (Section 2.2). The tertile and quintile cut points were applied to the neighborhood factor values for each of the FLASHE buffer configurations to generate tertile and quintile neighborhood-factor classes. Figure 13 gives the distribution of neighborhood factor tertiles averaged across all buffer configurations for FLASHE home and school locations.

Figure 13 – Distribution of neighborhood factor tertiles for FLASHE home and school locations



Neighborhood factor 1				
High density	Home		School	
Tertile	Count	Percent	Count	Percent
1:Low	619	36.4%	494	31.9%
2:Medium	612	36.0%	574	37.1%
3:High	470	27.6%	481	31.1%
	1,701		1,550	

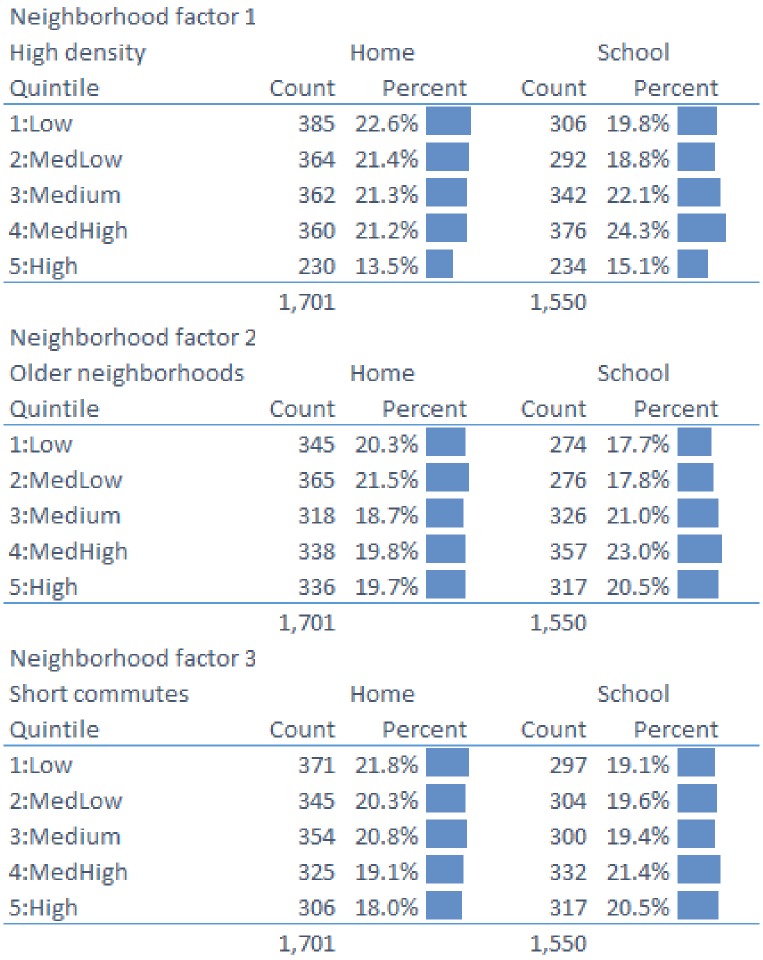
Neighborhood factor 2				
Older neighborhoods	Home		School	
Tertile	Count	Percent	Count	Percent
1:Low	584	34.3%	465	30.0%
2:Medium	562	33.0%	530	34.2%
3:High	555	32.6%	555	35.8%
	1,701		1,550	

Neighborhood factor 3				
Short commutes	Home		School	
Tertile	Count	Percent	Count	Percent
1:Low	598	35.1%	497	32.1%
2:Medium	582	34.2%	511	33.0%
3:High	521	30.6%	542	35.0%
	1,701		1,550	

For neighborhood factor 1 (high-density areas), slightly fewer FLASHE home locations are in the highest tertile (27.6%). The distribution of neighborhood factor 1 is fairly even for FLASHE school locations. The distribution of neighborhood factors 2 and 3 are fairly even for both the FLASHE home and school locations.

Figure 14 gives the distribution of neighborhood factor quintiles averaged across all buffer configurations for FLASHE home and school locations.

