LAIM3.0 Methods

Introduction

There is more to housing affordability than the size of your monthly rent or mortgage payment. Transportation costs are the second-biggest budget item for most families, but until recently there hasn't been an easy way for people to fully factor transportation costs into decisions about where to live and work. The goal of the Location Affordability Portal (LAP), launched in November 2013 by the U.S. Department of Housing and Urban Development (HUD) and the U.S. Department of Transportation (DOT), is to provide the public with reliable, user-friendly data and resources on combined housing and transportation costs to help consumers, policymakers, and developers make more informed decisions about where to live, work, and invest.

The LAP connects users to the Location Affordability Index, a robust, standardized data set containing household housing and transportation cost estimates at the Census block-group level for all 50 states and the District of Columbia. These estimates are generated using the Location Affordability Index Model (LAIM) Version 3.0 (LAIM3.0), a combination of statistical modeling and data analysis primarily using data from a number of federal sources. LAIM3.0 uses the same modeling methodology and data sources as Version 2.0 but uses the updated vintage and constructed at the Census tract level using the 2012-2016 American Community Survey (ACS) 5-year estimates as the primary dataset.

This document is a description of the methods used to develop the LAIM3.0. This document contains, as Appendices A-G that were written while developing this method. It will require that the reader have access to a set of files of Python scripts, R scripts, SQL queries, and PostgreSQL tables.

The following is a list of the files delivered with this document.

Table 1: List of Files

File Name	Type	Description
make_variable_definitions.sql	SQL create query	This makes the make_variable_definitions table structure in PostgreSQL
variable_definitions.csv	Comma Separated Variable File (CSV)	File of the U.S. Census American Community Survey (ACS) variables used to create LAIM3; this is the content used to fill the make_variable_definitions table
fill_acs_data.py	Python script	Loops over all ACS variables and loads the values into a new table with the actual values, for every tract.
make_tracts_acs_2016.sql	SQL create query	Makes tracts_acs_2016 table structure in PostgreSQL
tracts_acs_2016.csv	CSV File	Values in the tracts_acs_2016 table
build_il_odometer_table.sql	SQL select into query	Creates a summary table of the Illinois odometer reading.
make_il_tract_vmt_table.sql	SQL select into	Creates il_tract_vmt table

File Name	Type	Description
	query	
block_tract_xtab.sql	SQL select into query	Performs the GIS matching of 2010 blocks to 2016 tracts
make_block_tract_xtab.sql	SQL create query	This makes the make_block_tract_xtab table structure in PostgreSQL
block_tract_xtab.csv	Cross tab of 2010 blocks to 2016 tracts	Data in the block_tract_xtab table
LODESTechDoc7.3.pdf	Census Documentation	LODES technical documentation[1]
wac_tracts.sql	SQL select into query	Aggregates WAC data from LODES into tracts
wac_counties.sql	SQL select into query	Aggregates WAC data from LODES into counties
wac_states.sql	SQL select into query	Aggregates WAC data from LODES into states
wac_tracts.csv	CSV File	Aggregated job for all Census tracts from the Census block LODES data.
wac_counties.csv	CSV File	Aggregated job for all counties from the Census block LODES data.
wac_state.csv	CSV File	Aggregated job for all states from the Census block LODES data.
make_simple_job_density.sql	SQL select into query	Calculated a simple employment density for all jobs and retail jobs.
run_job_density.py,	Python script	Calculates the simple employment density for all jobs and retail jobs.
run_gravity.py	Python script	Loops over all tracts in a county or state (per an input variable) and calculates the employment gravity indexes.
gravity_measures.csv	CSV file	Gravity employment values calculated with the 2014 (with 2013 Wyoming) LODES data.
run_median_commute_d.py	Python script	Calculates the median commute distance for each Census tract from the LODES data.
run_tract_uza_dist.py	Python script	Calculates the distance from every Census tract to every UZA.
run_model.py	Python script	Loops over counties and Eight Control Households and runs the LAIM3.0
sem_avar.r	R script	Calibrates the LAIM3
model_sem_fit.r	R script	Resource file – functions etc. - for the LAIM3.0 calibration.
laim_beta.csv,	CSV file	Matrix of SEM derived coefficients for LAMI3.0, for the exogenous variables
laim_gamma.csv	CSV file	Matrix of SEM derived coefficients for LAMI3.0, for the endogenous variable interactions

[1] https://lehd.ces.census.gov/data/lodes/LODES7/LODESTechDoc7.3.pdf

File Name	Type	Description
laim_inv_omg.csv	CSV file	Inverse matrix of the Unity matrix minus the SEM derived coefficients for LAMI3.0, for the endogenous variable interactions
vmt_fit.r	R script	Calibrates the VMT OLS
lai_sem_model.sql	SQL make function query	Makes a function in PostgreSQL that calculates the endogenous variables from the SEM fit.
lai_model.sql	SQL make function query	Makes a function in PostgreSQL that calculates the dependent variables from the OLS fit.
run_costs.py	Python script	Calculates the cost of auto ownership, auto use, and transit by tenure, and combines these and the housing cost into the overall LAI
run_income_pctile.py	Python script	Calculates the income percentile that each of the Eight Control Household for every Census tract

Creating the Data

This document assumes that the user has access to raw data from the US Census (2012-2016 ACS 5-year data) and U.S. Census Longitudinal Employment-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) and knows how to query these datasets.

ACS Data

There are many variables that are obtained from the ACS data. The list of these variables is stored in PostgreSQL table - make_variable_definitions – that lists all the ACS variable names, and what the name assigned to each. Create a copy of this table by following these steps in PostgreSQL:

1. Run the make_variable_definitions.sql query;
2. Use the SQL command “copy into” to load the variable_definitions.csv data into this table.

The following screen shot show a few sample entries in this table:

Figure 1: Screen Shot of " variable_definitions" PostgreSQL Table

	ndx integer	var_name character varying	acs_field_name character varying	description character varying	tab_order integer
1	1	population	b01003 001	Total Population	1
2	2	households	b25003 001	Total Number of Households	2
3	8	avg hh size all	b25010 001	Average HH Size both Renters and Owners	9
4	9	avg hh size owners	b25010 002	Average HH Size Owners	10
5	10	avg hh size renters	b25010 003	Average HH Size Renters	11

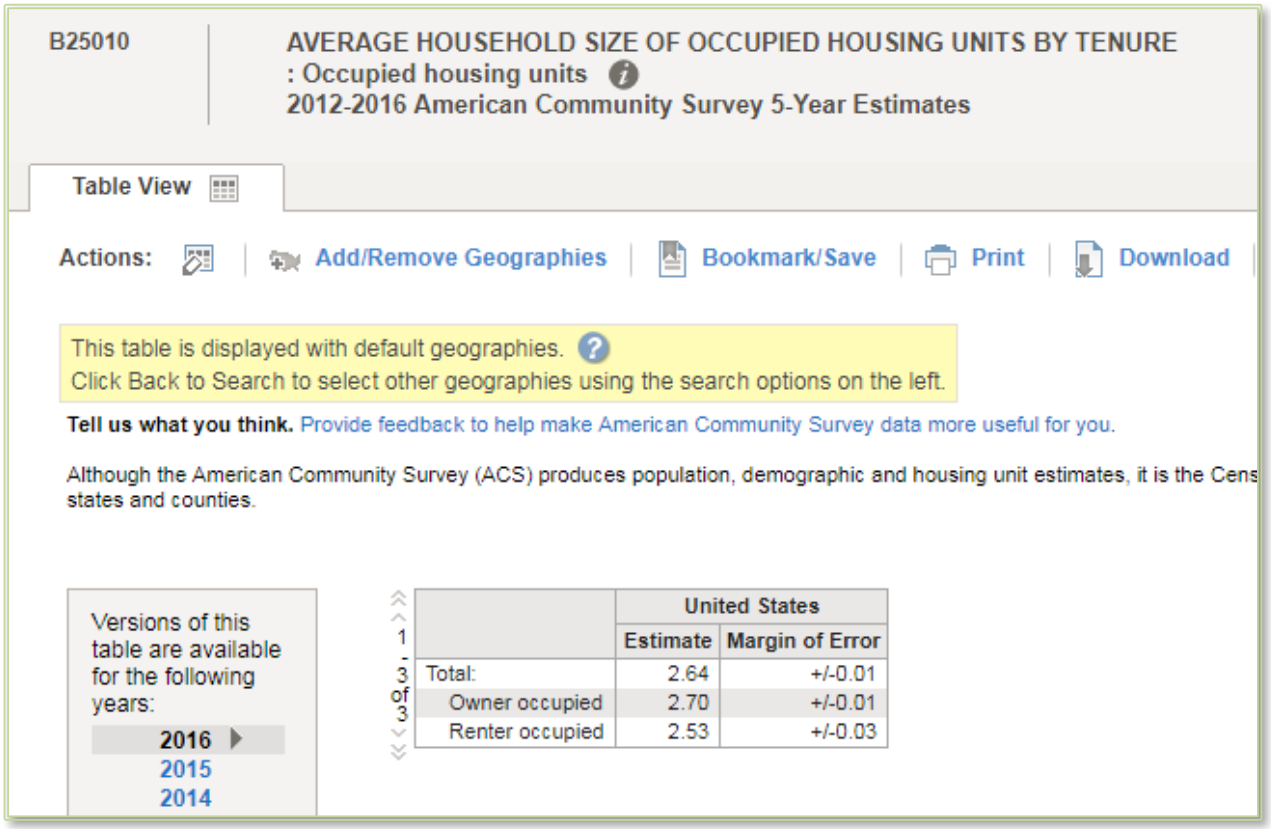
The columns are defined as:

- ndx – index, a serial variable that keeps track of each entry
- var_name – the name of the variable to be used in model development

- acs_field_name – the name of the field in the 5-year ACS data, the structure of this is defined by the census
- description – optional field to remind the user what this is
- tab_order – a variable that is used to sort the variables into a convenient order.

As an example, the highlighted row (#4) has ndx = 9, var_name is avg_hh_size_owners, the acs_field_name is b25010_002, description is "Average HH Size Owners" and the tab_order is 10. The acs_field_name of b25010_002 comes from the ACS variables definitions – the screenshot below is from the FactFinder[2] site for the B25010 variable for the entire country:

Figure 2: Screen Shot of US Census FactFinder Website



Note that the "Owner occupied" is the second on the list, hence the _002 on the acs_field_name.

Once there is a good variable_definitions table, the next step is to load the data from your ACS data set into a table for the values of each of these variables for all census tracts – tracts_acs_2016. The Python[3]

[2] https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t

[3] In this and all Python scripts, CNT's database address needs to be included; however, to not give out the credentials, the user, password and host have been replaced by the placeholders' foo, bar, and foobar respectively, in the database connection line – see screen shot below:

script fill_acs_data.py performs this using CNT's ACS archive to fill this table; presumably this could be done using the ACS's API to obtain the data, but that was not the approach used.

Once this is accomplished the tracts_acs_2016 is full, the following screen shot is of the first several records showing only the first six columns (out of the total of 79 columns).

Figure 3: Screen shot of a part of the"tracts_acs_2016" PostgreSQL Table

	stfid character varying(11)	population numeric	households numeric	avg_hh_size_all numeric	avg_hh_size_owners numeric	avg_hh_size_renters numeric	med num
1	01001020100	2010	740	2.72	2.58	3.12	
2	01001020200	2196	838	2.41	2.6	2.09	
3	01001020300	3136	1223	2.56	2.64	2.39	
4	01001020400	4563	1776	2.57	2.59	2.47	
5	01001020500	10529	4289	2.41	2.55	2.26	
6	01001020600	3742	1327	2.82	2.67	3.34	
7	01001020700	3047	1167	2.56	2.55	2.58	
8	01001020801	3025	1080	2.8	2.62	3.58	
9	01001020802	10743	3997	2.69	2.63	2.97	
10	01001020900	5912	2109	2.79	2.74	3.02	
11	01001021000	2899	1007	2.87	2.85	2.93	
12	01001021100	3247	1247	2.57	2.7	1.98	

VMT from Illinois Odometer Readings

Because of the lack of a ubiquitous source for driving information the LAI has used a dataset from Illinois where there are odometer reading from the smog-check program. Illinois Department of Natural Resources shared the odometer reading from 2013-2015; the smog-check program is applied in Illinois the Chicago and St. Louis metro areas. Using these data linking autos via the vehicle identification numbers (VIN) number and zip code the distance autos are driven is calculated. The validity of this data set for the entire country is ensured by examining national driving records from the National Household Travel Survey.

Data were obtained geographically identified with ZIP+4™ (9-digit zip code) and then assigned to Census tracts. Automobiles were matched using the vehicle VIN, and the total distance driven was determined over the time between inspections.

Data available included:

- "VIN" Vehicle Identification Number – a unique number for every vehicle in the world,
- "ZIPCODE" 9-digit zip code;
- "ODOMETER" Odometer reading at time of test rounded to the 1,000th place;
- "TEST_DATE" Date the test was performed;
- "MAKE" Make of the Vehicle;
- "MODEL" Model of Vehicle;

```
#
# set up connection to the avar_lai
#
conn_lai = psycopg2.connect(database='avar_lai', user='foo', password='bar', host='foobar', port='5432', sslmode='require')
```

- "MY" Model Year of Vehicle and
- "PURCHASE_DATE" date of purchase of the vehicle at the 9-digit zip code.

The query "build_il_odometer_table.sql" sorts out the correct combination of vehicles and setting up the data to run "make_il_tract_vmt_table.sql" to make the final "il_tract_vmt" table used in the modeling VMT per household.

LEHD and LODES Data

As stated above, this section relies on the fact that the user has downloaded the LEHD data sets for the most recent year. For LAIM3.0 this was 2014, with the caveat that Wyoming did not submit its data that year but did for 2013. Therefore, the need to merge the Wyoming 2013 data into the rest of the country's 2014 data. There are three components for these data, one is to calculate gravity measures to calculate employment intensity, and the second is to measure local employment density and the final use is to find the median commute distance.

To calculate the employment gravity and density it is necessary to get the jobs using the Worker Area Characteristics (WAC) data files for every state. These tables are produced by Census Block which are then be aggregated up to tracts; the blocks used are from 2010, and since tracts have changes somewhat in the intervening years, we produced a lookup table from 2010 blocks to 2016 tracts using GIS rather than the Census Codes – see table block_tract_xtab, and the script block_tract_xtab.sql that fills it as well as the table definition in make_block_tract_xtab.sql. The screen shot below shows the first several rows that result from a query on the block_tract_xtab where the tract codes (tract_from_block) as calculated, by using the first 11 characters, from the block (stfid) do not match the 2016 tract (tract_stfid) that contains the centroid of the 2010 block (for all 11 million plus blocks).

Figure 4: Screen Shot of "block_tract_xtab" PostgreSQL Table

	stfid [PK] character varying(15)	tract_from_block text	tract_stfid character varying(11)
1	010030114014056	01003011401	01003011403
2	010030114014058	01003011401	01003011403
3	010030114014085	01003011401	01003011403
4	010030114014086	01003011401	01003011403
5	010030114014087	01003011401	01003011403
6	010150011001000	01015001100	01015002000
7	010150011001001	01015001100	01015002000

The SQL select into query wac_tracts.sql creates a table - wac_tracts – that contains the total of jobs within the tract, by using the above files to aggregate the WAC jobs into tracts. Similarly, there are wac_counties.sql and wac_states.sql that make wac_county and wac_state table that aggregate the LODES data to counties and states.

Calculating Job Density

The Job density is calculated using all the jobs within a Census tract divided by the land area of that tract. The make_simple_job_density.sql query does this calculation using the wac_tracts table built above. Similarly, retail density is calculated using the retail jobs within a Census tract divided by the land area of that tract.

Calculating Gravity Variables

There are three model input variables that depend on a "gravity" calculation. They are Employment Access Index and Retail Employment Access Index; they are constructed using a gravity model that factors in the quantity of, and distance to, all employment destinations, in relation to any given tract. Using an inverse-square law, the indexes are calculated by summing the total number of jobs (or retail jobs in NAICS sector 44-45 specifically) divided by the square of the distance to those jobs. This method provides more information than a simple job density measure, in that it includes the accessibility to jobs outside a given Census tract. In addition to measuring access to jobs and retail jobs, it also provides a measure of economic activity created by those jobs. The Employment Access Index is calculated as:

$$E \equiv \sum_{i=1}^{n} \frac{p_i}{r_i^2}$$

Where:

- E is the Employment Access for a given Census tract
- n is the total number of Census tracts
- p_i is the number of jobs in the ith Census tract
- r_i is the distance (in miles) from the center of the given Census block group to the center of the ith Census tract

The proximity of jobs to the Census tract determines their contributive value to the Employment Access Index. For example, one job a mile away adds one, but one job 10 miles away adds 0.01. The measure includes all jobs in all US Census tracts by employing the following parameters to accelerate the calculation[4]:

- State totals and distance to states when the state is not the same as the given block group and is more than 250 miles away,
- County totals and distance to counties when the county is not the same as the given block group and is more than 33 miles away, and
- Census tract totals and distance for tracts not in a state or county that satisfies the above requirements.

[4] These distance thresholds were developed using the average distance between the geographic entities and factor determined such that the calculation remains consistent with using the block groups for a small representative sample.

The Python script run_gravity.py is then ran, this script does the gravity calculations and allows the user to complete one county or state at a time. This is a necessary feature since this process is very computer time intensive, allowing the user to run several state/counties in parallel to expedite the process. The output of this program is a PostgreSQL table names gravity_measures, which is created the first time the run_gravity.py script is run (per a flag set by the user).

Make Model Variable Fitting Table

There are now three files that contain all the information to make the model inputs: tracts_acs_2016, tract_simple_job_density and gravity_measures. These are then run combined using the query make_modeling_data.sql, this results in a table fitting_data that is then used to calibrate the various models.

As stated in "Appendix C: Define Eight Control Households and Run Model" there are some variables that are missing because of data suppression in the ACS, and there are several strategies for fixing them. There is a flag, "fixes," in the data table that defines how these missing variables were filled, and the following table shows the possible values for "fixes":

Table 2: Codes Used in "fixes" Field Indicating how ACS Suppressed Measures were estimated

	Algebra	Model	Full Value	Other Tenure	Neighbors Average
Fraction of AMI Owners	1	2	3	4	5
Fraction of AMI Renters	10	20	30	40	50
Household Size Owners	100	200	300	400	500
Household Size Renters	1000	2000	3000	4000	5000
Commuters per HH Owners	10000	20000	30000	40000	50000
Commuters per HH Renters	100000	200000	300000	400000	500000
Rooms per HU Owners	1000000	2000000	3000000	4000000	5000000
Rooms per HU Renters	10000000	20000000	30000000	40000000	50000000

The final "fixes" value is the sum of all the fix codes used. So for example, if the value for "fixes" is 13000045 (and there are four tracts with this value) that means the Fraction ***of AMI Owners*** was fixed using the "Neighbor Average", the ***Fraction of AMI Renters*** was fixed using "Other Tenure", ***Rooms per HU Owners*** was fixed using "Full Value" and ***Rooms per HU Renters*** was fixed using "Algebra" but ***Household Size Owners, Household Size Renters, Commuters per HH Owners*** and ***Commuters per HH Renters*** did not need fixing. All 72,241 tracts could be fixed except for one (Isle Royal in the middle of Lake Superior); the most common value, of the 100 various combinations, for "fixes" are listed below:

Table 3: Top 10 Fixed - Including no fix Necessary

"fixes"	Number of Tracts	"fixes"	Number of Tracts
0 (none needed)	67,293	1000001	224

10	2,809	10000010	170
1	550	1000101	46
20	400	1010001	36
1010101	298	10000020	35

Assigning every tract to an Urbanized Area (UZA) so the transit cost can be estimated required one of three strategies outlined in "Appendix E: Transit fare allocation strategy." The fields "alpha_beta_uzas" (list of UZA Census Identifier – stfid) and "uza_dist" (distance to closest valid UZA) are used to identify which strategy was used. Recall the three strategies from Appendix E:

1. Assign to α and β of the UZA where the tract only intersects one UZA where α and β can both be calculated i.e. UZAs with good NTD data (49,039 tracts);
2. Take weighted average of α and β from the multiple UZAs with good NTD data a tract intersects (247 tracts intersect 2 UZAs and 1 intersect 3 UZAs, zero tracts intersect more than 3 UZAs);
3. Use α and β from closest UZAs with good NTD data if the tract does not intersect any UZAs with good NTD data (22,954 tracts).

The distance from every tract to every UZA is calculation using the Python script run_tract_uza_dist.py. If "uza_dist" is zero, and there is only one UZA in "alpha_beta_uzas" then #1 strategy was used, if uza_dist is zero, and there is more than one UZA in "alpha_beta_uzas" then #2 strategy was used, and if "uza_dist" is greater than zero then #3 was used and the UZA in "alpha_beta_uzas" is the UZA used to assign the α and β for transit cost calculations.

Calibrating the Models

Simultaneous Equations Model (SEM)

The SEM used in LAIM3.0 consists of six nested equations, each drawing from a pool of 18 exogenous variables that predict six interrelated endogenous variables. The final version of the SEM model is captured in the R script sem_avar.r which produces the final fit and outputs the coefficients etc. into various PostgreSQL tables; sem_avar.r makes use of several of the modules in model_sem_fit.r which is library of functions developed by CNT.

Vehicle Miles Traveled (VMT model)

VMT is not included in the SEM due to data limitation and is instead modeled using OLS regression. The regression model was fit using data on the total number of miles households that drive their autos, calculated from odometer readings from the Chicago and St. Louis metro areas for 2013 through 2015, obtained from the Illinois Environmental Protection Agency. The final version of the VMT model is captured in the R script vmt_fit.r which produces the final fit and outputs the coefficients etc. into various PostgreSQL tables; vmt_fit.r makes use of several of the modules in model_sem_fit.r which is library of functions developed by CNT.

Other models used

There are eight other regression models that are used to fix missing data – see "Appendix C: Define Eight Control Households and Run Model."

Running the Models

The transportation behavior and housing cost SEM model and the VMT OLS model need to be run.

SEM model

Running the SEM model is not as simple as running an OLS. When calibrating the SEM model, note that all variables were normalized by subtracting the mean value for all tracts and dividing by the standard deviation, this normalization if completed before the linearization, note that this gives a fit where the average intercept is zero, so in the calibration all the intercepts were set to zero. Consider the following matrix equation representation of a general SEM for each jth endogenous variable (Y_j):

$$Y_j = \sum_{i=1}^{n} \beta_{ij} x_i + \sum_{k=0}^{m} \gamma_{kj} y_k + \zeta_j$$

Where:

- n – the number of exogenous variables
- m – the number of endogenous variable
- xi – the ith exogenous variable
- βijj – the coefficient for the ith exogenous variable to predict the jth endogenous variable
- γijj – the coefficient for the ith exogenous variable to predict the jth endogenous variable
- yik – the measured values for the kth endogenous variable
- ζj – the intercept used to predict the jth endogenous variable
- Yj – the modeled values for the jth endogenous variable

Or in matrix notation:

$$\bar{Y} = \mathrm{B} \cdot \bar{x} + \Gamma \cdot \bar{y} + \bar{\zeta}$$

Where:

- B (Capitol Beta) is the matrix of exogenous coefficients
- Γ (Capitol Gamma) is the matrix of endogenous coefficients
- $\bar{x}$ – the vector of measured exogenous variables
- $\bar{y}$ – the vector of measured endogenous variables
- $\bar{\zeta}$ – the vector of intercepts for each endogenous variable (set to zero for this fit)
- $\bar{Y}$ – the vector of modeled endogenous variables

To find the predicted $\bar{Y}$ one assumes:

$$\bar{Y} = \bar{y}$$

And then solve, assuming $\bar{\zeta}$ is zero:

$$\bar{Y} - \Gamma \cdot \bar{Y} = \mathrm{B} \cdot \bar{x}$$

or

$$(U - \Gamma) \cdot \bar{Y} = \mathrm{B} \cdot \bar{x}$$

therefore

$$\bar{Y} = (U - \Gamma)^{-1} \cdot (\mathrm{B} \cdot \bar{x})$$

Where U is the unitary matrix (a matrix of order m, with 1 for all diagonal elements and zero for every other element think of it as multiplying by one), and $(\mathrm{U} - \Gamma)^{-1}$ – the "inverse of one minus gamma" – being the inverse (solved for and saved in R to the PostgreSQL database table inv_omg) of the $(\mathrm{U} - \Gamma)$ matrix.

The following are the product of the final fit:

Table 4: First half of Beta Matrix

	Fraction of AMI Owners	Area Median Income	Household Size Owners	Block Density	Commuters per HH Owners	Gross HH Density	Employment Access Index	Median Commute Distance	Rooms per HU Owners
Owner Auto Ownership	0.151	0.089	0.127	-0.190	0.259	-0.127	-0.119	0.058	0.061
Renter Auto Ownership	0.000	0.000	0.000	-0.097	0.000	-0.086	-0.153	0.000	0.000
Renter Housing Cost	0.000	0.326	0.000	0.000	0.000	0.000	-0.094	0.000	0.000
Owner Housing Cost	0.561	0.539	0.216	0.000	-0.208	0.000	0.189	0.064	0.000
Owner Transit Commute Share	0.000	0.067	0.109	0.000	0.068	0.255	0.356	0.049	0.044
Renter Transit Commute Share	0.000	0.086	0.000	0.000	0.000	0.101	0.306	-0.045	0.000

Table 5: Second half of Beta Matrix

	Fraction of Single Family Detached HU	Fraction of Rental HU	Retail Employment Access Index	Fraction of AMI Renters	Household Size Renters	Commuters per HH Renters	Local Job Density	Rooms per HU Renters	Local Retail Jobs Density
Owner Auto Ownership	0.262	0.175	-0.190	0.000	0.000	0.000	0.000	0.000	0.000
Renter Auto Ownership	0.102	0.000	-0.081	0.177	0.064	0.326	0.055	0.115	0.000
Renter Housing Cost	0.000	0.000	0.321	0.375	0.153	-0.055	0.000	0.107	0.000
Owner Housing Cost	-0.108	-0.061	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Owner Transit Commute Share	-0.083	-0.107	-0.233	0.000	0.000	0.000	-0.132	0.000	0.057
Renter Transit Commute Share	-0.086	0.000	-0.146	0.000	0.075	0.051	-0.096	0.073	0.000

Table 6: Gamma Matrix

	Owner Auto Ownership	Renter Auto Ownership	Renter Housing Cost	Owner Housing Cost	Owner Transit Commute Share	Renter Transit Commute Share
Owner Auto Ownership	0	0.149695	0	0.048123	0	0
Renter Auto Ownership	0.188324	0	0.14341	0	0	0
Renter Housing Cost	0	0	0	0.309351	0	0
Owner Housing Cost	0	0	0.091598	0	0	0
Owner Transit Commute Share	-0.30422	0	0	0.085218	0	0.247179
Renter Transit Commute Share	0	-0.27406	-0.00934	0	0.433431	0

Table 7: Inverse of Unitary Minus Gamma Matrix

	Owner Auto Ownership	Renter Auto Ownership	Renter Housing Cost	Owner Housing Cost	Owner Transit Commute Share	Renter Transit Commute Share
Owner Auto Ownership	1.029009	0.154037	0.027403	0.057996	0	0

	Owner Auto Ownership	Renter Auto Ownership	Renter Housing Cost	Owner Housing Cost	Owner Transit Commute Share	Renter Transit Commute Share
Renter Auto Ownership	0.193787	1.029009	0.152752	0.05658	0	0
Renter Housing Cost	0	0	1.029162	0.318372	0	0
Owner Housing Cost	0	0	0.094269	1.029162	0	0
Owner Transit Commute Share	-0.36531	-0.13055	-0.01459	0.07335	1.11999	0.276837
Renter Transit Commute Share	-0.21145	-0.33859	-0.0578	0.013313	0.485439	1.11999

The lai_sem_model.sql builds a function that performs these matrix calculations in PostgreSQL.

