Measured versus Reported Distances in the American Housing Survey

Kwame Donaldson

Matthew Streeter

11/30/2011

SEHSD Working Paper Number 2011-30

U.S. Census Bureau

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, and technical issues are those of the author and not necessarily those of the U.S. Census Bureau.

Contents

Abstract

The American Housing Survey (AHS) collects data on 26 neighborhood quality and amenity questions. These survey questions are primarily intended to appraise the condition and value of the respondent's neighborhood. For example, the AHS asks respondents "Are any railroads, airports, or highways with at least 4 lanes within a half block of your home?" and "Is your public elementary school within one mile of your home?" We used administrative shapefiles from sources such as the U.S. Geological Survey, National Center for Education Statistics, and Environmental Protection Agency in conjunction with the coordinates of the units in the AHS sample to determine the impact of replacing these survey responses with data obtained through Geographic Information Systems (GIS). We individually analyzed three questions along with the effectiveness and availability of each administrative source. These questions asked whether the respondent was within one mile of their public elementary school, 300 feet of the nearest body of water, and 300 feet of the nearest airport, four lane highway, or railroad. For public elementary schools, we found 80.2% of respondent answers agreed with the GIS measurements with a kappa coefficient of “moderate” agreement at 0.56. For bodies of water, the percent agreement was 82.3% with a kappa coefficient at 0.23 or “fair” agreement. Finally, for airports, four lane highways, and railroads, we found 82.5% agreement with a 0.22 kappa coefficient showing “fair” agreement.

Background

Distance measurements to neighborhood amenities are useful in a variety of ways, as neighborhood walkability and community health have become an important topic for residential development in recent years. Within the past decade there has been significant research using distance to parks, public transportation, open spaces, grocery stores, and schools (Tomer, Kneebone and Puentes; Kaczynski, Potwarka and Saelens; Tilt, Unfried and Roca; Moore, Diez Roux and Brines; Green, Smorodinsky and Kim). Additionally, neighborhood proximity questions, whether self-reported or objective, have played a vital role in hedonic pricing model research (Sirmans, Macpherson and Zietz). Thus, increasing the accuracy and decreasing overall respondent burden when collecting these data plays an important part in improving the American Housing Survey.

The American Housing Survey (AHS) is a longitudinal survey examining housing unit and neighborhood quality. Since 1985, the national AHS has interviewed the same housing units every other year. In addition to housing and neighborhood quality, the AHS asks a variety of questions covering a wide range of subjects including household demographics, income, mortgages, and recent mover status.

Many questions asked in the 2009 AHS focused on the proximity of the housing unit to nearby landmarks, buildings, and geographic features. Boolean (yes or no) questions were asked for the presence of features ("Are any of the following features included in your community?") or in reference to a specified distance measurement ("Is that public elementary school within one mile of here?"). This study focuses on identifying the differences between self-reported distances in the AHS and objective measures using GIS in order to evaluate the effects of using GIS as a replacement for items asked in the questionnaire. Because this type of evaluation requires answers based on a uniform distance, only the questions framed using a specific unit of analysis will be used in this analysis.

Most proximity questions in the AHS were removed for the 2011 survey along with the neighborhood quality and observation modules. These questions are currently being considered for reintroduction in future survey supplements. With a consistently increasing national sample, reaching a size of 62,135 units in 2009 and 186,083 units in 2011, each question in the AHS presents a public burden. This burden can be reduced through the incorporation of publicly available GIS shapefiles to obtain the data.

For our purposes, we analyze the impact of replacing these survey responses with administrative data based on three criteria – reliability of the data source, sameness with the self-reported measure, and definitional differences between the question intent and GIS sources used.

First, we analyze the reliability of the data sources used in the potential replacement. This includes research into the completeness of the data, how often it is updated, and how it is used in research. We also address the specific need for updated data and the ease with which it can be used. Two of the three AHS questions we refer to in this research require the use of multiple GIS sources to verify.

The second approach we take in this analysis is to look at how similar the self-reported data are to our GIS measurements. Because the perception that respondents have of their neighborhoods in and of itself can be a useful statistic for some researchers, we want to quantify cases where respondents answer similarly to the distances measured using GIS. Though much of the research shows a lack of agreement between perceived and objective distance measures (McCormack, Cerin and Leslie; Macintyre, Macdonald and Ellaway), we hope to show that implementing GIS technology is an accurate and stable alternative to collecting the data through computer-assisted questionnaires.

Finally, we examine the definitional differences between the question and the GIS sources. The GIS sources used to answer neighborhood proximity questions may not exactly align with the intent of the question. It is important to ensure that the differences between what the question is asking and what the GIS sources are showing will not significantly affect data quality.

Methodology

For each of our analyses, we used a GIS shapefile containing the points sampled in the 2009 American Housing Survey sample. The Census Bureau maintains a Master Address File (MAF) containing geographic information for every housing unit in the United States. The Census 2010 canvassing operation updated this file with GPS coordinates taken by field representatives through the use of hand-held devices. We matched the 2009 national AHS sample to the MAF using the address fields in order to obtain these coordinates. Of the 62,135 cases in the 2009 sample, 52,469 (84.4%) were matched to a record containing coordinates on the MAF.

We then geocoded the remaining 9,666 cases that were missing coordinates after the MAF match. Street maps from the 2010 Topologically Integrated Geographic Encoding and Referencing (TIGER) system were used in conjunction with respondent addresses to approximate coordinates for remaining cases. We created address locators for each state using these street maps, and all cases with valid addresses were matched to these files. Because the TIGER files contain address ranges along street segments, the unit locations were interpolated along each segment. Not all houses are uniformly distributed across a street segment, so coordinates retrieved using this method are not as exact as those obtained using the coordinates on the MAF. This process improved coverage by providing approximated coordinates for 6,894 additional cases.

Out of the 62,135 number of units in our survey, 59,363 were ultimately plotted for this analysis. Many of the cases that were not matched were invalid units identified as "Type B" and "Type C" units – a designation given by the AHS to structures that no longer fit the definition of a housing unit[1]. We then collected the GIS sources used to verify each variable and calculated the distance between the shapefile of our sample units and the source shapefiles.

To perform distance calculations, we used ArcGIS version 9.3. We created a table with an observation representing each unit in the sample. These tables contained the distance, in meters, to the nearest feature on the administrative shapefile. We were then able to import these tables into SAS version 9.2 and merge them with the respondent's answers to our geographic proximity questions.

Public Elementary Schools

AHS Question: Is [the public elementary school for this address] within one mile of here?

Methodology

Public elementary school locations were extracted from the 2008-2009 school year Common Core of Data (CCD) provided by the National Center for Education Statistics (NCES). These data were provided in SAS format and contained latitude and longitude variables for each public school, which we imported into ArcGIS.

[1] Type B cases have a chance to become housing units and come back into the sample, such as a housing unit being converted to a store. Type C units have no chance of coming back into sample, because they were completely destroyed or demolished.

We then removed all observations that were not public elementary schools from the shapefile. The schools dataset initially contained 103,829 schools with coordinates, 54,774 of which contained coordinates that were classified as public elementary schools. In the 2009 AHS survey, 11,347 units had valid responses to the school distance question, of which 10,953 had coordinates.

