COMMUTING PATTERNS AND THE HOUSING STOCK

November 20, 2005

The research presented here was performed under the “AHS Analytical Support” Contract issued by the Office of Policy Development and Research at the U.S. Department of Housing and Urban Development (HUD).

The work was performed by ICF Consulting (http://www.icfconsulting.com) and Econometrica, Inc. (http://www.econometricainc.com). Kevin S. Blake and William Cowart of ICF Consulting were the principal authors. ICF Consulting staff who worked on this report included Linda Bailey, Kevin S. Blake, Mariana Carrera, William A. Cowart, Liisa Ecola, Joshua Leftin, and Aleksandra Simic, Ph.D.

The authors would like to thank David A. Vandenbroucke of HUD, Gregory J. Watson of The Moran Company, and Frederick J. Eggers, Ph.D. of Econometrica, Inc. for their oversight and comments.

The views and findings presented here represent those of the authors only, and should not be construed to necessarily reflect those of HUD.

Table of Contents

Executive Summary

The prime objective of this research was to explore the applicability of transportation and commute-related variables in the American Housing Survey (AHS) to analyzing the relationship between the housing stock and commuting patterns. Particular attention was given to analyzing the usefulness of the AHS data in testing the spatial mismatch hypothesis.

We determined that while the AHS in its current form may contribute marginally to the spatial mismatch discussion, it would be more important to have improved commuting data. These improved data could enable the AHS to contribute significantly to other important discussions – i.e., public policy discussions related to the nexus of housing, transportation, and urban form.

The chief reason that the AHS may be only marginally useful for transportation-related research, including the spatial mismatch hypothesis, is that better data are available elsewhere. For example, the National Household Transportation Survey (NHTS) is designed to allow trip chaining whereas the AHS is not. Also, the NHTS is based on a travel diary, as opposed to a survey format, which means the resulting data on commuting should be both richer in detail and more accurate.

We did explore the AHS data in order to understand its commuting pattern variables, both its strengths and limitations. We used demographic and income variables from the Metropolitan AHS to stratify the metropolitan population and identify low-skill and or low-wage workers and potential workers. We also investigated the differences between this group and the general population with respect to vehicle ownership, employment, commuting modes, and length of commute (as measured by time and distance).

Because the AHS is a rich source for data on American households, we explored linking the AHS with the NHTS. Linking the two datasets is possible because the NHTS is a national-level dataset like the AHS.

As might be expected, however, there were barriers to establishing effective linkages with the AHS. We fast discovered that incompatibilities in the geography variables between the two datasets prevented using the Metropolitan AHS, as was originally anticipated. We also discovered that minorities are under-represented in the NHTS.

We proceeded to test merging the NHTS with the National AHS. While we identified three types of merging (i.e., one-to-one merging, merging by proxy, and synthetic merging), only one was relevant to this exercise and that was synthetic merging.

In a synthetic merge, cohorts are linked on the basis of variables common to the two datasets. Common variables that are related to the topic of interest (e.g., spatial mismatch) should be selected. This methodology is best if interested in a limited number of the population's characteristics, as it is problematic to identify a finite set of variables that will identify groups similar in a broad number of general characteristics. (Our discussion includes an explanation of the different types of data merging and what could facilitate or hinder a synthetic data merge.)

We were able to explore how well the surveys fit together under a variety of merging approaches, and suggested other options for creative use of the NHTS information, including special runs at the Census Bureau. We also included an explanation of what information in the AHS could enrich the discussion of the spatial mismatch hypothesis.

Introduction

This research explores the applicability of the American Housing Survey (AHS), specifically its transportation- or commute-related variables, to research into the spatial mismatch hypothesis. We begin by summarizing how past research has used the AHS, and, if applicable, where the analyses could have benefited from the use of AHS. We detail several model specifications used by these researchers.

We then develop a methodology to stratify the AHS and identify cohorts of low skill/wage workers. Our purpose was to explore workers' characteristics within low skill/wage cohorts as well as to compare them to other population cohorts. This analysis utilized both the person-level AHS data as well as the household-level data included in the "flattened" AHS file. Summary statistics were generated for the variables included in the AHS commuting module.

Lastly, we discuss the issues associated with merging the AHS with other datasets. The National Household Transportation Survey (NHTS) is the focus of the merging discussion because it is a national-level dataset with a similar sample size and is otherwise theoretically comparable to the AHS.

The rest of this report is organized into a Background section, which is followed by a Summary of Past Research and a discussion of Potential Data Sources. The section titled AHS Analyses presents our research and analytic findings from the AHS. We discuss the results of our synthetic merging in the Joint AHS / NHTS Analyses section. The Conclusions section is followed by three sets of Appendices for our references, a literature review, and the results from our matching process.

Background

The spatial mismatch hypothesis was first put forward by John Kain in a 1968 paper, although it did not acquire the name until later. In it, he speculated that part of the reason for high unemployment rates for lower-skilled blacks living in central cities was that most jobs requiring their skill levels were created in suburban areas, thus making it harder for blacks to learn about and hold such jobs.

Using data from Chicago and Detroit, he tested three specific hypotheses:

1) Residential segregation affects the distribution of black employment;

2) Residential segregation increases black unemployment; and

3) The impacts of residential segregation are magnified by the decentralization of jobs.

Kain concluded that the housing discrimination that led to the segregation of blacks significantly constricted the employment opportunities of blacks living in central cities.

The spatial mismatch hypothesis has gone in and out of vogue in the ensuing years. After a flurry of critical attention in the late 1960s, the issue was not much studied in the 1970s and 1980s. Toward the end of the 1980s, interest by researchers again picked up. Ihlanfeldt, in a 1994 paper, attributed this renewed interest to three factors:

1) Worsening of urban problems such as crime, poverty, and unemployment;

2) Research by non-economists, such as the sociologist William Julius Wilson; and

3) Anecdotal evidence of high job vacancy rates at suburban employers.

Much of the literature has had, as a primary or secondary issue, questions of the role of race. This matter has confounded and complicated analyses, because of the correlations between racial segregation, housing discrimination, and job discrimination. In more recent years' data, the substantial growth of other racial/ethnic minorities complicates attempts to incorporate all appropriate racial/ethnicity issues into a jobs-housing spatial mismatch analysis.

While we also address the issue of race, the scope of our analysis focuses our efforts on the mismatch of affordable housing to lower-skill jobs. Thus we concentrate on the hypothesis of a jobs-housing imbalance, incorporating race and other factors as explanatory rather than the explicit focus of testing.

A search of the academic literature on the spatial mismatch hypothesis found literally dozens of papers investigating whether the hypothesis could be proved, as well as a number of reviews of those individual studies. However, the key problem since Kain's publishing of the hypothesis nearly 40 years ago continues to be how to prove that the link exists, given the data difficulties and multiple correlated factors involved.

Summary of Past Research

We reviewed 18 papers, including those mentioned in the work plan and others identified as a result of online and library database searches. While there have been dozens of papers written on the spatial mismatch hypothesis, the review generally found very limited use of the AHS data in this field, and only one published attempt to combine it with another data source. (The complete literature review is included as an Appendix.)

The four general methodologies, identified by Ihlanfeldt and Sjoquist in a 1998 paper, used to examine the hypothesis are:

1) Racial comparisons of commuting time or distance. These studies look at whether the average commuting time and or distance varies between blacks and whites, on the grounds that if blacks live further from available jobs they will have longer commutes.

2) Wages, employment, or labor force participation correlated with job accessibility. These studies look at whether measures of employment for blacks are related to the number of jobs within a given geographic area. If a spatial mismatch exists, blacks should have lower accessibility and lower wages or employment rates.

3) Comparisons of suburban and city labor market outcomes. These studies compare blacks living in the suburbs to those living in the central city to see if employment rates are similar, on the grounds that blacks with similar educational and or skill levels in the suburbs should be more likely to find employment if a spatial mismatch exists.

4) Differences in labor market tightness between cities and suburbs. These studies compare wages and the level of job vacancies for similar types of jobs in the suburbs and the central city, postulating that central city neighborhoods should have lower wages and lower vacancy rates than suburbs. These studies hypothesize that if spatial

mismatch exists, then suburban employers should have a harder time filling jobs, and consequently they will pay higher wages and or experience more vacancies.

The methodologies described below are given in order of most basic to most complex. We also identify and describe the data sources for each of the following methodologies.

Potential Data Sources

While the AHS is a dataset rich in detail, it has not typically been used to explore the spatial mismatch hypothesis. Many of the studies on the spatial mismatch hypothesis rely on one of three national datasets:

U.S. Census, generally the Public Use Microdata Sample (PUMS). PUMS matches specific housing units to the characteristics of the occupants using the decennial census.

Panel Study of Income Dynamics, a dataset collected by the University of Michigan that has followed the same families since 1968. Data are currently collected in odd-numbered years (until 1999 the study was conducted annually). Topics include earnings, employment, and housing.

National Household Transportation Survey (NHTS), a dataset collected by the Bureau of Transportation Statistics. It combines the previous National Personal Transportation Survey (NPTS) on commuting and other daily travel and the American Travel Survey on long-distance travel. The NHTS survey was conducted in 2001; NPTS surveys were conducted in 1995, 1990, 1983, 1977, and 1969.

Claritas Urban Place Type supplementary data, a dataset developed by Miller and Hodges of Claritas, which is a private firm. The Claritas dataset distinguishes Census block groups into one of five place types: urban, second city, suburban, town/exurban, and rural. The classification of each block group is based on a combination of its own density and the density of neighboring areas, as well as the density of a nearby population center. The Claritas data were developed to work in conjunction with both the Census and NPTS/NHTS datasets.

A number of local (MSA-specific) studies are based largely on locally generated data, often from a metropolitan planning organization or other regional council of governments, although often these sources incorporate Census and NHTS datasets.

Why the AHS and NHTS?

Among the available datasets, we believe there is a good opportunity to merge travel information in the NHTS with housing and household data from the AHS. The NHTS can estimate average travel expenses, based on vehicle miles traveled as well as imputed travel time costs, for work trips as well as other trips.[1] The AHS presents information on education

[1] In addition to the travel information available in the NHTS, the survey purchases some geographic information from Claritas, Inc. to enhance the understanding of the residence and workplace of each respondent. Residential density both at home and at the workplace are included at the Census tract and Census block group level.

and income, as well as specific housing and neighborhood characteristics. Merging the two datasets can provide estimates of the total impact of housing and travel on the family budget, for example.

The AHS and the NHTS each collect information that can, theoretically, inform analyses of the spatial mismatch hypothesis. This hypothesis depends on information both about where people live and their ability to reach various destinations for suitable employment, meaning a dataset to evaluate it would ideally include both residential, job skill, and transportation information.

The NHTS data are collected through a travel diary as opposed to a questionnaire. The travel diary approach produces a high level of accuracy about all types of trip-making made by the respondent. The quality of journey to work data is also much more accurate, including time of departure, one or more modes of transportation for travel, and actual distance to work. The NHTS staff also compute the great circle distance between the respondent's home and place of work using geographic information systems (GIS). In addition to journey to work data, the NHTS contains complete information on other travel, including such categories as lunchtime travel, dropping off and picking up children at school, shopping, and social visits. It also includes data regarding trips with multiple stops and or purposes, known as trip chaining.

The AHS is a survey, as opposed to a diary, conducted by the U.S. Census Bureau for HUD. It currently consists of national surveys conducted in every odd-numbered year with metropolitan surveys in even-numbered years. Data are collected on household characteristics as well as several commuting variables. Researchers need to decide both what level (i.e., national versus metropolitan) is most pertinent and whether to conduct their analyses at the household or person level. The commuting variables, such as vehicle ownership, are available at the household level whereas commuting time and distance are at the person level. Further, the AHS data do not allow for trip chaining as its commuting variables are limited to one mode of transportation and one commute time/distance.

AHS Analyses

At first, the AHS might seem to have limited usefulness when compared to the NHTS and its detail. What this ignores is that the AHS remains the key source for information on the U.S. housing stock. We highlight in this section what data are available from the AHS and what stratifications of these data could help begin to inform research into the spatial mismatch hypothesis.

Using the 2002 AHS Metropolitan survey, we first separated the person data file from the larger AHS data file. We used this file to first assess what could be a "low-wage" and or "low-skill" worker. Our first stratification was to limit our analysis only to workers aged 18 or higher.

We then created a series of skill stratifications based on income and education. We refined our initial stratifications to account for skill and work experience. When entering the labor force, the main determinant of a worker's wage is their education-level. But as workers accumulate relevant work experience, their education level becomes relatively less important as compared to their work experience. We use AGE as a proxy for experience. We expect that workers with college degrees and no years of experience, all other things equal, would earn less than workers without college degrees and 20 years of experience.