VMT model

Since the VMT model is an OLS it is very straight forward to calculate the modeled outcomes. The lai_model.sql query builds a function that performs these calculations in PostgreSQL.

Calculating Costs and Producing the Indexes

Finally, all of this can come together for each of the control households and an LAI is produced for every census tract, for each of the eight control households. The Python script run_costs.py does this by looping through the eight control household tables and updating the all the costs. Then it uses the weighted average by tenure to calculate the combined LAI in addition to the tenure specific values. Once this is completed the LAI is finished.

Appendix A: Fitting Data

lai_fitting_data.zip contains a data file (fitting_data.csv), that contains a row for every tract in the US (excluding Puerto Rico) and has associated with it all the data from the 2016 ACS, 2014 Lodes (except for Wyoming who did not submit data for 2014 – so we used 2013 data for Wyoming only), and the Illinois VMT data using the odometer readings from 2013-2015 (only for the Illinois tracts where there was data). For the employment density, using jobs and land area only in the tract itself (simple), for the tract plus ¼ mile buffer (qmile), the tract plus ½ mile buffer (hmile) and the union of a ½ mile buffer around the centroid the tract itself. Table 8 shows the variable names along with its description.

Table 8: Description of Variables

Variable Name	Description
Stfid	Census id
Households	Number of households
owner_occupied_hu	Number of owner occupied housing units
renter_occupied_hu	Number of renter occupied housing units
pct_transit_j2w_renters	Percent of commuters living in rental households using transit for their journey to work
pct_transit_j2w_owners	Percent of commuters living in owner households using transit for their journey to work
median_smoc_mortgage	Median selected monthly ownership costs
median_gross_rent	Median gross rent
autos_per_hh_renters	Autos per household for renter households
autos_per_hh_owner	Autos per household for owner households
commuters_per_hh_renters	Average number of commuters per household in renter households
commuters_per_hh_owners	Average number of commuters per household in owner households
avg_hh_size_renters	Average number of people in renter households
avg_hh_size_owners	Average number of people in owner households
area_income_renter_frac	Fraction of area median household income for renters in this tract relative to the regional median household income
pct_hu_1_detached	Percent of single family detached housing units
median_rooms_per_renter_hu	Median number of rooms in renter households
median_rooms_per_owner_hu	Median number of rooms in owner households
gross_hh_density	Number of households per land acre
pct_renters	Percent of rental housing units
area_income_owner_frac	Fraction of area median household income for owners in this tract relative to the regional median household income

Variable Name	Description
area_median_hh_income	Regional median household income
block_density	Census blocks per acre
avg_block_acres	Average block size in acres
job_density_simple	Jobs per land acre simple
job_density_qmile	Jobs per land acre qmile
job_density_hmile	Jobs per land acre hmile
job_density_union	Jobs per land acre union
retail_density_simple	Retail jobs per land acre simple
retail_density_qmile	Retail jobs per land acre qmile
retail_density_hmile	Retail jobs per land acre hmile
retail_density_union	Retail jobs per land acre union
job_gravity	Total jobs for every tract in the US divided by its distance from the centroid squared
retail_gravity	Retail jobs for every tract in the US divided by its distance from the centroid squared
median_commute	Median distance of commuters in the tracts using the centroid of the employment block
veh_count	Number of vehicles that were used to determine VMT from Illinois odometer reading (for appropriate Illinois tracts only)
avg_vmt	Average VMT from Illinois odometer reading (for appropriate Illinois tracts only)
std_dev_vmt	Standard deviation for VMT from Illinois odometer reading (for appropriate Illinois tracts only)
area_type	'county' or 'cbsa' depending on the location of the tract
area_stfid	Census id for either county of cbsa for this tract's location

Included in the .zip file is also the data from the national transit database (all the files in the subdirectory ntdb) for 2014, and the average gas prices by EPA defined regions for 2014 (gas_prices.csv).

Appendix B: Re-Fit the New Data at Census Tract Level

Remake the Layout for the Fitting Program for SEM for owner/renter for Housing Cost, Autos/HH and %Transit Journey to Work (J2W)

CNT uses the R software environment for statistical computing. CNT has built an internal web-based fitting platform that allows for perusing the complex variable input/output processing for SEM and for multivariate OLS modeling. To implement the HUD LAI SEM model required accessing the data that was developed in Task 1 from the PostgreSQL data base. To make sure that everything in this complex environment is the same as the previous work CNT first accessed and recreated the previous generation of the LAI/SEM model. Once that was completed the new data was loaded into the package so fitting can proceed.

Variable Transformations

Similar to LAIM Version 2.0, SEM variables are transformed to allow for better fits for non-linear relationships in LAIM 3.a. The approach is to apply a series of transformations to each of the endogenous and exogenous variables and pick the transformation that produces the most normal distribution for each one (i.e., the distribution that maximizes the R2 value when compared with a normal distribution). This transformed variable is then standardized by subtracting the mean of the transformed distribution and dividing by the standard deviation:

$$Z = \frac{f - \bar{f}}{StDev}.$$

Where f is the transformed variable, $\bar{f}$ is the mean of the distribution of f and *StDev* is the standard deviation of the distribution of f.

This standardization was applied for all the variables in the SEM function as listed in Table 9 to handle the wide variation in values.

The Original Model – LAIM2.0

This subsection recalls (using the current modeling code) the original SEM model for LAI – referred to the LAIM2.0 model.

Table 9: Model SEM variables including the linearization transformation function

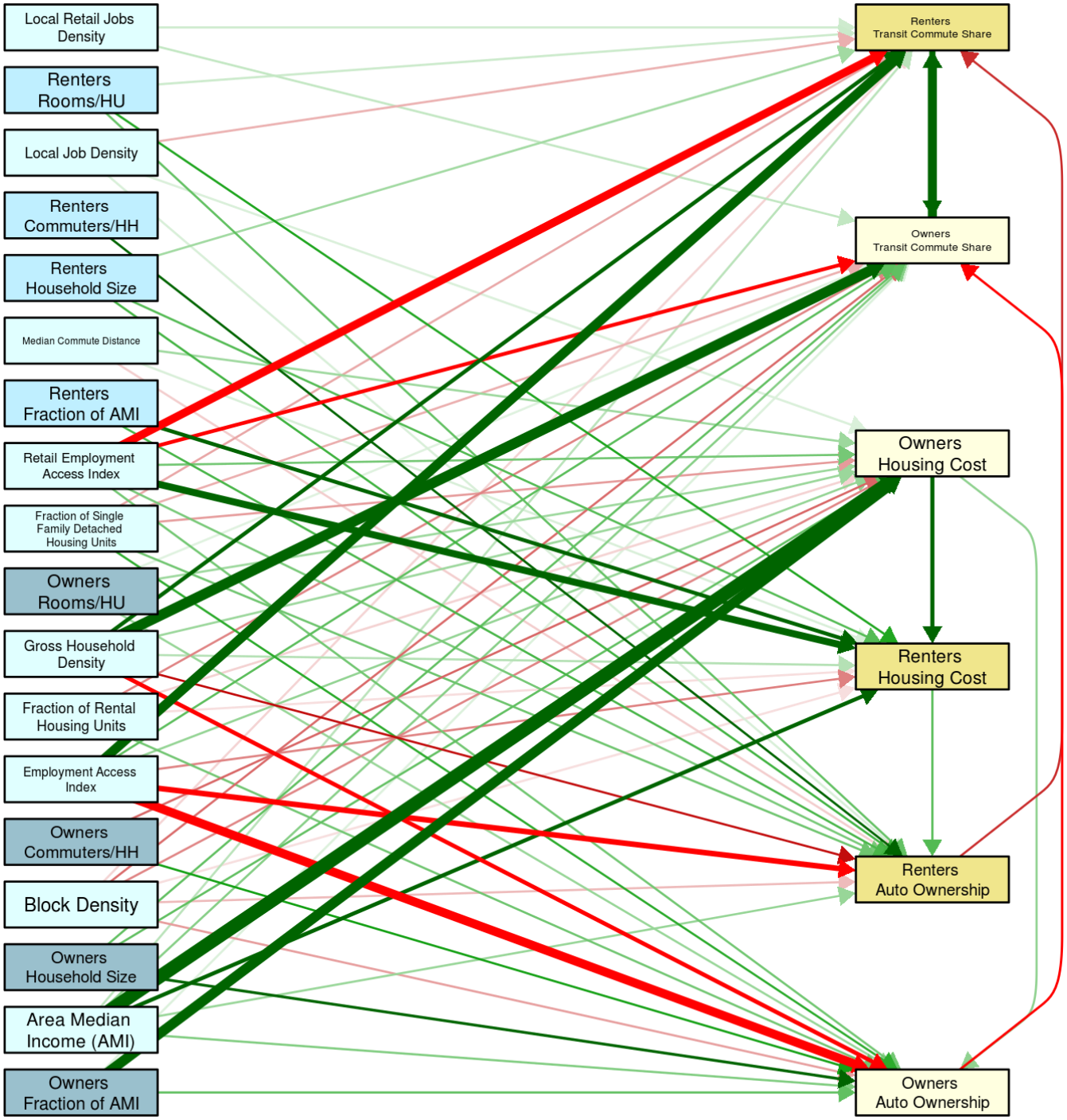
Exogenous Variable	Trans./Formula
Owners Fraction of AMI	ln(x)
Renter Fraction of AMI	ln(x)
Area Median Income (AMI)	ln(x)
Median Commute Distance	ln(x)
Owners Average Household Size	ln(x)
Renter Average Household Size	ln(x)
Block Density	√ x

Endogenous Variable	Trans./Formula
Owner Auto Ownership	x
Renter Auto Ownership	x
Renter Housing Cost	ln(x)
Owner Housing Cost	ln(x)
Owner Transit Commute Share	x
Renter Transit Commute Share	x

Owners Commuters/HH	x
Renter Commuters/HH	x
Employment Access Index	ln(x)
Fraction of Rental HU	√ x
Gross HH Density	√ x
Local Job Density	√ x
Local Retail Jobs Density	√ x
Owners Rooms/HU	x
Renter Rooms/HU	x
Fraction of Single Family Detached Housing Units	x
Retail Employment Access Index	ln(x)

The following path diagram is from the SEM for the LAIM2.0 model showing the interrelationships of the exogenous and endogenous variables.

Figure 5: Path Diagram for LAIM2.0



Recall that red means a negative correlation (example: Employment Access increases, autos/households decreases) and green is a positive correlation (example: Area Median Income increases housing costs increase) and that the width of the line represents the strength of the correlation.

Another way to look at the output of the fit is to examine the strength of the coefficient numerically; the following table lists these. Again, the colors relate the same way as the path diagram.

Table 10: Direct effects for endogenous variables and direct and total effects for exogenous variables for LAIM2.0

Endogenous Variable	Owner Auto Ownership		Renter Auto Ownership		Renter Housing Cost		Owner Housing Cost		Owner Transit Commute Share		Renter Transit Commute Share	
Owner Auto Ownership									-.215			
Renter Auto Ownership											-.172	
Renter Housing Cost			.142									
Owner Housing Cost	.092				.263							
Owner Transit Commute Share											.397	
Renter Transit Commute Share									.321			
Exogenous Effects	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**
Owners Fraction of AMI	.130	.169		.016		.112	.425	.425	.029	-.01		-.007
Renter Fraction of AMI			.149	.186	.256	.256				-.012		-.037
Area Median HH Income (AMI)	.106	.153	.085	.141	.259	.395	.519	.519	.098	.086	.057	.067
Median Commute Distance		.007	-.041	-.033	.032	.054	.081	.081		0		.006
Owners Average HH Size	.227	.239		.005		.034	.129	.129	.129	.089		.035
Renter Average HH Size			.156	.176	.141	.141				.017	.076	.053
Block Density	-.080	-.089	-.056	-.064	-.027	-.054	-.100	-.1	-.122	-.127	-.036	-.076
Owners Commuters per HH	.193	.181		-.005		-.033	-.127	-.127		-.044		-.017
Renter Commuters per HH			.213	.213						-.013		-.042
Employment Access Index	-.392	-.384	-.302	-.31	-.104	-.081	.088	.088	.133	.417	.408	.628
Fraction of Rental Housing Units	.116	.113		-.005	-.027	-.037	-.037	-.037	-.101	-.143		-.056
Gross HH Density	-.264	-.257	-.200	-.189	.061	.082	.082	.082	.434	.666	.256	.553
Local Job Density		.003	.032	.033		.008	.030	.03		-.027	-.065	-.081
Local Retail Jobs Density									.050	.072	.040	.068
Owners Rooms per Housing Unit	.084	.092		.003		.021	.081	.081	.026	.007		.002
Renter Rooms per Housing Unit			.128	.154	.183	.183				.006	.044	.02
Fraction of Single Family Detached HU	.174	.166	.118	.115		-.022	-.082	-.082	-.066	-.151	-.072	-.151
Retail Employment Access Index	.116	.126	.111	.164	.343	.372	.113	.113	-.251	-.466	-.372	-.585

Quoting from CNT's paper – *Untangling Housing Cost and Transportation Interactions: The Location Affordability Index Model—Version 2 (LAIM2.0)*

> *"The LAIM2.0 consists of six nested equations, each drawing from a pool of 18 exogenous variables that predict six endogenous variables—monthly housing cost, auto ownership, and transit commute share for both homeowners and renters. These allow LAIM2.0 to model housing costs and transportation choices, by tenure, for households in urban, suburban, and rural settings. In addition to these relationships, the SEM allows for interaction among the endogenous variables. Figure 1 shows the path diagram for the LAIM2.0. This diagram is presented to simply show the predictive interactions as well as the general strength and sign of those interactions. The details on the interactions are provided in Table 1.*
>
> *The endogenous variables all depend on one another, and this interaction is revealed in the LAIM2.0 structure. Whereas it was assumed that causality can go both ways, in the actual implementation, it was found that once causality is explained in one direction, the other direction is either not statistically significant or markedly less significant, and the overall model goodness of fit is reduced. For example, having monthly housing costs in the auto ownership equation obviates the need to put auto ownership into the monthly housing costs equation. The one exception to this finding is the interaction between owner and renter transit commute share; in these cases, both interactions were found to be important and thus, both were included in the final model."*

Table 11: Goodness-of-fit (GOF) parameters with their values from LAIM2.0

Parameter	Value	Description[5]	Cut-off for Good Fit
Confirmatory Fit Index (CFI)	0.97	Not very sensitive to sample size. Compares the fit of a target model to the fit of an independent, or null, model.	CFI ≥.90
Root Mean Square Error of Approximation (RMSEA)	0.053	A parsimony-adjusted index. Values closer to 0 represent a good fit.	RMSEA < 0.08
(Standardized) Root Mean Square Residual (SRMR)	0.013	The square-root of the difference between the residuals of the sample covariance matrix and the hypothesized model. If items vary in range (i.e. some items are 1-5, others 1-7) then RMR is hard to interpret, better to use SRMR.	SRMR <0.08

[5] These descriptions and cut-offs are copied from a handout from Cornell University's Statistical Consulting Unit. https://www.cscu.cornell.edu/news/Handouts/SEM_fit.pdf - note that this document suggests CFI, RMSEA, SRMS and Chi-Square, but has the caveat that Chi-Square does not work well for large sample sizes. The calculated p-value of these fits is always very significant.

Table 82:R^2s for the 6 endogenous variables for the LAM2.0

Endogenous Variable	R^2 Value
Owner Auto Ownership	54.99%
Renter Auto Ownership	46.98%
Renter Housing Cost	57.81%
Owner Housing Cost	61.05%
Owner Transit Commute Share	62.8%
Renter Transit Commute Share	63.01%

Run the fitting program in R and make sure the distributions and goodness-of-fit all look good

Using the same structure as the LAIM2.0 and the new data and using, for consistency sake, the old algorithm for employment densities (LAM3.a). The same structure was again used this time using the new measure for employment densities (LAIM3.b) The following tables give the goodness of fit and R^2s:

Table 93: GOF for LAIM 2.0/3.a/3.b

Parameter	LAIM2.0 Value	LAIM3.a Value	LAIM3.b Value	Cut-off for Good Fit
CFI	0.97	0.946	0.946	CFI ≥.90
RMSEA	0.053	0.086	0.086	RMSEA < 0.08
SRMR	0.013	0.016	0.015	SRMR <0.08

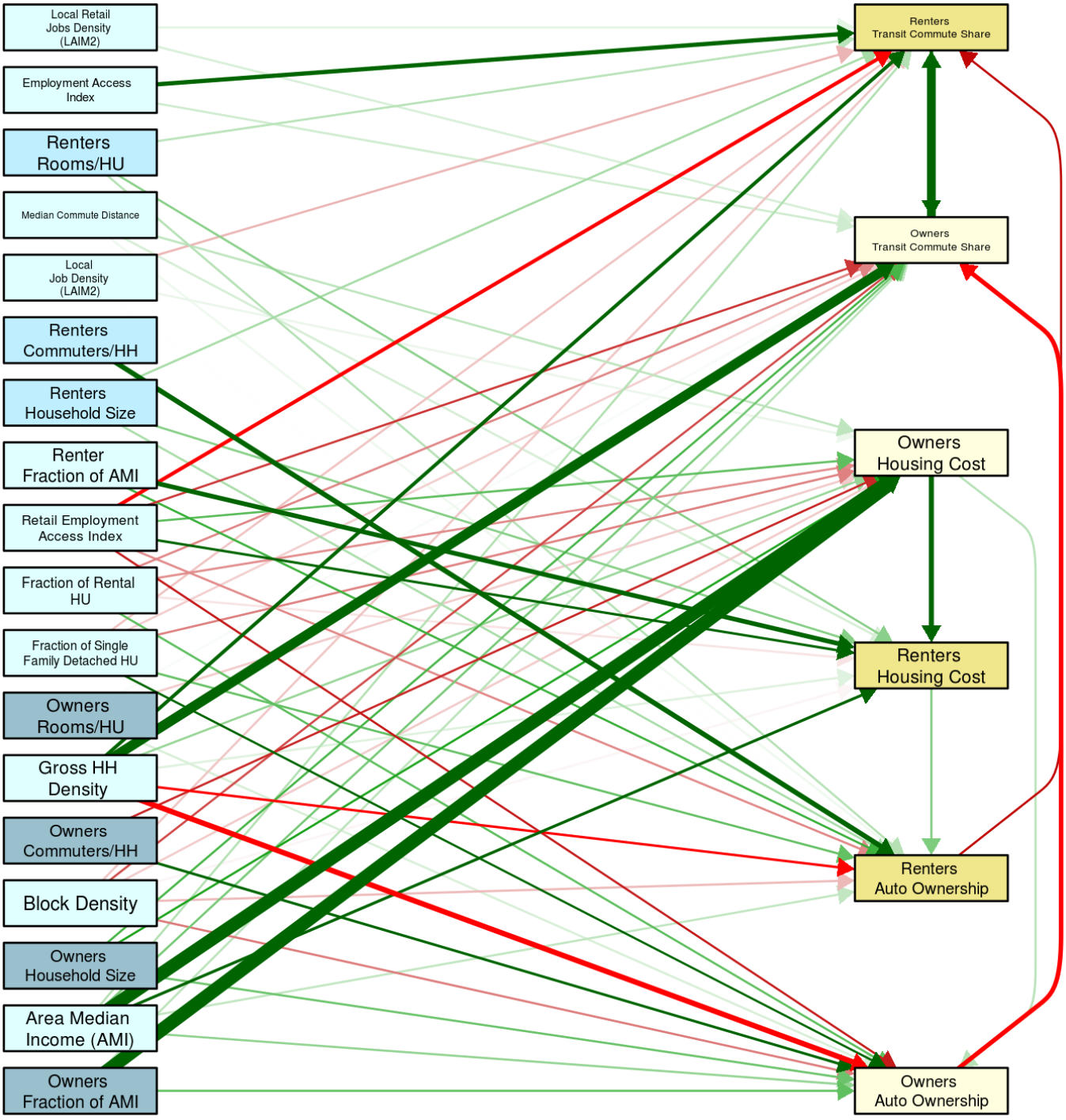
Table 104: R^2s for the 6 endogenous variables for the LAM2.0/3.a/3.b

Endogenous Variable	LAIM2.0 R^2 Value	LAIM3.a R^2 Value	LAIM3.b R^2 Value
Owner Auto Ownership	54.99%	67.78%	67.78%
Renter Auto Ownership	46.98%	66.83%	66.83%
Renter Housing Cost	57.81%	75.4%	75.4%
Owner Housing Cost	61.05%	76.97%	76.97%
Owner Transit Commute Share	62.8%	71.24%	71.12%
Renter Transit Commute Share	63.01%	76.41%	76.39%

While the overall goodness-of-fit parameters are not as robust as in LAIM2.0 the R^2 are all larger – probably related to the fact that by using tracts rather than block groups some of the variation is ignored by the averaging of the variables over the larger area – these models are not optimal, and in fact the RMSEA measure in both LAIM3.a and LAIM3.b do not meet their cut-off for a good fit. In LAM3.a one

of the fit coefficients (how the old measure of employment density relates to Renter Auto Ownership) is not statistically significant, and others are much less significant than in LAIM2.0, and in LAIM3.b two are not statistically significant. Figure 6 is the path diagram for LAIM3.a (LAIM3.b is essentially the same), which is like the original path diagram from LAIM2.0 (note that some of the variables have been reordered) but, the strength of many of the relationships have changed.

Figure 6: Path Diagram for LAIM2.0 structure but new data – LAIM3.a



Adjust linear transformations for any variables that do not look good

To optimize the linear transformations for LAIM2.0 a variety of transformations were tested to see which one had the normal distribution for every variable. That seemed to work well, but we have found that that method does not always give the best fit. So, to optimize the transformation for each variable 6 OLS models were used for each of the endogenous variables using the appropriate exogenous ones and cycling through each variable using one of 6 transformation formulae. The following table shows an example of this for the Employment Access Index showing that the previous functional form of natural log (ln(x)) was on average not the optimal transformation, and by changing it to the square root ($\sqrt{x}$) would on average improve the fits.

Table 115: R^2 for OLS fit using all variables, but changing the transformation function for x=Employment Access Index; showing that on average the square root transformation gives the best R^2

Endogenous Variable Function	Owner Auto Ownership	Renter Auto Ownership	Renter Housing Cost	Owner Housing Cost	Owner Transit Commute Share	Renter Transit Commute Share	Average
x	69.39%	67.90%	70.72%	77.51%	59.82%	60.43%	**67.63%**
sqrt(x)	69.39%	67.84%	70.73%	77.56%	60.89%	62.29%	**68.12%**
ln(x)	69.47%	67.63%	70.68%	77.74%	57.05%	58.53%	**66.85%**
ln(1+x)	69.47%	67.63%	70.68%	77.74%	57.05%	58.53%	**66.85%**
1/x	69.21%	67.46%	70.70%	77.83%	57.50%	58.72%	**66.90%**
1/(1+x)	69.21%	67.46%	70.70%	77.83%	57.51%	58.73%	**66.91%**

After proceeding through all the endogenous variables a few of linearization formula were changes to the optimal form. The following table shows these changes that are implemented in a new SEM LAIM3.c:

Table 126: Linearization transformation functions from LAIM2.0 and the optimized functions

Variable	**LAIM2.0/3.a/3.b**	**LAIM3.c/3.d/3.e**	**Changed**
Owners Fraction of AMI	ln(x)	ln(x)	
Renter Fraction of AMI	ln(x)	ln(x)	
Area Median Income (AMI)	ln(x)	ln(x)	
Owners Average Household Size	ln(x)	x	X
Renter Average Household Size	ln(x)	x	X
Block Density	√ x	√ x	
Owner Commuters/HH	x	x	
Renter Commuters/HH	x	x	
Gross HH Density	√ x	x	X

Local Job Density	√ x	√ x	
Employment Access Index	ln(x)	√ x	X
Median Commute Distance	ln(x)	ln(x)	
Owners Rooms/HU	x	x	
Renter Rooms/HU	x	x	
Fraction of Single Family Detached HU	x	√ x	X
Fraction of Rental HU	√ x	x	X
Local Retail Jobs Density	√ x	ln(1+x)	X
Retail Employment Access Index	ln(x)	ln(x)	
Owner Auto Ownership	x	x	
Renter Auto Ownership	x	x	
Renter Housing Cost	ln(x)	ln(x)	
Owner Housing Cost	ln(x)	ln(x)	
Owner Transit Commute Share	x	x	
Renter Transit Commute Share	x	x	

Using the same structure as LAIM2.0/3.a/3.b but with these new transformations (LAIM3.c) we get the following goodness-or-fit and R^2s:

Table 137: GOF for LAIM 2.0/3.a/3.b/3.c

Parameter	LAIM2.0 Value	LAIM3.a Value	LAIM3.b Value	LAIM3.c Value	Cut-off for Good Fit
CFI	0.97	0.946	0.946	0.944	CFI ≥.90
RMSEA	0.053	0.086	0.086	0.089	RMSEA < 0.08
SRMR	0.013	0.016	0.015	0.015	SRMR <0.08

Table 148: R^2s for the 6 endogenous variables for the LAM2.0/3.a/3.b/3.c

Endogenous Variable	LAIM2.0 R^2 Value	LAIM3.a R^2 Value	LAIM3.b R^2 Value	LAIM3.c R^2 Value
Owner Auto Ownership	54.99%	67.78%	67.78%	67.55%
Renter Auto Ownership	46.98%	66.83%	66.83%	67.31%
Renter Housing Cost	57.81%	75.4%	75.4%	75.41%
Owner Housing Cost	61.05%	76.97%	76.97%	76.82%
Owner Transit Commute Share	62.8%	71.24%	71.12%	75.02%
Renter Transit Commute Share	63.01%	76.41%	76.39%	75.4%

While this looks about the same as before, there is only one relationship that is not statistically significant. The conclusion to draw from this is that the SEM (and the OLS) models are not very sensitive to small changes in the linearization transformations.