Source Analysis

The CCD survey is updated annually through State Education Agencies (SEAs) and collects data about all public elementary and secondary schools in the United States. Multiple other surveys use the CCD as a sampling frame, such as the National Assessment of Education Progress, the Early Childhood Longitudinal Study, and the Schools and Staffing Survey. After receiving the data from the SEAs, the Census Bureau appends latitude and longitude onto the CCD file before it is published (Hoffman and Young).

The CCD file is updated annually for every school year. The frequency of data collection is useful for analyzing schools in particular due to regular school additions, conversions, and demolitions. This file has been used to measure school distance at local levels in other published research (Zandbergen, Levenson and Hart).

Definitional Analysis

The intent of the school proximity question ("Is that public elementary school within one mile of here?") differs from the CCD derived shapefile in a couple of ways. The AHS school distance question is asked as a follow-up to the question "Is the public elementary school for this address satisfactory?" The question is specifically aimed at the public elementary school that services the respondent's address. The closest elementary school is not always the one zoned for the address, but due to the difficulty of compiling school zoning information on a national scale, the closest elementary school was used in this analysis.

Additionally, the question is asked only of those who have at least one child age 13 or younger – a group which would be more familiar with school distance. Though we calculated the distance to the nearest elementary school for all units with coordinates, we were only able to analyze those which fit into this universe and gave a valid response to the question.

Similarity Analysis

Table 1: AHS/GIS cross tabulation of public elementary schools

		AHS Response		
		Yes	No	Total
GIS Measurement	Yes	15,948,485 56.8%	3,339,395 11.9% *(type II error)*	19,287,880 68.7%
	No	2,208,387 7.9% *(type I error)*	6,577,916 23.4%	8,786,303 31.3%
	Total	18,156,872 64.7%	9,917,311 35.3%	28,074,183 100.0%

In Table 1, we cross-tabulate the answers that AHS respondents gave to the question “Is that public elementary school within one mile of here?” with the responses that we derived using GIS. We see that 56.8%[2] of AHS respondents answered “yes” to this question and were less than one mile away from the closest elementary school according to the GIS measurement. Additionally, we find that 23.4% of AHS respondents answered “no” to this question and were GIS-measured to be more than one mile away from the closest elementary school. Combining these estimates, 80.2% of AHS responses agree with GIS measurements for this question. Throughout this report, we will refer to the sum of “yes/yes” and “no/no” AHS response/GIS measurement combinations as the percent agreement.

Table 1 also shows that the percent of eligible respondents who reported in the AHS that they are within one mile of their elementary school is 64.7% and the percent of respondents who

[2] Unless otherwise noted, all percentages and counts in this analysis are weighted to account for population units that are sampled with different selection probabilities or varying response rates.

are measured to be within one mile of their closest elementary school is 68.7%. We find the kappa coefficient for this question is 0.56, which indicates a moderate level of agreement between AHS responses and GIS measurements using conventional ranges (Landis and Koch).[3]

Table 1 also shows that the percent of AHS respondents who reported that they are less than one mile away from their public elementary school but are measured to be more than one mile away from the closest school (type I error) is smaller than the percent of respondents who reported that they are more than one mile away from the school but are measured to be less (type II error). As previously noted, the AHS specifically asks the respondent to consider "the public elementary school for this address" which might not be the closest school if the respondent's elementary school is in a different school zone or district. We also note that the GIS-measured distance is a straight-line distance measurement, not the indirect driving distance that the respondent may be contemplating. Both of these considerations imply that the GIS answer for this question will sometimes be "yes" when the AHS respondent says "no" and help explain why the type II error exceeds the type I error.

In Figure 1, we show how the overall percent agreement varies in relation to the distance from the closest elementary school. In this analysis, we group AHS respondents into deciles based on the distance from their residence to the nearest public elementary school along the horizontal axis, and we plot the percent agreement for each decile along the vertical axis. Figure 1 shows that 93.0% of AHS respondents who were 342 meters or less (i.e., the first decile of the GIS straight-line distance) from the closest elementary school said "yes" in response to this

[3] Kappa coefficients, which are also known as Cohen's kappa coefficients, measure the level of agreement between two respondents who give binary responses by taking into account the level of agreement that occurs by chance. This measurement discounts one-sided situations in which both respondents give the same single response in the vast majority of cases (i.e., both respondents almost always answer "yes" or both respondents almost always answer "no"). Conventional ranges are: Poor: Less than 0; Slight: 0.00 – 0.20; Fair: 0.20 – 0.40; Moderate: 0.40 – 0.60; Substantial: 0.60 – 0.80; Almost Perfect: 0.80 – 1.00.

question, and 91.0% of respondents who were 5,150 meters or more (the tenth decile) from the closest elementary school said “no.” However, only 54.2% of respondents (slightly more than the rate that we would expect by chance) who were between 1,360 and 1,843 meters (the seventh decile) provided an answer that agreed the GIS findings.

Figure 1: Percent agreement grouped by GIS-measured distance to nearest public elementary school