Figure 14 – Distribution of neighborhood factor quintiles for FLASHE home and school locations



Neighborhood factor 1

High density	Home		School	
Quintile	Count	Percent	Count	Percent
1:Low	385	22.6%	306	19.8%
2:MedLow	364	21.4%	292	18.8%
3:Medium	362	21.3%	342	22.1%
4:MedHigh	360	21.2%	376	24.3%
5:High	230	13.5%	234	15.1%
	1,701		1,550	

Neighborhood factor 2

Older neighborhoods	Home		School	
Quintile	Count	Percent	Count	Percent
1:Low	345	20.3%	274	17.7%
2:MedLow	365	21.5%	276	17.8%
3:Medium	318	18.7%	326	21.0%
4:MedHigh	338	19.8%	357	23.0%
5:High	336	19.7%	317	20.5%
	1,701		1,550	

Neighborhood factor 3

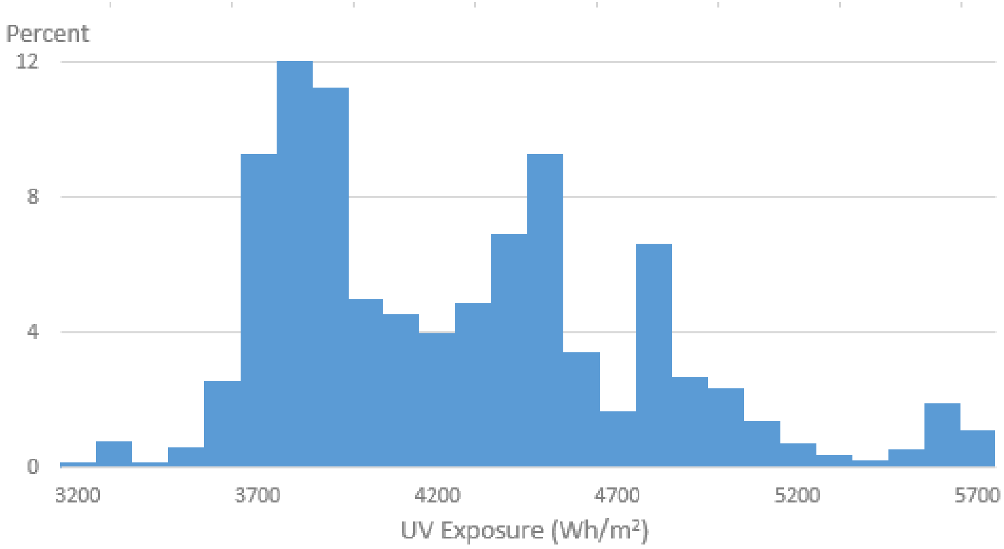
Short commutes	Home		School	
Quintile	Count	Percent	Count	Percent
1:Low	371	21.8%	297	19.1%
2:MedLow	345	20.3%	304	19.6%
3:Medium	354	20.8%	300	19.4%
4:MedHigh	325	19.1%	332	21.4%
5:High	306	18.0%	317	20.5%
	1,701		1,550	

For neighborhood factor 1 (high-density areas), slightly fewer FLASHE home and school locations are in the highest quintile (13 to 15%). The distribution of neighborhood factors 2 and 3 are fairly even for both the FLASHE home and school locations.

2.7 UV Exposure Measures

An estimate of exposure to ultraviolet (UV) radiation is provided for each of the FLASHE home and school buffer configurations. These values describe the average daily global solar radiation defined as the total amount of direct and diffuse solar radiation in units of watt-hours per square meter (Wh/m^2) between the years 1961-1990. More details are available in Tatalovich et al. 2006. Values range from 3,200 to 5,733. Figure 15 provides a histogram of UV exposure values for FLASHE home and school locations.

Figure 15 – Histogram of estimated UV exposure values

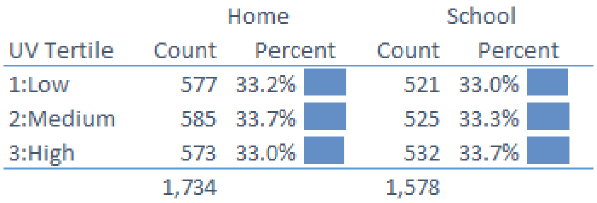


The UV exposure values do not vary significantly at the neighborhood scale. For a particular FLASHE home or school location, the values are similar for each of the different buffer configurations. Values across buffer configurations are within five Wh/m^2 for 93% of the FLASHE home and school locations. The maximum range of values across buffer configurations is 24 Wh/m^2. Because there is not much variation in UV exposure for different buffer configurations, researchers may wish to treat home or school UV exposure as a constant. For this purpose, a mean UV exposure value across all buffer configurations is provided for the home (UV_H_Mean) and school (UV_S_Mean) locations.