Optimize final fit

Since we are confining ourselves to the variables used in LAIM2.0 (except for a slight change in the two local employment densities), and using the finding that the linearizing transformation functions can be optimized, we have four choices

1. Keep the structure of the model the same as LAIM2.0 and refit using the new linear transformation functions – this is LAIM3.c
2. Keep the endogenous linkages the same as LAIM2.0 but allow for optimal combinations of exogenous variables to go into the fit – LAIM3.d
3. Optimize the structure keeping all statistically significant interactions of exogenous and endogenous variables – LAIM3.e
4. Optimize the structure keeping all statistically significant interactions of exogenous variables but only using the interactions of the endogenous that link like tenure and transportation choice – LAIM3.f. The following summarizes the goodness of fit for all the models discussed in this appendices:

Table 159: GOF for LAIM 2.0/3.a/3.b/3.c/3.d/3.e/3.f

Parameter	LAIM2.0 Value	LAIM3.a Value	LAIM3.b Value	LAIM3.c Value	LAIM3.d Value	LAIM3.e Value	LAIM3.f Value
CFI	0.97	0.946	0.946	0.944	0.957	0.991	0.974
RMSEA	0.053	0.086	0.086	0.089	0.084	0.039	0.056
SRMR	0.013	0.016	0.015	0.015	0.014	0.006	0.014

Table 20: R2s for the 6 endogenous variables for the LAM2.0/3.a/3.b/3.c/3.d/3.e/3.f

Endogenous Variable	LAIM2.0 R^2 Value	LAIM3.a R^2 Value	LAIM3.b R^2 Value	LAIM3.c R^2 Value	LAIM3.d R^2 Value	LAIM3.e R^2 Value	LAIM3.f R^2 Value
Owner Auto Ownership	54.99%	67.78%	67.78%	67.55%	69.17%	73.22%	72.38%
Renter Auto Ownership	46.98%	66.83%	66.83%	67.31%	67.83%	68.49%	71.27%
Renter Housing Cost	57.81%	75.4%	75.4%	75.41%	75.89%	76.84%	75.64%
Owner Housing Cost	61.05%	76.97%	76.97%	76.82%	77.14%	78.31%	78.48%
Owner Transit Commute Share	62.8%	71.24%	71.12%	75.02%	74.57%	61.04%	73. 37%
Renter Transit Commute Share	63.01%	76.41%	76.39%	75.4%	77.04%	78.04%	77.85%

Figure 7: Path Diagram for LAIM3.d

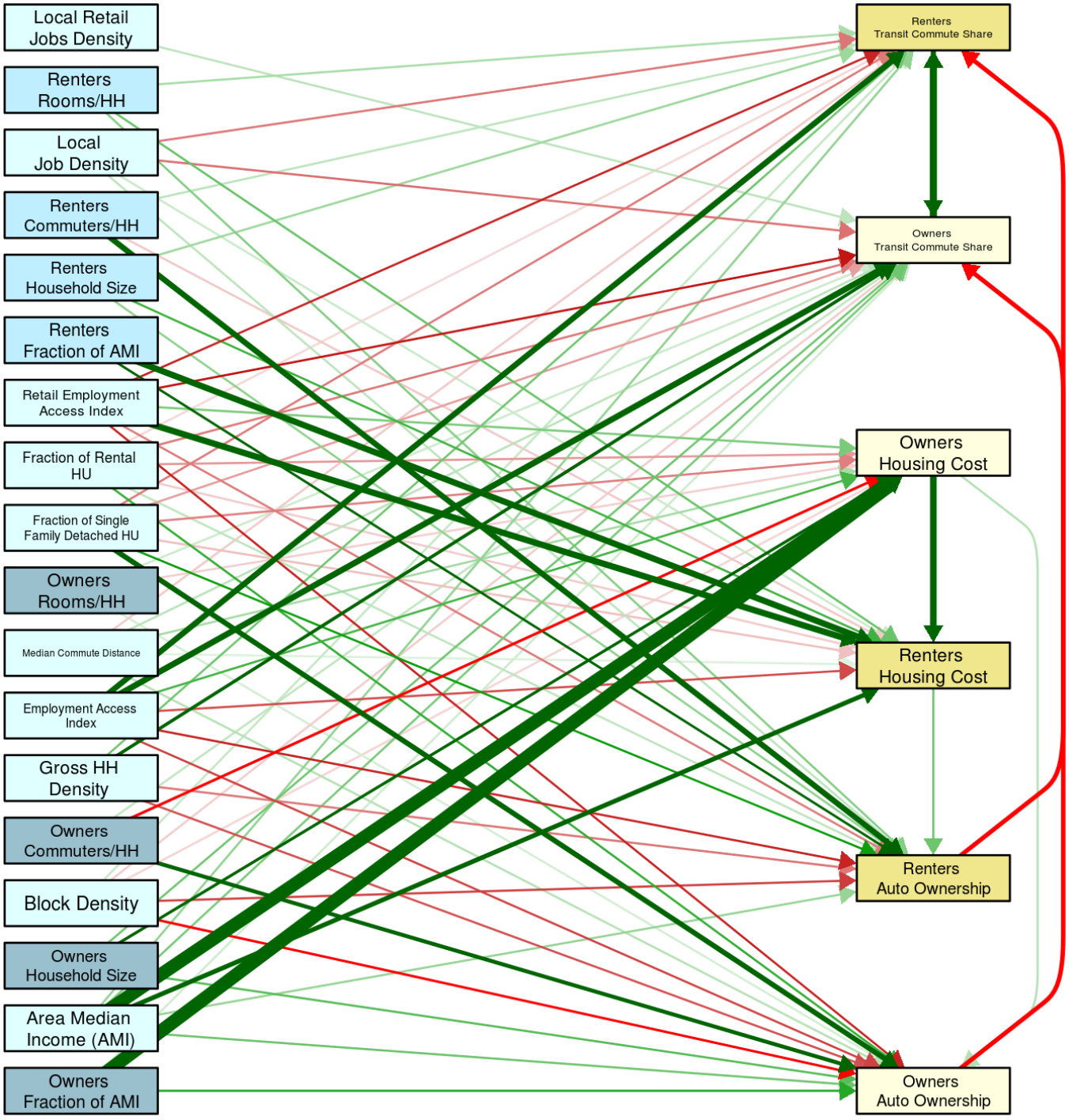


Table 161: Coefficient strength LAMI3.d

Endogenous Variable	Owner Auto Ownership		Renter Auto Ownership		Renter Housing Cost		Owner Housing Cost		Owner Transit Commute Share		Renter Transit Commute Share	
Owner Auto Ownership									-.279			
Renter Auto Ownership											-.289	
Renter Housing Cost			.139									
Owner Housing Cost	.061				.359							
Owner Transit Commute Share											.363	
Renter Transit Commute Share									.327			
Exogenous Effects	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**
Owner Income/AMI	.181	0.218		0.03		0.217	.604	0.604	.039	-0.028		-0.019
Median Income/AMI			.183	0.232	.347	0.347				-0.025		-0.076
Area Median Income (AMI)	.117	0.153	.063	0.133	.291	0.502	.588	0.588	.104	0.09	.091	0.086
Owner HH Size	.122	0.136		0.012		0.083	.230	0.23	.112	0.083		0.027
Renter HH Size	.025	0.025		0.022	.155	0.155				0.02	.080	0.081
Block Density	-.217	-0.219	-.157	-0.158		-0.011	-.030	-0.03	-.041	0.057	.049	0.115
Owner Commuters/HH	.270	0.256		-0.011		-0.081	-.226	-0.226	.050	-0.024		-0.005
Renter Commuters/HH			.383	0.377	-.043	-0.043				-0.021	.053	-0.063
Gross HH Density	-.142	-0.142	-.107	-0.107					.241	0.373	.116	0.283
Simple Job Density		-0.001	.044	0.049	.042	0.033	-.024	-0.024	-.111	-0.178	-.126	-0.205
Job Gravity	-.152	-0.142	-.164	-0.177	-.146	-0.088	.161	0.161	.329	0.565	.342	0.598
Median Commute Distance	.058	0.062	.027	0.034	.027	0.053	.073	0.073	.056	0.024	-.043	-0.044
Owner Rooms/HH	.039	0.036		-0.002		-0.015	-.041	-0.041	.038	0.032		0.012
Renter Rooms/HH			.131	0.148	.122	0.122				0.01	.070	0.031
Percent SFD	.306	0.3	.207	0.194	-.051	-0.089	-.106	-0.106	-.076	-0.242	-.108	-0.252
Percent Renters	.166	0.161	.067	0.057	-.041	-0.069	-.079	-0.079	-.107	-0.192	-.035	-0.121
Simple Retail Density									.052	0.069	.026	0.051
Retail Gravity	-.182	-0.176	-.100	-0.045	.351	0.39	.108	0.108	-.182	-0.215	-.186	-0.251

Figure 8: Path diagram LAIM3.e

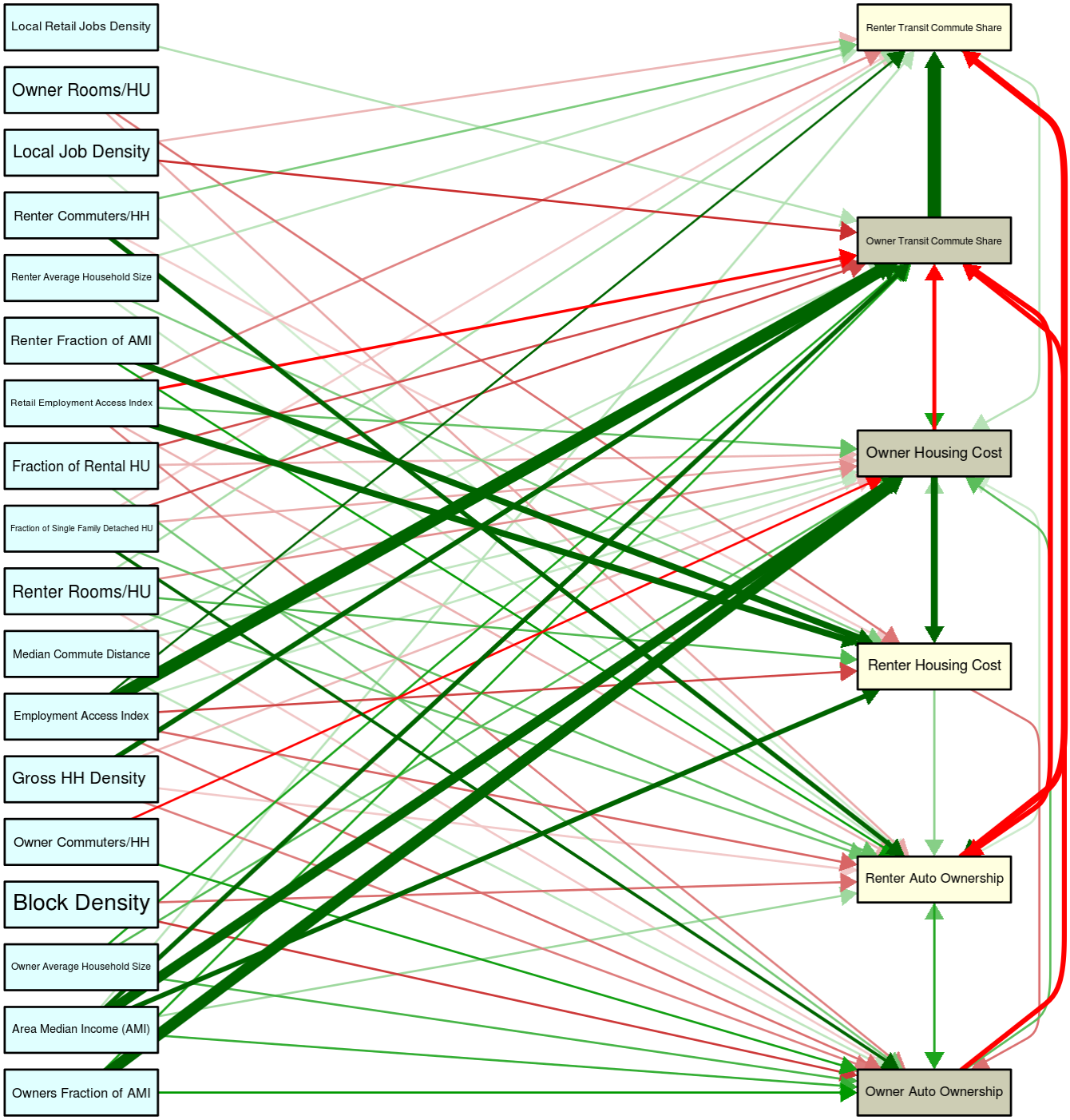
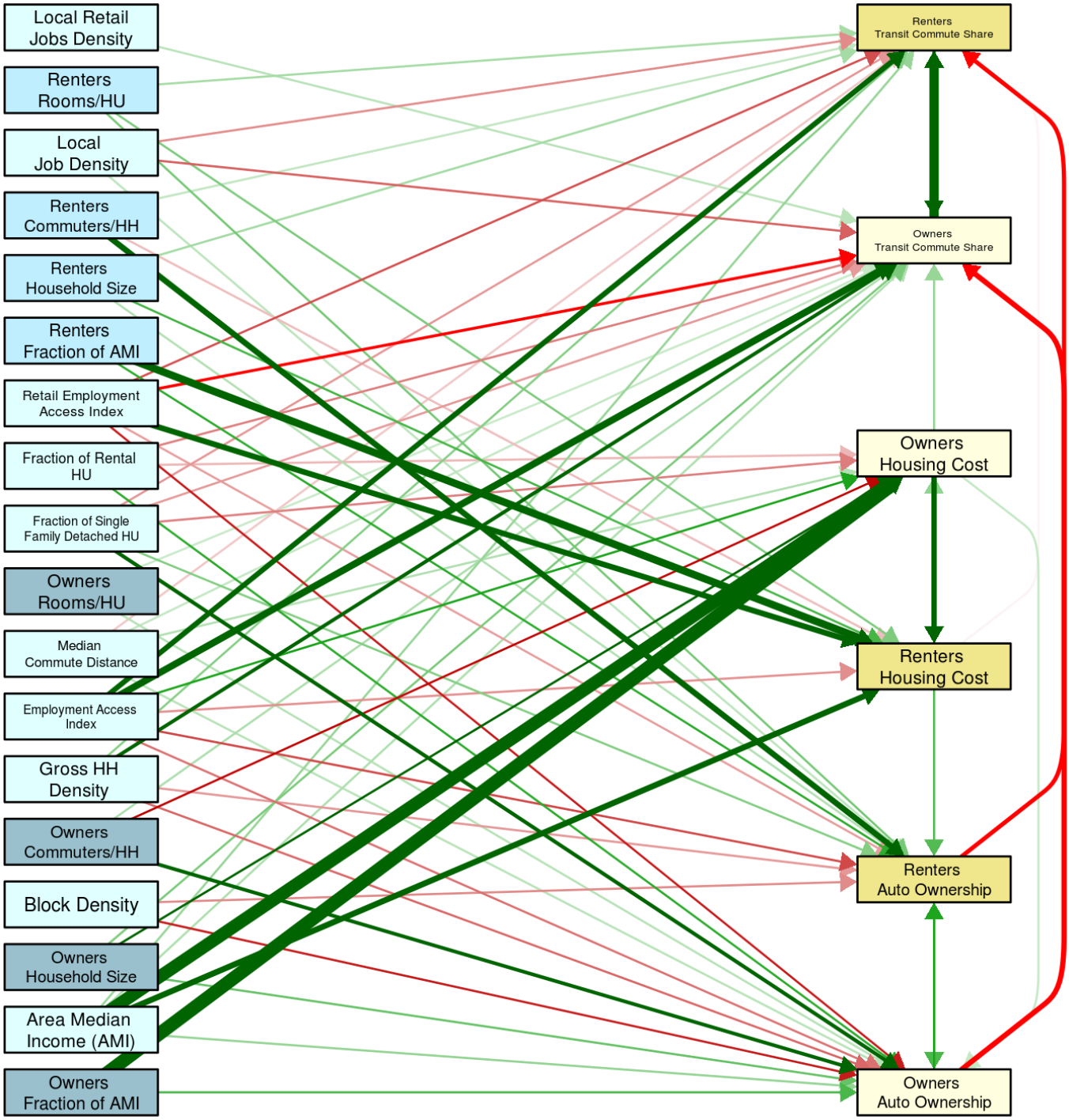


Table 172: Coefficient strength LAMI3.e

Endogenous Variable	Owner Auto Ownership		Renter Auto Ownership		Renter Housing Cost		Owner Housing Cost		Owner Transit Commute Share		Renter Transit Commute Share	
Owner Auto Ownership			.141				.150		-.320			
Renter Auto Ownership	.209						.045				-.378	
Renter Housing Cost	-.132		.111				.087					
Owner Housing Cost					.379				-.29			
Owner Transit Commute Share			-.324				.218				.582	
Renter Transit Commute Share			.25				.066					
Exogenous Effects	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**
Owner Income/AMI	.235	0.217		0.053		0.214	.517	0.565	.214	-0.019		-0.031
Median Income/AMI		0.002	.227	0.248	.363	0.377		0.035		-0.011		-0.1
Area Median Income (AMI)	.187	0.149	.089	0.154	.318	0.535	.472	0.573	.309	0.095	.067	0.064
Owner HH Size	.169	0.161		0.013		0.082	.159	0.216	.212	0.098		0.052
Renter HH Size		0	.056	0.078	.117	0.122		0.015		-0.004	.062	0.03
Block Density	-.182	-0.215	-.134	-0.163		-0.007		-0.017		0.074		0.105
Owner Commuters/HH	.218	0.233		0.024		-0.08	-.235	-0.21		-0.014		-0.017
Renter Commuters/HH		0.072	.310	0.32	-.048	-0.043		0.014		-0.027	.118	-0.019
Gross HH Density	-.120	-0.146	-.049	-0.124		0.002	-.065	0.004	.321	0.367		0.26
Simple Job Density		0.014	.044	0.055		-0.019		-0.05	-.192	-0.183	-.068	-0.195
Job Gravity	-.134	-0.162	-.142	-0.2	-.171	-0.104	.055	0.176	.556	0.557	.234	0.634
Median Commute Distance	.060	0.058		0.007		0.029	.059	0.076	.063	0.023		0.011
Owner Rooms/HH		-0.001	-.081	-0.088	-.128	-0.133		-0.013		0.004		0.036
Renter Rooms/HH	-.053	-0.04	.135	0.143	.159	0.129	-.103	-0.081		0.036	.079	0.045
Percent SFD	.266	0.312	.142	0.196		-0.036	-.078	-0.096	-.172	-0.244	-.048	-0.264
Percent Renters	.127	0.141		0.045		-0.036	-.068	-0.096	-.166	-0.183		-0.123
Simple Retail Density		-0.003		-0.011		0.006		0.017	.070	0.066		0.043
Retail Gravity	-.119	-0.18	-.073	-0.039	.366	0.398	.142	0.084	-.253	-0.22	-.114	-0.227

Figure 9: Path diagram LAIM3.f



To choose the exogenous variables interactions for LAIM 3.f the following logic was applied:

1. Allow interactions between variables for the same tenure – i.e. among renter variables, or owner variables

2. Allow interaction between variables that measure the same thing but for different tenures i.e. housing cost, auto ownership and commute choice
3. Only use the endogenous interactions in one direction – i.e. use the more significant of "renter autos/HH with renter housing cost," or "renter housing cost with renter autos/HH," except when the behavior is the same i.e. use both renter housing cost with owner housing cost and owner housing cost with renter housing cost.
4. Only use the exogenous interactions that are statistically significant and improve the goodness of fit.

Table 183: Coefficient strength LAMI3.f

Endogenous Variable	Owners Auto Ownership		Renters Auto Ownership		Renters Housing Cost		Owners Housing Cost		Owners Transit Commute Share		Renters Transit Commute Share	
Owners Auto Ownership			.188						-.305			
Renters Auto Ownership	.150										-.274	
Renters Housing Cost			.144				.092				-.009	
Owners Housing Cost	.048				.309				.085			
Owners Transit Commute Share											.434	
Renters Transit Commute Share									.246			
Exogenous Effects	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**	**Direct**	**Total**
Owners Fraction of AMI	.151	0.188		0.061		0.179	.561	0.577		-0.014		-0.024
Renters Fraction of AMI		0.038	.177	0.239	.375	0.386		0.035		-0.028		-0.081
Area Median Income (AMI)	.089	0.132		0.098	.326	0.507	.539	0.585	.067	0.101	.086	0.098
Owners Household Size	.128	0.144		0.037		0.069	.216	0.222	.109	0.092		0.029
Renters Household Size		0.014	.064	0.09	.153	0.158		0.014		0.01	.075	0.053
Block Density	-.190	-0.211	-.097	-0.137						0.082		0.073
Owners Commuters/HH	.258	0.253		0.038		-0.066	-.208	-0.214	.067	-0.034		-0.025
Renters Commuters/HH		0.049	.326	0.327	-.055	-0.056		-0.005		-0.028	.051	-0.05
Gross HH Density	-.127	-0.144	-.086	-0.114					.255	0.371	.101	0.293
Local Job Density		0.008	.055	0.057					-.132	-0.181	-.096	-0.19
Employment Access Index	-.119	-0.137	-.153	-0.184	-.094	-0.036	.189	0.186	.356	0.562	.306	0.6
Median Commute Distance	.058	0.064		0.015		0.02	.064	0.065	.049	0.026	-.045	-0.038
Owners Rooms/HU	.061	0.063		0.012					.044	0.027		0.009
Renters Rooms/HU		0.021	.115	0.135	.107	0.11		0.01		0.003	.073	0.036
Fraction of Single Family Detached HU	.262	0.279	.103	0.15		-0.034	-.108	-0.111	-.083	-0.233	-.086	-0.228
Fraction of Rental HU	.175	0.177		0.03		-0.019	-.061	-0.063	-.107	-0.189		-0.09
Local Retail Jobs Density									.057	0.064		0.028
Retail Employment Access Index	-.190	-0.199	-.081	-0.071	.321	0.33		0.03	-.234	-0.227	-.146	-0.228

Run OLS for VMT/HH fit for the Illinois tracts

Since odometer readings are only collected in Illinois, the VMT model was done as a sperate OLS fit to these data. There is also no tenure information associated with each odometer reading so the model ignores tenure to produce an overall household VMT model.

The same variables and transformation functions used in the SEM described above are used except where tenure is involved, so rather that use Owners Commuters/HH and Renters Commuters/HH we use the average Commuters/HH to replace those independent (exogenous) variable. Table 24 list the independent variables and the transformation used in the fit to dependent variable ***average household VMT***:

Table 194: Independent variables for the VMT OLS model

Variable	Transformation Function
Fraction of AMI	ln(x)
Area Median Income	ln(x)
Household Size	x
Block Density	sqrt(x)
Commuters/HH	x
Gross HH Density	x
Local Job Density	sqrt(x)
Employment Access Index	sqrt(x)
Median Commute Distance	ln(x)
Rooms/HU	x
Fraction of Single Family Detached HU	sqrt(x)
Fraction of Rental HU	x
Local Retail Jobs Density	ln(1+x)
Retail Employment Access Index	ln(x)

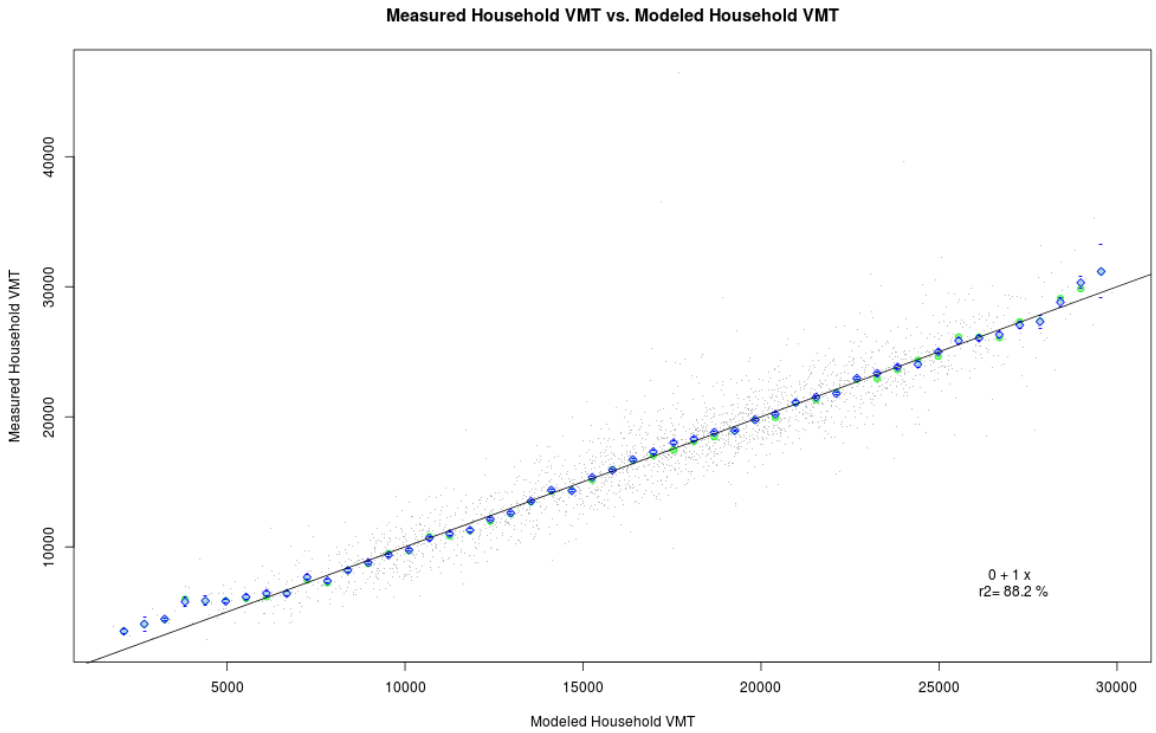
Using the same flexible form used in LAIM2.0 and keeping only the variables and interaction combinations that are statistically significant we get the following coefficients:

Table 205: Final fit coefficients, with error estimate, statistical significant, individual R^2, the incremental improvement in R^2 by adding this variable, and the variance inflation factor (VIF) measuring multicollinearity