We then defined the bottom quartile to be both low-wage and low-skill. Note that this threshold is defined using *salary/wage* information as opposed to *income*. This is important because at the person-level, we do not have household income and thus are not accounting for any types of assistance that may be received by both individuals and the household.

Comparing summary statistics from the low-wage population with the high-wage or rest of the population as well as with the entire population yields some interesting results. The key statistics are outlined in Table 1 on the next page.

Table 1. Key Summary
Statistics from the AHS Person File (Percentages)

Variable	Low Skill, Low Wage	High Skill, High Wage	Entire Pop.		Diff. Between Low Skill & Entire Pop.
Yes, I worked last week	29.1	48.6	45.8		-16.7
Citizenship					
Nat., US born	73.5	84.3	82.7		-9.2
For. born, not a US cit.	17.2	8.7	9.9		7.3
Gender					
Male	31.8	51.6	48.7		-16.9
Female	68.2	48.4	51.3		16.9
Race					
White	70.2	74.6	73.9		-3.7
Black	11.3	10.5	10.6		0.7
Amer. Indian et al	0.7	0.6	0.6		0.1
Asian/Pac. Islander	7.1	4.5	4.9		2.2
Other	10.7	9.9	10.0		0.7
Education					
Less than High School Education	17.7	16.9	17.1		0.6
High School Diploma	24.8	26.6	26.2		-1.4
Some College/Assoc. Degree	30.3	29.4	29.6		0.7
College or Higher Degree	27.3	27.1	27.1		0.2

Source: ICF Consulting analysis of AHS data.

The variable WLINEQ asks whether or not you worked at all during the past week and the data would suggest that if you are low-wage, then in 29 percent of people's cases, they did work during the previous week. This is almost 17 percentage points lower than the population as a whole and highlights something that the AHS does not have – i.e., variables tracking how much each person works each week, whether the position is a full-time, salaried position or a part-time, seasonal position. These are very different kinds of work and data related to these factors currently cannot be teased from the AHS.

The citizenship variable yielded a result that is in accord with general labor economics literature results showing race as a significant determinant of wage. AHS data on citizenship indicate that a significantly larger percentage of foreign-born workers are low-wage workers than workers born in the U.S. AHS data indicate that low-wage workers are more likely to be women, the result is also supported by the income gap analysis well documented in a number of studies.

The relevant literature indicates that race is central to the spatial mismatch hypothesis. The AHS data, however, indicate that there is no obvious or large difference between the low-wage workers and the population at large. There is less than a percentage point difference between the low-wage worker and the population at large for black. White and Asian are the only race categories where there is a difference greater than a percentage point between the population and the low-wage worker.

Education is also included in our summary statistics. Across the education levels, there does not appear to be a large difference between the low-wage worker and the population at large.

We acknowledge that looking only at characteristics of workers, such as education and experience, is unlikely to fully explain the wage differential, as some individuals may not be able to find a suitable job (i.e., one matching their education and experience level) near the area in which they prefer and or can afford to live. Adding a spatial dimension to the research (i.e., analyzing location of housing versus location of employment centers) may provide further insight into the issue.

Therefore, we extended our analysis to the “flattened” AHS file, which allows us to access a number of transportation and commuting variables, including vehicle and mode of transportation information. [2] Once one moves from the person file to the flattened AHS file, however, the analytic focus effectively shifts to the household level.

We continued to use the same stratifications we created during our person-level analysis. We also continued to limit our focus to workers aged 18 or greater. A new constraint, however, was limiting the analysis to the householder and if one was present, the spouse. This is an important point because there may be more than two people contributing to total household income. We did not attempt to identify and include other workers within a household into our analysis.

We also extended our focus to those workers who reported no salary income and one-person households. The inclusion of spouse, no wage income, and one-person households were designed to isolate and test whether or not there were obvious differences in commuting and transportation choices, as well as a limited number of demographic variables, between these stratification levels.

[2] An implicit assumption to our research is that commuting long distances is not a “preferred” option for low-wage workers – i.e., it is driven by economic necessity rather than choice. For example, the lack of affordably housing near local centers of employment means low-wage workers must commute longer distances.

The possible permutations for the flat file stratification are the following:

Householder	Spouse	Sample Size
High Wage (HW)	HW	12,925
HW	Low Wage (LW)	4,774
HW	No Wage (NW)	4,252
LW	HW	750
LW	LW	272
LW	NW	101
NW	HW	5,052
NW	LW	644
NW	NW	3,195
Single HW	--	9,848
Single LW	--	645
Single NW	--	4,610

Source: ICF Consulting analysis of AHS data.

The small sample sizes associated with the low-wage samples (e.g., LW, LW; LW, NW) limit the analyses that could be reliably performed using such stratifications. Care should be taken when interpreting any results from such analyses.

Tables 2 through 5 present the key summary statistics from our analysis, with select points from each being discussed. Because there are a total of 12 different cohorts, the tables highlight the major ones for two-person households with a gray background. These are the HW, HW; LW, LW; and NW, NW categories.

Table 2. Summary of Gender and Racial Charactistics, By Stratification Level

	Two-Person Households									One-Person Households		
Variable	**HW, HW**	**HW, LW**	**HW, NW**	**LW, HW**	**LW, LW**	**LW, NW**	**NW, HW**	**NW, LW**	**NW, NW**	**HW**	**LW**	**NW**
Gender												
Head of Household (HOH) is Male …	71.8	92.6	92.2	31.0	69.5	74.9	59.4	64.3	77.2	54.8	50.7	31.9
Race												
HOH												
White	79.5	75.4	75.5	83.5	73.0	70.6	86.6	74.1	89.2	80.7	80.2	86.9
Black	7.5	5.2	4.7	5.8	8.2	1.1	4.4	7.6	4.3	10.9	10.6	8.2
Asian/Pac. Islander	4.7	6.3	6.3	4.2	10.4	13.6	3.4	7.2	3.0	2.8	4.4	2.3
Other	7.8	12.6	13.1	5.5	8.0	14.2	5.1	10.3	3.2	5.0	3.6	2.1
Spouse												
White	78.8	73.5	74.1	82.6	73.2	71.8	85.9	72.1	88.3			
Black	7.3	5.3	4.8	6.1	6.9	1.1	4.3	7.3	4.1			
Asian/Pac. Islander	5.2	7.5	7.4	4.8	9.8	10.7	3.8	8.7	3.5			
Other	8.2	13.2	13.3	6.1	9.7	15.5	5.6	11.1	3.7			
Education												
HOH												
Less than HS	13.2	15.7	18.8	1.5	0.3		19.7	22.0	21.8	10.1	0.4	24.9
HS	22.4	19.1	20.5	14.1	13.4	8.3	25.0	22.4	25.0	21.1	12.8	28.5
Some college/Assoc. Degree	31.0	27.0	26.4	31.4	28.7	32.9	27.7	25.9	25.9	33.0	32.5	27.4
College or Higher Degree	33.5	38.2	34.5	53.0	57.6	58.7	27.6	29.8	27.3	35.8	54.3	19.2
Spouse												
Less than HS	13.9	15.6	21.5	11.0	3.9	10.3	19.3	18.6	22.5			
HS	27.2	25.0	28.5	19.8	18.9	26.1	29.0	27.0	31.2			
Some college/Assoc. Degree	29.1	26.1	25.5	27.6	31.0	28.0	25.5	26.0	25.6			
College or Higher Degree	29.9	33.2	24.5	41.6	46.2	35.6	26.3	28.5	20.7			

Source: ICF Consulting analysis of AHS data.

Table 2 is a summary of gender, education and racial characteristics for different cohorts. Most householders in two-person households are male; the proportion is closer to half in single-person households. An even split is expected due to gender distribution but the wage levels indicated that women are heads of households more typically in low-wage households.

The data associated with education levels across wage/skill cohorts for both the householder and spouse are not detailed enough to capture whether individuals are currently full-time students. This matters because such respondents should not necessarily be considered low-wage for the purpose of this research.

The "No Wage" categories may seem to be an odd category to include. However, these categories highlight that the analysis is focused on wage information and households may receive income from other sources. We did not present information on household income due to data concerns with many households reporting no wage but incomes (ZINC) in excess of $100,000.[3]

The same trends for the race variable identified in the person-level summary statistics were evident in these data for the three major cohorts as well as the single-person household cohorts. The other cohorts are difficult to assess and no apparent trend has been found.

Table 3 is the first table to focus on transportation related variables, in this case vehicle ownership. The interesting points are related to wealth with higher wage categories tending to own relatively more vehicles – e.g., 21 percent of the high-wage category own three vehicles as opposed to 15 and 11 percent for the low-wage and no wage categories. The issue of population mobility is another critical part of the spatial mismatch hypothesis, thus seeing vehicle ownership apparently tied to wealth is an expected result.

Table 4 focuses on the mode of transportation, including whether or not a member of the household uses public transportation. We would expect that reliance on public transportation be inversely related to income. This was evident by comparing the high-wage with the low-wage cohorts for both the one- and two-person household groups.

The choice of driving to work was high across cohorts, but lower rates were evident for those with low-wages or no wages. The frequency of respondents who walked to work increased between the high- and low-wage workers. Whether or not this highlights a finding typical of spatial mismatch would require additional analysis – e.g., differentiating between those who work in urban areas and those in suburban areas.

[3] We explored both our code and the AHS data in order to isolate where the data issues were. We found that these numbers (i.e., no salary reported but ZINC exceeding $100,000) were reported in the raw data. This meant that there was not a coding issue nor was there an issue with how the file flattener was handling data from the NEWHOUSE file. This may be an issue to discuss with U.S. Census.

Table 3. Summary of Household Vehicle Ownership, By Stratification Level

Variable	Two-Person Households									One-Person Households		
	HW, HW	HW, LW	HW, NW	LW, HW	LW, LW	LW, NW	NW, HW	NW, LW	NW, NW	HW	LW	NW
# of Cars Owned												
0	12.0	16.7	17.0	14.2	16.1	12.4	13.6	20.4	13.8	22.0	29.5	33.7
1	39.9	42.2	43.3	41.8	44.3	55.6	49.1	42.7	53.2	59.9	52.2	58.7
2	36.1	30.5	30.2	34.4	28.5	25.8	30.9	27.7	28.1	14.8	14.5	6.5
Total %	88.0	89.4	90.5	90.4	88.9	93.8	93.6	90.8	95.1	96.7	96.2	98.9
# of Trucks Owned												
0	37.7	36.1	38.2	35.9	52.5	44.6	51.3	38.9	56.6	66.9	73.1	83.2
1	42.7	43.4	42.3	44.2	36.7	41.9	36.1	36.6	32.3	27.1	21.6	12.9
2	16.6	17.5	16.6	17.8	7.7	11.6	10.9	19.0	9.5	5.4	4.6	3.6
Total %	97.0	97.0	97.1	97.9	96.9	98.1	98.3	94.5	98.4	99.4	99.3	99.7
Total Vehicles Owned												
0	0.6	1.4	1.5	0.9	5.6	3.7	2.5	4.0	3.4	5.9	15.5	25.5
1	10.0	14.5	17.4	10.4	23.7	27.3	27.8	21.9	35.2	60.6	54.1	58.4
2	56.7	53.9	52.7	60.1	47.5	47.9	50.3	43.0	45.1	24.2	21.7	11.3
3	21.0	20.6	19.8	19.1	14.6	13.3	13.1	16.7	10.7	7.0	5.7	3.5
Total %	88.3	90.4	91.4	90.5	91.4	92.2	93.7	85.6	94.4	97.7	97.0	98.7

Source: ICF Consulting analysis of AHS data.

Table 4. Summary of Transportation Modes, By Stratification Level

Variable	Two-Person Households									One-Person Households		
	HW, HW	HW, LW	HW, NW	LW, HW	LW, LW	LW, NW	NW, HW	NW, LW	NW, NW	HW	LW	NW
Yes, someone in the Household uses Public Transportation	12.4	16.1	16.0	13.9	24.3	25.9	10.2	15.9	9.2	13.2	22.6	17.2
Yes, drives to work alone …												
HOH	90.2	91.2	91.1	92.3	85.2	80.3	89.4	82.0	83.5	94.1	92.5	90.8
Spouse	89.2	88.5	83.6	92.5	87.3	78.0	92.0	84.6	85.9			
Mode of Transportation												
HOH Car/Truck	87.0	86.8	86.6	75.9	78.6	85.2	63.3	66.5	65.5	86.5	76.2	66.6
Spouse Car/Truck	85.1	72.6	64.4	86.2	73.0	66.5	85.9	67.6	65.3			
HOH Walked	1.4	1.1	1.2	3.0	4.9	1.7	2.8	2.0	1.9	2.6	6.4	3.4
Spouse Walked	1.3	2.6	3.5	1.1	5.0	8.9	1.3	4.0	3.4			
Spouse Worked at Home	3.1	12.5	21.9	2.9	8.8	14.5	4.0	17.3	22.3			

Source: ICF Consulting analysis of AHS data.