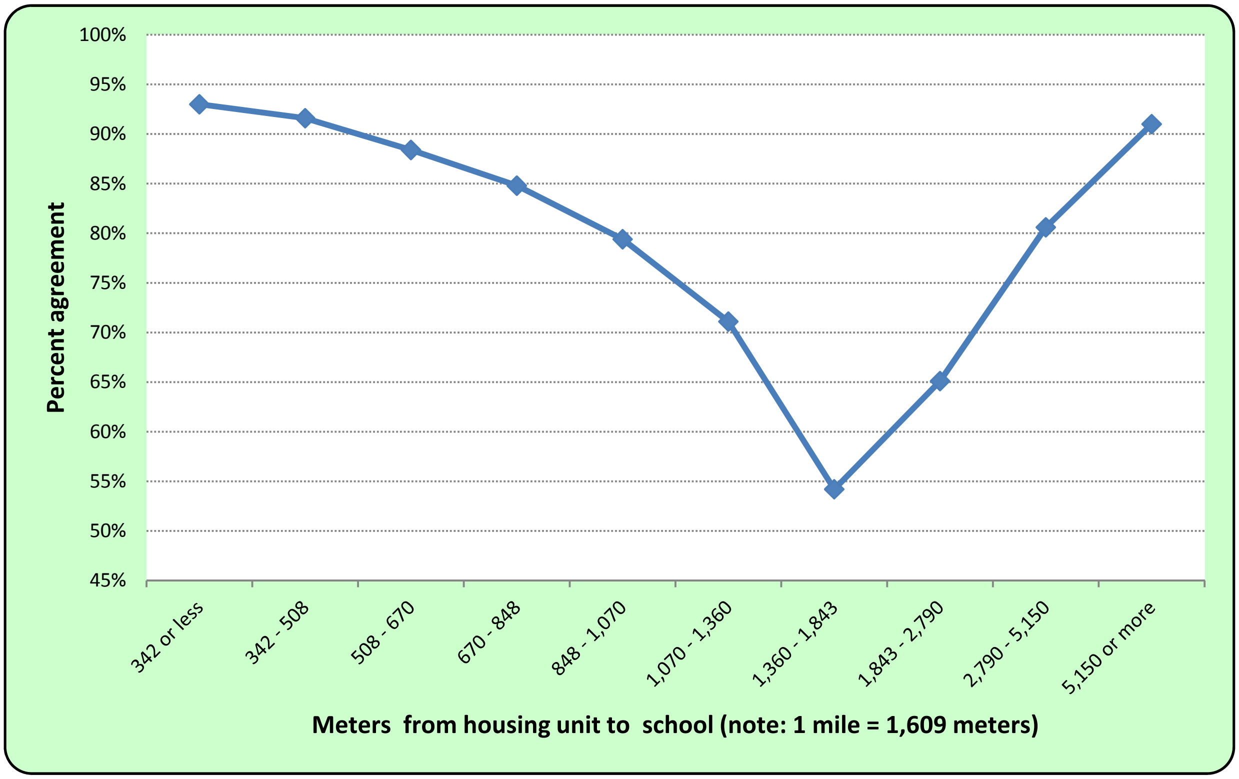
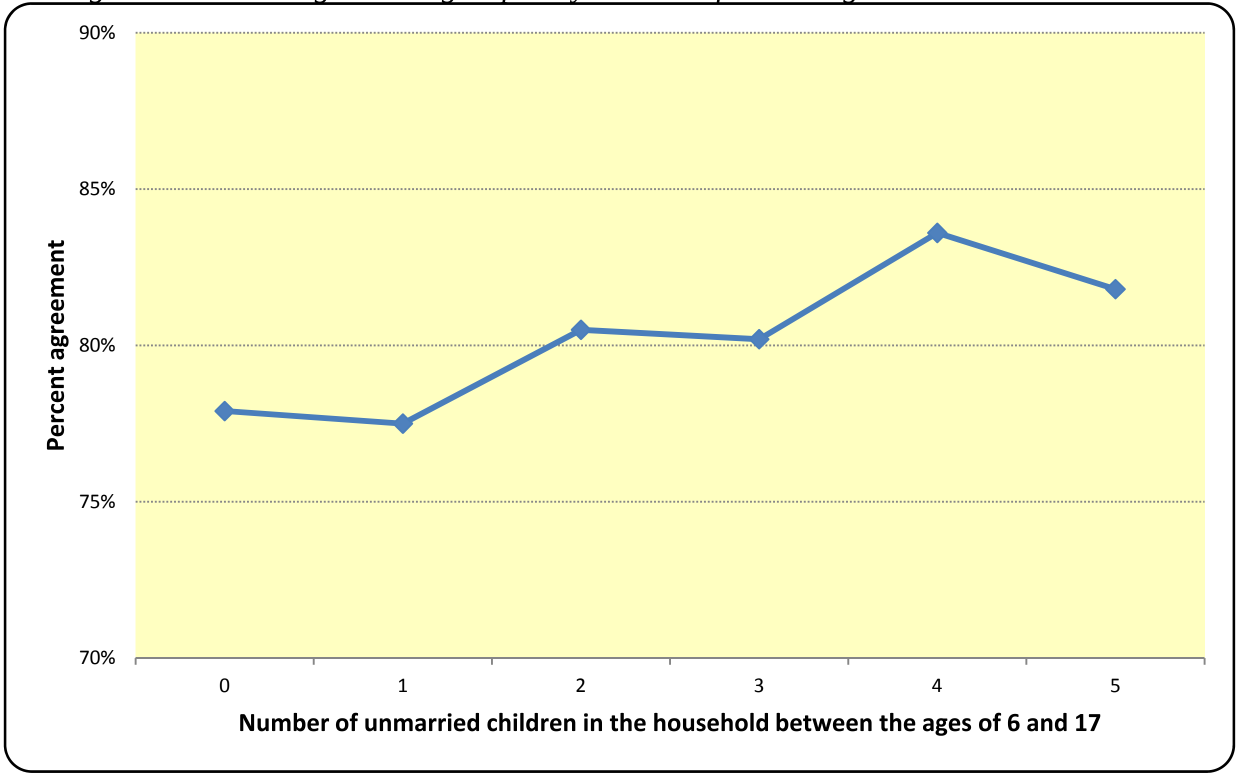


Figure 1 illustrates a clear pattern: the percent agreement reaches its maximum value in the first decile; then it falls and lands at its minimum value in the seventh decile; and then it climbs back to near-peak levels in the final decile. Since one mile equals 1,609 meters (a point within the seventh decile’s range), we can conclude that the type I and type II errors are mainly caused by a discrepancy between the respondent’s interpretation of the one mile cutoff and the GIS straight-line distance measurement.

A second important pattern is illustrated in Figure 2 where we find that households with more children are also more likely to provide AHS responses that match our GIS measurements.

Seventy-eight percent of AHS households with no school-aged children provided responses that match our GIS measurements compared with 84% of households with four school-aged children. The Pearson's correlation between these two series is 0.866. If we assume that respondents in families with more children can more reliably gauge the distance to the elementary school (because members of these households have made the trip to the school more regularly and frequently), then this finding suggests that the GIS measurement is a more reliable source of this distance measurement than the typical AHS respondent.[4]

Figure 2: Percent agreement grouped by number of school-aged children in household



[4] Other reasons that are unrelated to this more-frequent-trips-to-school hypothesis can explain why the percent agreement increases for families with more school-aged kids. For example, we see from Table 1 that the percent agreement for respondents who are less than one mile from the closest elementary school (i.e., the "yes/yes" respondents divided by all GIS "yes" respondents, 82.7%) is greater than the percent agreement for respondents who are more than one mile away (i.e., the "no/no" respondents divided by all GIS "no" respondents, 74.8%). If families with school-aged kids are also more likely to locate within one mile of an elementary school, then the pattern illustrated in Figure 2 might simply be a result of this neighborhood preference. A more thorough, multi-variable analysis is required to make a more conclusive case.

Bodies of Water

AHS Question: Are there any bodies of water, such as ponds, lakes, rivers, or the ocean within a half block [300 feet] of [your home]?

Methodology

We used three separate water boundary sources to derive distance measurements for bodies of water – Streams and Waterbodies of the United States from the United States Geological Survey (USGS), U.S. Water Bodies from Environmental Systems Research Institute (ESRI), and U.S. Rivers and Streams from ESRI.

For each housing unit, we created distinct variables representing the distance to the nearest feature on each of these data sources. The source with the smallest distance was kept and used for the analysis. Of the 62,135 cases in sample, 43,250 of them were occupied and had both a set of coordinates and a valid answer to the water distance question. Though this question was asked of respondents representing usual residence elsewhere (URE) and vacant units as well, we excluded those units from our analysis.

Source Analysis

The Streams and Waterbodies of the United States file includes two separate files for polygons and lines. The features covered include major streams and rivers, canals, aqueducts, lakes, reservoirs, marshes, glaciers, bays, oceans, waterfalls, dams, and channels. Glaciers, marshes, and any streams referred to as “dry” were removed for our analysis. The resulting polygon shapefile contained 14,093 distinct features while the line shapefile contained 76,610. USGS created this file by combining state level hydrography files from Digital Line Graph data[5].

The U.S. Rivers and Streams file is a linear file published by ESRI containing water features that comprise the surface water drainage system of the United States. The file contains 2,955,059 distinct features and is meant for use at a scale of 1:24,000. The U.S. Water Bodies file is a polygon file published by ESRI containing major lakes, reservoirs, rivers, lagoons, and estuaries. For both files, the version we used in our research was created in 2004 under the coordination of USGS, the Environmental Protection Agency, and ESRI. The file contains 463,591 distinct features, of which 393,440 we determined were relevant bodies of water excluding features such as marshes and glaciers.

Though information we collected from these shapefiles dated back to 2003 and 2004, natural water features as a whole do not change as often as other man-made features such as roads and schools, so we would not need to update this dataset quite as frequently. USGS follows the United States National Map Accuracy Standards outlined by the U.S. Bureau of the Budget (U.S. Geological Survey).