For two dyads, the original UV exposure values for a few of the buffer configurations were missing. This occurred when the dyad home or school location was at the edge of the UV exposure spatial dataset used to create the values. In these cases, the missing values were replaced with imputed values based on the values for the non-missing buffers. These two dyads are identified in the Geo_Notes field.

For categorical analysis, tertile variables have been included based on the distribution of exposure values within the FLASHE home and school locations. Although UV exposure values do not vary significantly for different buffer sizes, there is still a small chance that the tertile cutoff values might be within the range of values for a given home or school location. In these cases, some buffers might be in a different tertile even though their values are fairly similar. Thirteen of the 1,736 home locations and fifteen of the 1,580 school locations have different UV exposure tertiles for different buffer sizes. For researcher wishing to avoid these tertile differences, tertile variables for the mean UV exposure values for the home (UVTert_H_Mean) and school (UVTert_S_Mean) have been provided. Figure 16 gives the distribution of UV exposure tertiles averaged across all buffer configurations for FLASHE home and school locations.

Figure 16 – Distribution of UV exposure tertiles for FLASHE home and school locations



	Home		School	
UV Tertile	Count	Percent	Count	Percent
1:Low	577	33.2%	521	33.0%
2:Medium	585	33.7%	525	33.3%
3:High	573	33.0%	532	33.7%
	1,734		1,578	

As expected, the home and school buffer configurations are evenly distributed between the UV tertiles.

3. GeoFLASHE Variable Reference

This section provides detailed information about each of the variables in the GeoFLASHE dataset. We divide GeoFLASHE variables into two main categories. Geographic variables provide information about the geographic locations of the home and school and the process of geocoding that information. Neighborhood variables provide information about the characteristics of the areas around the home and school locations. There are multiple instances of each of the neighborhood variables corresponding to difference buffer configurations.

The key or index variable for the GeoFLASHE data is the DYADID. To link GeoFLASHE measures with the FLASHE survey databases, the DYADID variable should be used. All GeoFLASHE data are associated with the dyad rather than the individual parent or adolescent respondent.

3.1 Geographic Variables

In this section, we describe the set of GeoFLASHE variables that provide basic geographic information about the geocoding results for the home and school locations.

Table 9 describes the geographic variables by three main topics: geocoding process, location information, and variables describing distance.

Table 9 - Geographic variables by topic

Geocoding Process	Location Information	Distance
Dyad Identifier	Urban/rural category-home	Home-to-school straight line distance (miles)
Geocoding status-home	Urban/rural category-school	Home-to-school street-network distance (miles)
Geocoding status- school		Ratio of street-network to straight-line distance
Level of geocoding accuracy-home		
Level of geocoding accuracy-school		
Geographic notes for special cases		

Geocoding Process:

Dyad Identifier

- SAS/SPSS variable: DYADID
- Description: Unique dyad identifier.
- Usage notes: Integer key used to link to survey response data in other FLASHE datasets.

Geocoding status – home location

- SAS/SPSS variable: Home_Geo_Status
- Description: Indicates whether geographic information is available for the home.
- Usage notes: Can be used to identify dyads with geographic information based on the home location (n=1,736) and dyads who indicated a preference for geographic anonymity (n=173). See Section 1.1 for additional information.

Geocoding status – school location

- SAS/SPSS variable: School_Geo_Status
- Description: Indicates whether geographic information is available for the school.
- Usage notes: Can be used to identify dyads with geographic information based on the school location (n=1,580). This includes 103 dyads where the adolescent is home schooled. In these cases, the school and home locations are the same. Researchers can exclude these cases if desired (School_Geocode_Level = "8-HomeSchool"). See Section 1.1 for additional information.