The last table, Table 5, summarizes the commute information available using the DISTJ and TIMEJ variables from AHS.[4] We would expect that low-wage workers both travel further and have to spend longer times commuting than those with higher wages. This would support the idea of limited mobility and options for the low-wage workers. The data do not clearly indicate this for either the heads of household or the spouses. We find similar results for the single-person households.

We see that a relatively higher percentage of people in wage cohort LW, LW and LW, NW have commuting times of 45-60 minutes relative to other wage cohorts (10 and 14 percent respectively). Travel times cannot be entirely explained by the distance traveled because only nine percent of these individuals travel more than 20 miles during their commute.

The variables here may indicate that there are underlying data issues as much as the problems with survey data for these types of variables. In a larger sense, this is why researchers often seek multiple data sources – i.e., augment weaker data in one source with stronger data from another source.

This is why in the next section we begin to explore how one could merge the AHS with the NHTS, which is considered the best national-level source of data on transportation.

[4] We found a number of odd high values for commute distance (DISTJ) and confirmed that these values were also present in the raw data. After referring to the old AHS codebook, we saw that 996 denoted those who worked at home. The current variable is listed in the AHS codebook as a numeric with values from 0-997, with 998 denoting 998 miles or more. We wondered if either the current AHS codebook is incorrect or if there is an issue with Census' coding of the variable.

Table 5. Summary of Commuting Time and Distances, By Stratification Level

	Two-Person Households									One-Person Households		
Variable	**HW, HW**	**HW, LW**	**HW, NW**	**LW, HW**	**LW, LW**	**LW, NW**	**NW, HW**	**NW, LW**	**NW, NW**	**HW**	**LW**	**NW**
Commuting Time												
HOH												
15 Minutes or Less	38.6	38.4	38.8	48.0	45.3	39.6	38.4	35.9	34.4	47.9	48.8	43.4
15-30	36.7	36.2	36.4	27.7	30.6	34.3	24.9	24.6	26.3	33.9	32.7	24.7
30-45	14.1	13.4	13.3	11.7	7.4	6.2	7.7	10.6	9.8	10.5	8.8	7.2
45-60	4.6	6.0	5.6	2.2	10.2	14.0	2.3	4.6	3.7	3.3	3.3	1.5
Total %	94.0	94.0	94.1	89.6	93.5	94.1	73.3	75.7	74.2	95.6	93.6	76.8
Median Time Traveled	20.0	20.0	20.0	17.0	20.0	20.0	25.0	25.0	25.0	18.0	15.0	20.0
Spouse												
15 Minutes or Less	42.0	50.1	41.6	35.7	52.1	41.0	38.6	42.7	42.1			
15-30	35.7	25.6	24.1	35.0	27.9	44.5	33.9	25.6	21.7			
30-45	12.5	7.5	8.1	14.8	8.3		13.7	9.9	8.3			
45-60	4.8	2.6	2.7	6.7	3.0		6.1	3.0	3.8			
Total %	95.0	85.8	76.5	92.2	91.3	85.5	92.3	81.2	75.9			
Median Time Traveled	20.0	15.0	20.0	20.0	19.0	17.5	25.0	25.0	25.0			
Commuting Distance (Miles)												
HOH												
0-5	21.5	21.4	22.1	31.5	29.1	23.5	6.5	22.2	20.5	27.3	35.8	26.3
5-10	21.8	21.9	22.1	17.3	17.5	19.3	21.6	13.0	12.8	24.7	22.2	17.4
10-15	17.6	16.7	16.2	14.7	14.0	15.6	14.7	14.7	14.7	16.9	14.2	13.0
15-20	12.1	12.1	12.4	9.3	11.9	11.7	12.0	7.8	8.4	10.2	7.8	5.2
20-25	6.8	7.4	7.3	6.1	5.6	3.5	9.2	3.8	4.1	5.1	2.0	5.1
25-30	5.4	5.1	5.0	3.9	3.0	1.0	4.0	1.7	1.4	3.7	2.7	1.7
Total %	85.2	84.6	85.1	82.8	81.1	74.6	68.0	63.2	61.9	87.9	84.7	68.7
Median Dist. Traveled	12.0	12.0	12.0	10.0	10.0	10.0	14.0	15.0	15.0	10.0	7.0	10.0
Spouse												
0-5	23.2	30.3	20.0	20.8	38.0	22.5	22.4	27.3	24.6			
5-10	22.7	21.9	19.8	20.7	17.8	3.8	21.8	14.7	16.0			
10-15	17.7	12.0	10.7	18.1	12.0	20.6	15.1	12.6	14.7			
15-20	12.6	8.2	8.4	12.2	5.8	29.6	10.6	5.8	3.0			
20-25	6.6	4.0	5.4	5.9	6.2		6.9	5.0	3.4			
25-30	5.0	2.6	2.1	5.2	5.3		6.7	5.4	3.2			
Total %	87.8	79.0	66.4	82.9	85.1	76.5	83.5	70.8	64.9			
Median Dist. Traveled	11.0	10.0	12.5	12.0	9.0	12.0	12.0	12.0	12.0			

Note: Only medians are presented because the mean values were skewed by outliers, hence medians are more informative.
Source: ICF Consulting analysis of AHS data.

Joint AHS / NHTS Analyses (Data Merging)

As was previously discussed, there is no single “perfect” dataset available for researchers. There are always additional data that would be desirable, but all too often the marginal costs associated with gathering those data are too high. This is why researchers very commonly attempt to merge different datasets.

But merging different datasets is something to be undertaken with great care. It is all too easy to incorrectly assume that variable definitions and collection methods for similarly named variables are the same across datasets, for example.

There are three primary methods to merge datasets:

- **One-to-one merging.** A unique identifier or control variable is present in both datasets that researchers can use to link the datasets. Where possible, this is the best option. This opportunity rarely occurs across datasets however, with confidentiality and survey fatigue both being problematic.

- **Merging by proxy (or merging by many).** Several different variables, all of which are defined the same way, are used to sort and merge two or more datasets by linking cohorts defined by the results for the several different variables. All variables common to the two datasets can then be examined with greater sample sizes.

- **Synthetic merging.** A set of variables common to each dataset is identified and these are used to create different cohorts within each dataset; these cohorts are then used to identify or link these groups. The characteristics for the variables missing in one dataset and present in the other are imputed to extend the range of variables covered in both datasets, as well as the sample size of the merged dataset.

In a synthetic merge, cohorts are linked on the basis of variables common to the two datasets. Common variables that are related to the topic of interest (e.g., spatial mismatch) should be selected. This methodology is best if interested in a limited number of the population’s characteristics, as it is problematic to identify a finite set of variables that will identify groups similar in a broad number of general characteristics.

The synthetic merging process proceeds as follows: if dataset A has 25 variables A1 - A25 and dataset B variables B1 - B30, then, for example, 5 key variables (for simplicity, A1-A5 and B1-B5) are selected that are common to the two datasets and relevant to the topic of interest. The combined set of selected variables must be both precise enough to insure quite similar cohorts are pulled out of each dataset, but large enough to allow for statistically robust imputation (described below). Similar cohorts are thus identified between the two datasets. Some additional variables may also be in common (perhaps A6-A10 and B6-B10), which should be retained as they are. For the remaining variables, the information for variables A11-A25 may be randomly assigned to the records in dataset B who belong to the same cohort, while the information for variables B11-B30 may similarly be randomly assigned to the records in dataset who belong to the same cohort. Thus, one emerges with a combined dataset (albeit, with applicability limited to a small set of topics) that both has a larger sample size and broader set of characteristics to draw from.

Based on the findings from our literature review, we focused our efforts on assessing how the AHS and NHTS datasets could be merged using the synthetic merging methodology.

Merging the AHS and NHTS

We first assessed which methodology would be suitable for these two datasets. Because no unique identifier exists in these two datasets to link them, we could not do one-to-one merging. We then assessed whether there were sufficient similar variables and basic data structure to allow us to do merging-by-many. We did not believe that the two datasets were sufficiently comparable overall to allow this.

There were a sufficient number of common variables – i.e., variables having comparable values and from comparable universes – to allow us to test the synthetic merging approach.[5] (Due to each dataset's distinct weighting procedures, the universes will be similar but with slight differences.) These common variables were then used to impute characteristics from a group in one dataset to a similar group in the other dataset.

The common variables selected for matching each have a discrete number of possible values, or ranges of values that could be made comparable.[6] The total number of possible permutations, or cohorts, to be merged is the number of possible responses for all matching variables multiplied together. Each of these possible combinations is referred to as a "cell." The matching procedure will, on a cell-by-cell basis, apply the characteristics from one dataset to another.

Challenges Encountered

The synthetic merging of these two datasets uncovered a number of challenges for researchers. We list below the key items that we identified during our work but it is expected that other researchers will uncover others, similar to or extensions of these. They may also identify challenges we did not encounter as they extend our current analytic efforts.

Dataset Comparability

The NHTS is updated every five or six years, while the AHS is updated either every other year or every six years, respectively for the National and Metropolitan AHS.

Since the latest NHTS dataset was from 2001, we planned to use the 2002 AHS for the specific analyses in order to test the use of the ZONE variable. The ZONE variable is a very useful element of the AHS, especially since under special circumstances the NHTS data can include a geographic variable even as fine as zip code. However no combination or manipulation of the two datasets' finer-level geographic variables were found that were suitable for comparisons or

[5] Theoretically, this merging approach would have allowed us to test the use of ZONE data, which is the finest level of detail available from the AHS public use files. The ZONE data are only available from the Metropolitan AHS. Unfortunately, there was not a comparable variable to ZONE in the NHTS data. The specific issues will be described in greater detail in the next section.

[6] If the number of discrete values were not identical (e.g., one dataset had a variable with values of 0, 1, or 2+ and the other had 0, 1, 2, or 3+), we would truncate the discrete values (e.g., combining the 2 and 3+ into a 2+) so that both datasets were identical. Similarly, if one dataset had a continuous variable (income), it would be transformed to match the structure of the dataset with a discrete version (e.g., income quintile).

merging – most often because the variables represented both different size areas and disjointed sets of coverage.

After better understanding the geographic variables' limitations, we decided to use the 2001 National AHS. This actually provided for improved comparability between the AHS and the NHTS. The chief reason being that we were using two *national* level datasets, datasets that have comparable sample sizes as well as comparable national numbers.

The data collection for each dataset was conducted in the same year – i.e., 2001. This means we have improved comparability due to contiguous time frames – i.e., responses are closer to one another in time.

Geographic Level of Detail

The geography variables included in the National AHS did not address the level of detail we desired for our research. Specifically, we could analyze geographic location based on the very broad area of the country (REGION), but we could not achieve a local, subcounty geographic level (similar to ZONE).

The NHTS data includes some information about metropolitan statistical areas of respondents, as well as general residential density near each household. Urbanization level is coded through two different methods, one using Census definitions of urbanized areas and the other a proprietary method of density classification from Claritas, Inc., which uses a roughly 4 mile by 4 mile grid of the U.S. which is then mapped to Census block groups.

In addition to the publicly available data for NHTS, the U.S. Department of Transportation (DOT) may be able to provide researchers with finer geographic detail files. But these data would be available on a limited, case-specific basis. DOT is not covered under the same statute as the Bureau of Census and has the flexibility to divulge data for legitimate research purposes. These additional data could possibly allow NHTS data to be categorized by the AHS-defined zones within metropolitan areas. However, even with these additional details and data, the number of NHTS respondents in a given zone may be too low to allow for statistically valid manipulations.

Geographic Scale

Geographic scale is, in a related manner, a major issue for the examination of the spatial mismatch hypothesis. Typically, commuting and other travel takes place in a somewhat limited area near each person's home (with 12.1 miles being the national average commute distance, albeit with considerable variability).

Although even long commutes typically stay within the metropolitan statistical area, Census tracts may still be too broad a geography to use to understand built environment patterns (on both the residential and employment side) and their effect on travel. However, smaller geographic areas typically reduce sample sizes to the extent that analyses are made much more difficult.

Population Under-Representation in the NHTS

Because spatial mismatch is thought to be particularly important for some African-Americans and lower-income households, ideally the data used to examine the issue would have good representation of these populations. However, African-Americans are under-represented in the unweighted sample by more than a factor of two (five percent of the sample; 12 percent of the population).

Other minorities are also under-represented. The NHTS survey was conducted with a Spanish language option in 2001, but only 1.2 percent of respondents took part in Spanish. Decennial Census figures show that of the population five years and older, 5.2 percent speak Spanish at home and do not speak English fluently (defined here as "less than very well"). Including Spanish-speakers, a total of 8 percent of the five and over population does not speak English fluently.[7]

Variable Comparability: AHS and NHTS

The comparability of the individual variables within the AHS and NHTS datasets is critical when conducting a synthetic merging. We have broken our discussion of this issue into two parts. The first is a brief summary of the demographic variables used during our data merging. The second is a discussion of geographic variables and their importance to data merging.