Definitional Analysis

Unlike the school distance question, water distance was asked of all housing units that were interviewed. The question text for water features includes examples such as ponds, lakes, rivers, and the ocean. Respondents are asked to exclude features such as swimming pools, bird baths, and temporary pools of water. There are many types of water features that the question

[5] The Digital Line Graph is a digital map file produced by USGS at varying scales.

did not address specifically to be included or excluded. For instance, some respondents may consider local common fountains as a body of water, and these are not included in the sources.

Additionally, though we excluded obviously out-of-scope features such as glaciers and marshes, many of the streams and rivers included in the sources were intermittent, meaning that while the features carry water for a significant amount of time, they are dry during part of the year. Some of these water features may be dry at the time of the interview, or the respondent may consider it temporary if it is dry more often than not. Table 2 shows the distribution of water feature types found closest to each unit in the 2009 AHS sample using descriptions provided by the shapefiles. On this table, we see that intermittent and perennial streams, rivers, and ponds are the closest water feature to 71.6% of units. By definition, some of these features are not always filled with water. Additionally we see 9.2% of respondents are closest to artificial paths and 5.1% are closest to a canal or ditch. Some respondents may not consider these features as bodies of water.

Table 2: Water Feature Types Closest to AHS Units

Feature Type	**Percent of Cases**	**Feature Type**	**Percent of Cases**
Perennial Stream/River	33.23%	Bay or Estuary or Ocean	0.33%
Intermittent Stream/River	24.04%	Water Storage Reservoir	0.23%
Perennial Lake/Pond	14.28%	Treatment Reservoir	0.20%
Artificial Path	9.19%	Not Controlled Inundation Area	0.13%
Stream	6.27%	Controlled Inundation Area	0.05%
Canal/Ditch	5.07%	Aqueduct Pipeline	0.04%
Shoreline	3.57%	Apparent Limit	0.04%
Right Bank	0.96%	Siphon Pipeline	0.04%
Intermittent Lake/Pond	0.49%	Reservoir	0.03%
Connector	0.49%	Aquaculture Reservoir	0.03%
Canal	0.46%	Intracoastal Waterway	0.02%
Lake	0.42%	Dam	0.02%
Left Bank	0.38%	Aqueduct	0.01%

Similarity Analysis

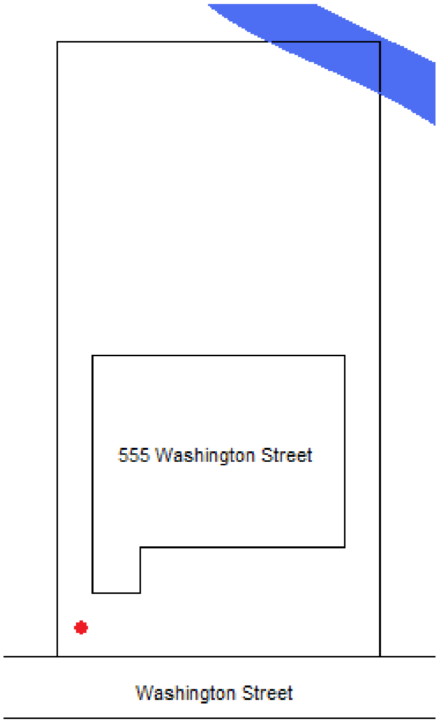
Table 3: AHS/GIS cross tabulation of bodies of water

		AHS Response		
		Yes	No	Total
GIS Measurement	Yes	4,602,637 4.3%	5,820,556 5.4% *(type II error)*	10,423,193 9.7%
	No	13,107,306 12.3% *(type I error)*	83,441,608 78.0%	96,548,914 90.3%
	Total	17,709,943 16.6%	89,262,164 83.4%	106,972,107 100.0%

In Table 3, we cross-tabulate the answers that AHS respondents gave to the question “Are there any bodies of water within a half block [about 300 feet] of your home?” with the responses that we derive using GIS data and tools. We find that 4.3% of AHS respondents answered “yes” to this question and were less than 300 feet from a body of water according to the GIS measurement. We also see that 78.0% of AHS respondents answered “no” to this question and were more than 300 feet away from the nearest body of water. Combining these estimates, we calculate that the overall percent agreement for this question is 82.3%.

While the AHS response and GIS measurement for this question agree in nearly five out six cases, this agreement is primarily because the vast majority of AHS respondents are both self-reported and GIS-measured to be more than 300 feet from a body of water. However, Table 3 shows that the percent of respondents who reported being less than 300 feet from a body of water (16.6%) is greater than the proportion of respondents that GIS measures to be within 300 feet of a coastline, shoreline or bank (9.7%). Because of this difference, the kappa coefficient for this question is 0.23, which indicates a fair level of agreement between AHS responses and GIS measurements.

Figure 3: Difference between AHS response and GIS measurement



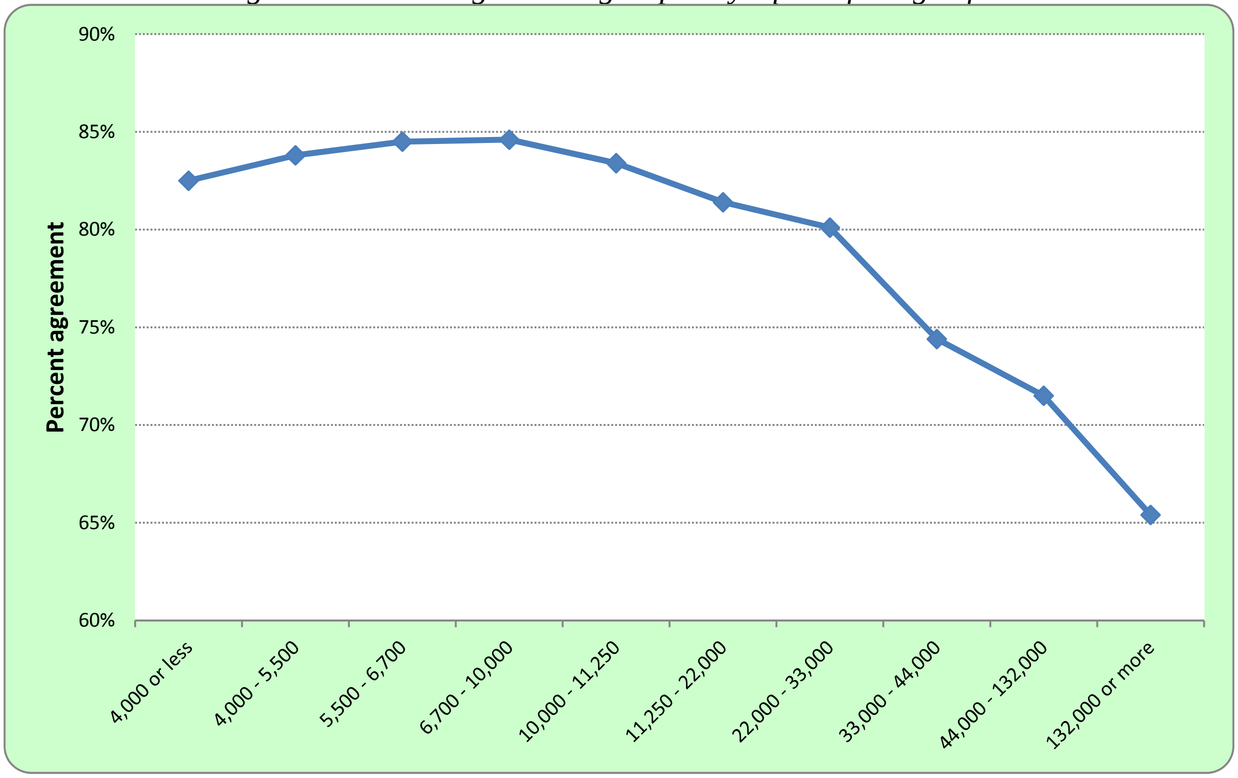
We explain this fair level of agreement with the finding in Table 3 that 12.3% of respondents reported that they are within 300 feet of a body of water but our GIS measurements indicate otherwise (type I error). Some of this type I error is understandable. In GIS, the location of AHS respondents is represented by points and many smaller rivers and streams are symbolized with lines. Neither points nor lines have width. In effect, our GIS tools are measuring from a single point usually near the front of the respondent's property to the closest point on a line near the middle of a river or stream. By contrast, AHS respondents may be thinking about the distance from the border of their property to the nearest edge of the shore in high tide. In such cases, the AHS respondent might reasonably report that their residence is less than 300 feet away from a body of water even though GIS reports otherwise. This discrepancy helps explain why the type I error exceeds the type II error for this question.