Level of geocoding accuracy – home

- SAS/SPSS variable: Home_Geocode_Level
- Description: Provides a qualitative assessment of the accuracy of the geocoded home location information. See Table 3 for values.
- Usage notes: Point and street address geocodes are the most accurate. Intersection level geocodes are reasonably accurate. Researchers may wish to exclude observations with just street name or ZIP code levels of accuracy. See Section 1.2 for additional information.

Level of geocoding accuracy – school

- SAS/SPSS variable: School_Geocode_Level
- Description: Provides a qualitative assessment of the accuracy of the geocoded school location information. See Table 4 for values.
- Usage notes: Because the geographic information available for the schools was limited to the street intersection, only results for geocoding levels of "3-Intersect" and "8_HomeSchool" have been retained and used to generate GeoFLASHE data for the school locations. Researchers can exclude home schooled cases if desired. See Section 1.2 for additional information.

Geographic notes for special cases

- SAS/SPSS variable: Geo_Notes
- Description: Used to identify cases with particular characteristics that might be of interest to researchers. Current possible values are:

    - “1-HomeSch?” (n=2): the response to the school intersection questions indicated the adolescent is home-schooled but the school-type response (TSCHLTYPE) indicated a public or private school. These cases are treated as home schooled for geographic variables: the home and school locations are the same and the distance between them is zero.
    - “2-SchDist?” (n=1): a questionable distance between the home and the school – the school seems to be an out-of-state boarding school (n=1) even though one of the eligibility criteria for FLASHE participation was that the parent and adolescent had to live in the same household at least 50% of the time.
    - “3-ImputeUV” (n=2): the UV exposure values for some buffer sizes were missing and were imputed from non-missing buffer values. See Section 2.7 for details.
- Usage notes: Researchers may wish to exclude these observations or handle them in particular ways. For example, for a study of home schooled adolescents, a researcher may want to include the two cases with Geo_Notes = “1-HomeSch?” as well as those with TSCHLTYPE = 3.

Location Information:

Urban/rural category - home

- SAS/SPSS variable: UrbRurlGeo_home
- Description: Urban/rural environment of the home based on Census 2010 categorizations (see https://www.census.gov/geo/reference/urban-rural.html). Values are categorized using the National Center for Education Statistics (NCES) urban-centric categories:
    - 1: City (in an Urban Area and a Principal City)
    - 2: Suburb (in an Urban Area but not a in a Principal City)
    - 3: Town (in an Urban Cluster)
    - 4: Rural (not in an Urban Area or Urban Cluster)
- Usage notes: Characterizes the urban/rural nature of the location of the home. Additional urban/rural variables are available for the neighborhoods around the home – see Section 2.4 for details.

Urban/rural category - school

- SAS/SPSS variable: UrbRurlGeo_school
- Description: Urban/rural environment of the school based on Census 2010 categorizations (see https://www.census.gov/geo/reference/urban-rural.html). Values are categorized using the National Center for Education Statistics (NCES) urban-centric categories:
    - 1: City (in an Urban Area and a Principal City)
    - 2: Suburb (in an Urban Area but not a in a Principal City)
    - 3: Town (in an Urban Cluster)
    - 4: Rural (not in an Urban Area or Urban Cluster)
- Usage notes: Characterizes the urban/rural nature of the location of the school. Additional urban/rural variables are available for the neighborhoods around the school – see Section 2.4 for details.

Distance:

Home-to-school straight-line distance (miles)

- SAS/SPSS variable: SLDistMi
- Description: The straight-line distance (as the crow flies) from the home to the school in miles. Calculated for dyads where we have both a home and school location.
- Usage notes: Can be used to assess the general proximity of the school from the home. Has a value of zero if the adolescent is home-schooled or if the school is in the same block as the residence. See Section 1.3 for additional information.

Home-to-school street-network distance (miles)

- SAS/SPSS variable: NETDistMi
- Description: The shortest distance along the street network from the home to the school in miles. Calculated for dyads where we have both a home and school location.
- Usage notes: Can be used to estimate how far the adolescent travels (and how long it takes) to get to and from school each day. Has a value of zero if the adolescent is home-schooled or if the school is in the same block as the residence. See Section 1.3 for additional information.