Variable	Function	Value	Est. Error	Pr(>|t|)	Individual R^2	Incremental R^2	VIF
Intercept	NA	12373	1767	0.00%	NA	NA	NA
Commuters/HH | Fraction of Single Family Detached HU	x | √ x	-584	104	0%	68.6%	68.6%	91.8
Block Density | Rooms/HU	√ x | x	-1811	172	0%	26.1%	80.6%	16.5
Median Commute Distance	ln(x)	-4756	825	0%	44.3%	83.3%	**75.1**
Commuters/HH | Rooms/HU	x | x	555	81	0%	50.2%	84.3%	20.3
Retail Employment Access Index | Retail Employment Access Index	ln(x) | ln(x)	-178	21	0%	41.5%	85.8%	**55.2**
Fraction of Rental HU | Fraction of Rental HU	x | x	.3	.1	0.16%	49.5%	86.3%	15.3
Median Commute Distance | Retail Employment Access Index	ln(x) | ln(x)	810	110	0%	13.8%	86.6%	**46.7**
Household Size | Fraction of Single Family Detached HU	x | √ x	117	26	0%	59.9%	86.7%	24.7
Fraction of AMI | Fraction of AMI	ln(x) | ln(x)	-1242	169	0%	11.6%	87.1%	2.5
Block Density | Block Density	√ x | √ x	4127	986	0%	48.5%	87.2%	18.6
Commuters/HH | Gross HH Density	x | x	-52	19	0.61%	22.6%	87.3%	28.5
Block Density | Gross HH Density	√ x | x	77	15	0%	8.2%	87.5%	23.8
Gross HH Density | Median Commute Distance	x | ln(x)	-28	13	3.32%	16.9%	87.6%	**61.8**
Commuters/HH | Fraction of Rental HU	x | x	-40	7	0%	29.2%	87.6%	17.4
Area Median Income | Commuters/HH	ln(x) | x	964	110	0%	20.8%	87.7%	**63.2**
Area Median Income | Fraction of Single Family Detached HU	ln(x) | √ x	76	12	0%	63%	87.7%	**59.1**
Fraction of AMI | Local Job Density	ln(x) | √ x	155	86	7.19%	4.7%	87.8%	4.9
Commuters/HH | Employment Access Index	x | √ x	-12	2	0%	19.1%	87.8%	**28.9**
Employment Access Index | Employment Access Index	√ x | √ x	.016	.002	0%	28.3%	88%	12.9
Local Job Density | Median Commute Distance	√ x | ln(x)	-278	63	0%	16%	88.1%	21.4
Household Size | Local Job Density	x | √ x	440	84	0%	17.3%	88.1%	**43.4**
Commuters/HH | Local Job Density	x | √ x	-479	135	0.04%	14.1%	88.2%	**36.8**
Fraction of AMI | Commuters/HH	ln(x) | x	455	198	2.16%	21.3%	88.2%	6.8

The final fit gives and R2 = 88.24 with the modeled vs measured plot shown here:

Figure 10: Measured vs Modeled household VMT



Summary

The following charts display the goodness of fit variables for all the SEM models, and the R^2 for each Endogenous Variable:

Figure 11: CFI for each of the models, note that all satisfy the cut-off criteria

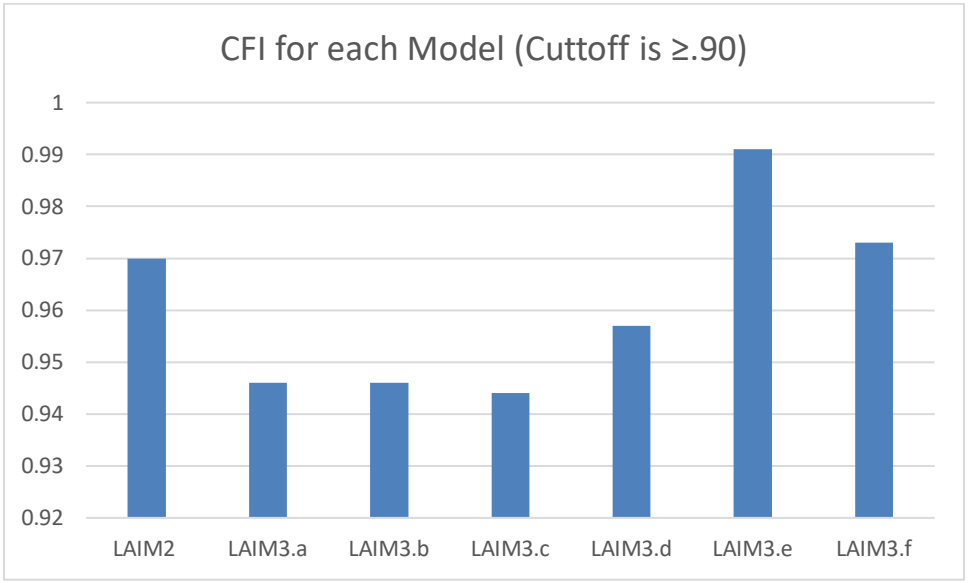


Figure 12: RMSEA for each of the models, note that besides the original, only LAIM3.e and LAIM3.f satisfy the cut-off criteria

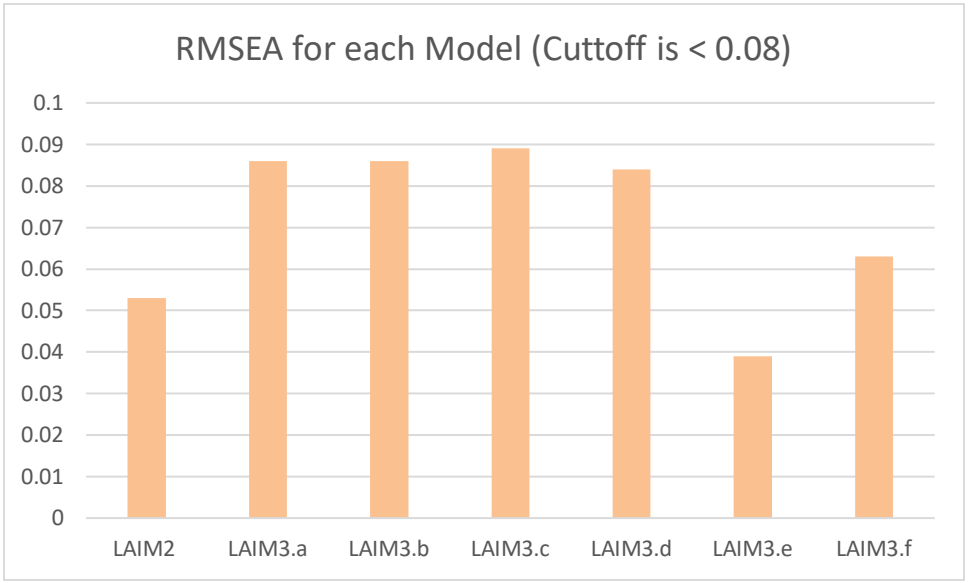


Figure 13: SRMR for each of the models, note that all satisfy the cut-off criteria

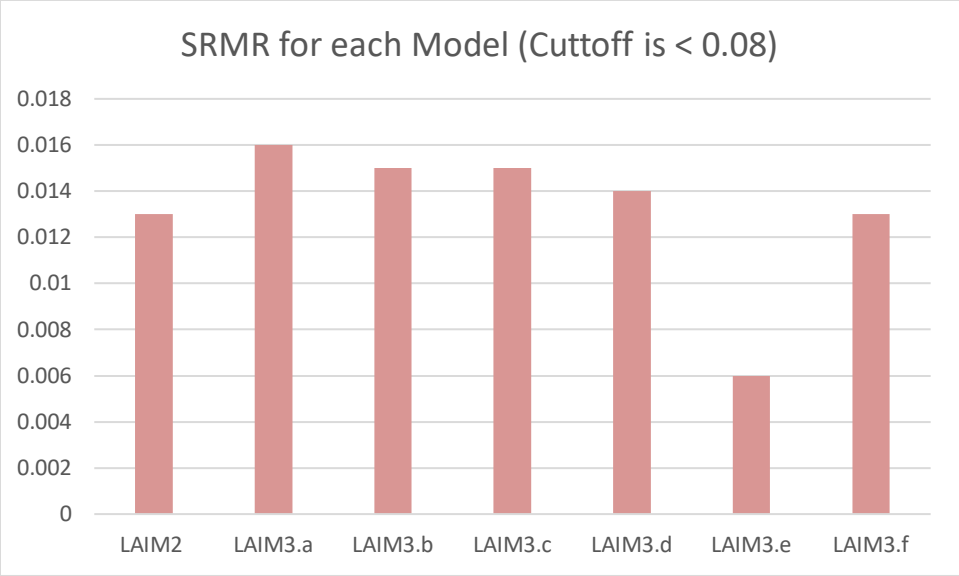
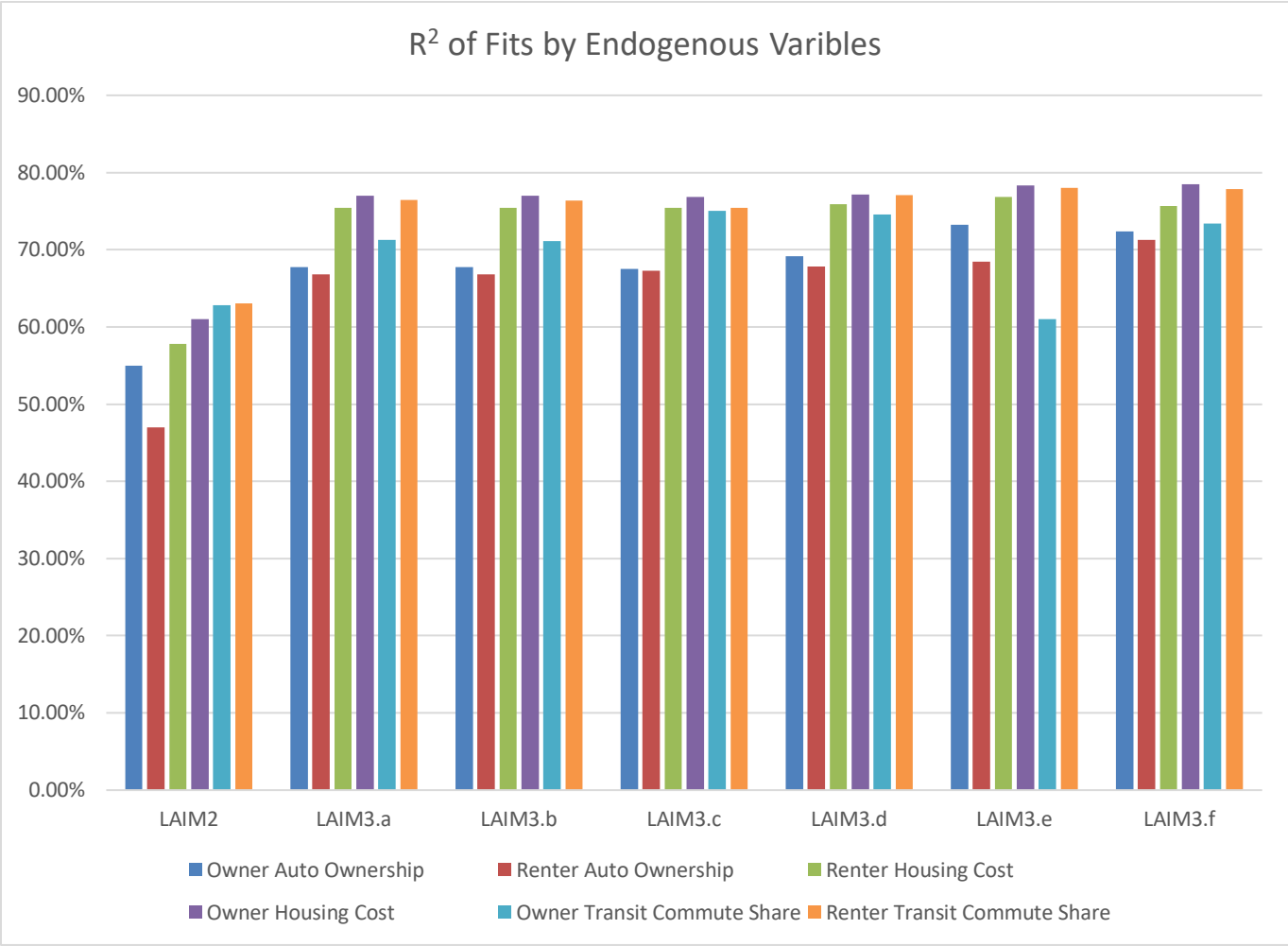


Figure 14: R^2 for each endogenous variable from all the models



The two new models that meet all the goodness of fit measures are LAIM3.e and LAIM3.f. Even though LAIM3.e out performs LAIM3.f, because of the non-intuitive way the endogenous variables interact **the suggested optimal model in LAIM3.f, this model will now be referred to as LAIM3.0 for the rest of this document**. The new VMT model uses the same variables as LAIM2.0 however using different combinations to optimize the fit that produces a very robust R^2.

Appendix C: Define Eight Control Households and Run Model

Control Households

To isolate the built environment's influence on the balance between transportation and housing costs, the exogenous household variables (income, household size, and commuters per household) are set at fixed values (i.e., the "household profiles") in the Model's outputs to control for any variation they might cause. By establishing and running the model for a "household profile," any variation observed in housing and transportation costs can be attributed to aspects of the built environment (including location within the metropolitan area), rather than household characteristics. The Location Affordability Index, as the name suggests, is a spatial index, and so as such needs to control for household characteristics to make for a good index. Obviously, one cannot have an index that uses as its control household the "local" household – the typical households that live in any given tract, since that would show that most places are affordable for the people who live in a given location – in other words wealthy households live in expensive areas, and low-income households live in lower cost areas. Thus, the control household used really does not matter much if it is the same everywhere where comparisons will be made. However, it is useful to model a variety of households to answer different questions, so HUD has chosen eight different households to compare. The model was run for the eight household types in the LAI, each characterized by income, household size, and number of commuters (the same built environment inputs were used each time). They are listed in Table 26.

Table 216: HUD's Eight Control Households

ID	Household Type	Income	Size	Number of Commuters
1	Median-income family	Area Median Household Income	4	2
2	Very low-income individual	National poverty line ($11,880 for a single person household in 2016)[6]	1	1
3	Working individual	50% of median Household Income	1	1
4	Single professional	135% of area median Household Income	1	1
5	Retired couple	80% of area median Household Income	2	0

[6] https://aspe.hhs.gov/computations-2016-poverty-guidelines

6	Single-parent family	50% of area median Household Income	3	1
7	Moderate-income family	80% of area median Household Income	3	1
8	Dual-professional family	150% of area median Household Income	4	2

Note that only household #2 can be compared on a national level, since all the others depend on the area median income as an input, but the others are good to compare within an area (CBSA for Metro/Micropolitan areas and County for Rural areas), but also to compare these areas. The model was run for both owner and renter tenure for each profile.

Running LAIM3.0 on Eight Control Households

To run the model over these household we created eight new tables each with all the exogenous and endogenous variables plus columns from α (average household's transit cost), β (journey to work transit trips), gas price and a column that indicate which income bracket (for the income bin for the auto ownership costs) the control household is in. These eight tables are then run through a program that uses the SEM calibrations results to calculate the endogenous variables (Owner Auto Ownership, Renter Auto Ownership, Renter Housing Cost, Owner Housing Cost, Owner Transit Commute Share, Renter Transit Commute Share) using the controlled exogenous input variables as defined in the SEM structure

Fixing missing data

Running the model of the eight households works well, except that there are many tracts that do not have certain statistics by tenure, for example in a tract where there are mostly home owners, and relatively few renters, the ACS often suppresses household income, size, commuters, and rooms/HU for renters, but does give it for owners. This causes a problem since when the SEM equation is implemented to solve for the endogenous variables, an answer for any of endogenous variables will be undefined when there is missing exogenous input variables (note that this does not change the fit, since these tracts are excluded from the SEM regression). CNT developed several ways to substitute reasonable values when there is missing data. The following table list the method (each method is described in the following set of bullet points in the order they are applied) used to estimate these missing values in the 72,241 tracts that have households, 67,293 have all the data filed in, of the remaining 4,948 tracts here is how they were fixed:

Table 22: Counts of problem census tracts and how they were fixed

Exogenous Variable	Total missing	Algebra fix	Model Fix	Full overwrite	Use other tenure overwrite	average of neighbor values
Owners Fraction of AMI	1381	1181	29	17	8	146
Renter Fraction of AMI	3683	3049	481	2	151	0
Owners Average Household Size	486	390	65	0	0	31
Renter Average Household Size	152	46	75	0	31	0
Owners Commuters/HH	479	382	83	0	14	0
Renter Commuters/HH	149	63	67	7	12	0
Owners Rooms/HU	791	623	38	10	1	119
Renter Rooms/HU	447	275	52	0	120	0

- **Algebra** method: The first fix is to use the other tenure and the overall measure along with the percent of rental occupied housing unit to estimate the missing variable, if the other two are available. Consider a tract with p Fraction of Rental HU, r value for the rental component (Renter Fraction of AMI for example), t value for the total component (Fraction of AMI for example) and o the value for the owner component that is missing because the ACS suppressed this (Owners Fraction of AMI for example), if we say that the total should be the weighted average of the two tenure components we have:

$$t = \frac{r \times p + o \times (100 - p)}{100}$$

Equation 1: Simple weighted average of the two tenure variables should approximate the total variable

In this case o is missing so solving for o in terms of p, r, and t we get:

$$o = \frac{t \times 100 - r \times p}{(100 - p)}$$

Equation 2: Solving Equation 1 for the missing variable o

In the alternative case where r is missing the we have good values for p, o and t we solve for r with:

$$r = \frac{t \times 100 - o \times (100 - p)}{p}$$

Equation 3: Solving Equation 1 for the missing variable r

The final step is to check that the estimated value is positive definite, if it is not this method is not used.

- **Model** method: If the **Algebra** method does not work, and the total, and other tenure variable have valid values, CNT developed a set of a regression equations to estimate the missing variable using the other tenure variable, the total variable and the percent of renters. This can be represented for the missing variable o by:

$$o \equiv C_0 + C_t \times t + C_r \times r + C_p \times p + C_{t^2} \times t^2 + C_{r^2} \times r^2 + C_{p^2} \times p^2 + C_{tr} \times t \times r + C_{tp} \times t \times p + C_{rp} \times r \times p$$

Equation 4: Flexible Form Regression Equation

Table 28: Regression Coefficients and R2 for model to fix missing tenure data (SI means statistically insignificant)

Exogenous Variable	R^2	C_0	C_t	C_r	C_p	C_{t2}	C_{r2}	C_{p2}	C_{tr}	C_{tp}	C_{rp}
Owners Fraction of AMI[7]	80.1%	1.64	.05	.56	-1.07	SI	SI	.187	.109	.412	-.362
Renter Fraction of AMI[7]	64.8%	1.06	3.10	-2.94	-.50	.15	.095	.061	-.26	-.47	.654
Owners Average Household Size	94.2%	-0.091	.794	.250	SI	SI	SI	5.31×10^{-5}	SI	0.02966	-0.03082
Renter Average	90.7%	.10	4.25	-3.32	SI	SI	-.092	SI	.103	-.0472	.0474

[7] These fits used "safe log" (ln(1+x)) as linearization formula for Fraction of Rental HU (p).

Household Size											
Owners Commuters/HH	92.4%	-.039	.810	.248	SI	SI	SI	$4.81x10^{-5}$	SI	.0288	-.03130
Renter Commuters/HH[7]	92.1%	-.037	11.04	-10.02	SI	SI	-.049	SI	SI	-2.310	2.329
Owners Rooms/HU	79.7%	1.34	.73	-.12	SI	-.029	-.039	.000282	.098	.0161	-.0182
Renter Rooms/HU[7]	62.1%	-2.7	7.02	-4.85	SI	SI	.091	.054	-.216	-1.119	1.07

- **Full overwrite** method: If either of the above two fixed did not work, the missing value is overwritten using the overall value (t), in the example above for missing Renter Fraction of AMI the value for Fraction of AMI is substituted
- **Use other tenure overwrite** method: If either of the above three fixed did not work, the missing value is overwritten using the value from the other tenure (r), in the example above for missing Renter Fraction of AMI the value for Owner Fraction of AMI is substituted
- **Average of neighbor values** method: finally, if none of the above works, the tenure household weighted average for the variable is calculated for all the tracts that touch the tract with the missing data. In the example above the Renter Fraction of AMI will be estimated form the neighboring tracts using the rental household weighted by average of Renter Fraction of AMI. In other words:

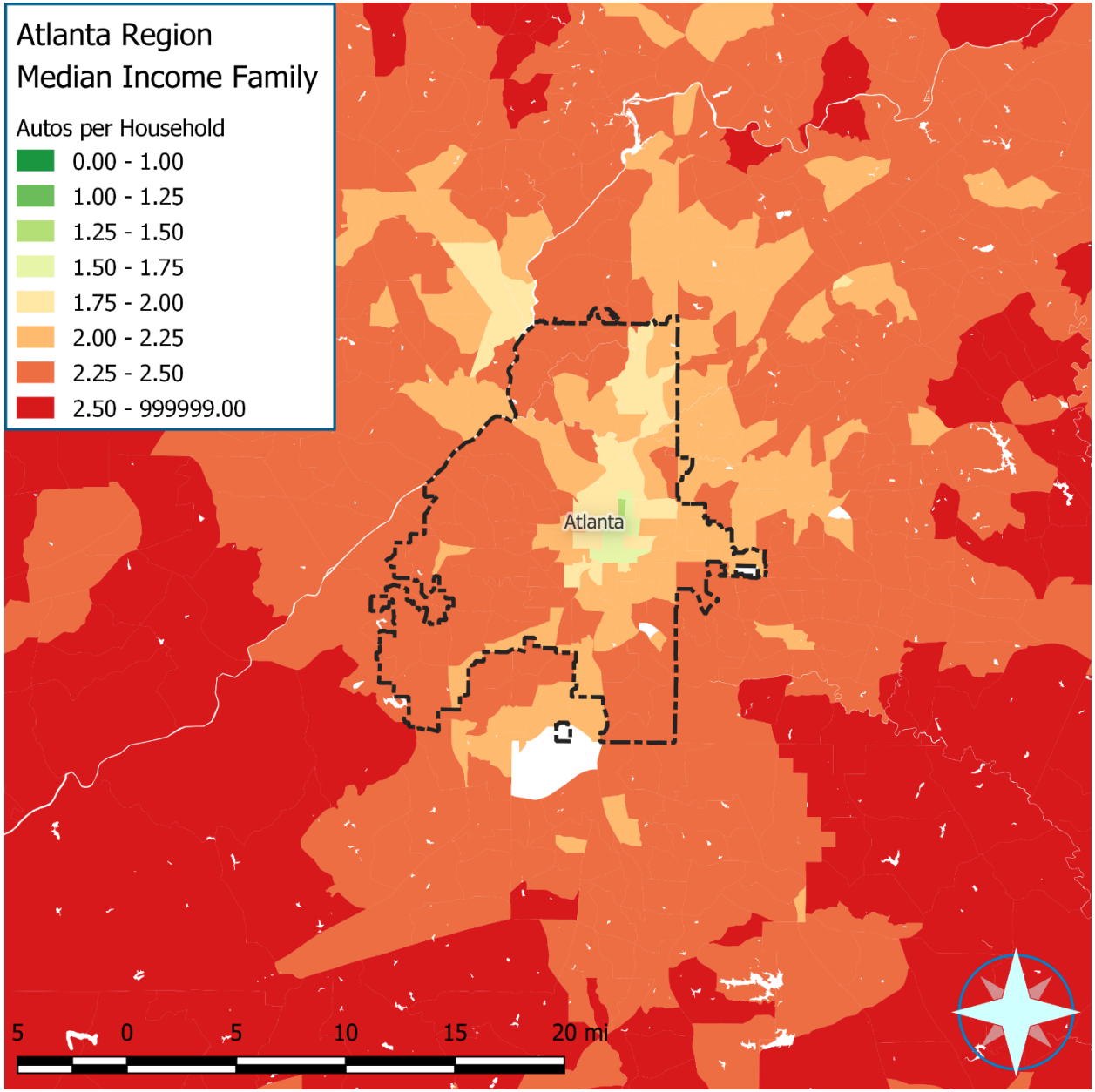
$$o = \frac{\sum_{i=1}^{n} OH_i \times o_i}{\sum_{i=1}^{n} OH_i}$$

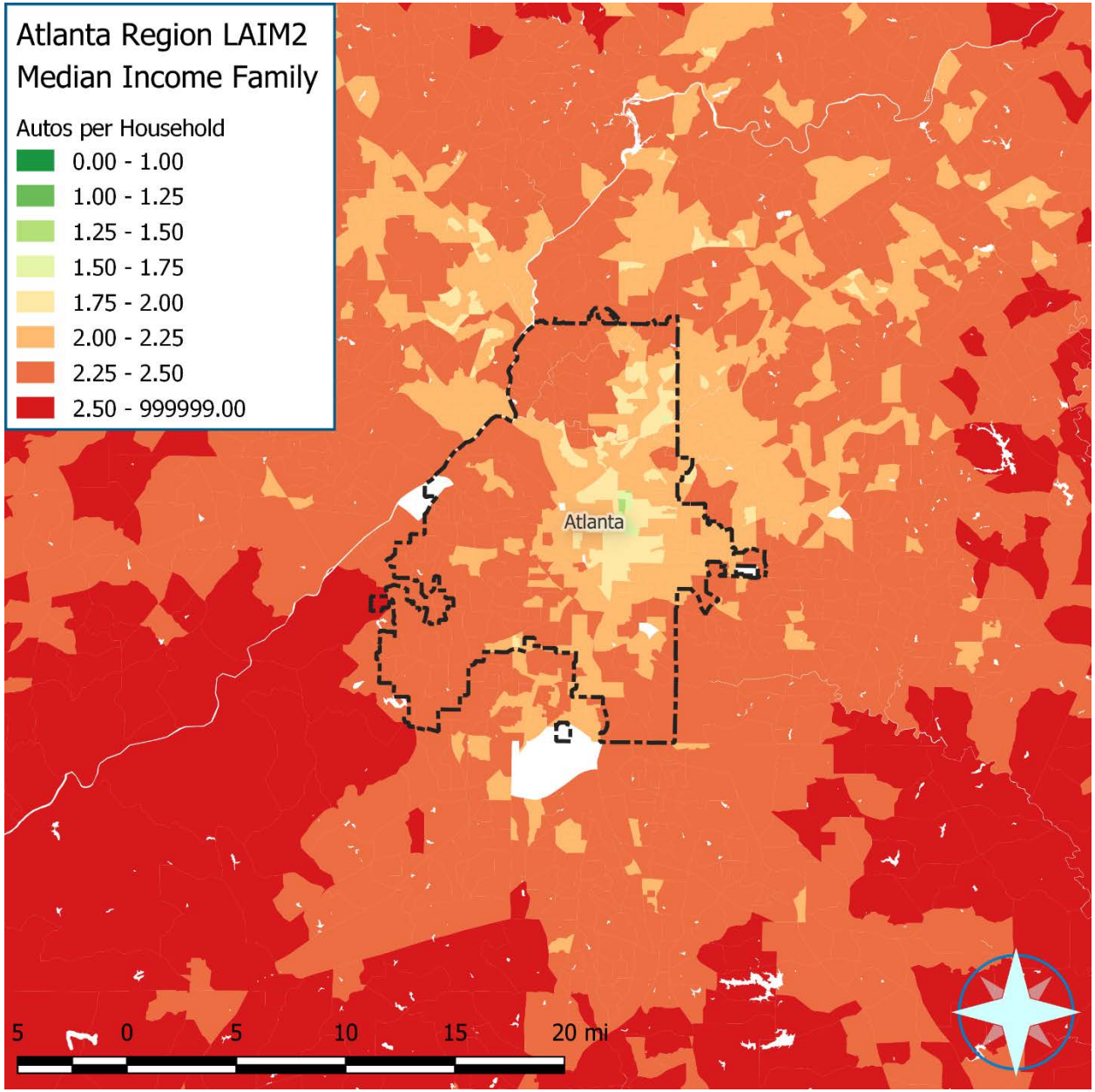
Where n is the number of neighboring tracts, OH_i is the owner-occupied household in the i^{th} tracts.