We illustrate this difference between the AHS respondent's answer and GIS measurement in Figure 3. The dot represents the coordinate where our GIS data have pinpointed the respondent's property near the front of 555 Washington Street. The shape at the rear of the property represents a river that runs through it. The AHS respondent at 555 Washington Street would certainly report that she is within 300 feet of a body of water, but the GIS measurement of the distance from the red dot to the closest point on the river might report otherwise.

It is obvious from Figure 3 that respondents in homes on the largest lots are most exposed to the AHS response/GIS measurement discrepancy described in the previous two paragraphs. If 555 Washington Street is an extremely large estate, then this error is nearly certain to occur. In Figure 4, we group respondents into deciles based on the square footage of their lots and plot the percent agreement for each decile. We see that the percent agreement among respondents with the smallest lots is uniformly high, but this agreement abruptly falls among respondents with residences on the largest lots. In the three highest deciles, 75.2% of this disagreement is due to AHS respondents reporting that they are close to a body of water and GIS measuring otherwise (type I error).[6]

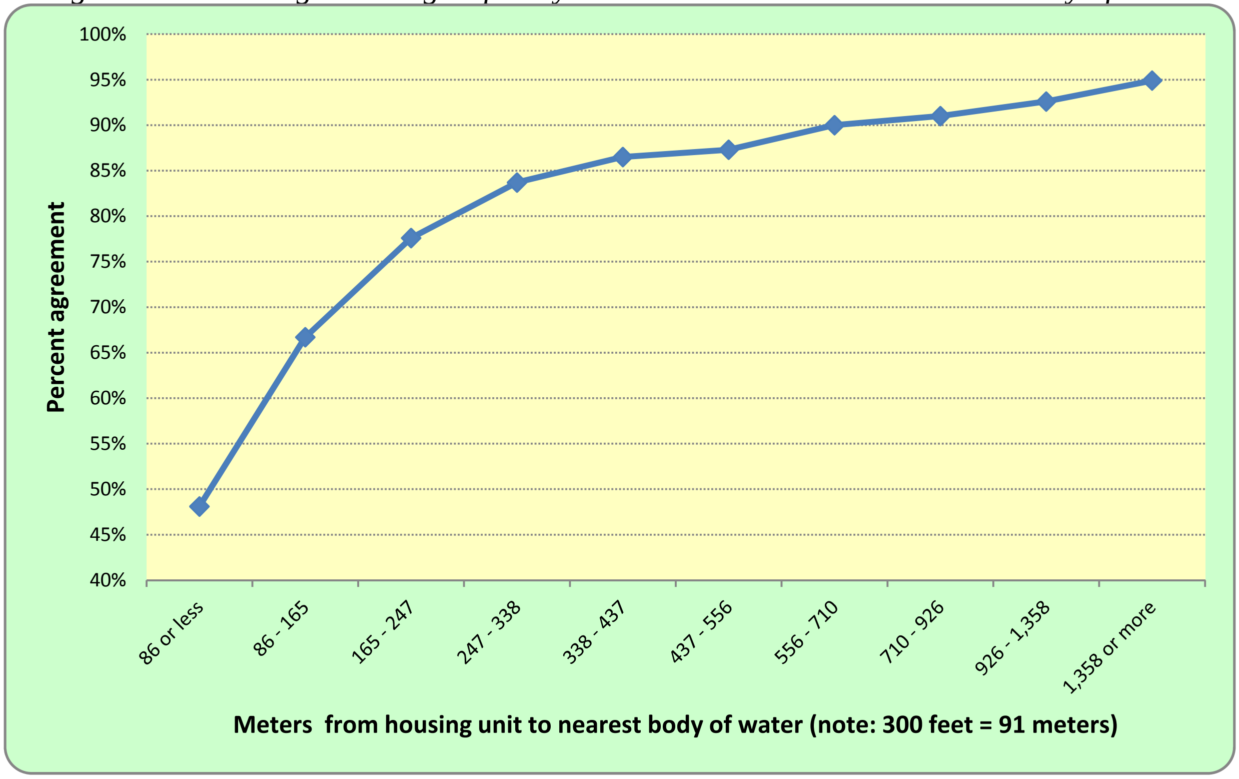
[6] For housing units on lots over 33,000 sq. ft, the AHS estimates that 4,724,741 households report that they are closer than 300 feet to a body of water and GIS measures them to be farther away, while 1,562,237 households report that they are farther than 300 feet from a body of water and GIS measures them to be closer.

Figure 4: Percent agreement grouped by square footage of lot



In Figure 5, we group respondents into deciles based on the GIS-measured distance from their residence to the closest body of water, and we plot the percent agreement for each decile. We see that the percent agreement is lowest in the first decile where, according to our GIS data, every respondent is within 300 feet (91 meters) of a body of water. However, less than half of AHS respondents in the first decile agree with this GIS measurement, which is an exceptionally low level of agreement. As mentioned previously, we include a broad range of water features in our definition of "body of water" – many AHS respondents may not recognize some of these features as legitimate bodies of water or might not realize that these water features exist near them (see Table 2).

Figure 5: Percent agreement grouped by GIS-measured distance to nearest body of water



Beyond the first decile in Figure 5, we also find the percent agreement is below average for AHS respondents who live between 86 and 338 meters (the second, third, and forth deciles) from the closest body of water. When there is disagreement between the AHS response and GIS measurement in these deciles, it is because AHS respondents reported that they are within 300 feet of a body of water and the GIS measurements reported otherwise. We have already shown that some of this discrepancy is due to limitations in the GIS data. By contrast, the percent agreement in the four deciles that are farthest from any body of water (556 meters or more based on the GIS measurement) is high – between 90% and 95%.

Railroads, Airports, and Highways

AHS Question: How about any railroads, airports, or highways with at least 4 lanes -- any of these within a half block of [your home]?

Methodology

Distance calculations for airports, railroads, and four lane highways were taken from three separate shapefiles – one for each type of feature. Features in the airports and railroads files were not deleted for the purpose of the calculation, but the four lane highway file was trimmed to exclude any roads with fewer than four lanes listed.