Ratio of street-network to straight-line distance

- SAS/SPSS variable: DistRatio
- Description: The ratio of the street-network distance between the home and school to the straight-line distance between the home and school. Calculated for dyads that have non-zero values for the straight-line distance.
- Usage notes: Because the network distance is always at least as long as the straight-line distance, the distance ratio is always greater than or equal to one. Can be used as a measure of connectivity: a network distance that is only marginally above the straight-line distance indicates that an efficient route exists between the home and school [Thornton et al., 2011]. See Section 1.3 for additional information.

3.2 Neighborhood Variables

In this section, we describe the set of GeoFLASHE variables that provide information about the characteristics of the areas around the home and school locations.

Table 10 contains a summary of the GeoFLASHE neighborhood variables by four main topics: demographic variables, socio-economic variables, built environment variables, and exposure variables.

Table 10 – Neighborhood variables by topic

Demographic Variables	Socio-Economic Variables	Built Environment Variables	Exposure Variables
Percent White (non-Hispanic)	Percent of persons living below poverty line	Percent of commutes by car, truck, or van	Ultraviolet radiation exposure
Percent Black (non-Hispanic)	Percent of persons living below 150% poverty line	Percent of commutes by public transportation	
Percent Asian (non-Hispanic)	Percent of persons living below 200% poverty line	Percent of commutes by walk or bike	
Percent Pacific Islander (non-Hispanic)	Percent of persons who are unemployed	Percent of commutes less than 20 minutes	
Percent American Indian Alaska Native (AIAN) (non-Hispanic)	Percent owner-occupied housing	Percent of commutes more than 35 minutes	
Percent two or more races (non-Hispanic)	Percent renter-occupied housing	Percent of housing structures that are a single detached unit	
Percent some other race (non-Hispanic)	Percent vacant housing units	Percent of housing structures with 5 or more units	
Percent Hispanic	Percent female headed households	Median number of rooms	
Population density	Percent of households on public assistance	Percent of units built before 1950	
Percent of population living in an urban area	Annual median household income	Percent of units built in 1970 or later	
Urban/rural category	Median house value	Median year structure built	
	Median rent	Neighborhood factors	
	SES Index		

Please note that the variables listed followed by an * indicate the addition of a buffer suffix as displayed in Table 3.

Demographic Variables:

Percent White (non-Hispanic)

- SAS/SPSS variable: PctWhite_*
- Calculation: Number of non-Hispanic people who are only one race (White)*100/total population
- Usage notes: Represents the racial makeup of the neighborhood. Hispanics are included in a separate variable (PctHisp_*).

Percent Black (non-Hispanic)

- SAS/SPSS variable: PctBlack_*
- Calculation: Number of non-Hispanic people who are only one race (Black)*100/total population

- Usage notes: Represents the racial makeup of the neighborhood. Hispanics are included in a separate variable (PctHisp_*).

Percent Asian (non-Hispanic)

- SAS/SPSS variable: PctAsian_*
- Calculation: Number of non-Hispanic people who are only one race (Asian)*100/total population
- Usage notes: Represents the racial makeup of the neighborhood. Hispanics are included in a separate variable (PctHisp_*).

Percent Pacific Islander (non-Hispanic)

- SAS/SPSS variable: PctPacIs_*
- Calculation: Number of non-Hispanic people who are only one race (Pacific Islander)*100/total population
- Usage notes: Represents the racial makeup of the neighborhood. Hispanics are included in a separate variable (PctHisp_*).

Percent American Indian Alaska Native (AIAN) (non-Hispanic)

- SAS/SPSS variable: PctAIAN_*
- Calculation: Number of non-Hispanic people who are only one race (AIAN)*100/total population
- Usage notes: Represents the racial makeup of the neighborhood. Hispanics are included in a separate variable (PctHisp_*).

Percent two or more races (non-Hispanic)

- SAS/SPSS variable: Pct2plus_*
- Calculation: Number of non-Hispanic people who are two or more races*100/total population
- Usage notes: Represents the multi-racial makeup of the neighborhood. Hispanics are included in a separate variable (PctHisp_*).

Percent some other race (non-Hispanic)

- SAS/SPSS variable: PctOther_*
- Calculation: Number of non-Hispanic people who are some other race (other)*100/total population
- Usage notes: Indicates how many people in the neighborhood do not identify as one of the standard race categories. Hispanics are included in a separate variable (PctHisp_*).