Demographic Variables for Dataset Merging

Seven demographic variables were explored as the basis for the merge between AHS and NHTS: household size, household adults, household workers, household vehicles, race, income tercile, and income quartile. Because of differences in the coding of these variables (or closely related variables from which they were derived), some manipulation was necessary to insure their consistency for merging purposes.

- Household size variables are defined virtually identically in AHS and NHTS. For both categories, large numbers are grouped together to prevent low cell sizes, with truncation occurring at all numbers greater than four treated as a single 5+ category.

- Household adults is also virtually identical in AHS and NHTS. For both categories, large numbers are grouped together to prevent low cell sizes, with truncation occurring at all numbers greater than five treated as a single 6+ category.

- Household workers is virtually identical in AHS and NHTS. For both categories, large numbers are grouped together to prevent low cell sizes, with truncation occurring at all numbers greater than three treated as a single 4+ category.

- Household vehicles must be summed for the AHS data from two variables (cars and trucks are listed separately). Large vehicle numbers are then grouped together to prevent low cell sizes, with truncation occurring at all numbers greater than four treated as a single 5+ category.

- Race and ethnicity were simplified to address the most important racial and ethnic issues of the spatial mismatch hypothesis as well as to avoid a large number of low cell sizes. It was simplified into the four race categories of white, black, Hispanic, and other.

[7] (Language Use and English-Speaking Ability: 2000, Census 2000 Briefs. Accessed at: http://www.census.gov/prod/2003pubs/c2kbr-29.pdf)

- Income terciles and income quartiles were generated from the income variables in AHS and NHTS in order to produce comparable variables that would both reflect income levels and maintain robust cell sizes.

Demographic Variables	AHS	AHS Values	NHTS	NHTS Values
Household Size	Per	(count, up to 30)	Hhsize	(count, up to 14)
Adults in Household	Zadult	(count, 0-10 with 11 denoting 11+)	Numadlt	(count, up to 10)
Workers in Household		(Calculated using SAL1-16)	Wrkcnt	(count, up to 10)
Vehicle Count	Cars, Trucks	(Vehicle calculated using cars, trucks, up to 5)	Hhvehcnt	(count, up to 19)
Race	Race1, Span1	(Calculated using Race1 and Span1, 1-4)	Hhr_race	(nominal, 1-17)
Income	Zinc	(count, 0-999,997)	Hhfaminc	(18 ranges, in $5,000 increments up to $80,000)

Source: ICF Consulting analysis of AHS data.

Geographic Matching

While our analysis ultimately has focused on the National AHS and was simplified by only using a Census REGION variable, our initial intent was to use the Metropolitan AHS and assess how the ZONE-level data could be brought to bear on the spatial mismatch hypothesis. What we found, as was discussed earlier, was that we could not use the ZONE-level data in the Metropolitan AHS and, therefore, used the National AHS.

Issues of geographic matching, though, are important issues to researchers and bear discussion here. This discussion will focus on the Metropolitan AHS as it compares to the NHTS.

Both the NHTS and the AHS Metropolitan publish some geographic information for individual respondents in their survey. The level of geographic detail that is publicly available with both datasets is limited by respondent confidentiality agreements. The AHS-MA does not publish locational information below an area with a population of at least 100,000. The National AHS does not publish geographic location below the Metropolitan Statistical Area (MSA) level. The NHTS does not publish geographic details for states below a certain sample size, nor for Metropolitan Statistical Areas (MSAs) below a certain size. Because of this, the list of MSAs shown in the NHTS data is significantly shorter than the list from the AHS.

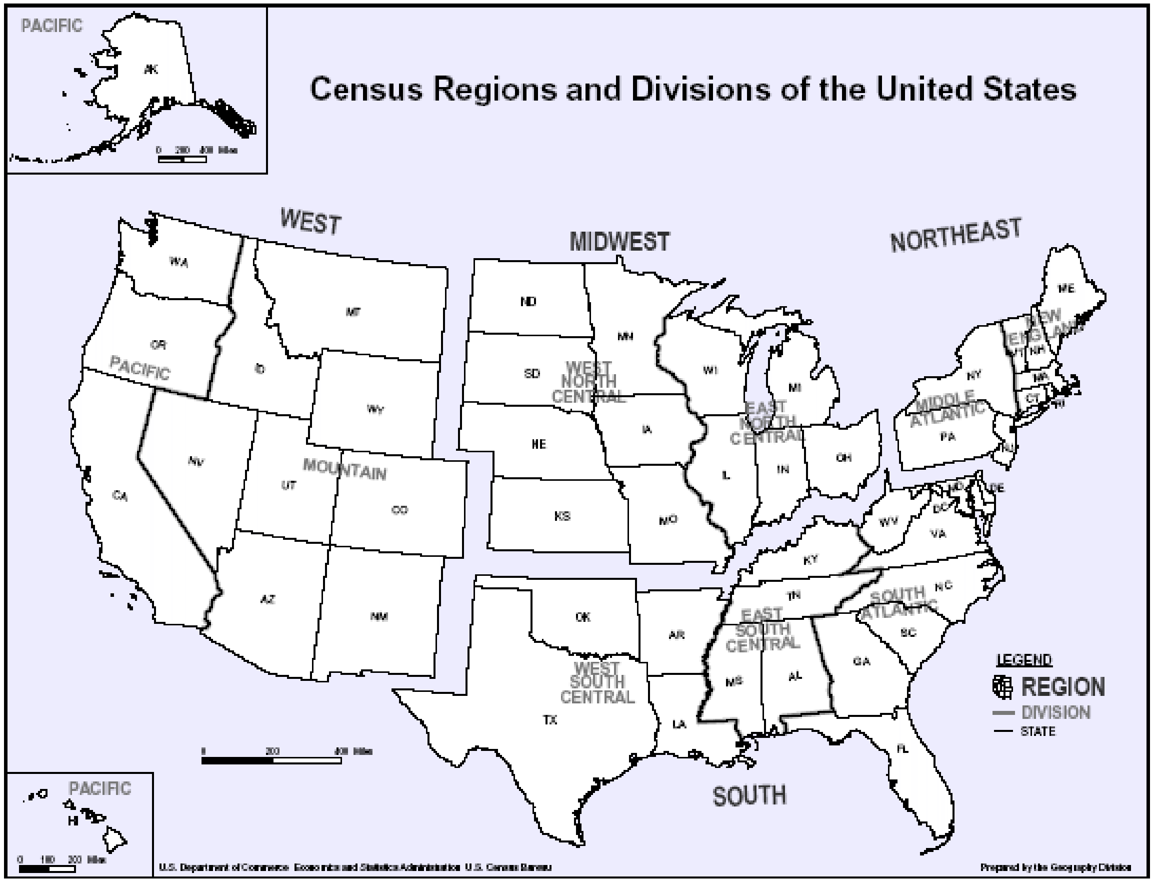
The following variables were considered for matching between AHS and NHTS data based on geographic location of each respondent:

Table 6. Variable Comparison, AHS Metropolitan Survey and NHTS

Variable	Description	AHS	NHTS
Census Region	There are 4 regions in the United States: West, South, Midwest, Northeast	[imputed based on STATE]	CENSUS_R
Census Division	There are 9 divisions in the United States: New England, Middle Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain, Pacific (see Map 1 below)	[imputed based on STATE]	CENSUS_D
Metropolitan Statistical Area (MSA)	MSAs are redefined based on each decennial census. Each MSA is made up of complete counties, with a few exceptions. In general, MSAs have been growing over time. AHS does not change their definition of MSAs for each new decennial census because it could breach confidentiality rules.	SMSA [1980 definition]	CMSA [2000 definition]
MA-Zone	Grouping of census tracts within an MSA based on various socioeconomic characteristics. Calculated by American Housing Survey.	ZONE	---
Urbanization Level within MSAs	Measure of urbanization for each respondent. AHS uses Census-based definitions of “Central City” to classify respondents as within the central city or a secondary central city, or in a suburb. Codes: Central city (1), Secondary cities (2-6), Suburban (7). NHTS uses a proprietary method developed by Claritas, Inc. to classify Census block groups and Census tracts by population density, both within each area and in relation to surrounding tracts/block groups. Codes: Urban (U), Secondary City (C), Suburb (S), Town (T), and Rural (R).	METRO	HBHUR (block group) HTHUR (tract)

Source: ICF Consulting analysis of AHS and NHTS datasets.

Map 1: Census Regions and Divisions



Limitations on Metropolitan Statistical Area (MSA) Definitions

A specific limitation in a geographic merge relates to the inclusion of certain counties under the NHTS definitions of metropolitan areas that were not included in the AHS SMSA definition.

The table below illustrates this issue by showing the counties in the 1980 and the 2000 definitions of Metropolitan Statistical Areas; some were added in the twenty years, and households classified as “Dallas” for NHTS in those counties would not be marked for Dallas by the AHS.

Table 7. Difference Between 1980 and 2000 Metropolitan Area Definitions in the Dallas-Forth Worth CMSA

Dallas PMSA				
		In 1980		**Percent of**
	County	**SMSA?**	**Population**	**2000 CMSA**
	Collin County		491,675	9.4%
	Dallas County		2,218,899	42.5%
	Denton County		432,976	8.3%
	Ellis County		111,360	2.1%
	Henderson County	**NEW**	**73,277**	1.4%
	Hunt County	**NEW**	**76,596**	1.5%
	Kaufman County		71,313	1.4%
	Rockwall County		43,080	0.8%
Fort Worth PMSA				
	Hood County		41,100	0.8%
	Johnson County		126,811	2.4%
	Parker County		88,495	1.7%
	Tarrant County		1,446,219	27.7%
	Total CMSA population:		**5,221,801**	**100%**

Source: ICF Consulting analysis of AHS and NHTS datasets.

As shown in Table 7, although the two datasets may differ in which counties are included in their MSA definitions, new counties are likely to include only a fraction of the total population of the MSA. In the Dallas-Forth Worth area, just three percent of the 2000 MSA population resided in the two counties added since 1980. However, this is not likely to severely bias the merge of the AHS and NHTS data.

Instead, a logistical issue is represented by the correlation between 1980 SMSA codes (from AHS) and the 2000 CMSA codes used by NHTS. In general, the 1980 SMSA designations have been refined for the 2000 designations, defining both a consolidated metropolitan statistical area (CMSA) and a primary metropolitan statistical area (PMSA). SMSAs roughly correspond to modern-day CMSAs or MSAs. However, the codes Census uses for CMSAs are slightly changed from the original SMSAs. In the Dallas-Fort Worth example in Table 2 above, the 1980 Dallas-Fort Worth SMSA had code ‘1920’. The 2000 Dallas-Fort Worth CMSA has the code ‘1922’, and is coded as ‘1922’ in the NHTS dataset. This can be easily remedied, but each MSA designation should be checked to make sure that it has not changed too drastically in the 20-year time period.

Another issue is that some SMSAs were merged to form a CMSA under the 2000 rules. The only example of metropolitan areas included both in the AHS and the NHTS data is the merging of San Francisco and San Jose. These two are both in the same CMSA under the 2000 Census designation.

Table 8 lists how metropolitan areas match between AHS and NHTS.