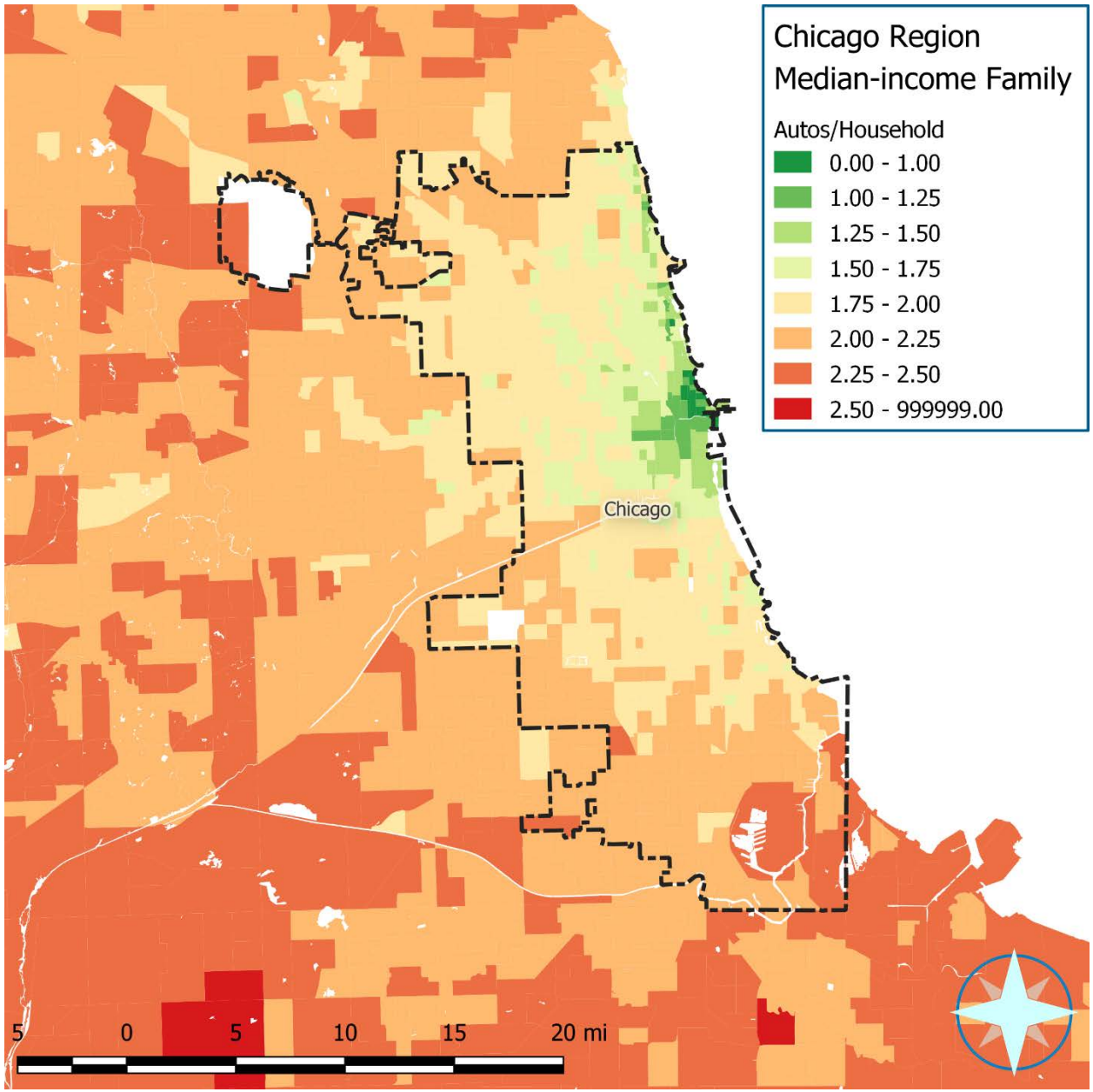
These fixes provide a reasonable estimate for all tracts that are lacking data, except for one – the tract that consists of Isle Royale National Park in the middle of Lake Superior.

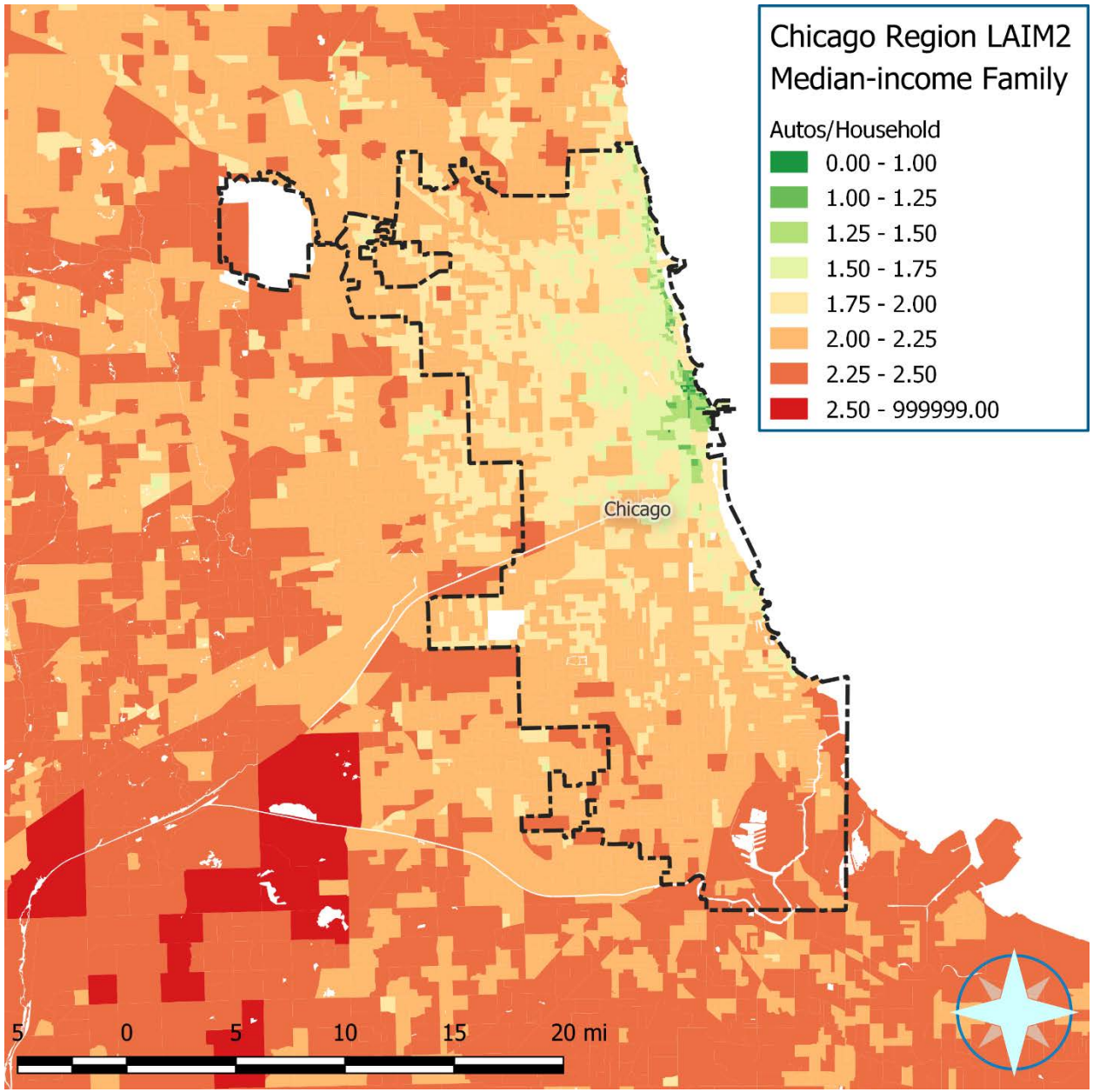
Results

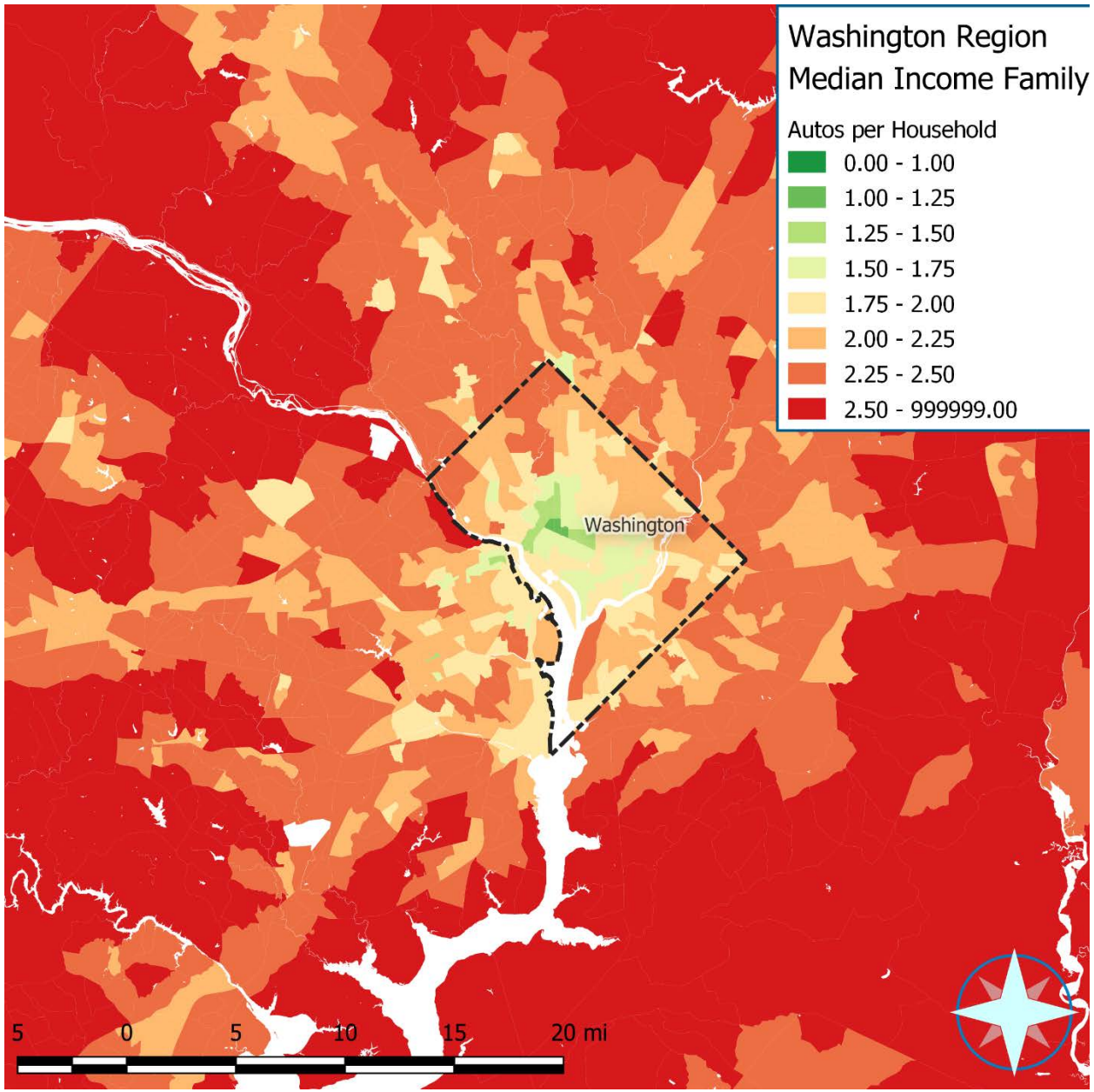
The model can now be run on all tracts that have households in them. And as a comparison to LAM2.0, maps were made for the Atlanta, Chicago, San Francisco and Washington DC Metro show autos per household for home owners using the Median-income family control household, for this new model – tract based – and LAIM2.0 – block group based. These maps all show very similar spatial distribution for the two models. Maps for other variables and control households will look very similar.

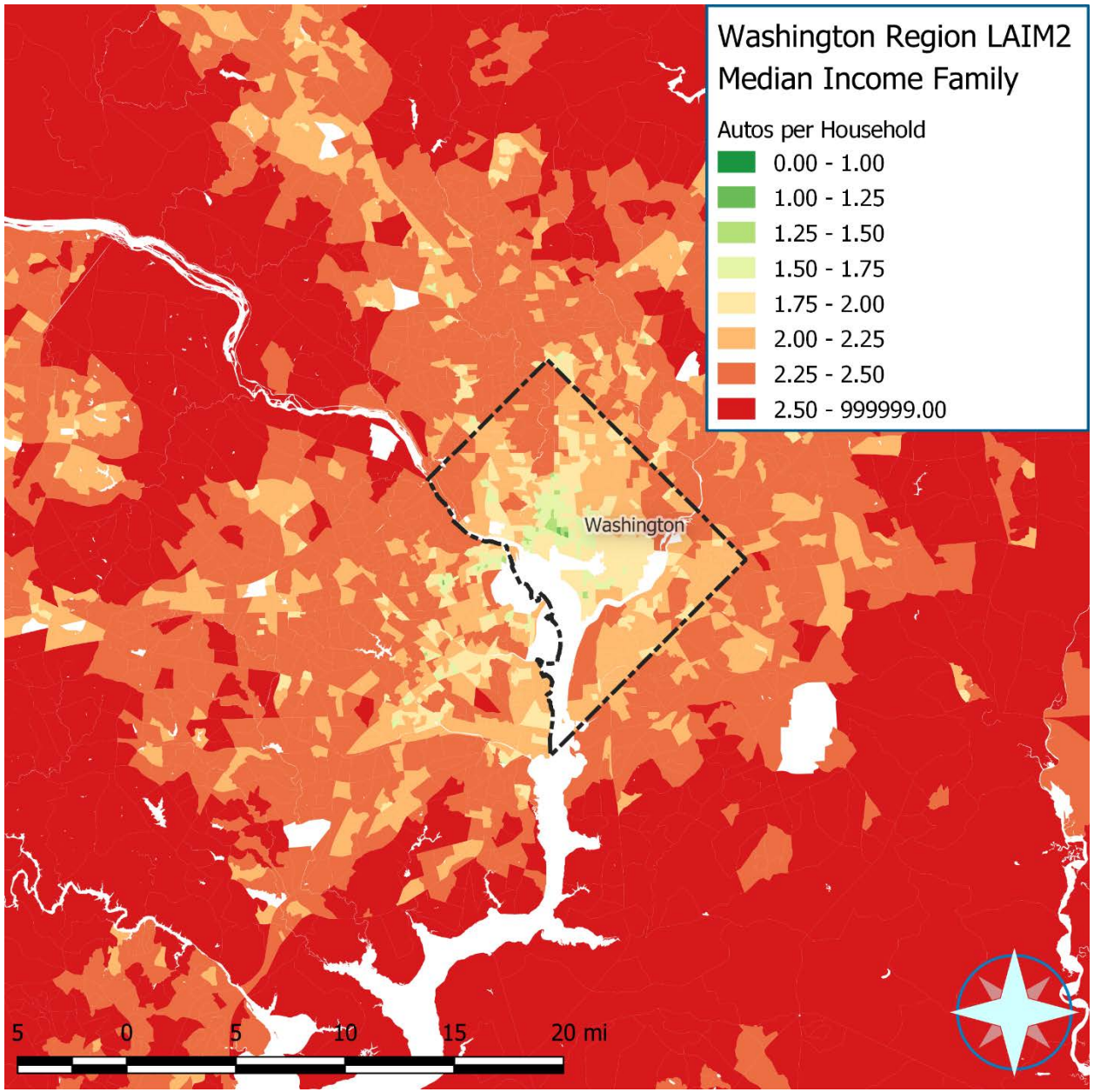

Atlanta Region
Median Income Family
Autos per Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
Atlanta
5 0 5 10 15 20 mi

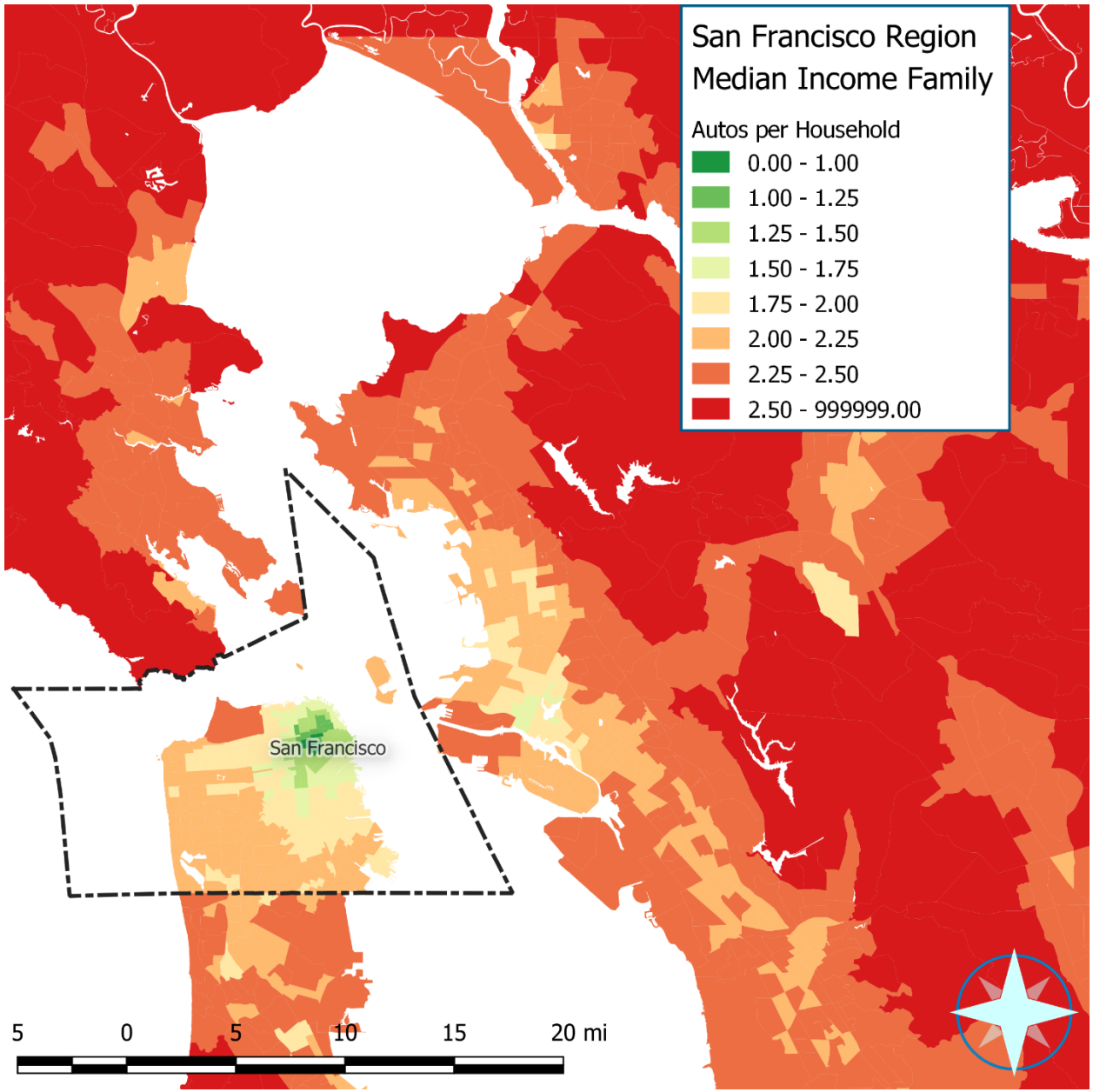

Atlanta Region LAIM2
Median Income Family
Autos per Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
Atlanta
5 0 5 10 15 20 mi

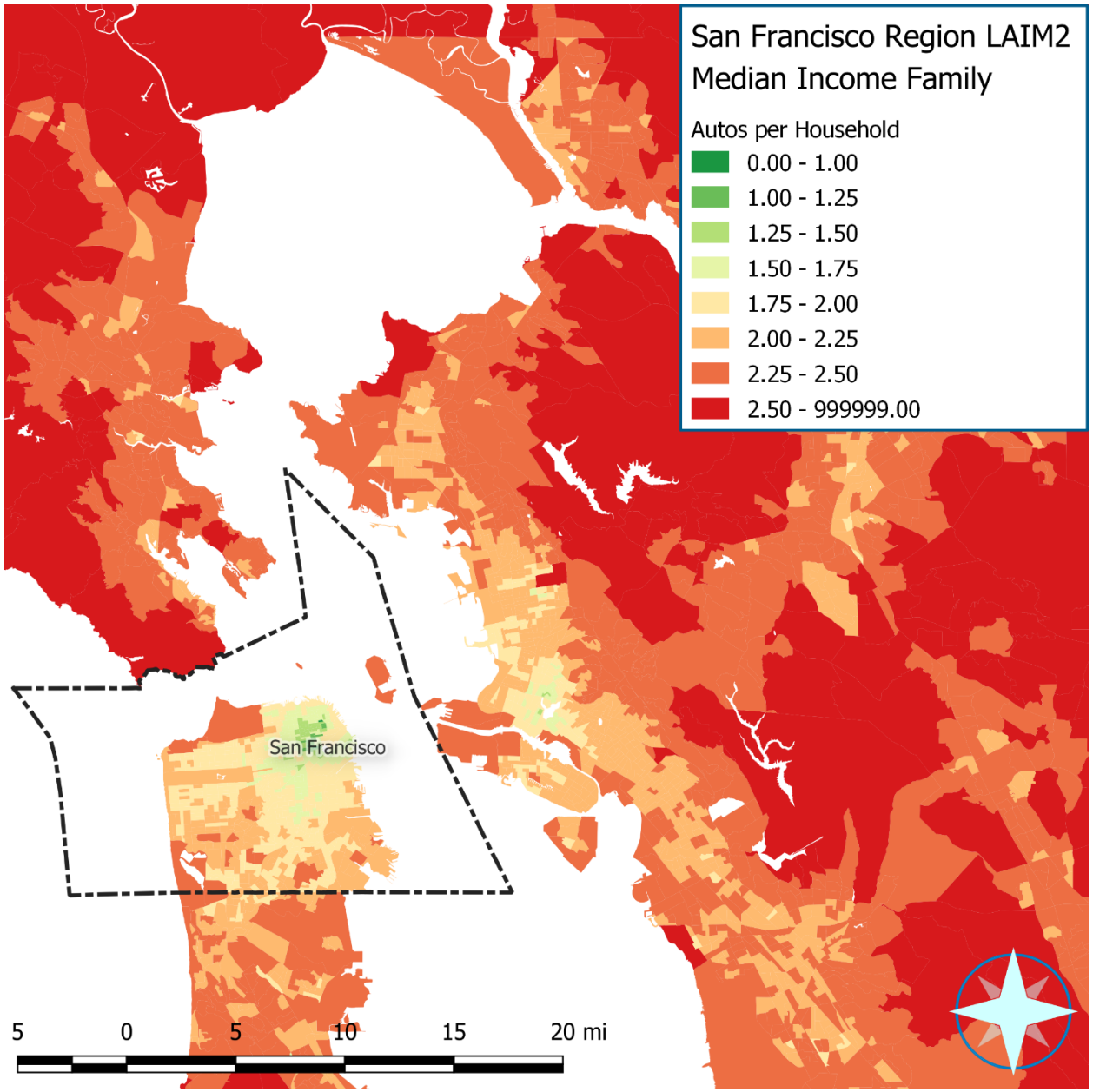

Chicago Region
Median-income Family
Autos/Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
Chicago
5 0 5 10 15 20 mi


Chicago Region LAIM2
Median-income Family
Autos/Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
Chicago
5 0 5 10 15 20 mi


Washington Region
Median Income Family
Autos per Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
Washington
5 0 5 10 15 20 mi


Washington Region LAIM2
Median Income Family
Autos per Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
Washington
5 0 5 10 15 20 mi


San Francisco Region
Median Income Family
Autos per Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
San Francisco
5 0 5 10 15 20 mi


San Francisco Region LAIM2
Median Income Family
Autos per Household
0.00 - 1.00
1.00 - 1.25
1.25 - 1.50
1.50 - 1.75
1.75 - 2.00
2.00 - 2.25
2.25 - 2.50
2.50 - 999999.00
San Francisco
5 0 5 10 15 20 mi

Appendix D: Apply Costs

Auto Ownership and Auto Use Costs[8]

The Consumer Expenditure Survey (CES) from the U.S. Bureau of Labor Statistics is the basis for the auto ownership and auto use cost components of the LAI Version 2.0. Research conducted by Diane Schanzenbach, PhD and Leslie McGranahan PhD, which included a range of new and used autos, examined expenditures based on the 2005-2010 waves of the CES. This research advanced the effort to overcome limitations of other measures that focused primarily on autos less than five years old. Based on the research, expenditures are represented in inflation-adjusted 2010 dollars using the Consumer Price Index for all Urban Consumers (CPI-U). Expenses are segmented by five ranges of household income <$20,000; $20,000-$39,999; $40,000-$59,999; $60,000-$99,999; and, $100,000 and above) and applied to the modeled autos per household and annual VMT for the appropriate income range. This new version of LAI uses an additional inflation factor of 1.105765[9] to adjust to 2016 dollars. Expenditures related to the purchase and operation of cars and trucks are divided into five categories:

- Average annual service flow value[10] from the time the vehicle was purchased to the time the consumer responded to the CES;
- Average annual finance charge paid;
- Ownership Costs: cost of continuing to own a purchased vehicle even if it is not driven;
- Drivability Costs: cost of keeping the vehicle in drivable shape, e.g. maintenance and repairs; and
- Driving Costs: cost of the fuel used to drive the vehicle.

Table 29: Per-Vehicle Costs by Income Group among Households with at Least One Vehicle

Income group number and range	Average Annual Service Flow (1)	Finance Charges (2)	Per vehicle (fixed) ownership costs (3)	Per vehicle (variable) drivability costs (4)	Per vehicle fuel costs (5)	Number of vehicles (6)	Average Ratio drivability to fuel costs (7)
1	$2,396	$73	$657.3	$400.8	$1,182.0	1.4	0.34

[8] This section is taken from "Data and Methodology: Location Affordability Index 2.0 Simultaneous Equation Model" pages 10 and 11 – and modified for 2016 data-set.

[9] http://www.bls.gov/data/inflation_calculator.htm

[10] Service flow is the average annual dollar amount of depreciation the vehicle has lost over the time of ownership.