Percent Hispanic

- SAS/SPSS variable: PctHisp_*
- Calculation: Number Hispanic or Latino*100/total population
- Usage notes: Represents the Hispanic ethnic makeup of the neighborhood. Hispanics are not included in previous racial variables.

Population density

- SAS/SPSS variable: PopDen_*
- Calculation: Population density in people per square mile = Number of people/ (Land area in square meters / 2,589,988). The number 2,589,988 is the conversion factor between square meters and square miles provided in the Census table documentation.
- Usage notes: Can be used as a measure of residential density. Related to urbanicity but some urban land uses have low population density such as industrial areas and urban parks. Also related to sprawl but sprawl measures usually include land use, employment location, and street accessibility.

Percent of population living in an Urban Area or Urban Cluster

- SAS/SPSS variable: PctUrban_*
- Calculation: (Number of people living in an Urban Area + number of people living in an Urban Cluster)*100/total population
- Usage notes: The methods used by the Census Bureau to identify urban areas include low-density urban land uses such as industrial areas and urban parks. See Section 2.4 for additional information.

Urban/rural category

- SAS/SPSS variable: UrbRurlCat_*
- Calculation: See Section 2.4 on calculation or urban/rural category
- Usage notes: A categorical measure based on PctUrban_* and using NCES categories. In most cases, researchers will want to collapse the nine categories into fewer categories. See Section 2.4 for details.

Socio-Economic Variables:

Percent of persons below poverty line

- SAS/SPSS variable: Pct100Pov_*
- Calculation: Number of people below poverty line*100/estimate of total population for whom poverty status is determined
- Usage notes: A measure of the degree of poverty in the neighborhood. Based on individual income over the last 12 months.

Percent of persons below 150% poverty line

- SAS/SPSS variable: Pct150Pov_*
- Calculation: Number of people below 150% poverty line*100/estimate of total population for whom poverty status is determined
- Usage notes: A measure of the degree of poverty in the neighborhood. Based on individual income over the last 12 months. Eligibility for some social programs require incomes below 150% of the federal poverty line.

Percent of persons below 200% poverty line

- SAS/SPSS variable: Pct200Pov_*

- Calculation: Number of people below 200% poverty line*100/estimate of total population for whom poverty status is determined
- Usage notes: A measure of the degree of poverty in the neighborhood. Based on individual income over the last 12 months. 200% of the federal poverty line is often considered "low income".

Percent of persons who are unemployed

- SAS/SPSS variable: PctUnempl_*
- Calculation: Unemployment rate for population 16 years and over (taken directly from the Census ACS 2010-2014 table)
- Usage notes: A measure of the degree of unemployment in the neighborhood.

Percent of owner-occupied housing

- SAS/SPSS variable: PctHomeOwn_*
- Calculation: People who own homes*100/estimate of occupied housing units
- Usage notes: A measure of the relative number of homes in the neighborhood that are occupied by the owner as opposed to being rented. The sum of this measure and PctRentOcc_* will equal 100.

Percent of renter occupied housing

- SAS/SPSS variable: PctRentOcc_*
- Calculation: People who rent homes*100/estimate of occupied housing units
- Usage notes: A measure of the relative number of homes in the neighborhood that are rented out as opposed to being occupied by the owner. The sum of this measure and PctHomeOwn_* will equal 100.

Percent of vacant housing units

- SAS/SPSS variable: PctVacant_*
- Calculation: Vacant homes*100/estimate of occupied housing units
- Usage notes: A measure of the relative number of vacant homes in the neighborhood. Often used with poverty and unemployment measures to assess economic vitality.

Percent of female-headed households

- SAS/SPSS variable: PctFemHead_*
- Calculation: Homes with a female head of household*100/estimate of total households
- Usage notes: A measure of the relative number of households where a woman is the primary source of income. Include both single mothers and married mothers.

Percent of households on public assistance

- SAS/SPSS variable: PctPubAsst_*
- Calculation: Households on public assistance*100/estimate of total households
- Usage notes: A measure of the relative number households that receive cash public assistance or participate in the SNAP (food stamp) program.

Annual median household income

- SAS/SPSS variable: MedHHInc_*
- Calculation: Annual median household income (taken directly from the Census ACS 2010-2014 table)
- Usage notes: A measure of the income level of households in the neighborhood. Units are in 2013 inflation adjusted dollars.