Table 8. Metropolitan Area Code Matching, AHS and NHTS

NHTS	AHS	NHTS: 2000 CMSA/MSA Titles	AHS: 1980 SMSA Titles
Complete Matches			
520	520	Atlanta, GA	Atlanta, GA
1840	1840	Columbus, OH	Columbus, OH
3280	3280	Hartford, CT	Hartford, CT
3480	3480	Indianapolis, IN	Indianapolis, IN
3760	3760	Kansas City, MO--KS	Kansas City, MO-KS
4920	4920	Memphis, TN--AR--MS	Memphis, TN-AR-MS
5120	5120	Minneapolis--St. Paul, MN--WI	Minneapolis-Saint Paul, MN
5560	5560	New Orleans, LA	New Orleans, LA
5880	5880	Oklahoma City, OK	Oklahoma City, OK
6200	6200	Phoenix--Mesa, AZ	Phoenix, AZ
6280	6280	Pittsburgh, PA	Pittsburgh, PA
6480	6480	Providence--Fall River--Warwick, RI--MA	Providence, RI
6840	6840	Rochester, NY	Rochester, NY
7040	7040	St. Louis, MO--IL	Saint Louis, MO-IL
7160	7160	Salt Lake City--Ogden, UT	Salt Lake City-Ogden, UT
7240	7240	San Antonio, TX	San Antonio, TX
7320	7320	San Diego, CA	San Diego, CA
8280	8280	Tampa--St. Petersburg--Clearwater, FL	Tampa-Saint Petersburg-Clearwater, FL
Codes Mismatch; Check Scope			
1122	1120	Boston--Worcester--Lawrence, MA--NH--ME--CT	Boston, MA
1602	1600	Chicago--Gary--Kenosha, IL--IN--WI	Chicago, IL
1642	1640	Cincinnati—Hamilton, OH--KY--IN	Cincinnati, OH-KY-IN
1692	1680	Cleveland--Akron, OH	Cleveland, OH
1922	1920	Dallas--Fort Worth, TX	Dallas, TX
2082	2080	Denver--Boulder--Greeley, CO	Denver, CO
2162	2160	Detroit--Ann Arbor--Flint, MI	Detroit, MI
3362	3360	Houston—Galveston--Brazoria, TX	Houston, TX
4472	4480	Los Angeles--Riverside--Orange County, CA	Los Angeles-Long Beach, CA
5082	5080	Milwaukee--Racine, WI	Milwaukee, WI
5602	5600	New York—Northern New Jersey--Long Island, NY--NJ--CT—PA	New York City, NY
6162	6160	Philadelphia--Wilmington--Atlantic City, PA--NJ--DE--MD	Philadelphia, PA-NJ
6922	6920	Sacramento--Yolo, CA	Sacramento, CA
7362	7360	San Francisco--Oakland--San Jose, CA	San Francisco, CA
7602	7600	Seattle--Tacoma--Bremerton, WA	Seattle, WA
8872	8840	Washington--Baltimore, DC--MD--VA—WV	Washington, DC-MD-VA

Table 8. Metropolitan Area Code Matching, AHS and NHTS

NHTS	AHS	NHTS: 2000 CMSA/MSA Titles	AHS: 1980 SMSA Titles
In AHS-MA sample; not in NHTS			
	2800		Fort Worth-Arlington, TX
	5775		Oakland, CA

Source: ICF Consulting analysis of AHS and NHTS datasets.

Last, when using MSA-level data in the NHTS data, researchers need to be aware that the sample is not designed for statistical significance at the MSA level. The sample is designed to be significant at the level of Census Division and MSA type. This combination of variables is given in the survey as a single variable: CDIVMSAR. MSA type is either served by rail transit or not; MSAs are then categorized as above or below one million; and non-MSA households are in another category.

Results

We tested a number of variable combinations in order to assess how well the two datasets (i.e., AHS and NHTS) matched one another. The variables used in these comparisons are the following:

- Census Region
- Race (black/white/other)
- Number of vehicles in household
- Number of adults in household
- Number of workers in household
- Number of persons in household
- Income
- Household size

Some of these variables were used interchangeably (such as number of adults vs. workers in household), while some were used in combination. For each combination, a cross-tabulation was created with the total number of weighted households for AHS and NHTS in each cell.

For example, the total number of households in each cell of a [Region x Household size] matrix is close to identical for AHS and NHTS. In comparison, the matrix showing cells by [Region x Race x Vehicles in Household] varies significantly between AHS and NHTS.

This difference is quantified by measuring the difference between the AHS and the NHTS cell size, then showing that relative to the sum of the two. A simple example shows how this is calculated:

Weighted Count of Households in Cell	AHS	NHTS	Weighted Difference/Sum
Region=1; Race=1; Vehicles=1	100,000	120,000	9.1%
Region=1; Race=1; Vehicles=2	120,000	100,000	9.1%
Region=1; Race=1; Vehicles=3	10,000	8,000	11.1%
Etc.			

Source: ICF Consulting analysis of AHS and NHTS datasets.

We then count the number of cells for this combination of three variables where the weighted difference is greater than 10 percent, 15 percent, and 20 percent.

The table below shows the percentage results for some variable combinations that are empirically compelling for use in spatial mismatch research. Greater detailed results are presented in Appendix C.

	Weighted Difference/Sum Greater than…		
	10%	15%	20%
Region x Race x Vehicles	57% of cells	48% of cells	37% of cells
Region x Race x Vehicles x Adults in Household	55% of cells	45% of cells	37% of cells
Region x Race x Vehicles x Workers in Household	65% of cells	54% of cells	42% of cells

Source: ICF Consulting analysis of AHS and NHTS datasets.

Based on the literature review of spatial mismatch, we selected Region, Race, Vehicles in Household, Income, and Adults/Workers in Household, to be the most important comparably available variables for matching travel and residential cohorts between these two datasets.[8] Our analysis reveals some differences between the AHS and NHTS on the populations estimated in each cell of these matrices, but the matches are still statistically tenable with the appropriate combinations. No determinative statistical measure exists to select which variables to use for merging. This is because, for example, different variables or different groups of cells and cohorts will be of interest to different researchers and for different specifications of testing the spatial mismatch hypothesis. Thus, for example, some cells or cohorts of the merging process may turn out to be statistically weak matches, but if they are of little or no interest and importance to the researcher, then their presence has little impact on the selected analysis of other, strongly matched cells and cohorts.

Potential Spatial Mismatch Analyses Using Merged NHTS and AHS Data

Synthesizing a matched AHS and NHTS dataset opens the door to several analytical routes that will lead to a better understanding of the combination of housing and transportation characteristics affecting spatial mismatch. Spatial mismatch is the basic premise that some population cohorts live far from work and from potential job sites. Symptoms of this problem include a higher rate of unemployment, higher cost burden in money and time for travel to work, disproportionately poor job/commute options for lower-income populations with less housing

[8] There are other factors – such as urbanization level, educational attainment, job skill level, and others – that would also be useful, but which are not available or comparable in both datasets. However, the merging process can make forms of these data available for analysis for these and other variables.

options, and a higher overall share of household budget going to transportation expenses than the average. Some research questions that might be answered using the matched dataset are discussed below.

1. What is the financial and time burden of the commute relative to household expenses for lower-income households?

   This question would be answered using the information on time and distance of the commute, converted using per-mile cost estimates to dollars, from the NHTS. AHS data showing home expenses and size of the home (to control for crowding as a solution to high-priced housing) would be used to show relative cost of travel to work. Built characteristics of the household environment from AHS could additionally be used to control for other influences on home cost.

2. What is the correlation between distance to work, residence in a low-income neighborhood, and income?

   Generally, the theory of spatial mismatch holds that low-income households are stuck in lower-income neighborhoods because of the cost of housing elsewhere and, as a result of the location of jobs far from low-income housing, must travel further to their work. While some of this analysis can be conducted using AHS alone, the NHTS adds exact figures for distance to work (both great circle distance and reported miles driven). In addition, general information about the density of residential development near each respondent's workplace is available in NHTS. Workplace density can be used as a proxy for suburban/urban character of the workplace environment – including public transportation access and levels of service.

3. Analysis of total transportation time and money expenses for households in different types of environments. The combination of AHS housing data with NHTS spatial location type data can make this more robust.

4. Analysis of transportation time and money burden on low-income households that do and do not commute between urban, lower-income neighborhoods and suburban jobs (reverse commuters)

Thus, the merging process allows many new analyses to be conducted of the spatial mismatch hypothesis. For example, these analyses would allow a closer look at the borderline cases, those that presumably are currently paying the maximum price the market will bear in terms of transportation burden in order to hold a job in the suburbs while living in low-income, urban neighborhoods. Similar to the first analysis, the AHS would form the basis of the population to be examined, using demographic and geographic information about where each household resides. NHTS information on trip-making over the course of a day would be assigned to each household to see the time and money costs these households undertake as part of a long reverse-commute lifestyle. Effects may include longer commute times, shorter hours at home and higher expenses for gasoline or public transport as a portion of total income. These would be compared to similar households living in urban, low-income neighborhoods who work closer to home, and to those with one or more non-working adults of working age.

Conclusions

The NHTS, formerly known as the NPTS (National Personal Travel Survey), is conducted every five to six years. Typically, the NHTS survey is changed relatively dramatically each cycle. This makes it more flexible than the AHS, which is a longitudinal survey. While this may hamper the ability to look back using the NHTS, it means that NHTS is also more flexible as time goes forward in adapting to new trends and a growing understanding of transportation behavior. AHS could take advantage of this fact by working with NHTS – either with more comparable geographic and demographic variables that could greatly facilitate merging, or by producing amalgamated data that would be even more helpful than joining the data together post hoc.

An example of a good way to match the two datasets would be for NHTS to deliver AHS with special data runs based on the metropolitan area zones that AHS has created. In order to preserve confidentiality, NHTS would likely have to provide AHS with a synthesized version of the dataset for each zone, but one that would still be better matched than the post hoc synthesis. Another simple way to ease comparisons would be for NHTS to include a center city variable consistent with AHS, in addition to the Claritas urban measures that are currently included.

In general, the currently envisioned data matching seems to have been as effective as a previous matching effort done with the 1995 NPTS dataset. The similarity of merging results point to a consistently similar variable distribution between the two datasets. The matching exercise tested here would provide a useful set of data for examining the spatial mismatch hypothesis, although it has certain limitations based on geographic information disparities (particularly regarding comparable measures of urbanization between the two datasets).

Several important research questions regarding the spatial mismatch hypothesis could be answered using the merged dataset. Further, other research efforts at the nexus of housing, transportation, and urban form could also be explored. And although this merging effort was focused on variables thought important for the spatial mismatch hypothesis, other variables could be used for merging to explore other research areas, such as questions regarding aging, income, or vehicle ownership and their relationships to housing and transportation. Thus, the merging process has been shown to be useful for spatial mismatch and potentially other areas. With better coordination of just a few variables between AHS and NHTS, even more statistically robust synthetic merging could be conducted that would be analytically useful across a wide range of questions involving transportation and housing.

Appendix A. References

Gabriel, Stuart and Stuart Rosenthal. 1996. “Commute, Neighborhoods Effects, and Earnings: An Analysis of Racial Discrimination and Compensating Differentials.” Journal of Urban Economics, Vol. 40.

Kain, John. 1968. Housing Segregation, Negro Unemployment, and Metropolitan Decentralization. *Quarterly Journal of Economics*, Vol. 82, No. 2.

Ihlanfeldt, Keith. 1994. “The Spatial Mismatch Between Jobs and Residential Locations Within Urban Areas.” *Cityscape: A Journal of Policy Development and Research*, Vol. 1, Issue 1 (August).

Ihlanfeldt, Keith and David Sjoquist. 1998. “The Spatial Mismatch Hypothesis: A Review of Recent Studies and Their Implications for Welfare Reform.” *Housing Policy Debate*, Vol. 9, Issue 4.

Nelson, Arthur C. and Thomas Sanchez. 1997. “Exurban and Suburban Households: A Departure from Traditional Location Theory?” *Journal of Housing Research*, Vol. 8, Issue 2.

O’Hare, W. 1983. “Racial Differences in the Journey to Work: Evidence from Recent Surveys.” Annual Meeting of the American Statistical Association. August.

Sanchez, Thomas and Casey Dawkins. 2001. “Distinguishing City and Suburban Movers: Evidence from the American Housing Survey.” *Housing Policy Debate*, Vol. 12, Issue 3.

South, Scott and Glenn Deane. 1993. “Race and Residential Mobility: Individual Determinants and Structural Constraints.” *Social Forces*, Vol. 71, No. 1 (September).

Spencer, James. 2000. “Why Spatial Mismatch Still Matters.” *Critical Planning*, Spring.

Taylor, Brian and Paul Ong. 1995. “Spatial Mismatch or Automobile Mismatch: An Examination of Race, Residence, and Commuting in U.S. Metropolitan Areas.” *Urban Studies*, Vol. 32, No. 9.

Thurston, Lawrence, and Anthony Yezar. 1991. “Testing the Monocentric Urban Model: Evidence Based on Wasteful Commuting.” *AREUEA Journal*, Vol. 19, No. 1.

Appendix B. Literature Review

The below literature review is divided into three sections:

- History and State of Spatial Mismatch Hypothesis;
- Use of AHS Data to Investigate the Spatial Mismatch Hypothesis; and
- Papers Reviewed but Not Cited.

History and State of Spatial Mismatch Hypothesis

The spatial mismatch hypothesis was first put forward by John **Kain** in a 1968 paper, although it did not acquire the name until later. In it, he speculated that part of the reason for high unemployment rates for lower-skilled blacks living in central cities was that most jobs requiring their skill levels were created in suburban areas, thus making it harder for blacks to learn about and hold such jobs.

Using data from Chicago and Detroit, he tested three specific hypotheses:

1) Residential segregation affects the distribution of black employment;

2) Residential segregation increases black unemployment; and

3) The impacts of residential segregation are magnified by the decentralization of jobs.

Kain concluded that the housing discrimination that led to the segregation of blacks significantly constricted the employment opportunities of blacks living in central cities.