(<$20,000)							
2 ($20,000-$39,999)	$2,478	$133	$732.0	$421.1	$1,369.5	1.6	0.31
3 ($40,000-$59,999)	$2,586	$182	$755.6	$458.8	$1,494.2	1.9	0.31
4 ($60,000-$99,999)	$2,727	$211	$758.6	$477.6	$1,552.8	2.2	0.31
5 ($100,000 & above)	$3,139	$201	$836.6	$593.1	$1,635.6	2.5	0.36
Overall average	$2,717	$165	$752.5	$474.5	$1,460.9	1.9	0.32

The calculation of auto cost:

$$Cost = I_{2010/16}\{A \times (V_{sf} + V_{fc} + V_{fixed})\} + I_{2014/16}\left\{\left(\frac{VMT}{MPG}\right) \times G \times (1 + R)\right\}$$

Where

$I_{2010/16}$ = Inflation factor from 2010 to 2016[9] (1.105765)

A = Modeled autos per household

V_{sf} = Per vehicle service flow cost from

Table 29(1) – for the appropriate income group

V_{fc} = Per vehicle finance charge from

Table 29 (2) – for the appropriate income group

V_{fixed} = Per vehicle (fixed) ownership cost from

Table 29 (3) – for the appropriate income group

$I_{2014/16}$ = Inflation factor from 2014 to 2016[9] (1.011223)

VMT = the modeled annual household VMT

MPG = the national average fuel efficiency (21.6 mpg for 2014[11])

G = the cost of gas per gallon (average annual regional cost for 2014 - DOE[12])

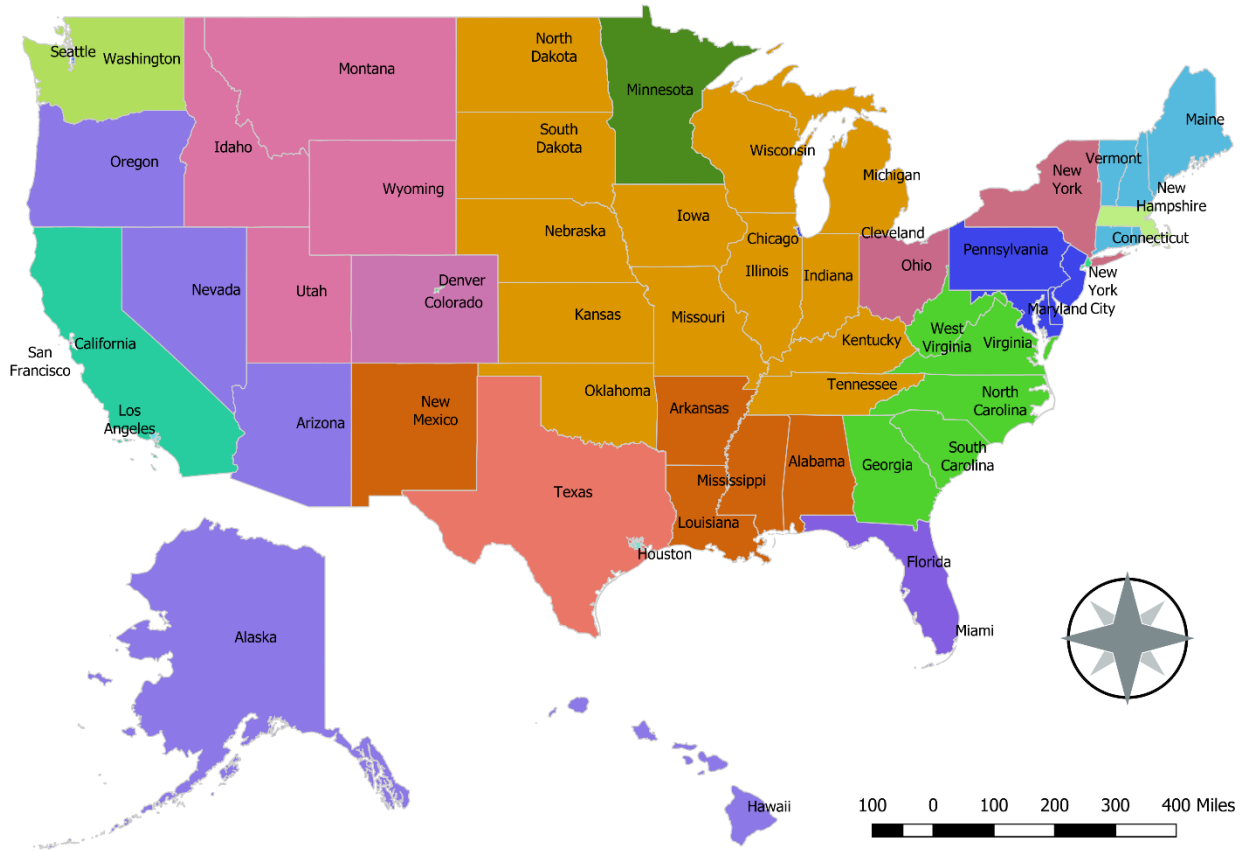
[11] https://cta.ornl.gov/data/tedbfiles/Spreadsheets/Table4_03.xls

[12] https://www.eia.gov/petroleum/gasdiesel/

R = the Average Ratio drivability to fuel cost from

Table 29 (7) – for the appropriate income group

Note that in the end all the costs are in 2016 dollars since that is what the housing cost will be in from the 2016 5-year ACS dataset. The following map shows the US-EPA's defined gas price regions.



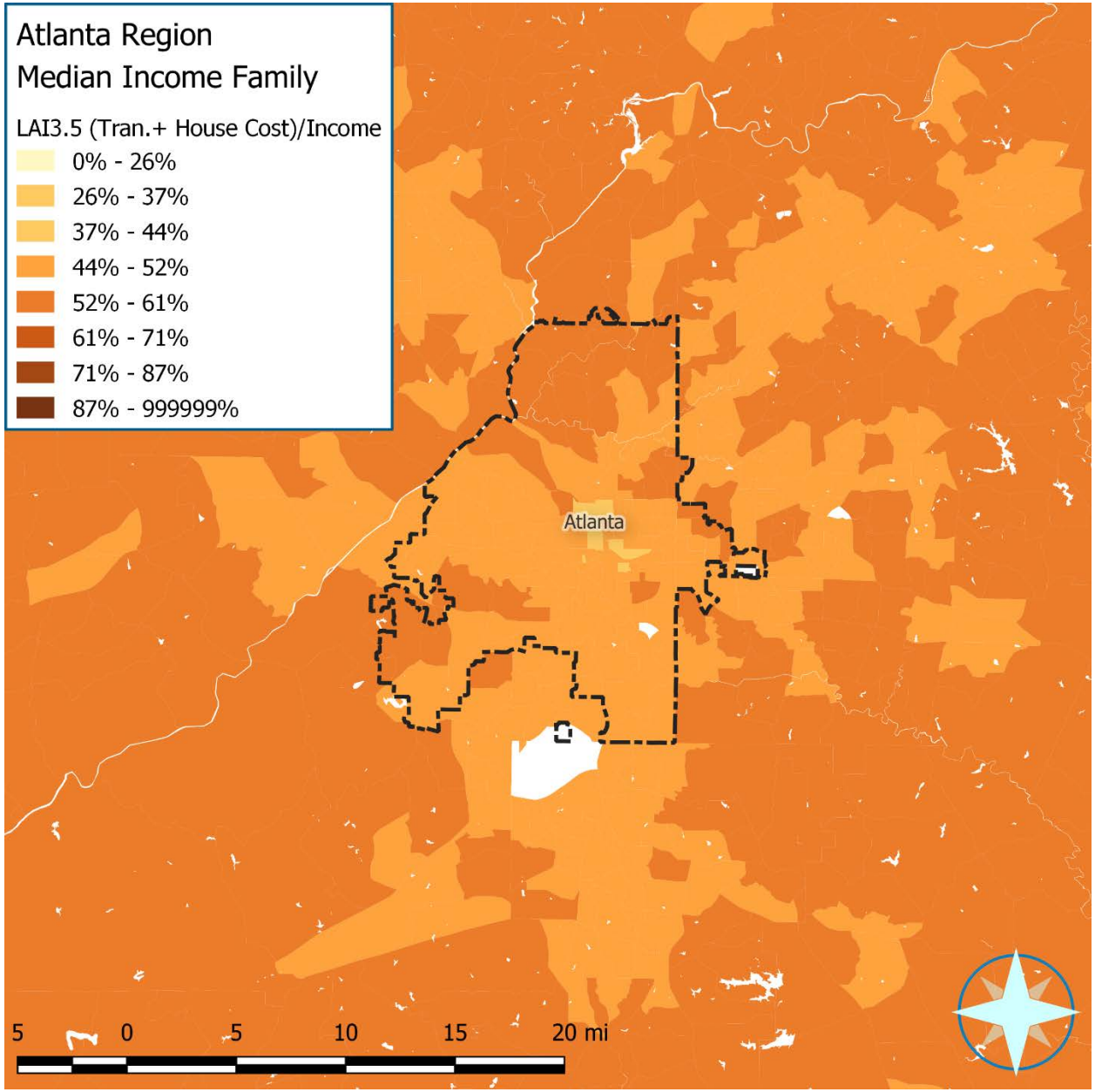
Gas Price Zones

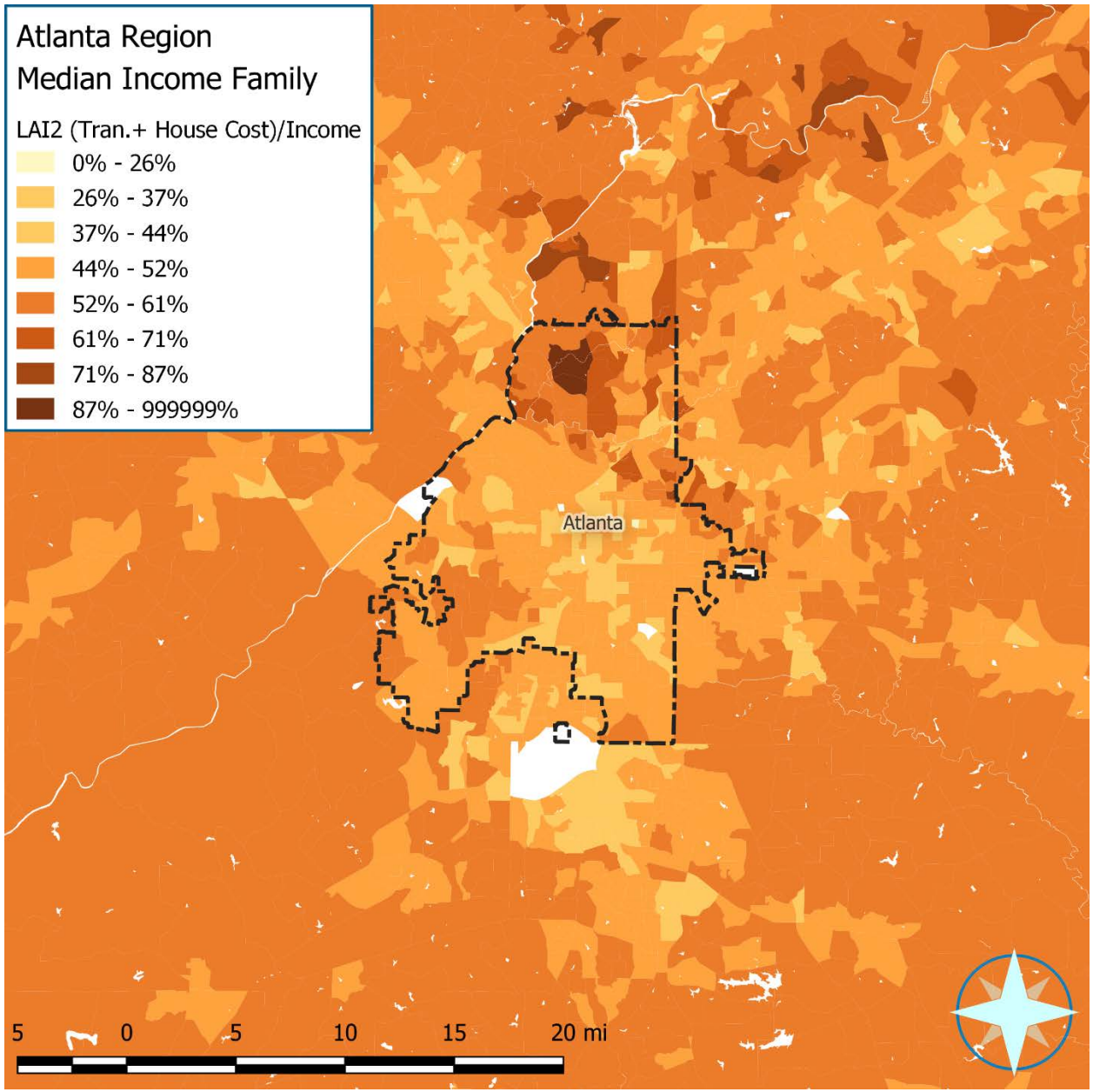
Regular All Formulations Retail Gasoline Prices Annual Average per Gallon

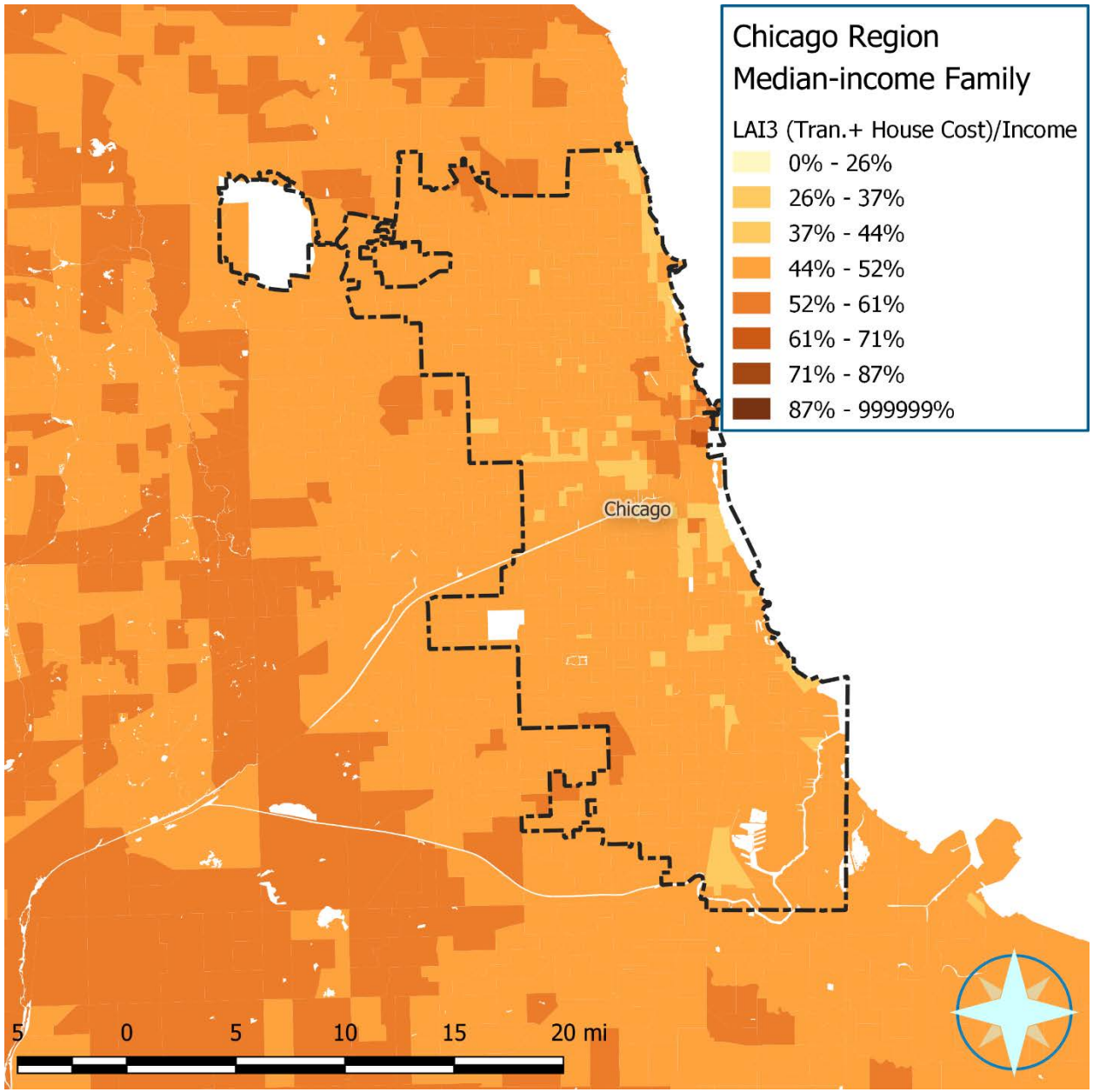
- Boston $3.39
- California $3.745
- Central Atlantic $3.433
- Chicago $3.569
- Cleveland $3.351
- Colorado $3.329
- Denver $3.323
- Florida $3.302
- Gulf Coast $3.136
- Houston $3.128
- Los Angeles $3.784
- Lower Atlantic $3.283
- Massachusetts $3.397
- Miami $3.468
- Midwest $3.303
- Minnesota $3.24
- New England $3.467
- New York City $3.439
- New York $3.598
- Ohio $3.324
- Rocky Mountain $3.335
- San Francisco $3.781
- Seattle $3.601
- Texas $3.149
- Washington $3.557
- West Coast $3.491

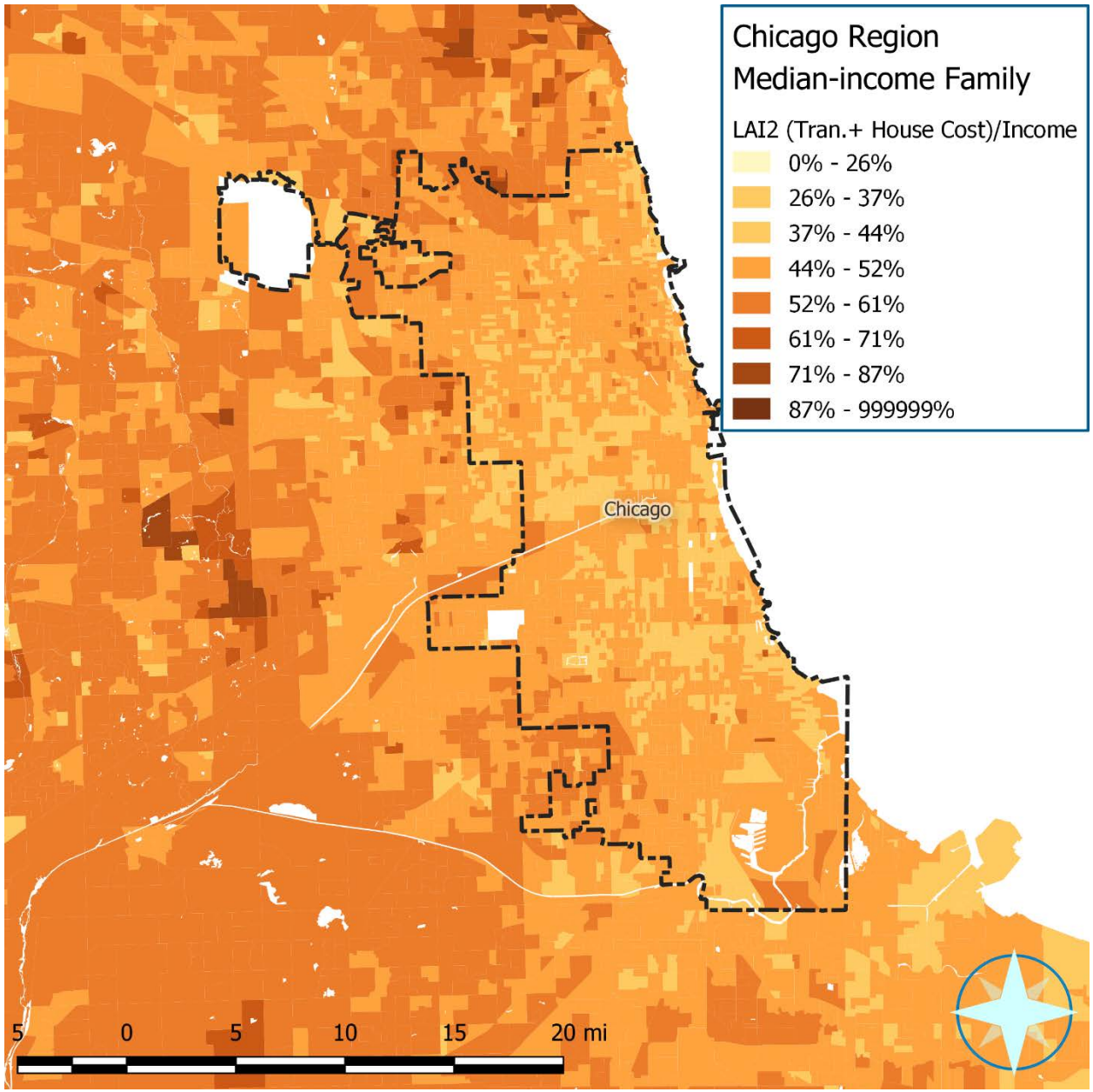
Results

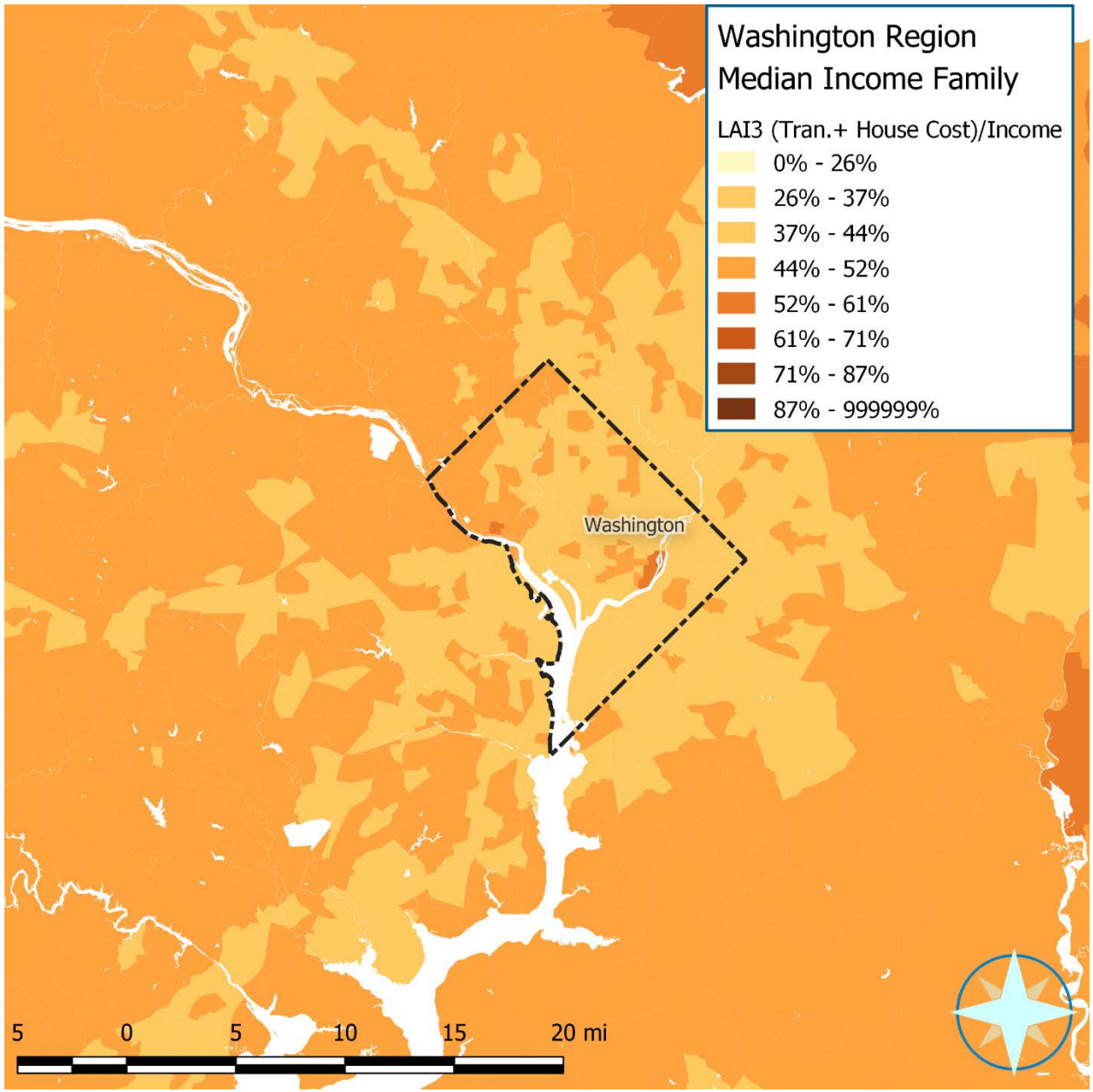
The model and costs have been run on all tracts that have households in them. And as a comparison to LAM2.0 maps were made for the Atlanta, Chicago, San Francisco and Washington DC Metros show autos per household for home owners using the Median-income family control household, for this new model – tract based – and LAIM2.0 – block group based. While in general the spatial trends in these maps look similar, there are differences, these differences are a result of a changing income environment in the US, as well as changes in transportation decisions.