Median house value

- SAS/SPSS variable: MedHousVal_*
- Calculation: Median house value (taken directly from the Census ACS 2010-2014 table)
- Usage notes: A measure of the value of homes in the neighborhood. Units are in 2013 inflation adjusted dollars.

Median rent

- SAS/SPSS variable: MedRent_*
- Calculation: Median rent (taken directly from the Census ACS 2010-2014 table)
- Usage notes: A measure of the monthly housing cost expenses for renters in the neighborhood. Includes both rent and utilities paid by the renter. Units are in 2013 inflation adjusted dollars.

SES index

- SAS/SPSS variables: SESIndex_*, SESTert_*, SESQuint_*
- Calculation: See Section 2.5 on calculation of SES index values
- Usage notes: SESIndex_* is a continuous measure of relative socio-economic status of the neighborhood derived from seven individual measures. SESTert_* and SESQuint_* are categorical measures based on the SESIndex_* value. A difference in categorical values among buffer configurations for a given location could be due to a relatively steep socioeconomic gradient in the neighborhood or because the buffer values are near the cutoff value between categories. See Section 2.5 for additional details.

Built Environment Variables:

Percent of commutes by car, truck, or van

- SAS/SPSS variable: PctCommuteAuto_*
- Calculation: Commutes by car, truck, or van/total estimate of workers aged 16 and over
- Usage notes: A measure of commuting transit mode preferences in the neighborhood. See also PctCommutePublic_* and PctCommuteWalkBike_*.

Percent of commutes by public transportation

- SAS/SPSS variable: PctCommutePublic_*
- Calculation: Commutes by public transportation/total estimate of workers aged 16 and over
- Usage notes: A measure of commuting transit mode preferences in the neighborhood. See also PctCommuteAuto _* and PctCommuteWalkBike_*.

Percent of commutes by walk or bike

- SAS/SPSS variable: PctCommuteWalkBike_*
- Calculation: Commutes by walk plus commutes by bike/total estimate of workers aged 16 and over
- Usage notes: A measure of commuting transit mode preferences in the neighborhood. See also PctCommuteAuto _* and PctCommutePublic _*.

Percent of commutes less than 20 minutes

- SAS/SPSS variable: PctCommuteLT20_*
- Calculation: Commutes less than 5 minutes plus commutes 5-9 minutes plus commutes 10-14 minutes plus commutes 15-19 minutes/ total estimate of workers 16 and over who did not work at home
- Usage notes: A measure of the percentage of people with relatively short commutes in the neighborhood. See also PctCommute35plus_*.

Percent of commutes greater than or equal to 35 minutes

- SAS/SPSS variable: PctCommute35plus_*
- Calculation: Commutes 35-39 minutes plus commutes 40-44 minutes plus commutes 45-59 minutes plus commutes 60-89 plus commutes 90 minutes or more/total estimate of workers 16 years and over who did not work at home
- Usage notes: A measure of the percentage of people with relatively long commutes in the neighborhood. See also PctCommuteLT20_*.

Percent of housing structures that are a single detached unit

- SAS/SPSS variable: PctNumUnits1dtchd_*
- Calculation: Detached units/total housing units
- Usage notes: A measure of the proportion of single-family detached homes in the neighborhood. See also PctNumUnits5plus_*. Larger numbers of single-family detached homes are usually associated with less dense development and lower neighborhood walkability.

Percent of housing structures with five or more units

- SAS/SPSS variable: PctNumUnits5plus_*
- Calculation: Housing units with 5-9 attached plus units with 10-19 attached plus units with 20-49 attached plus units with 50 or more attached/estimate of total units
- Usage notes: A measure of the proportion of housing units that are in relatively large apartment buildings. See also PctNumUnits1dtchd_*. Larger apartment buildings are usually associated with more dense development and better walkability.

Median number of rooms

- SAS/SPSS variable: MedianRooms_*
- Calculation: Median number of rooms (taken directly from the Census ACS 2010-2014 table)
- Usage notes: A measure of the relative size of housing units in the neighborhood. Fewer rooms per unit is usually associated with more dense development and better walkability.