After researching government highway standards outlined by the Federal Highway Administration (FHWA), we assumed 3.65 meters as a minimum lane width for highways and railroads (Stein and Neuman). Assuming the line files were in the center of the highway or railroad, we created a 1.825 meter buffer surrounding our railroads file, and created a buffer surrounding the highway file equal to 1.825 multiplied by the number of lanes.

The distance to the nearest feature for each file was calculated separately and extracted as separate variables. For our analysis, we used the distance that was smallest of the three. Of the 62,135 cases in the 2009 national AHS sample, 42,491 of them were occupied and had both a set of coordinates and a valid answer to the transportation distance question. Though this question was asked of respondents representing URE and vacant units as well, we excluded those units from our analysis.

Source Analysis

To calculate distances to airports, we used the U.S. Airports shapefile published by ESRI. The airport locations were compiled by Tele Atlas North America, Inc. and ESRI, and included data corresponding to February 2007. This file includes polygon boundaries of airports and their runways. This file was preferable to us over the Airports of the United States file compiled by the National Transportation Atlas Database, Bureau of Transportation Statistics, and USGS because its features are polygons rather than points. With features as potentially large as airports, we wanted to run our distance calculation accounting for the full boundary rather than using a point in the middle of the airport that could be over 300 meters from the airport boundary. The file contains 5,745 airports.

The railroads file we used was the Railway Network provided by the 2010 National Transportation Atlas Database. This dataset was collected by the Federal Railroad Administration (FRA). Because these data are for 2010, there is a small chance that some of the newest railroads were not complete at the time of interview in 2009. These data are published annually, and are available on-line starting in 2010. For this reason, this is a stable dataset that can easily be updated for future survey years. The railroad file contains 172,888 railroad features.

The roads file we used was the National Highway Planning Network from the 2010 National Transportation Atlas Database. This shapefile was created by the Federal Highway Administration (FHWA). It contains highways, rural and urban arterials, and routes from the National Highway System. It is used by the FHWA for highway planning and policy analysis. The metadata provided with the shapefile states that the locations of the railroads are accurate within 80 meters. After we finished trimming the shapefile for the purposes of our research, we made use of 75,755 highway segments.

Definitional Analysis

In the case of airports, railroads, and four lane highways, there aren’t any significant definitional differences between the source files and the intent of the question apart from the fact that the respondent will most likely consider the distance from their property to the edge of the nearest railroad or highway. We attempted to remedy this using 1.825 meter buffers around the line files for highways and railroads, though in practice this distance may not be the same for all types of highways and railroads.

Similarity Analysis

In Table 4, we cross-tabulate the answers that AHS respondents gave to the question "Are railroads, airports or 4-lane highways within a half block [about 300 feet] of your home?" with the responses that we derive using GIS. We see that 3.6% of AHS respondents answered "yes" to this question and were less than 300 feet away from a major transportation mode according to the GIS measurement. We also find that 78.9% of AHS respondents answered "no" to this question and were more than 300 feet away from the closest railroad, airport or highway. Combining these estimates, we see that the overall percent agreement for this question is 82.6%.

Table 4: AHS/GIS cross tabulation of closest railroad, airport or highway

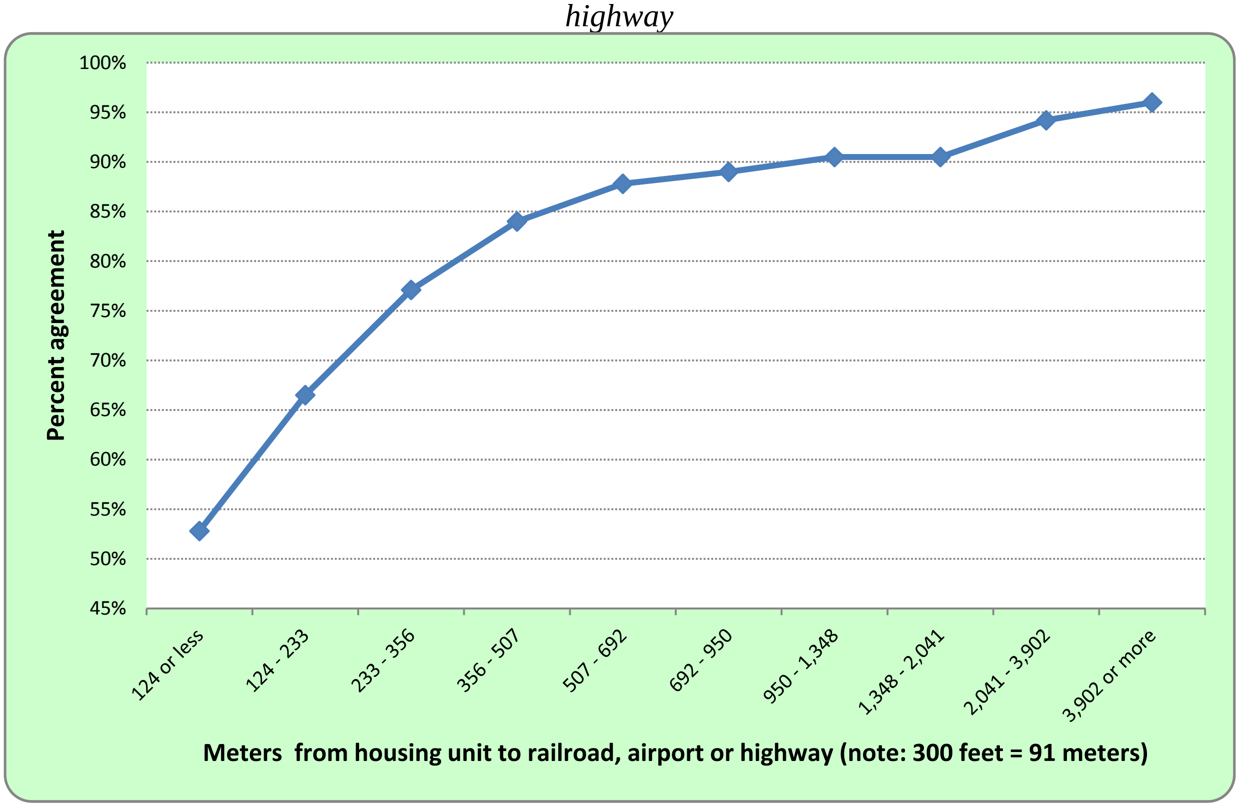
		AHS Response		
		Yes	No	Total
GIS Measurement	Yes	3,796,062 3.6%	3,264,573 3.1% *(type II error)*	7,060,635 6.7%
	No	15,049,944 14.3% *(type I error)*	82,814,524 78.9%	97,864,468 93.3%
	Total	18,846,006 18.0%	86,079,097 82.0%	104,925,103 100.0%

While the overall percent agreement is high, most of this agreement is because the overwhelming majority of AHS respondents are both self-reported and GIS-measured to be more than 300 feet from a railroad, airport or highway. However, Table 4 shows that the percent of respondents who reported that they are less than 300 feet from a railroad, airport or highway (18.0%) is greater than the percent of respondents that GIS measures to be less than 300 feet from these transport modes (6.7%). Because of this difference, the kappa coefficient for this question is 0.22, which indicates a fair level of agreement between AHS responses and GIS measurements.