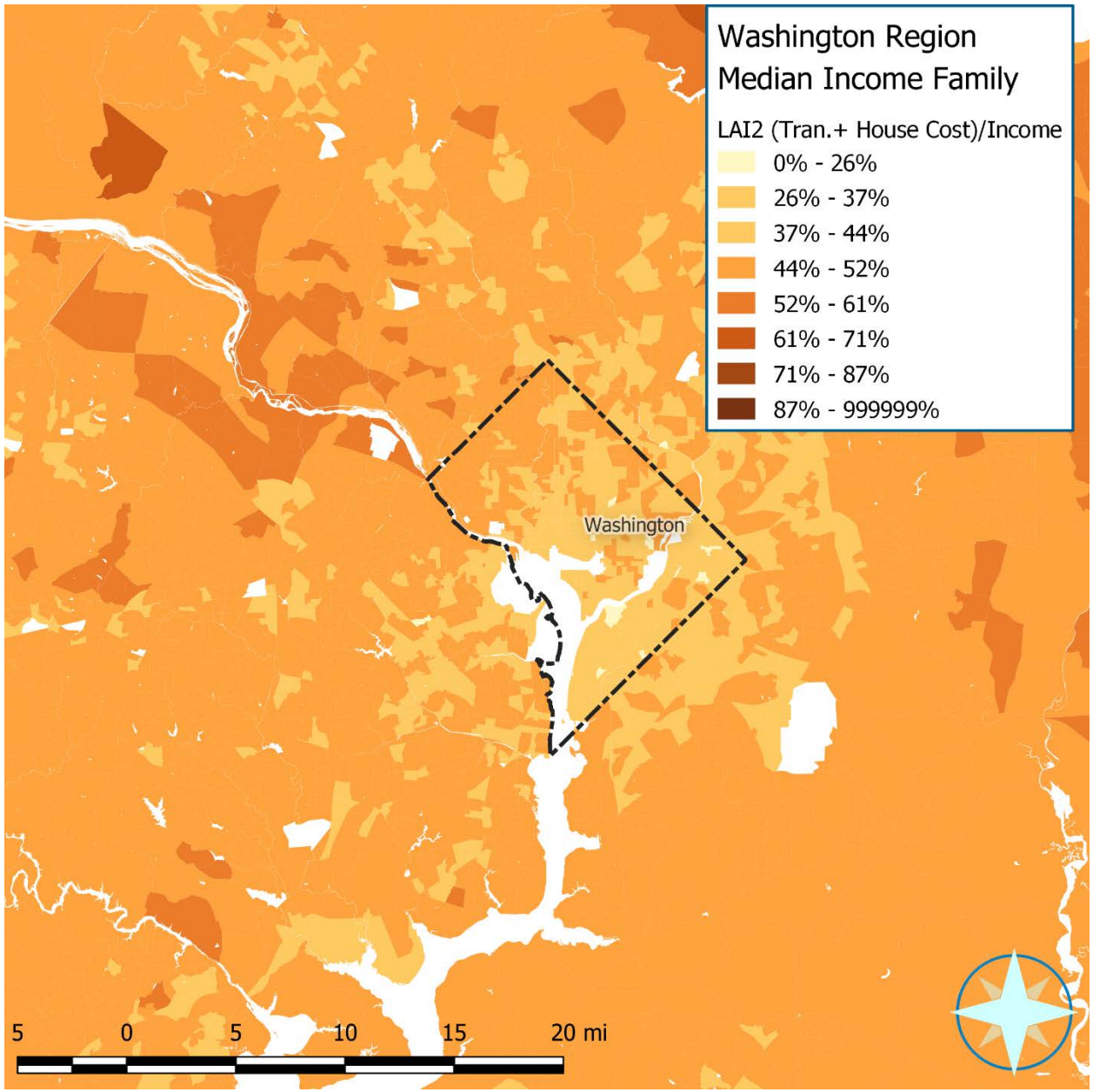

Atlanta Region
Median Income Family
LAI3.5 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
Atlanta
5 0 5 10 15 20 mi

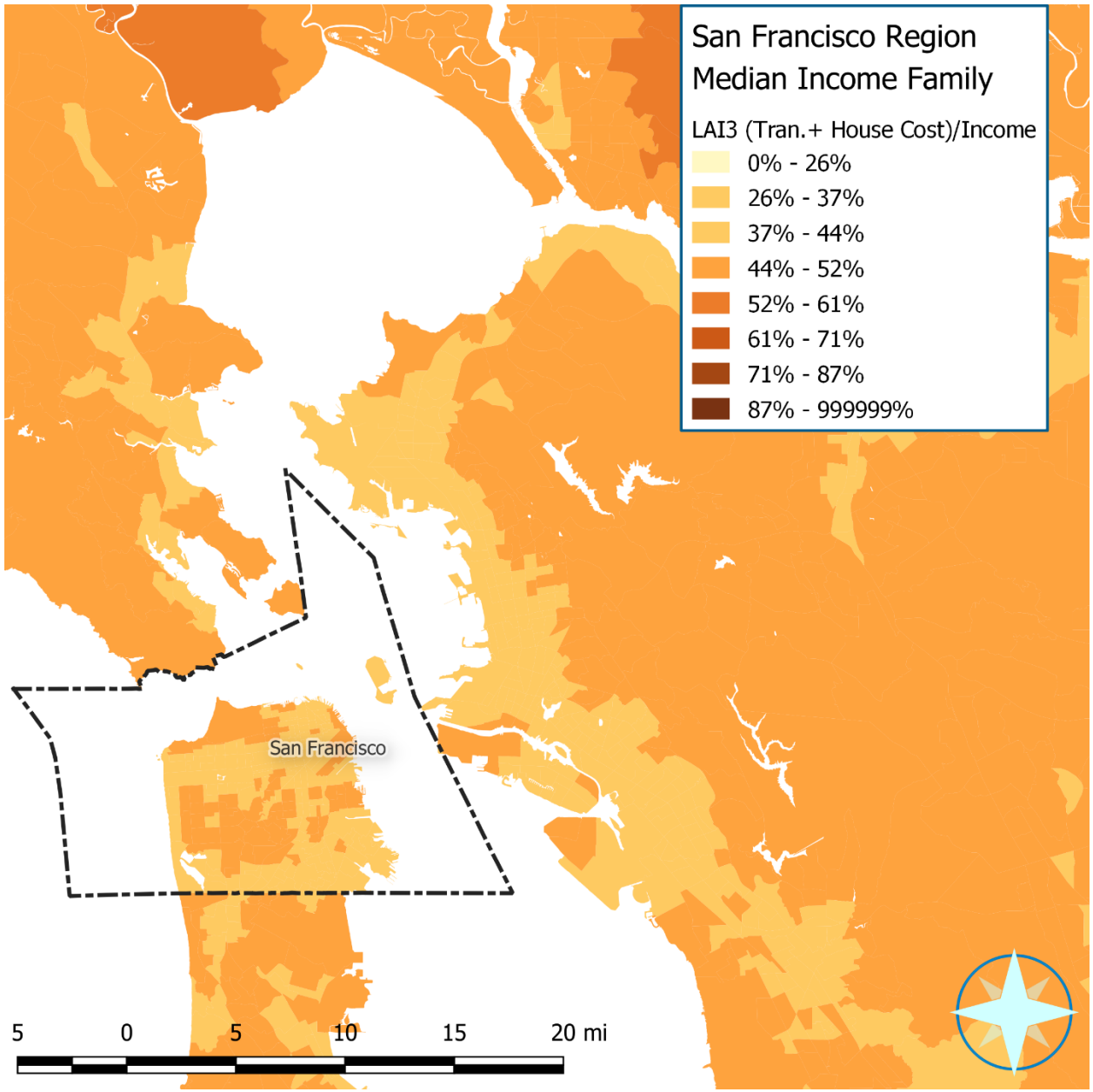

Atlanta Region
Median Income Family
LAI2 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
Atlanta
5 0 5 10 15 20 mi

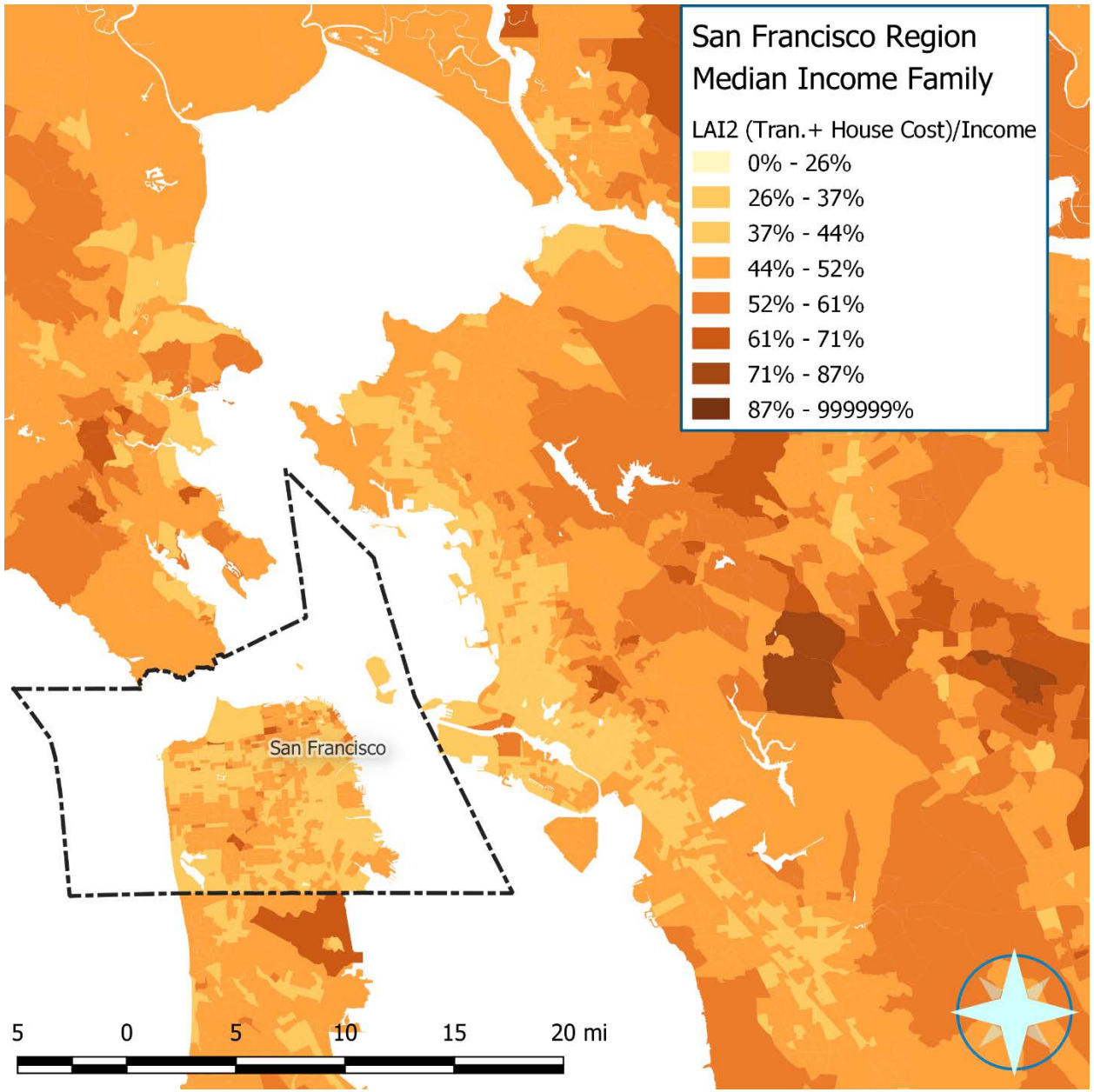

Chicago Region
Median-income Family
LAI3 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
Chicago
5 0 5 10 15 20 mi


Chicago Region
Median-income Family
LAI2 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
Chicago
5 0 5 10 15 20 mi


Washington Region
Median Income Family
LAI3 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
Washington
5 0 5 10 15 20 mi


Washington Region
Median Income Family
LAI2 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
Washington
5 0 5 10 15 20 mi


San Francisco Region
Median Income Family
LAI3 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
San Francisco
5 0 5 10 15 20 mi


San Francisco Region
Median Income Family
LAI2 (Tran.+ House Cost)/Income
0% - 26%
26% - 37%
37% - 44%
44% - 52%
52% - 61%
61% - 71%
71% - 87%
87% - 999999%
San Francisco
5 0 5 10 15 20 mi

Appendix E: Transit fare allocation strategy

Transit cost data were obtained from the 2014 National Transit Database (NTD). In the NTD there is a primary urbanized area (UZA) assigned to every transit agency, as well as other non-primary urbanized areas and non-urbanized areas (for example "Texas Non-UZA"). There are 849 transit agencies in the NTD, most (657) have service in only one urbanized area, and some (280) of those also serve non-urbanized areas. However, there are also some (191) that serve multiple urbanized areas, and of those most (155) also serve non-urbanized area. Only one agency serves only non-urbanized areas.

NTD transit agencies by service area(s)	Number
One UZA	377
One UZA and adjacent non-urbanized areas	280
Multiple UZAs	36
Multiple UZAs and adjacent non-urbanized areas	155
Non-urbanized areas only	1
TOTAL	**849**

To allocate the revenue from the total fares collected and total transit rides to geographical areas without there being some ambiguity, unless there were some means of reallocating those revenue and rides by station count, and/or frequency of service. Given that there is no federal data on station location and service frequency, the allocation must be done at the urbanized area geography. Where agencies provide service to non-urbanized areas, these areas will not be included in the full revenue allocation since it is impossible to divide the revenue/trips across multiple areas served. Also, this method assumes that the non-primary urbanized areas will average out. Since most transit is provided and used in urban settings there should be a minimal amount of error associated with this allocation method.

The idea is to come up with two normalization factors (α and β) to estimate the average household's transit cost and trips on transit given the fraction of commuters in a given Census tract using transit for their journey to work[13]. To calculate β, defined as the factor to leverage per household transit trips T_i in a given Census tract i from percent of commuters in that tract using transit for their journey to work P_i times the number of commuters per household E_i,

[13] Equation 1 assumes that fraction of transit used in the journey to work is a good surrogate for all transit use.

$$T_i = \beta P_i E_i \ \therefore \beta = \frac{T_i}{P_i E_i}$$

Equation 5

we use the following defining equation:

$$Q = \beta \sum_{i=1}^{N} f_i P_i C_i$$

Equation 6

Where Q is the total number of transit trips taken in the urbanized area (UZA), N is the number of census tracts intersecting the UZA, P_i is the fraction of commuters using transit for their journey to work in the "i^{th}" census tract, C_i is the total number of commuters in the "i^{th}" census tract and f_i is the fraction of the "i^{th}" census tract's area that is within the UZA. Since f_i P_i C_i gives the total journey to work trips taken on transit for any given tract, β is a factor that connects journey to work transit trips to all transit trips in a given UZA, assuming that 1) P_i and C_i accurately reflect commuting activity in tracts that is then realized by transit agencies as total trips and farebox receipts and 2) the number of a tract's commute trips C_i accruing to a UZA is proportional to the fraction of that tract's area intersecting with that UZA (f_i).[14]

Note that if every trip on transit was for the journey to work, then β would equal one, however, we know that this is not the case (many trips are for non-work reasons and they must get home from work as well), so we need to solve for β:

$$\beta = \frac{Q}{\sum_{i=1}^{N} f_i P_i C_i}$$

Equation 7

Then if we write Equation 6 in terms of our desired outcome variable – the per household transit trips in the "i^{th}" census tract (T_i) - and using the number of households in the "i^{th}" census tract (H_i) we get the following equations:

[14] The first assumption is worth articulating because the NTD and ACS are two totally different data sources and there are many reasons why either or both could not accurately reflect the "true" underlying phenomena of commuting and transit usage; for the ACS for example, the journey to work variable has been shown to be questionable. The second assumption could be tested against actual data for the small minority of tracts that intersect more than one UZA.

$$Q = \sum_{i=1}^{N} f_i \, T_i H_i$$

Equation 8

Therefore:

$$\sum_{i=1}^{N} f_i \, T_i H_i \; = \; \beta \sum_{i=1}^{N} f_i P_i \, C_i$$

Equation 9

Assuming that within the i^{th} census tract the elements in the sum are equal and solving for T_i, the following is derived:

$$T_i = \frac{\beta P_i C_i}{H_i} = \; \beta P_i E_i$$

Equation 10

Where E_i is the average number of commuters per household in the "i^{th}" census tract. Once β is obtained for each UZA, the calculation the average number of trips per household for the control household in every census tract (T_i) in the UZA is simply Equation 10 therefore deriving Equation 5– using the modeled P_i and the control household's commuters per household E_i. For tracts that are split between (i.e. overlap with) multiple UZAs, a weighted average of β is used. To summarize this process, β is derived from the ratio of all journey to work trips in a UZA divided by the total number of trips provided by all transit agencies in the UZA. The sum of all trips taken in a tract is β times the number of journey to work trips for each household (P_i x E_i) – as shown in Equation 10. So, the variations in transit trips is really driven by the factors that go into estimating P_i, that is derived using the SEM model or LAIM – considering urban form, job access, and all the other factors folded into the model.

Likewise, for farebox revenue, it can be shown that:

$$\alpha = \frac{R}{\sum_{i=1}^{n} f_i P_i C_i}$$

Equation 11

and

$$F_i = \frac{\alpha P_i C_i}{H_i} = \alpha P_i E_i$$

Equation 12

Where R is the total farebox revenue in the UZA. Once α is obtained for each UZA, the calculation the average cost of transit per household for the control household in every census tract (F_i) in the UZA is simply Equation 12 – using the modeled P_i and the control household's commuters per household E_i. Note that for all census tracts within the UZA the value for α will be the same – however on the edge of UZAs, where a fraction of a tract is in both a weighted average of β is used. To summarize this process, α is derived from the ratio of all journey to work trips in a UZA divided by the total transit revenue by all transit agencies in the UZA. The transit revenue in a tract is α times the number of journey to work trips for each household ($P_i e$) – as shown Equation 12. So, the variations in transit cost is really driven by the factors that go into estimating P_i, that is derived using the SEM model or LAIM3.0 – considering urban form, job access, and all the other factors folded into the model.

The following table gives how (not using the AllTransit database[15]) the variables above will be calculated for any given urbanized area.

Table 30: Variable sources

Variable	Description	Method
Q	Total number of transit trips taken in the UZA	Sum all the unlinked passenger trips for all agencies that assigned this UZA as their primary urbanized area – from NTD
R	Total farebox revenue in the UZA	Sum all the farebox revenue for all agencies that assigned this UZA as their primary urbanized area – from NTD
N	Number of tracts in UZA	Assign tracts that **intersect** the geography of the UZA – urbanized areas do not conform to tracts since they are constructed from block groups.
f_i	Fraction of the "i^{th}" census tract's area that is within the UZA	Use GIS to calculate this fraction

[15] See appendix A for a review of LAIM2's method

Variable	Description	Method
P_i	Percent of commuters using transit for their journey to work in the "i^{th}" census tract	This is the measured percent from the ACS. Because this is sometimes suppressed in the ACS for privacy reasons, the total is scaled up by the ratio of all households in the UZA divided by the households in tracts where this percentage is reported.
C_i	number of commuters in the "i^{th}" census tract	Use the total number of employed workers in the tract minus the ones that work at home households in the census tract from ACS.
E_i	number of commuters per household in the "i^{th}" census tract	Use the total number of employed workers in the tract minus the ones that work at home divided by the number of households in the census tract from ACS.

There is however, one small wrinkle in all of this, and that is that in the NTD, not all transit agencies give their revenue, so sometimes it is impossible to calculate alpha. When this occurs, it is impossible to calculate α and β for those UZAs, so instead we use the closest UZA as a surrogate as listed below. For all tracts (total of 72,241 tracts that have people living in them) we assign α and β using one of the following:

4. Assign to α and β of the UZA where the tract only intersects one UZA where α and β can both be calculated i.e. UZAs with good NTD data (49,039 tracts);
5. Take weighted average of α and β from the multiple UZAs with good NTD data a tract intersects (247 tracts intersect 2 UZAs and 1 intersect 3 UZAs, zero tracts intersect more than 3 UZAs);
6. Use α and β from closest UZAs with good NTD data if the tract does not intersect any UZAs with good NTD data (22,954 tracts).

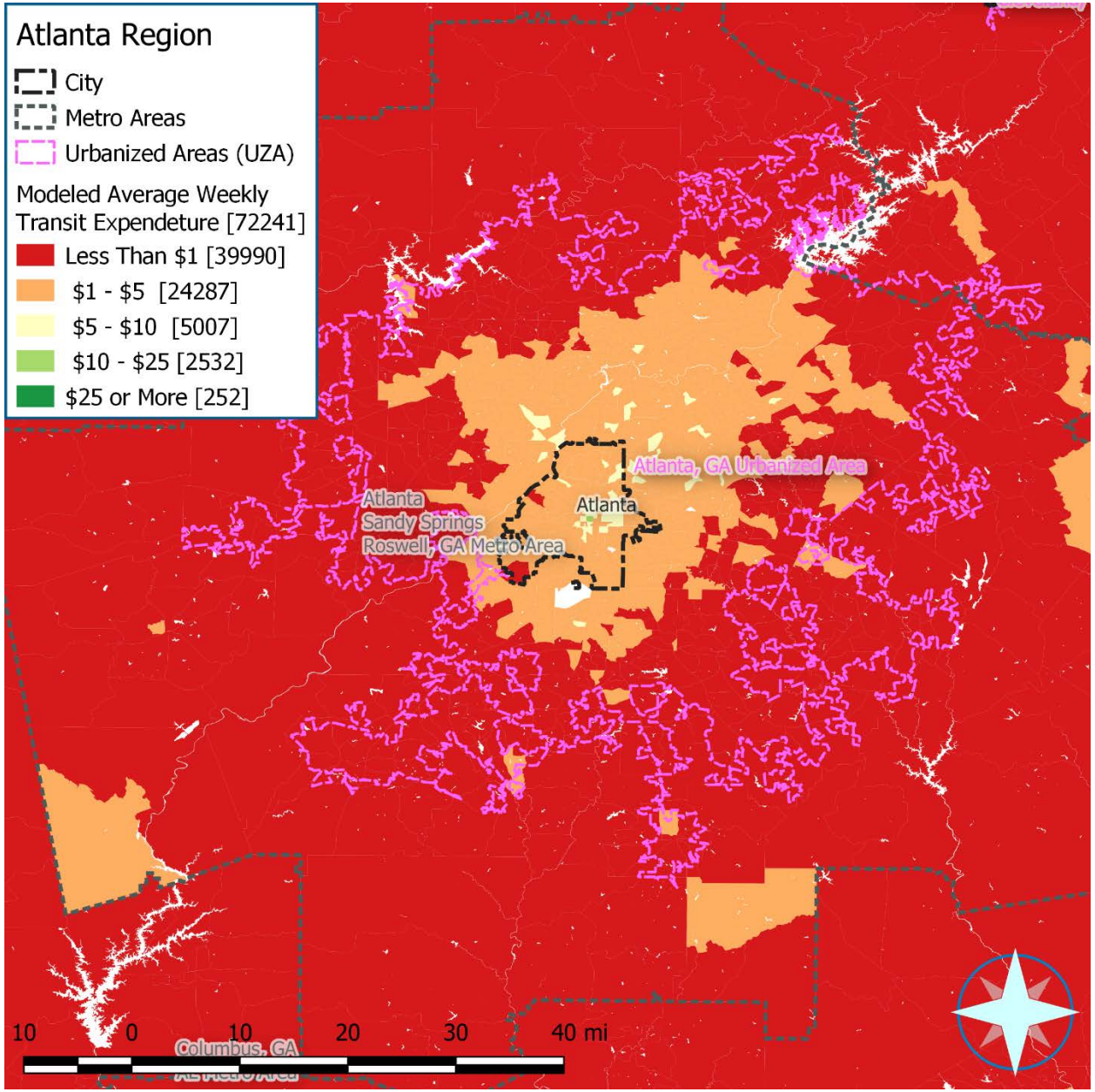
Note that the tracts in option 3. that do not intersect any UZAs are mostly in rural areas, where no one takes transit to work, therefore the accuracy is not that important; as stated above, mostly the variation in transit costs comes from the variability in the use of transit driven largely by access to transit, urban form, access to jobs etc., and these are all captured by the SEM model which shows that transit is not used much in rural locations. The following table summarizes what method of allocation is used by tracts and showing the number of households in those tracts.