The spatial mismatch hypothesis has gone in and out of vogue in the ensuing years. After a flurry of critical attention in the late 1960s, the issue was not much studied in the 1970s and 1980s. Toward the end of the 1980s, interest by researchers increased again. **Ihlanfeldt**, in a 1994 paper, attributed this renewed interest to three factors:

1) Worsening of urban problems such as crime, poverty, and unemployment;

2) Research by non-economists, such as the sociologist William Julius Wilson; and

3) Anecdotal evidence of high job vacancy rates at suburban employers.

Much of the literature has had, as a primary or secondary issue, questions of the role of race. This matter has confounded and complicated analyses because of the correlations between racial segregation, housing discrimination, and job discrimination. In more recent years' data, the substantial growth of other racial minorities would further complicate attempts to incorporate all appropriate racial issues into a jobs-housing spatial mismatch analysis.

While we plan to address the issue of race, the scope of our analysis focuses our efforts on the mismatch of affordable housing to lower-skill jobs. Thus we expect to concentrate on the hypothesis of a jobs-housing imbalance, incorporating other factors as explanatory rather than the subject of testing.

A search of the academic literature on the spatial mismatch hypothesis found literally dozens of papers investigating whether the hypothesis could be proved, as well as a number of reviews of those individual studies. However, a key problem continues to be how to prove that the link exists.

Ihlanfeldt and Sjoquist, in a 1998 paper, reviewed several dozen studies and found that there were four general methodologies used to examine the hypothesis:

> 1) Racial comparisons of commuting time or distance. These studies look at whether the average commuting time and or distance varies between blacks and whites, on the grounds that if blacks live further from available jobs they will have longer commutes.
>
> 2) Wages, employment, or labor force participation correlated with job accessibility. These studies look at whether measures of employment for blacks are related to the number of jobs within a given geographic area. If a spatial mismatch exists, blacks should have lower accessibility and lower wages or employment rates.
>
> 3) Comparisons of suburban and city labor market outcomes. These studies compare blacks living in the suburbs to those living in the central city to see if employment rates are similar, on the grounds that blacks with similar educational and or skill levels in the suburbs should be more likely to find employment if a spatial mismatch exists.
>
> 4) Differences in labor market tightness between cities and suburbs. These studies compare wages and the level of job vacancies for similar types of jobs in the suburbs and the central city, postulating that central city neighborhoods should have lower wages and lower vacancy rates than suburbs. These studies hypothesize that if spatial mismatch exists, then suburban employers should have a harder time filling jobs, and consequently they will pay higher wages and or experience more vacancies.[9]

In their review, Ihlanfeldt and Sjoquist reached the conclusion that while spatial mismatch appears to exist, its effects are not so cut-and-dried. They point out five subtleties in the findings from the spatial mismatch hypothesis literature:

> 1) The size of the metropolitan area makes a difference in the effect, with larger areas experiencing greater mismatch.
>
> 2) While job inaccessibility has been shown to increase unemployment, it is not always clear why that should be the case. While it may stem from commuting problems, it may also be a lack of information on job openings, discrimination against blacks by suburban employers, or a fear among black job seekers that they will not be accepted at suburban employers.

[9] Other explanations are also possible, especially considering the effect of a minimum wage. For example, the result may be similar wage rates for both (a surplus of low-skilled workers) and a higher unemployment rate in the central city (lower access to suburban jobs and more limited lower-skill central city jobs).

3) Spatial mismatch can affect lower-skilled workers of all races.

4) Since many studies of spatial mismatch are of male youth, on the assumption that their residential location is due to their parents' decision and job needs are unaffected by restrictions such as child care, it is less clear how it affects adults and women.

5) Many other factors play a role in determining black unemployment rates; even the studies that find the strongest impact find that spatial mismatch accounts for only one-half of racial differences in unemployment.

Another review by **Spencer** emphasizes many of the same points, and added several other complicating factors:

1) Skill mismatch, meaning that workers' skills do not match those required by employers, may be exacerbating the problem as the structure of the American economy changes away from a manufacturing base.

2) Commuting distances themselves are less a problem than lack of automobile ownership.

3) Not all lower-skilled workers living in ethnic enclaves experience high unemployment; many immigrant neighborhoods have created jobs for residents.

4) Employment discrimination may be a more important factor in black unemployment than lack of accessibility.

5) Job accessibility is an impediment only to obtaining an income, while the more important measure may be wealth accumulation, of which income is only one factor.

Use of AHS Data to Investigate the Spatial Mismatch Hypothesis

In the course of this literature review, we identified three papers that used American Housing Survey (AHS) data to investigate the spatial mismatch hypothesis. With the exception of the O'Hare study, none of the studies reviewed here used AHS data in conjunction with another dataset. These papers and their conclusions are summarized below.

O'Hare (1983)[10] used data from the 1975 Annual Housing Survey data, 1977 National Personal Transportation Survey, and 1980 Census to compare black and white commuting patterns.

O'Hare compares the commuter burden for blacks and whites. The commuter burden is defined as a ratio: hours spent commuting: hours spent at work. If a person works eight hours and commutes one hour, the commuter burden is .125 (1/8). The higher the number, the greater the burden. He also compares descriptive statistics such as average commute time and distance for various subsets of commuters.

[10] O'Hare, W. 1983. "Racial Differences in the Journey to Work: Evidence from Recent Surveys." Annual Meeting of the American Statistical Association. August.

The study used the journey-to-work supplements from the 1975 and 1979 Annual Housing Surveys, as well as population data from the 1980 Census.

He found that the "commuting burden" – the ratio of time spent commuting to time spent at work – is over twice as high for poor blacks than for poor whites. For blacks, the commuting burden decreases with income, but for whites it increases. In general, compared to whites, blacks travel shorter distances to work but longer times, because they are more likely to use public transportation.

Also, the data showed that for blacks, the majority of commutes that cross central city boundaries are from the central city to the suburb ("reverse commuting"), while the majority of white commutes between central city and suburb are from the suburb to the city. O'Hare's use of AHS data is limited to looking at responses to questions about transit use, using summary statistics rather than econometric analysis or combining it with other datasets for more quantitative analysis.

Taylor and Ong (1995)[11] used data from the 1977-8 and 1985 AHS to look at whether commuting patterns differed between blacks and whites, on the assumption that if the spatial mismatch hypothesis is correct, blacks should have longer commuting distances than whites.

Taylor and Ong group commuters into three types of neighborhoods: white, mixed, or minority, and compares the average commute time and distance for both years. They then used linear regression to control for commute mode, residential area type, income, education, age, and gender. Both methods were also used on a subset of low-skilled workers (since occupation is not given in AHS, they used years of education and income as a proxy; if a worker had no post-high school education and earned less than $8,000 in 1977/8 or $13,200 in 1985, she/he was classified as low-skilled).

They also "performed a series of logistic estimations" of how likely people who were employed in 1978/79 were to leave work by 1985. The study used data from the 1977/1978, and 1985 American Housing Surveys (the 10 metropolitan areas surveyed in both years with data on commuters by race).

They found that commuting distances were converging for all races, although commuting times were not. The fact that times varied was thought to be because of blacks' greater reliance on transit, which for most trips has a slower travel time than driving. They coined the term "automobile mismatch," theorizing that the difference in commute times would be reduced if all races have equivalent access to cars.

They concluded that the mismatch was less one of space, since most employees experienced a spatial mismatch in some form, and more one of mobility and accessibility.

[11] Taylor, Brian and Paul Ong. 1995. "Spatial Mismatch or Automobile Mismatch: An Examination of Race, Residence, and Commuting in U.S. Metropolitan Areas." Urban Studies, Vol. 32, No. 9.

However, **Gabriel and Rosenthal** (1996)[12] also reviewed AHS data from 1985 and 1989 (using a subset of 680 urban housing units and interviews with 10 neighboring units for each).

Gabriel and Rosenthal group households into different neighborhoods and estimate a commute time equation that controls for neighborhood fixed effects. Their empirical model is as follows:

$$\text{Log}(t_{ij}) = a_j\beta_a + p_j\beta_p + y_{ij}\beta_y + s_{ij}\beta_s + r_i\beta_r + R_i\beta_R + d_i\beta_d + e_{ij}$$

Where:

t = Commute time

i = Household-specific variables

j = Neighborhood-specific variables

a = Neighborhood characteristics, except race

p = Quality-adjusted house prices within a given neighborhood

y = Wage rates

s = Travel speed

r = Household race

R = Neighborhood race

d = Demographic variables (age, education, and marital status)

e = Error term

If all markets are perfectly competitive, the coefficients of d_i and r_i should be zero. While the data do not allow for an estimate of $\beta_{R,}$ (the impact of neighborhood race on commute time), a main goal is to find an unbiased and consistent estimate of β_r (the impact of household race on commute time).

If you include dummy variables for the neighborhoods (which Gabriel and Rosenthal suggest "is convenient, since one could never specify the complete vector of neighborhood amenities or obtain perfectly accurate measures of quality-adjusted home prices), the new equation is:

$$\text{Log}(t_{ij}) = \gamma_j + p_j\beta_p + y_{ij}\beta_y + s_{ij}\beta_s + r_i\beta_r + d_i\beta_d + e_{ij}$$

γ = Neighborhood fixed effects

[12] Gabriel, Stuart and Stuart Rosenthal. 1996. "Commute, Neighborhoods Effects, and Earnings: An Analysis of Racial Discrimination and Compensating Differentials." Journal of Urban Economics, Vol. 40.

The sample was restricted based on two criteria: first, the household head must be employed away from home, and second, "earning and commute times are both part of the workers' equilibrium package, which suggested that earnings are endogenous. This equation was estimated by two-stage least squares to control for possible simultaneity between earnings and commute times."

If the error term is positive, it means that workers are under compensated for their commutes, while a negative error term means that workers are overcompensated. Among under compensated workers, the likelihood of moving should increase with e_i.

To evaluate these arguments, a mobility equation is used:

$$I_{ij} = \delta_j + y_{ij}\theta_y + s_i\theta_s + r_i\theta_r + d_i\theta_d + z_{ij}\theta_z + N_{ij}e_{ij}\theta_{ne} + p_{ij}e_{ij}\theta_{pe} + \omega_{ij}$$

I = Latent index underlying the discrete decision to move

δ_j= Neighborhood fixed effects term

z = Housing attributes

ω_{ij} = Error term

Since e_i is the error term from the first equation, N_{ij} and p_{ij} equal 1 if e_i is positive or negative, respectively, and zero otherwise.

Consistent estimates for this equation can be obtained using a fixed linear probability model that controls for neighborhood effects with dummy variables. While a simple probit model would probably produce consistent estimates of θ_{ne} and θ_{pe}, other coefficients would probably be inconsistent.

The study used data from subsets of the 1985 and 1989 AHS. In 1985, AHS selected 680 urban housing units at random, and then surveyed up to 10 of the unit's "closest neighbors." These units were resurveyed in 1989, allowing the researchers to estimate the household move equation. However, commute time data were available only for 1985, not 1989. Data were eliminated if they could not be linked between 1985 and 1989; if the household race was other than black, white, or Asian; and if the household head was not employed outside the home.

They determined that educated black workers had longer commutes than their white or Asian counterparts. However, black workers with less than a high school education had similar commutes to whites and Asians at the same education level. They found that one-third of the estimated difference in commute time is offset by price differential (less expensive housing) and other neighborhoods amenities. However, even when controlling for neighborhood characteristics, blacks had longer commutes. Also, they found that blacks were less likely to move than whites, even though they were under-compensated for their commutes. (Note that the AHS 1989 dataset did not contain commute times.)

Four other studies we reviewed used AHS data to investigate related questions involving household locations and commuting:

> **Nelson and Sanchez** (1997) used data from the 1984 and 1985 AHS to look at why people move to exurban locations, with respect to whether the trend represents an extension of previously observed suburbanization patterns, or a new or distinct phenomenon.
>
> In comparing exurbanites and suburbanites, they found that exurbanites are more likely to have families, to be blue collar (both skilled and unskilled) and have larger houses and lots. However, in contrast to their hypotheses, exurbanites are similar to suburban residents in their age and income, their commuting behavior (they are not more likely to work from home or less likely to work in central city), and the proportion of income spent on housing and commuting.
>
> They concluded that exurbanization is an extension of regular suburbanization.
>
> **Sanchez and Dawkins** (2001) used AHS data from 1989 to 1991 to look at relocation patterns from center cities to suburbs and vice versa.
>
> They found that the racial and educational profile of both groups of movers was similar; suburban movers more likely to be married, have higher incomes, and be home-owners. Both groups cited job and household formation as main reasons to move. For those moving from suburbs to city, 13 percent cited commuting reasons. However, many respondents listed their main reasons for moving as “other,” which may mean that the survey does not list the most important factors.
>
> They conclude that “…both blacks and white are equally mobile in both directions.”
>
> **Thurston and Yezer** (1991) look at the average distance of residents and jobs from the city center, compute an “optimal commute distance,” then compare it to actual commutes as reported in the AHS (while the exact year is not given, most other data is from the early 1970s).
>
> They use two monocentric models; the first assumes all households and jobs are homogeneous, while the second assumes they are heterogeneous (as defined by occupation type, not race). Of the 14 cities studied, some have “wasteful” commutes (longer commutes than the model predicts), while others actually have shorter commutes. They conclude that “a semi-strong version of the monocentric model with heterogeneous households appears to account for actual commuting quite well.”
>
> **South and Deane** (1993) use 1979 and 1980 AHS data to look at the differences in household mobility between whites and blacks.
>
> While both groups move at the same rate (approximately 22 percent of people change households each year), blacks of similar socio-demographic background to whites more proportionally less. White homeowners are less likely to move than renters, but black homeowners move at similar rates to renters, possibly because the quality of housing is not as high as that of whites’. Blacks are less likely to move than whites if they are dissatisfied with their neighborhood, and high levels of segregation decrease blacks’ mobility.