We can attribute this fair amount of agreement to the finding in Table 4 that 14.3% of respondents reported that they are within 300 feet of a railroad, airport or highway but are measured to be more than 300 feet away (type I error). As with proximity to bodies of water, some of this error can be explained because our GIS data symbolize the location of AHS respondents using points, which lack dimensions. Our GIS tools are measuring from a single point usually near the front of the respondent's property to a point on the airport's property or the path of the highway or railroad. By contrast, the AHS respondent may be contemplating the distance from the border of their property to the nearest edge of the highway or railroad. In such cases, the AHS respondent might understandably report that their residence is less than 300 feet away from a transportation mode when GIS measures them to be more 300 feet away.

Some of this inconsistency is illustrated in Figure 6, where we group respondents into deciles based on their distance to the nearest railroad, airport or highway, and we plot the percent agreement for this question for each distance decile. We see that the percent agreement is lowest in the first decile, where only 52.8% of AHS respondents provided a response that matched the answer that we measured using GIS. This first decile includes all of the AHS respondents who are measured to be less than 300 feet away from a railroad, airport or highway (type II error), and it also includes many respondents who reported that they are less than 300 feet away from these modes of transportation, but GIS measures them to be slightly farther than that (type I error). The second decile, where the percent agreement is 66.5%, also includes many AHS response/GIS measurement combinations in the type I error category. As we have seen before, the percent agreement increases when the respondent is farther away from the threshold distance in the survey question. Among respondents in each of the four highest deciles, where residences are GIS-measured to be more than 950 meters (0.6 miles) away from a railroad, airport or highway, the percent agreement exceeds 90%.

Figure 6: Percent agreement grouped by GIS-measured distance to nearest railroad, airport or highway



Despite some inherent issues with our GIS datasets, we have other data which suggests that respondents may be overestimating how close they are to railroads, four-lane highways or airports. In Table 5, we estimate the land area of urban areas and these transit modes within urban areas in the United States. Urban areas in the United States are core census block groups or blocks that have a population density of at least 1,000 people per square mile and surrounding census blocks that have an overall density of at least 500 people per square mile. These densely settled areas occupy 244,000 sq. km or approximately 3% of the land area in the United States and contain 79.2% of the occupied housing units.[7] Within urban areas, we estimate that railroads, airports, and four-lane highways occupy a total of 3,361 sq. km or just 1.4% of the land area. These land area estimates suggest that the GIS measurement of respondents who are within 300 feet of a major transportation mode (6.2%) is probably more accurate than the AHS "yes" response to this question (18.0%).

Table 5: Size of urban areas, railroads, airports, and highways in the United States

Approximate total land area of urban areas in the United States	244,000 sq. km
Approximate area of railroads within urban areas in the United States	161 sq. km[8]
Approximate area of airports within urban areas in the United States	2,358 sq. km
Approximate area of four-lane highways within urban areas in the United States	842 sq. km[9]
Percent of land area within U.S. urban areas that is occupied by railroads, airports or four-lane highways.	**1.4%**

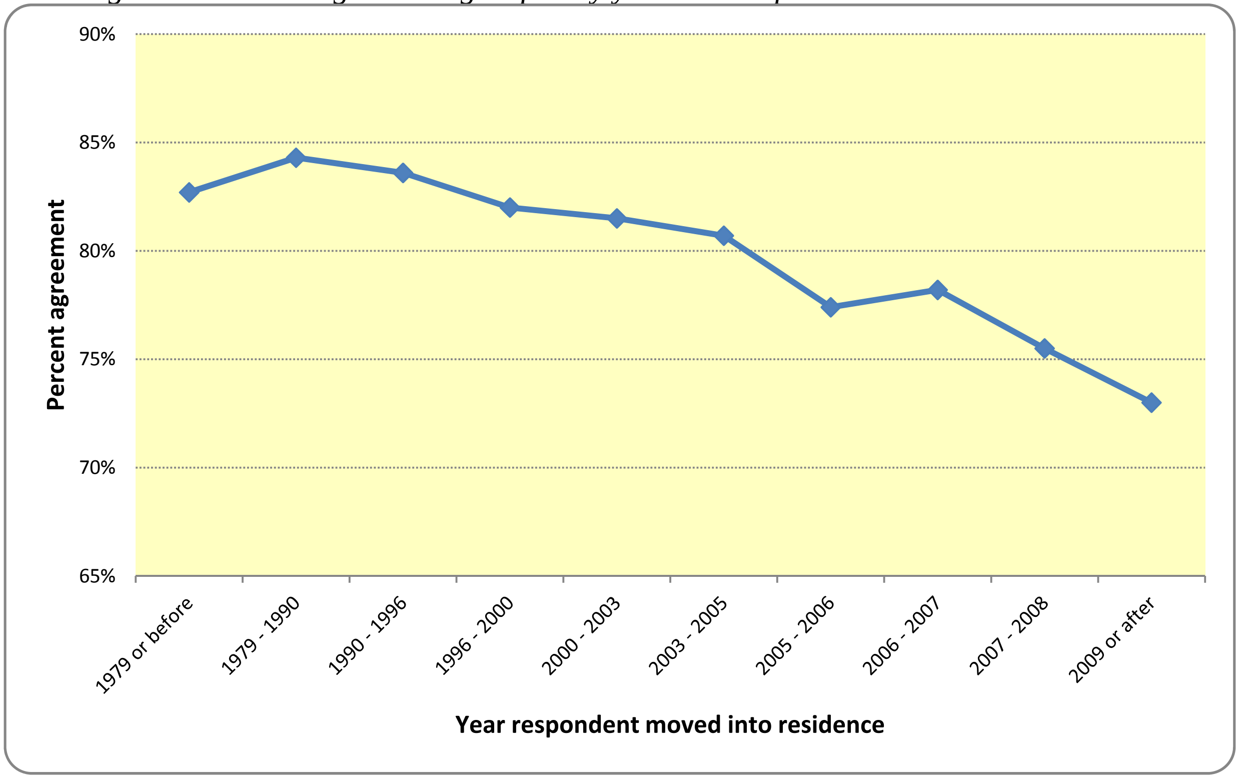
In Figure 7, we categorize respondents into deciles based on the year the respondent moved into the housing unit and plot the percent agreement for these move-year deciles. Figure 7 presents a clear pattern: AHS respondents who have lived in their residence for longer periods provided responses that match our GIS findings at higher rates. For example, respondents in the second decile (who moved into the residence between 1979 and 1990) provided matching responses 84.3% of the time, whereas respondents in the tenth decile (who moved into their residence in 2008 or later) gave matching answers 73.0% of the time. The Pearson's correlation between the percent agreement and the decile rank in Figure 7 is -0.933. If we assume that respondents who have lived in their residence for longer periods give more reliable responses about their community and surroundings, then the pattern in Figure 7 provides evidence that the GIS measurement for this question is more reliable than the typical respondent's answer.

[7] U.S. Census Bureau, Table GCT-H1 – Total and Occupied Housing Units for Urban/Rural and Metropolitan/Nonmetropolitan Areas: 2000

[8] This calculation assumes railroads are 3.65 m (12 ft) wide.

[9] This calculation assumes that one highway lane is 3.65 m (12 ft) wide.

Figure 7: Percent agreement grouped by year that respondent moved into residence



Discussion

Using GIS derived data carries the potential of saving interview time in the field. In 2009, the AHS spent a total of 118,687 seconds (1.4 days) on the schools distance question, 331,012 seconds (3.8 days) on the water distance question, and 416,179 seconds (4.8 days) on the transportation distance question – this totaled to 865,878 seconds. All totaled, field representatives and survey respondents spent more than 10 days on just the three questions reviewed in this research. This is a small fraction of the potential time savings; there are 23 additional questions that we may be able to replace using GIS.