Percent of units built before 1950

- SAS/SPSS variable: PctBuiltBefore1950_*
- Calculation: Units built 1940-1949 plus units built 1939 or earlier/total housing units
- Usage notes: A measure of the proportion of older homes in the neighborhood. See also PctBuilt1970orLater_*. Older homes are usually associated with more compact urban design and better walkability.

Percent of units built in 1970 or later

- SAS/SPSS variable: PctBuilt1970orLater_*
- Calculation: Units built 2010 or later plus units built 2000-2009 plus units built 1990 to 1999 plus units built 1980 to 1989 plus units built 1970 to 1979/total housing units
- Usage notes: A measure of the proportion of newer homes in the neighborhood. See also PctBuiltBefore1950_*. Newer homes are usually associated with less compact urban design and lower walkability.

Median year structure built

- SAS/SPSS variable: MedianYearBuilt_*
- Calculation: Median year structure was built (taken directly from the Census ACS 2010-2014 table)
- Usage notes: A measure of the general age of the housing structures in the neighborhood. Older homes are usually associated with more compact urban design and better walkability.

Neighborhood factors associated with walkability

- SAS/SPSS variables: NeighFactor1_*, NeighFactor2_*, NeighFactor3_*
  NeighFactor1Tert_*, NeighFactor2Tert_*, NeighFactor3Tert_*
  NeighFactor1Quint_*, NeighFactor3Quint_*, NeighFactor3Quint_*
- Calculation: See Section 2.6 on calculation of walkability factors.
- Usage notes: NeighFactor1_*, NeighFactor2_*, and NeighFactor3_* are continuous measures of neighborhood characteristics often associated with the walkability of the neighborhood derived from 13 individual measures. Factor 1 can be interpreted as an indicator of high-density neighborhoods, factor 2 as older neighborhoods, and factor 3 as neighborhoods with short commutes. The *Tert_* and *Quint_* variables are categorical measures based on the NeighFactor*_* values. A difference in categorical values among buffer configurations for a given location could be due to a relatively steep factor gradient in the neighborhood or because the buffer values are near the cutoff value between categories. See Section 2.6 for additional details.

Exposure Variables:

Ultraviolet (UV) radiation exposure

- SAS/SPSS variables: UV_*, UV_H_Mean, UV_S_Mean
  UVTert_*, UVTert_H_Mean, UVTert_S_Mean
- Calculation: See Section 2.7 on calculation of UV exposure measures.
- Usage notes: UV_*, UV_H_Mean, and UV_S_Mean are continuous estimates of UV radiation exposure in units of watt-hours per square meter (Wh/m^2). The variables UVTert_*, UVTert_H_Mean, and UVTert_S_Mean are categorical measures based on the continuous UV exposure values. Because UV exposure values do not vary significantly at the neighborhood scale, variables with the mean value across all buffer configurations are provided. These variables can be used to avoid apparent UV exposure differences among buffer configurations for a given location because the buffer values are near the cutoff value between categories. See Section 2.7 for additional details.

References

Census 2010. U.S. Census Bureau urban and rural areas, https://www.census.gov/geo/reference/urban-rural.html, accessed 2/3/2017.

Hoehner CM, Handy SL, Yan Y, Blair SN, Berrigan D. Association between neighborhood walkability, cardiorespiratory fitness and body-mass index. *Soc Sci Med.* 2011 Dec;73(12):1707-16.

James P, Berrigan D, Hart JE, Hipp JA, Hoehner CM, Kerr J, Major JM, Oka M, Laden F. Effects of buffer size and shape on associations between the built environment and energy balance. *Health Place.* 2014 May;27:162-70.

NCES 2010. National Center for Education Statistics urban-centric categories, https://nces.ed.gov/ccd/rural_locales.asp, accessed 2/3/2017.

Thornton LE, Pearce JR, Kavanagh AM. 2011. Using Geographic Information Systems (GIS) to assess the role of the built environment in influencing obesity: a glossary. *Int J Behav Nutr Phys Act.* 2011 Jul 1;8:71.

Yost K, Perkins C, Cohen R, Morris C, Wright W. Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. *Cancer Causes Control.* 2001 Oct;12(8):703-11.

Yu M, Tatalovich Z, Gibson JT, Cronin KA. Using a composite index of socioeconomic status to investigate health disparities while protecting the confidentiality of cancer registry data. *Cancer Causes Control.* 2014 Jan;25(1):81-92.