We also reviewed additional papers that examined the spatial mismatch hypothesis, but did not use AHS data:

> **Stoll**[13] compares a measure of job sprawl to the dissimilarity index. Job sprawl is defined as the percentage of jobs over five miles from the city center. The dissimilarity index is calculated by comparing the black, white, and Latino populations to where jobs are located. The dissimilarity index ranges from 0 to 1; the higher the number, the greater the imbalance (for example, if the black population were distributed in exactly the same manner as jobs, the index would be 0).
>
> The dissimilarity index is calculated with the following equation:
>
> $$D = (½) \sum_{i} \left| \frac{\text{Black}_i}{\text{Black}} - \frac{\text{Employment}_i}{\text{Employment}} \right|$$
>
> Where black_i is the black population in Zip code i (where I = (1, 2…n) and indexes the Zip codes in a given area), and employment_i is the number of jobs in Zip code i.
>
> This equation can also be used to measure segregation between blacks and whites by substituting white population for employment.
>
> The study used population data from the U.S. Census 2000 (SF 1) and jobs data from the U.S. Department of Commerce 1999 Zip Code Business Patterns files.
>
> **Khattak et al**[14] used three different models:
>
> - weighted least square (WLS) regression model explaining commute time,
> - weighted least square regression model explaining commute distance; and
> - two-step model existing of (a) a probit model that estimates the probability that a person is employed and (b) a weighted least square regression model that estimates commute distance.
>
> Data used to perform the analysis are from the 1995 National Personal Transportation Survey (now known as the National Household Transportation Survey).
>
> **Weinberg**[15] finds support for the mismatch hypothesis using data from the 1980 Census PUMS (5% A sample and 1% B sample), STF 3C, and data on segregation from the

[13] Stoll, Michael A. Job Sprawl and the Spatial Mismatch between Blacks and Jobs. The Brookings Institution, February 2005.

[14] Khattak, Asad, Virginie Amerlynck, and Roberto Quercia. 2000. "Are Travel Times and Distances to Work Greater for Residents of Poor Urban Neighborhoods?" Transportation Research Record, 1718, TRB, National Research Council, Washington, D.C., pp. 73-82.

[15] Weinberg, Bruce A. 1998. "Testing the Spatial Mismatch Hypothesis Using Inner-City Variations in Industrial Composition." Unpublished; available online at http://economics.sbs.ohio-state.edu/pdf/weinberg/mismatch1.pdf (accessed April 28, 2005).

National Bureau of Economic Research. "An increase in jobs or a decline in black concentration in the central city increases black unemployment relative to whites. The effects are greatest in large [metropolitan areas] where the costs of working in a distant portion of the city are likely to be greatest." He finds that the impacts of spatial mismatch are larger on women, the less educated, and the young. He also finds that job access (as determined by both where blacks live as well as job locations) is more important than social connections (in learning about jobs): "A 10 percentage point increase in the share of jobs located in central cities would increase the employment of young non-college educated black men by 6 percentage points."

The paper by Weinberg focused "…on inter-metropolitan area (MA) variations developing instruments for job location. Our instruments exploit inter-MA variations in industrial composition….Industry-level differences in the importance of being centrally located and in space requirements generate cross-industry variation in job locations. Cross-city variation in industry employment parents interacted with industrial differences in job locations provide a source of cross-MA variation in job location which his unlikely to be affected by black labor market status.

"The mismatch hypothesis also implies that black concentration in the central city will reduce access to suburban jobs and increase competition for the jobs that exist in the central city. Thus, an increase in the fraction of blacks that live in the central city should decrease black employment….We instrument for black residential locations using lagged data on the age of the housing stock and black residential locations….

Instruments for Job Locations

"Our instruments for job locations exploit inter-city variations in industrial composition interacted with industrial difference in job locations….We estimate the demand for workers in central cities using a fixed coefficients demand index….the fraction of the workforce in MA *c* employed in central cities is:

$$\hat{f}_{CC|c} = \sum_i f_{CC|i} f_{i|c}$$

Where:

$f_{a|i}$ = Fraction of workers in industry i

CC = Center city

"We classified industries according to the 3-digit system used in the census (this classification has 232 industries). We also develop separate instruments for the demand for labor in the central city by gender and education…." (these are shown in the appendix).

Instruments for Black Residential Locations

"Our instrument for black residential locations is the center city-suburban is the fraction [sic] of the pre-1960 housing central city housing units that were built before 1940….The results from our first-stage regression are

$$\frac{\text{Black}^{15\text{-}64}{}_{CC}}{\text{Black}^{15\text{-}64}} - \frac{\text{White}^{15\text{-}64}{}_{CC}}{\text{White}^{15\text{-}64}} =$$

$$.139 + .473\,\frac{\text{Pre-1940 Units}_{CC}}{\text{Pre-1960 Units}_{CC} +} + .011\,\frac{\text{Pre-1940 Units}_{CC+S}}{\text{Pre-1960 Units}_{CC+S}} + 0006\log(\text{population})$$

(.135) (.190) (.225) (.009)

The Mismatch Hypothesis and Wages by Place of Work

"…we start our analysis by studying the effect of job locations on the relative wages of central city and suburban workers….we employ a two stage procedure to control for individual characteristics. In the first stage, log weekly wages are regressed upon individual worker characteristics:

$W_{cai} = \beta x_{cai} + \varepsilon_{cai}$

Where

W_{cai} = the log weekly wage of individual i working in area a of city c

x_{cai} = individual i's characteristics

"The wage in part a of city c is the mean log wage residual of the individuals working in that part of the city c:

$$W_{ca} = \frac{1}{n_{ca}} \sum_i \varepsilon_{cai}$$

"Second stage regressions are run to estimate the effect of job location on the wage of individual working in the central city relative to those working in the remainder of the MA….The second stage specification is:

$W_{cCC} – W_{cS} = Z_c\Gamma + \theta f_{CC|c} + v_c$

Where

Z_c = vector of MA characteristics

$W_{cCC} – W_{cS}$ = central city-suburban difference in long wage residuals

"The second stage regressions are weighted by the MA population size. Use of the central city-suburban wage difference controls for differences in the cost of living across MAs.

The Mismatch Hypothesis and Racial Outcomes

"….To control for differences in employment rates across MAs, we take the difference between the black and white employment rates as our dependent variable. To avoid selection, we calculate employment rates for all blacks and whites in an MA not just central city residents….we control for differences in observable individual characteristics using a two step procedure regressing individual employment status on the same controls in the first stage and using the mean residual of blacks and non-blacks as our measure of employment status."

The analysis described above was conducted using both weighted least squares as well as instrumental variables. Estimates were made of the effects of job locations on employment by gender, education, and age. Education is divided into fewer or more than 12 years of schools, and age is divided into three cohorts: 18-30, 31-50, and 51-65.

This paper uses data from the 1980 Census Public Use Microdata Samples (5% A sample and 1% B sample), STF 3C, and data on segregation from Cutler, Glaeser, and Vigdor (*The Rise and Decline of the American Ghetto*, National Bureau of Economic Research Working Paper No. 5881, January 1997).

Papers Reviewed but Not Cited

Lastly, we also reviewed an additional six papers that we cited in the work plan. None of these papers included information on the use of AHS data:

Kain, John F. 1992. "The Spatial Mismatch Hypothesis: Three Decades Later." *Housing Policy Debat*e, Vol. 3, No. 2.

> While there had been previous work in economics looking at trading off workplaces against housing location, Kain was the first to hypothesize that for blacks, housing location might be fixed (previous models had assumed you could live anywhere). He used origin/destination surveys for Detroit and Chicago to test his theory and found that blacks behave like whites, but subject to the effects of discrimination. In his later career, Kain used data obtained from the St. Louis Urban Renewal Agency to test his hypothesis there. He found that "ghetto" housing is more expensive than comparable other housing; blacks have access to poorer housing based on measure of housing quality; blacks in that study were 15 percentage points lower in terms of homeownership than white in equivalent circumstances; and black wealth creation is negatively affected by differential rates of homeownership. Kain is currently working on applying the spatial mismatch theory to schools in Dallas.

Schill, Michael H. and Susan M. Wachter, 1995. "Housing Market Constraints and Spatial Stratification by Income and Race." *Housing Policy Debate*, Vol. 6, No, 1.

> Schill and Wachter review a number of papers that look at the non-market forces that contribute to segregation by income and race. They find that while measures of racial segregation are slightly declining, economic segregation of minorities is increasing. According to their analysis, "Statistical studies of income and house values in suburban communities support the hypothesis that income homogeneity across communities is an outcome of local control over taxes, public services, and land use regulation by fiscally motivated jurisdictions." Local land use regulations restrict affordability by constraining the supply of land, and lowering the tax burden increases price. Another factor is the structure of the federal public housing program, which allows regions to let some municipalities opt out, leading to more concentrations of racial minorities and poverty. Most studies have found that public housing leads to more concentrations of poverty within a census tract. Federal mortgage assistance programs contributed to declining neighborhoods through systematically preferring mortgages in new neighborhoods. Finally, studies have shows that both blacks and Latinos experience housing discrimination, while there is a mixed record on studies of whether minorities have similar access to mortgage credit as whites.

Arnott, Richard. **1998**. "Economic Theory and the Spatial Mismatch Hypothesis." *Urban Studies*, Vol. 35, No. 7. June 1.

> This paper provides a foundation for the spatial mismatch hypothesis grounded in economic theory rather than data analysis. He argues that the hypothesis is causal (spatial mismatch causes blacks' higher unemployment rates), but that causality could be weak or strong causality, and that current ideas are conceptually incomplete (for example, in what respect does the spatial mismatch hypothesis represent a clear market failure?). Other problems with the hypothesis are that it does not: 1) distinguish between economic and racial segregation; 2) account for changing residential patterns; 3) explain

why unemployment increases rather than wages decrease 4) address skilled black workers. He develops an economic model to explain the mechanism by which the spatial mismatch plays out.

Kain, John F. **2001**.“A Pioneer’s Perspective on the Spatial Mismatch Literature.” Keynote presentation at *Understanding Isolation and Change in Urban Neighborhoods: A Research Symposium*. Chicago, April 11.

> The originator of the spatial mismatch hypothesis reviews the field. He finds while there was a fair amount of interest in the topic, and it was covered by Commissions looking into the 1960s riots and causes of black poverty, interest waned in the 1970s. While two sociologists helped revitalize the idea, two economists who studied it found no evidence (“race, not space”). Kain finds fault with several of the studies (Jencks and Meyers survey, and Masters) that criticize his work or find no support for spatial mismatch. He points out that some researchers confuse two predictions: that black employment would be higher if there were no segregation, and that blacks in suburban areas would have higher employment than blacks in average income (not poor) city neighborhoods. Proving the second doesn’t prove the first. Suburbanization does not necessarily mean desegregation, and author claims it has not improved job access. Kain does agree that the effects of spatial mismatch may vary depending on other labor market conditions (for example, if labor markets are tight).

McArdle, Nancy. 1999. “Outward Bound: The Decentralization of Population and Employment.” Joint Centre for Housing Studies, Harvard University.

> McArdle considers trends in decentralization, finding that outlying counties are gaining population at a faster rate than metro counties; she also analyzes the geographic aspect, with more overall growth in the South and West. Job growth is also continuing in non-metro counties. However, she does not comment on the implications of these trends for inner-city residents.

Kamer, Pearl. 1977. “The Changing Spatial Relationships Between Residences and Worksites in the New York Metropolitan Region: Implications for Public Policy.” *AREUEA Journal*, Vol. 5.

> Kamer asks, to what extent have workers’ residence reacted to changes in employment locations? Using Census data, for the New York metropolitan area, she finds that from 1960-70, 97 percent of new population growth and 89 percent of job growth was in suburban areas. There was more black population growth in the core, more white population growth in suburbs. For jobs, there was more white collar growth in the core with a decline in blue collar jobs; suburban growth was mostly white collar. She analyses over- and under-representation of various jobs and employee residential locations by county, finding that managerial positions over-represented in Manhattan, and managers over-represented in living in suburbs (suggesting a trade-off between job proximity and “spacious living”). She concludes that, “In the long run, over a period of 20 years, most occupational groups adjusted their place of residence so as to remain relatively close to their jobs.” However, she did not test the spatial mismatch theory because her spatial analysis was based on occupation, not race.