However, there were many limitations to this study. First, we calculated all of our distances as straight lines. This did not account for the use of common walking or driving routes, which could show substantial differences that could disproportionately affect certain subgroups of the population (McKenzie). This may have had some affect on the longer distance questions such as “Is your public elementary school within one mile of here?” However, this difference may not be as important for questions dealing with shorter distances, such as those focusing on amenities within 300 feet of the unit, and research that directly compares route distances to straight-line Euclidean distances have found that the results are similar (Sparks, Bania and Leete).

In addition, there are some instances where using line files for large features could introduce some amount of error when comparing to short range distance questions. While large features such as major rivers were represented as polygons, some features such as schools and smaller rivers and streams used lines and points with no width or area. The housing units we used in the distance calculations were points. Because of this, respondents with larger lot sizes could have more variation in their true distance because they may measure it from the end of their lot rather than the center of their unit.

Third, a more thorough analysis would compare previous iterations of the data against the older versions of the sources from the time of their publication. This is one way to evaluate the consistency of our GIS sources across time and see how they match up in different years. We could use this trend to help predict how stable and useful the source will be in future enumerations.

Finally, we did not research the potential value obtained from a perceived distance versus an objectively measured one. It stands to reason that perceived distance to various amenities could have an effect on other subjective variables, such as self-reported home value, and perception alone could be valuable to some researchers. We have the capability to assess this using 2009 data by comparing self-reported and GIS measured answers and their relationships with other characteristics such as value data.

Future Research

Distances to schools, water features, and transportation account for a small part of the potential of this research. Our methods can be applied to other questions in the AHS survey as well. The second phase of this research will focus on distances to parks, woods, farms, or ranches. We plan to use the 2006 National Land Cover Database provided by the Multi-Resolution Land Cover Consortium along with the Protected Areas Database from the Conservation Biology Institute to analyze distances to these open spaces.

The AHS also asks questions about the buildings within 300 feet -- specifically regarding the structure type and age of the surrounding buildings We also plan to extend our analysis to each unit's distance to the nearest detached homes, townhouses, apartment buildings, and mobile homes using the structure type variable of surrounding units found on the MAF. We will use MAF coordinates to locate the nearest unit of each structure type to the units in the AHS sample.

The 2009 AHS also included questions in the neighborhood quality module that asked whether the respondent has various neighborhood amenities in his or her community. These amenities include features such as clubhouses, golf courses, and jogging trails. In future research, we will use the AHS respondents' answers along with GIS datasets to determine how people define their communities in relation to these amenities. We will construct this definition by pinpointing the distance that maximizes the percent agreement between the AHS response and GIS measurement for each community amenity question. We will also apply this definition to various geographic, socioeconomic, and demographic subsets of the data to determine if some groups define “community” more broadly or narrowly than others. By providing a better understanding of what “community” means to various groups, this research can be used by policy makers and analysts who implement programs or conduct research that targets specific communities.

Our future research is not limited to the differences between measured and reported distances. Using GIS, we are able to find numeric distances rather than the binary responses recorded through self-reported measures. This gives us the flexibility to better model the relationships between distances to neighborhood features and other variables in the AHS survey. Though this information would not be accessible to the public, it could also provide unique research opportunities for internal data users in the future. We could even look into the potential of recoding distance variables into categories and publishing it for public use.

References

Canter, David and Stephen K. Tagg. "Distance estimation in cities." Environment and Behavior 7.1 (1975): 59-80.

Coppock, J. T. and D. W. Rhind. "The history of GIS." Geographical Information Systems: Principles and Applications 1 (1991): 21-43.

Gebel, Klaus, Adrian Bauman and Neville Owen. "Correlates of non-concordance between perceived and objective measures of walkability." Annals of Behavioral Medicine 37 (2009): 228-238.

Gebel, Klaus, et al. "Mismatch between perceived and objectively assessed neighborhood walkability attributes: Prospective relationships with walking and weight gain." Health & Place 17 (2011): 519-524.

Gilinsky, Alberta S. "Perceived size and distance in visual space." Psychological Review 58 (1951): 460-482.

Green, Rochelle S., et al. "Proximity of California Public Schools to Busy Roads." Environmental Health Perspectives 112.1 (2004): 61-66.

Hoffman, Lee M. and Beth Aronstamm Young. "Changes in the use of education administrative records: The common core of data." National Center for Education Statistics, U.S. Department of Education, 2001.

Kaczynski, Andrew T., Luke R. Potwarka and Brian E. Saelens. "Association of park size, distance, and features with physical activity in neighborhood parks." American Journal of Public Health 98.8 (2008): 1451-1456.

Landis, J. Richard and Gary G. Koch. "The Measurement of Observer Agreement for Categorical Data." Biometrics 33.1 (1977): 159-174.

Macintyre, Sally, Laura Macdonald and Anne Ellaway. "Lack of agreement between measured and self-reported distance from public green parks in Glasgow, Scotland." International Journal of Behavioral Nutrition and Physical Activity (2008).

McCormack, Gavin R., et al. "Objective versus perceived walking distances to destinations: correspondence and predictive validity." Environment and Behavior 40.3 (2008): 401-425.

McKenzie, Brian S. "Transit Access and Labor Market Outcomes across Segregated Neighborhoods." 2011.

Moore, Latetia V., Ana V. Diez Roux and Shannon Brines. "Comparing perception-based and geographic information system (GIS)-based characterizations of the local food environment." Journal of Urban Health: Bulletin of the New York Academy of Medicine 85.2 (2008).

Pundt, Hardy and Klaus Brinkkötter-Runde. "Visualization of spatial data for field based GIS." Computers & Geosciences 26 (2000): 51-56.

Sirmans, Stacy G., David A. Macpherson and Emily N. Zietz. "The composition of hedonic pricing models." Journal of Real Estate Literature 13.1 (2005): 3-43.

Sparks, Andrea L., Neil Bania and Laura Leete. "Comparative Approaches to Measuring Food Access in Urban Areas: The Case of Portland, Oregon." Urban Studies (June 2011): 1715–1737.

Stein, William J. and Timothy R. Neuman. "Mitigation Strategies for Design Exceptions." U.S. Department of Transportation, 2007.

Tilt, Jenna H., Thomas M. Unfried and Belen Roca. "Using objective and subjective measures of neighborhood greenness and accessible destinations for understanding walking trips and BMI in Seattle, Washington." American Journal of Health Promotion 21.4 (2007).

Tomer, Adie, et al. "Missed Opportunity: Transit and Jobs in Metropolitan America." Brookings Institution, 2011.

U.S. Geological Survey. "Map Accuracy Standards." USGS Fact Sheet 171-99. November 1999.

United States House of Representatives. Subcommittee on Information Policy, Census, and National Archives Committee on Oversight and Government Reform. "2010 Census: Master Address File, Issues and Concerns." 21 October 2009.

Wolf, Jean, Randall Guensler and William Bachman. "Elimination of the travel diary: An experiment to derive trip purpose from GPS travel data." Transportation Research Board (2001): 125-134.

Zandbergen, Paul A., Jill S. Levenson and Timothy C. Hart. "Residential proximity to schools and daycares: An empirical analysis of sex offense recidivism." Criminal Justice and Behavior 37.5 (2010): 482-502.