Appendix C: Results from Matching Process

Descriptions of Match

The following table has six columns, three that indicate how the variable combination affects the number of cells, and three that indicate the level of match between the two datasets, using weighting variables.

The first three columns describe the characteristics of the variable combination. **Total Cells** describes the number of combinations made by the variable combination. **Cells with N<30** describes the number of cells with low sample sizes, those less than 30. **Average Cell Size, AHS** describes the average sample size for each cell using the AHS sample, which is smaller than the NHTS sample. This gives a general idea of what the expected cell size should be for each variable combination.

The fourth, fifth and sixth columns describe how well the two datasets might be expected to match.

The fourth column is **High Difference Cells (Weighted)**. It is calculated by dividing the weighted absolute count difference by the sum of counts. This is a way to calculate absolute difference between the two without assigning one as the basis for comparison. The percentage in the table shows how many cells had a difference greater than 15 percent.

The fifth column, **High Difference, Small N**, shows the overlap between cells with a high difference and a small sample size. This gives some sense of how much of the difference between the two datasets is possibly caused by low sample size.

The sixth and last column, **High Difference, Medium/Large N**, shows how many cells with a reasonable sample size (over 30) still have a high difference between the AHS and NHTS projected population size.

Table C-1. Variable Descriptions

Variable Name	Variable Description
region	Census Region (Northeast, South, West, Midwest)
race	Race (white, black, Hispanic, other)
vehicle	Number of Vehicles in Household
zero_veh	Zero-vehicle Household
adult	Number of Adults in Household
worker	Workers in Household
inc20	Income in 5 categories: 0-15K; 15-30K; 30-40K; 40-60K; and 60K+
inc33	Income in 3 categories: 0-25K; 25-45K; 45K+
hhsize	Total Household Size

Source: ICF Consulting analysis of AHS and NHTS datasets.

Table C-2. Results of Data Matching Based on Tested Variable Combinations

Variable Combination	Total Cells	Cells with N<30	Average Cell Size, AHS	High Difference Cells (Weighted)	High Difference, Small N	High Difference, Medium/Large N
A Priori Preferred Variable Combinations						
region race vehicle adult	403	69%	105	47%	34%	13%
region race vehicle worker	320	55%	133	56%	38%	18%
region race vehicle inc20	399	59%	106	63%	45%	19%
region race adult inc20	412	66%	103	46%	31%	15%
region race worker inc20	319	51%	133	62%	40%	22%
region race vehicle	80	16%	531	50%	13%	38%
region race vehicle adult inc20	1,649	88%	26	50%	45%	5%
region race vehicle worker inc20	1,476	84%	29	57%	51%	6%
region vehicle adult worker	408	64%	104	35%	26%	9%
region vehicle adult inc33	320	57%	133	35%	22%	13%
region vehicle adult hhsize	376	56%	113	30%	19%	11%
region adult worker inc33	265	57%	160	34%	20%	14%
region adult worker hhsize	289	57%	147	27%	19%	8%
region worker inc33 hhsize	204	21%	208	33%	17%	17%
Non-Preferred Variable Combination in Order of Increasing Complexity						
region	4	0%	10,622	0%	0%	0%
race	4	0%	10,622	50%	0%	50%
vehicle	5	0%	8,497	0%	0%	0%
zero_veh	2	0%	21,244	0%	0%	0%
adult	6	33%	7,081	0%	0%	0%
worker	4	0%	10,622	0%	0%	0%
inc20	5	0%	8,497	0%	0%	0%
hhsize	5	0%	8,497	0%	0%	0%
region race	16	0%	2,655	44%	0%	44%
region vehicle	20	0%	2,124	5%	0%	5%
region adult	24	33%	1,770	0%	0%	0%
region worker	16	0%	2,655	0%	0%	0%

Table C-2. Results of Data Matching Based on Tested Variable Combinations

Variable Combination	Total Cells	Cells with N<30	Average Cell Size, AHS	High Difference Cells (Weighted)	High Difference, Small N	High Difference, Medium/Large N
region inc20	20	0%	2,124	0%	0%	0%
region hhsize	20	0%	2,124	0%	0%	0%
region zero_veh	8	0%	5,311	0%	0%	0%
race vehicle	20	0%	2,124	50%	0%	50%
race adult	24	33%	1,770	33%	0%	33%
race worker	16	0%	2,655	44%	0%	44%
race inc20	20	0%	2,124	50%	0%	50%
race hhsize	20	0%	2,124	40%	0%	40%
race zero_veh	8	0%	5,311	50%	0%	50%
vehicle adult	30	33%	1,416	13%	0%	13%
vehicle worker	20	0%	2,124	15%	0%	15%
vehicle inc20	25	0%	1,699	32%	0%	32%
vehicle hhsize	25	0%	1,699	8%	0%	8%
zero_veh adult	12	33%	3,541	0%	0%	0%
zero_veh worker	8	0%	5,311	0%	0%	0%
zero_veh inc20	10	0%	4,249	10%	0%	10%
zero_veh hhsize	10	0%	4,249	0%	0%	0%
adult worker	24	42%	1,770	17%	8%	8%
adult inc20	30	33%	1,416	13%	0%	13%
adult hhsize	24	42%	1,770	4%	0%	4%
worker inc20	20	0%	2,124	40%	0%	40%
worker hhsize	17	0%	2,499	18%	0%	18%
inc20 hhsize	25	0%	1,699	12%	0%	12%
region race zero_veh	32	9%	1,328	50%	6%	44%
region race adult	90	37%	472	36%	6%	30%
region race worker	64	3%	664	44%	3%	41%
region race inc20	80	1%	531	46%	1%	45%
region race hhsize	80	4%	531	39%	0%	39%
region vehicle adult	115	40%	369	19%	8%	11%
region vehicle worker	80	4%	531	23%	1%	21%
region vehicle inc20	100	9%	425	30%	6%	24%

Table C-2. Results of Data Matching Based on Tested Variable Combinations

Variable Combination	Total Cells	Cells with N<30	Average Cell Size, AHS	High Difference Cells (Weighted)	High Difference, Small N	High Difference, Medium/Large N
region vehicle hhsize	100	7%	425	11%	1%	10%
region adult worker	95	43%	447	19%	8%	11%
region adult inc20	117	36%	363	20%	4%	15%
region adult hhsize	92	42%	462	7%	1%	5%
region adult zero_veh	46	43%	924	13%	11%	2%
region worker inc20	80	1%	531	36%	1%	35%
region worker hhsize	68	7%	625	15%	7%	7%
region worker zero_veh	32	9%	1,328	13%	3%	9%
region inc20 hhsize	100	0%	425	22%	0%	22%
region inc20 zero_veh	40	15%	1,062	20%	10%	10%
race vehicle adult	113	46%	376	39%	12%	27%
race vehicle worker	80	20%	531	54%	15%	39%
race vehicle inc20	100	23%	425	54%	13%	41%
race vehicle hhsize	100	21%	425	48%	15%	33%
race zero_veh adult	46	43%	924	35%	9%	26%
race zero_veh worker	32	16%	1,328	50%	16%	34%
race zero_veh inc20	40	18%	1,062	53%	13%	40%
race adult worker	94	52%	452	40%	16%	24%
race adult inc20	118	44%	360	41%	10%	31%
race adult hhsize	92	43%	462	34%	4%	29%
race worker inc20	80	19%	531	59%	18%	41%
race worker hhsize	68	9%	625	51%	6%	46%
race inc20 hhsize	100	3%	425	51%	3%	48%
vehicle adult worker	115	49%	369	30%	14%	17%
vehicle adult inc20	141	43%	301	33%	11%	22%
vehicle adult hhsize	111	45%	383	21%	5%	16%
vehicle worker inc20	100	13%	425	50%	9%	41%
vehicle worker hhsize	85	11%	500	22%	8%	14%
vehicle inc20 hhsize	125	13%	340	36%	7%	29%
zero_veh adult worker	47	51%	904	23%	15%	9%
zero_veh adult inc20	58	45%	733	24%	10%	14%

Table C-2. Results of Data Matching Based on Tested Variable Combinations

Variable Combination	Total Cells	Cells with N<30	Average Cell Size, AHS	High Difference Cells (Weighted)	High Difference, Small N	High Difference, Medium/Large N
zero_veh adult hhsize	45	42%	944	9%	0%	9%
zero_veh worker inc20	40	20%	1,062	50%	13%	38%
zero_veh worker hhsize	34	9%	1,250	12%	3%	9%
zero_veh inc20 hhsize	50	16%	850	26%	8%	18%
adult worker inc20	117	50%	363	34%	12%	22%
adult worker hhsize	82	49%	518	23%	9%	15%
adult inc20 hhsize	116	44%	366	20%	3%	17%
worker inc20 hhsize	85	12%	500	39%	12%	27%
region race vehicle hhsize	399	58%	106	53%	39%	15%
region race zero_veh adult	171	57%	248	39%	22%	16%
region race zero_veh worker	128	38%	332	53%	28%	25%
region race zero_veh inc20	160	41%	266	59%	32%	27%
region race zero_veh hhsize	160	39%	266	49%	27%	23%
region race adult worker	327	66%	130	44%	30%	14%
region race adult hhsize	309	61%	137	41%	26%	16%
region race worker hhsize	272	47%	156	50%	32%	17%
region race inc20 hhsize	400	57%	106	58%	42%	16%
race vehicle adult worker	394	72%	108	48%	34%	14%
race vehicle adult inc20	498	72%	85	55%	41%	14%
race vehicle adult hhsize	364	71%	117	47%	35%	13%
race vehicle worker inc20	399	62%	106	67%	47%	20%
race vehicle worker hhsize	339	58%	125	60%	42%	18%
race vehicle inc20 hhsize	500	64%	85	61%	46%	15%
race zero_veh adult worker	172	65%	247	41%	24%	17%
race zero_veh adult inc20	216	62%	197	46%	27%	19%
race zero_veh adult hhsize	160	61%	266	41%	23%	19%
race zero_veh worker inc20	160	48%	266	69%	41%	29%
race zero_veh worker hhsize	136	42%	312	60%	32%	29%
race zero_veh inc20 hhsize	200	44%	212	60%	34%	26%
vehicle adult worker inc20	493	72%	86	46%	33%	13%
vehicle adult worker hhsize	338	64%	126	33%	22%	11%

Table C-2. Results of Data Matching Based on Tested Variable Combinations

Variable Combination	Total Cells	Cells with N<30	Average Cell Size, AHS	High Difference Cells (Weighted)	High Difference, Small N	High Difference, Medium/Large N
vehicle adult inc20 hhsize	463	65%	92	38%	26%	12%
zero_veh adult worker inc20	209	65%	203	39%	24%	14%
zero_veh adult worker hhsize	143	62%	297	29%	19%	10%
zero_veh adult inc20 hhsize	202	60%	210	35%	23%	11%
adult worker inc20 hhsize	354	64%	120	37%	23%	14%
region race vehicle adult worker	1,248	86%	34	48%	42%	5%
region race vehicle worker hhsize	1,282	83%	33	58%	52%	6%
region race vehicle inc20 hhsize	1,846	87%	23	57%	53%	5%
region race zero_veh adult worker	549	77%	77	41%	34%	7%
region race zero_veh adult inc20	707	79%	60	47%	38%	9%
region race zero_veh worker inc20	603	74%	70	55%	43%	11%
region race zero_veh worker hhsize	531	71%	80	52%	43%	9%
region race zero_veh inc20 hhsize	748	77%	57	57%	49%	8%
race vehicle adult worker inc20	1,501	88%	28	49%	44%	6%
race vehicle adult worker hhsize	1,056	85%	40	49%	43%	6%
race vehicle adult inc20 hhsize	1,454	87%	29	51%	45%	5%
race vehicle worker inc20 hhsize	1,554	87%	27	57%	53%	5%
vehicle adult worker inc20 hhsize	1,294	84%	33	47%	41%	6%
race zero_veh adult worker inc20	674	80%	63	47%	38%	10%
race zero_veh adult worker hhsize	482	78%	88	46%	35%	11%
race zero_veh adult inc20 hhsize	664	78%	64	46%	36%	10%
race zero_veh worker inc20 hhsize	644	75%	66	57%	46%	11%
zero_veh adult worker inc20 hhsize	573	76%	74	40%	31%	10%

Note: The number of corresponding permutations accounts for the number of foreseeable ratios between the other variables, such as adults per household or vehicles per adult.

Source: ICF Consulting analysis of AHS and NHTS datasets.