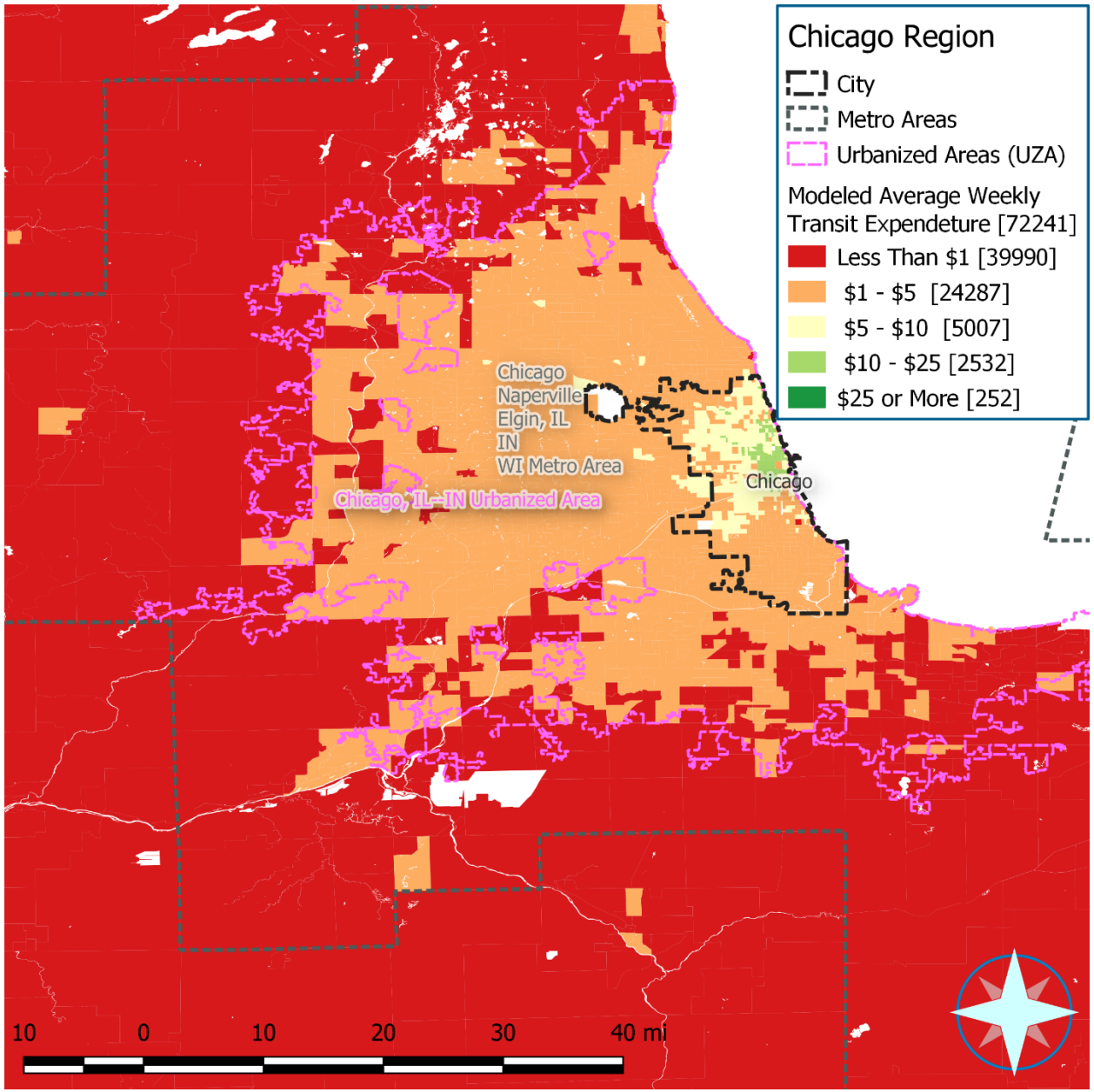
Table 23: Summary of how α and β is assigned to census tracts

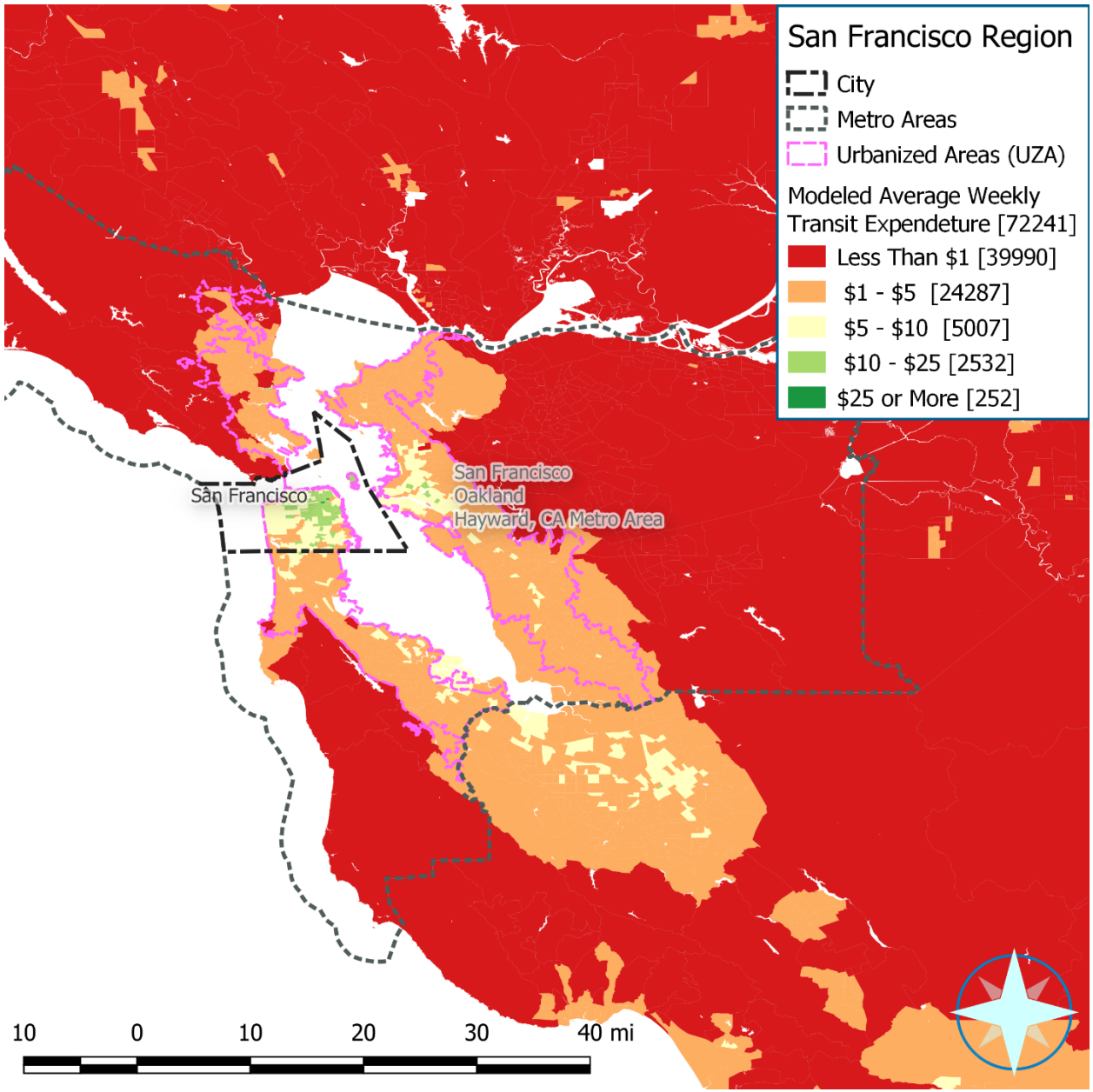
Method #	Tracts	Households	Trips/Week (Modeled)	Total Annual Auto Costs (Modeled)	Total Annual Transit Costs (Modeled)
1	49,039	81,324,594	246,783,719	931,069,410,397	11,608,404,967
2	248	488,983	445,950	7,127,298,215	20,050,129
3	22,954	35,902,660	10,948,265	459,179,096,406	577,559,175
Total	72,241	117,716,237	258,177,934	$1,397,375,805,018	12,206,014,272

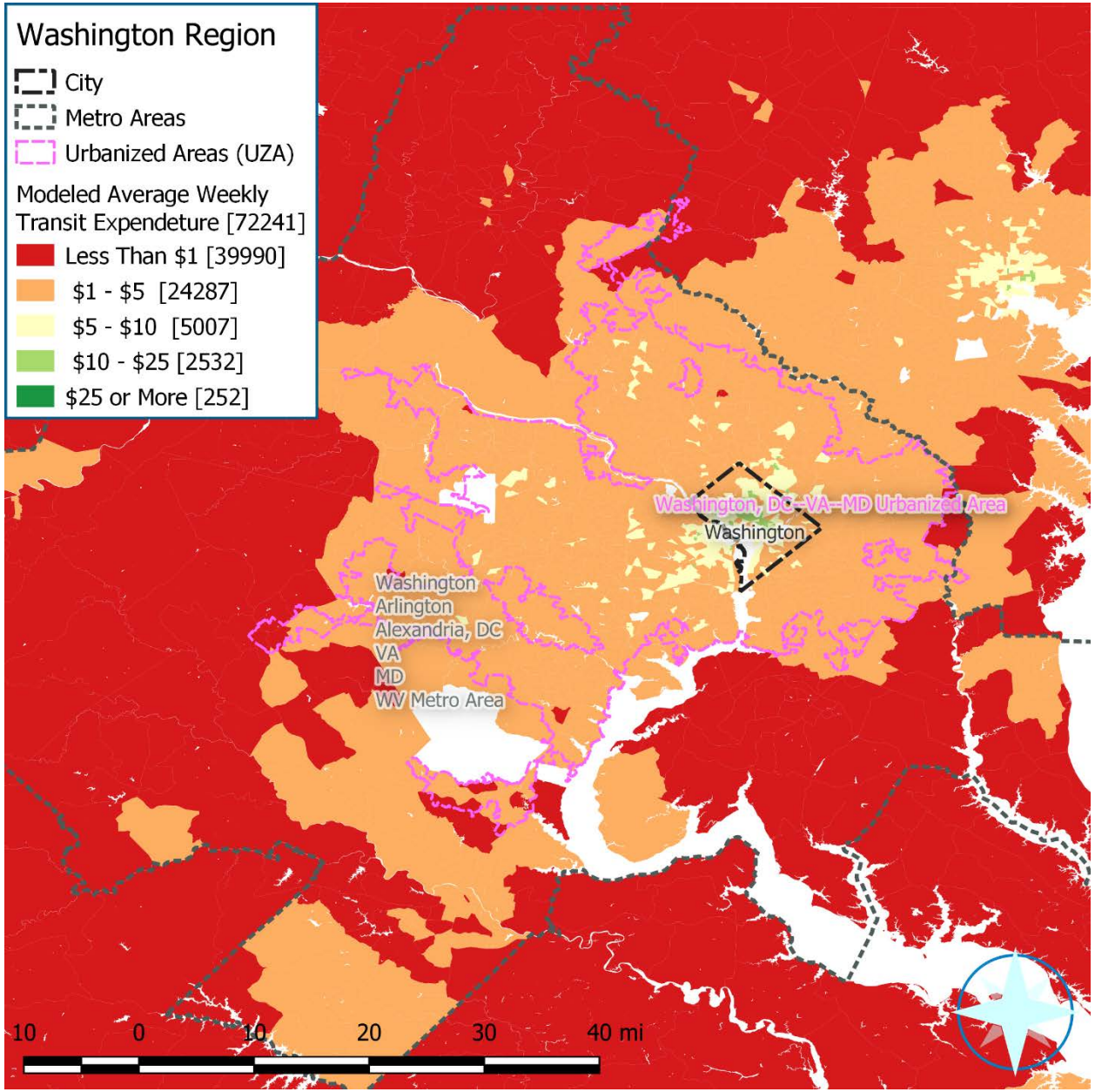
However, there are tracts that are in smaller places with transit, and some transit use, but with nothing else to go on, this method gives a reasonable cost estimates since often it is the neighboring urban area that provides the transit and is often the users destination.

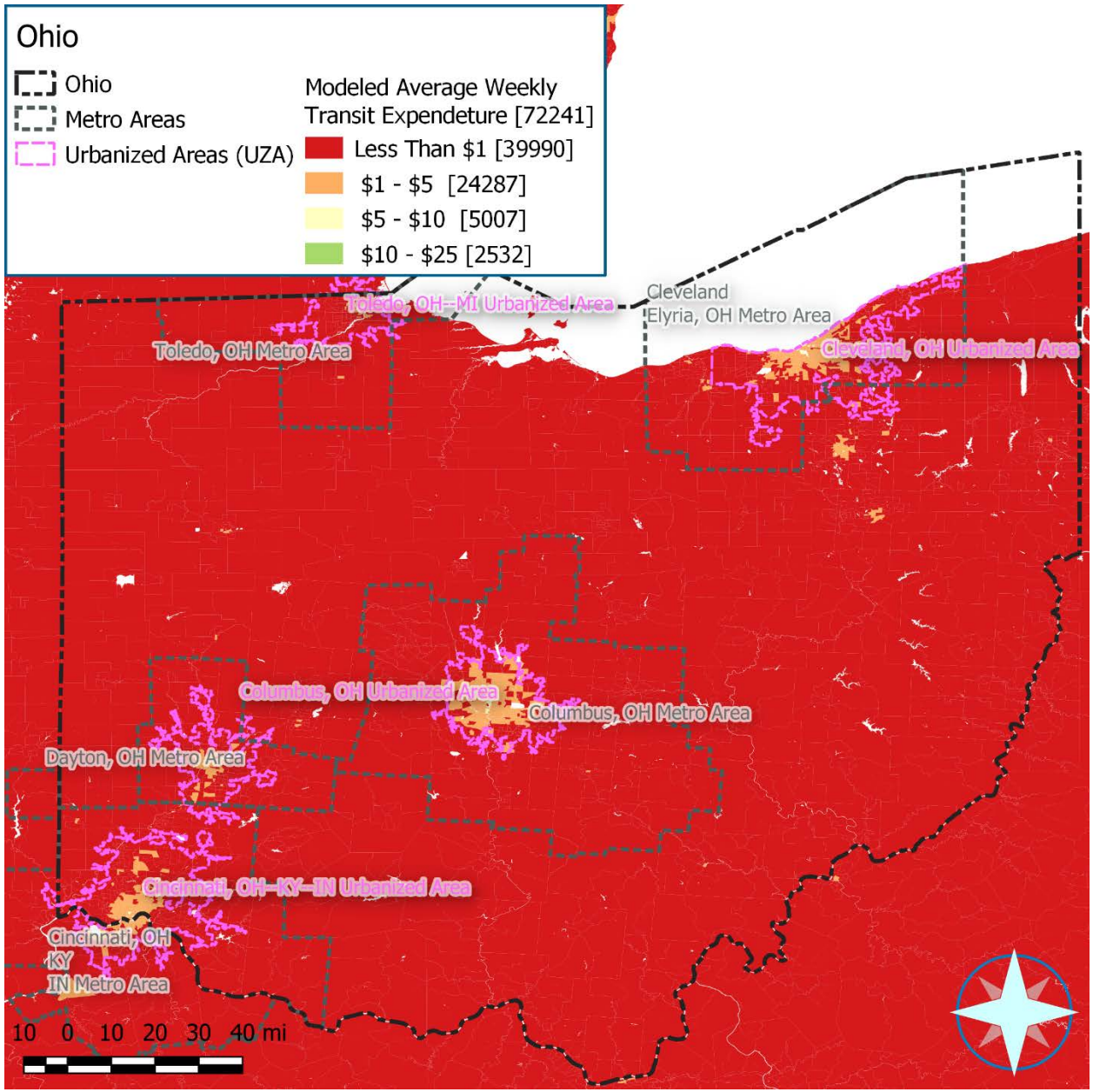
Here are a few maps of the results, with the various census geographies overlaid, for the Atlanta, Chicago, San Francisco and Washington DC Metro Areas as well as for the state of Ohio.


Atlanta Region
City
Metro Areas
Urbanized Areas (UZA)
Modeled Average Weekly
Transit Expendeture [72241]
Less Than $1 [39990]
$1 - $5 [24287]
$5 - $10 [5007]
$10 - $25 [2532]
$25 or More [252]
Atlanta, GA Urbanized Area
Atlanta
Atlanta
Sandy Springs
Roswell, GA Metro Area
10
0
10
20
30
40 mi
Columbus, GA


Chicago Region
City
Metro Areas
Urbanized Areas (UZA)
Modeled Average Weekly Transit Expendeture [72241]
Less Than $1 [39990]
$1 - $5 [24287]
$5 - $10 [5007]
$10 - $25 [2532]
$25 or More [252]
Chicago
Naperville
Elgin, IL
IN
WI Metro Area
Chicago, IL--IN Urbanized Area
Chicago
10 0 10 20 30 40 mi


San Francisco Region
City
Metro Areas
Urbanized Areas (UZA)
Modeled Average Weekly Transit Expendeture [72241]
Less Than $1 [39990]
$1 - $5 [24287]
$5 - $10 [5007]
$10 - $25 [2532]
$25 or More [252]
San Francisco
San Francisco
Oakland
Hayward, CA Metro Area
10 0 10 20 30 40 mi


Washington Region
City
Metro Areas
Urbanized Areas (UZA)
Modeled Average Weekly
Transit Expendeture [72241]
Less Than $1 [39990]
$1 - $5 [24287]
$5 - $10 [5007]
$10 - $25 [2532]
$25 or More [252]
Washington, DC--VA--MD Urbanized Area
Washington
Washington
Arlington
Alexandria, DC
VA
MD
WV Metro Area
10
0
10
20
30
40 mi


Ohio
Ohio
Metro Areas
Urbanized Areas (UZA)
Modeled Average Weekly
Transit Expendeture [72241]
Less Than $1 [39990]
$1 - $5 [24287]
$5 - $10 [5007]
$10 - $25 [2532]
Toledo, OH--MI Urbanized Area
Toledo, OH Metro Area
Cleveland
Elyria, OH Metro Area
Cleveland, OH Urbanized Area
Columbus, OH Urbanized Area
Columbus, OH Metro Area
Dayton, OH Metro Area
Cincinnati, OH--KY--IN Urbanized Area
Cincinnati, OH
KY
IN Metro Area
10 0 10 20 30 40 mi

Appendix E-1: LAIM2.0's method and how this is different:

LAIM2.0 Method

From "*Data and Methodology: Location Affordability Index 2.0 Simultaneous Equations Model*" dated June 15, 2014, pages 11 and 12:

"Transit Use Costs

The 2009 National Transit Database (NTD) served as the source for transit cost data. Specifically, directly operated and purchased transportation revenue are used.[16] The transit revenue, as reported by each of the transit agencies in the NTD, is assigned to agencies and related geographies where GTFS data were collected. This transit revenue is then allocated to the metropolitan areas served, based on the percentage of each transit agency's bus and rail stations within the primary versus surrounding metropolitan areas. For example, if a transit agency has a total of 500 bus stops and 425 of those stops are in the primary metropolitan area and 75 stops extend into a neighboring metropolitan area, the primary metropolitan area receives 85 percent of the transit revenue and the neighboring metropolitan area receives 15 percent.

To estimate average household transit costs, each metropolitan area's estimated transit revenue is then allocated to block groups based on the modeled value of the percentage of transit commuters and the total households within each block group. This is done by calculating the number of transit commuters for each block group, summing across block groups to estimate the total number of transit commuters in the metropolitan area, and then allocating the metro-wide transit revenue to block groups according to the proportion of the region's commuters living in each. The average household transit cost for each block group is then derived by dividing that block group's allocation of transit revenue by number of households.

This same method of allocating regional transit revenues to block groups is also used for allocating transit trips. Using the overall unlinked trip numbers also reported to the NTD, the average number of household transit trips for each block group is estimated by finding the total number of annual trips in each metropolitan area and allocating them proportionally to block groups based on number of households and the percent of journey to work trips.[17]

There are a few metropolitan areas for transit stop locations and/or no revenue listed in the NTD. The average from the allocation calculation described in the previous paragraph is used for these metropolitan areas. The average transit costs are then allocated to the

[16] Demand response revenue is not factored into this analysis.

[17] This normalization method carries the implicit assumption that the transit use for the journey to work is a good surrogate for overall transit use.

block group level based on the percentage of transit commutes and household commuter counts. The result is an average household transit cost at the block group level."

What is Different this Time:

Essentially this is the same process for both LAIM2.0 and this iteration, except in LAIM2.0, GTFS data (from AllTransit database) was used to allocate trips/revenue, now the allocation is less granular and relies on the NTD primary UZA to allocate the trips/revenue from each transit agency.

Appendix F: LAIM3.0 Results and Comparison

This section will examine the current model (LAIM3.f will use LAIM3.0 for this rest of this document) and compare it to the LAM2.0 results.

How to Compare

Since LAIM2.0 used the 2012 5-year ACS data collected and modeled at the Census Block Group level, there are some challenges in comparing the LAIM3.0 – based on the 2016 5-year ACS data collected and modeled at the Census Tract level. While the housing and transportation costs as well as income have presumably changed in the intervening years, we can normalize this by using the national numbers for the measured variables in these years from the ACS and then compare this to the difference in the LAIM2.0 to LAIM3.0 results. To consider, the different Census geography used, the LAIM2.0 results will be aggregated into tracts by using an tenured household weighted average (see Table 32 for the weighting used for each variable) from the constituent block groups, and then a OLS will be run between the two results to see how well they correlate. However, since we do not run the models on the actual households there remains some ambiguity, however using the Median-Income Family will probably minimize these issues.

Direct Model Output

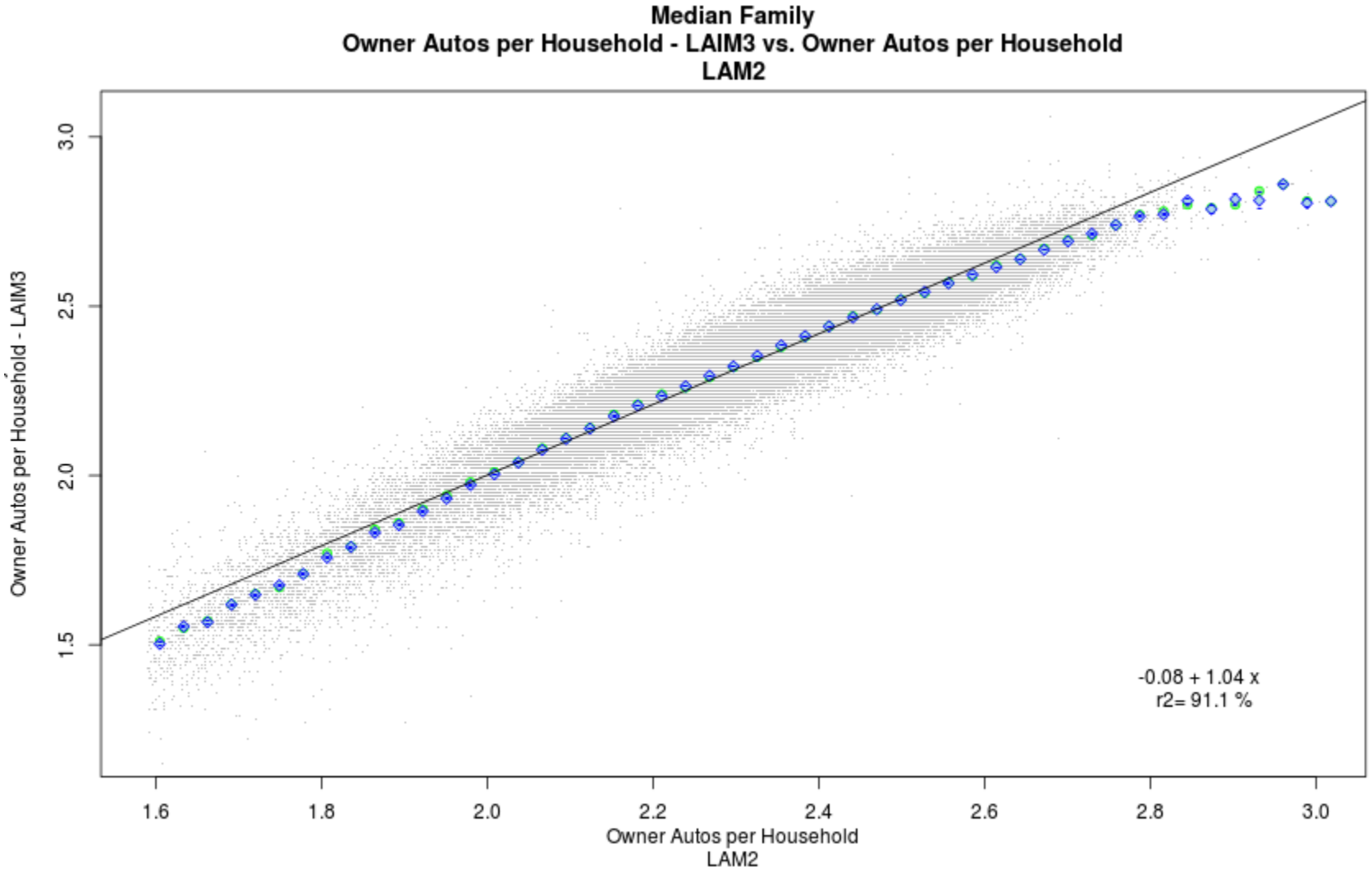
The following table summarized the comparison for six SEM endogenous variables and household VMT for the Median-income family control household, followed be a series of plots showing the strong correlation:

Table 24: Modeled Variables and How They Compare from LAIM2.0 to LAIM3.0

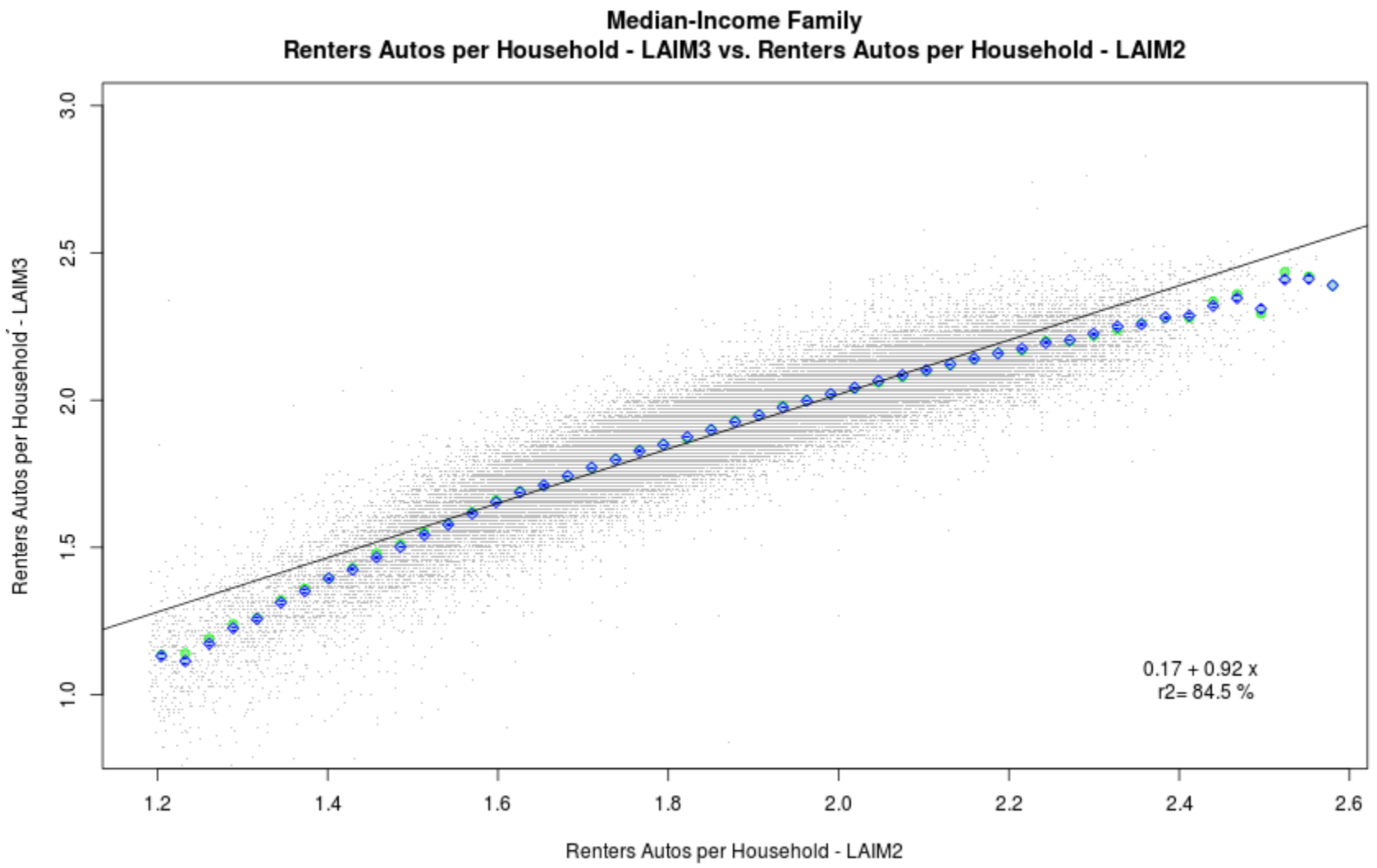
Variable	Weighting	LAIM2.0 Mean (St. Dev)	LAIM3.0 Mean (St. Dev)	National Value ACS 2012	National Value ACS 2016	Intercept	Slope	Inferred Value of LAIM3.0 from LAIM2.0 National Value	Ratio of 2016 inferred over 2016 ACS
Owner Auto Ownership	Owner Occupied Housing Units	2.31 (0.24)	2.32 (0.27)	2.04	2.06	-0.08	1.04	2.04	0.991
Renter Auto Ownership	Renter Occupied Housing Units	1.88 (0.23)	1.90 (0.24)	1.23	1.27	0.17	0.92	1.30	1.025
Renter Housing Cost	Owner Occupied Housing Units	$13,524 ($3,993)	$14,262 ($3,636)	10,668	11,388	$3,117	0.83	$11,971.44	1.051

Owner Housing Cost	Renter Occupied Housing Units	$16,913 ($4,970)	$15,644 ($4,478)	18,708	17,892	$1,490	0.84	$17,204.72	0.962
Owner Transit Commute Share	Owner Occupied Housing Units	5.13% (6.52%)	5.97% (8.61%)	2.91%	3.03%	-0.07%	1.10	3.13%	1.033
Renter Transit Commute Share	Renter Occupied Housing Units	6.48% (8.72%)	5.74% (10.46%)	9.51%	9.16%	-1.44%	1.01	8.17%	0.891
Household VMT	Households	25,938 (4,783)	28,269 (5,583)	NA	NA	1,386	1.04	NA	NA

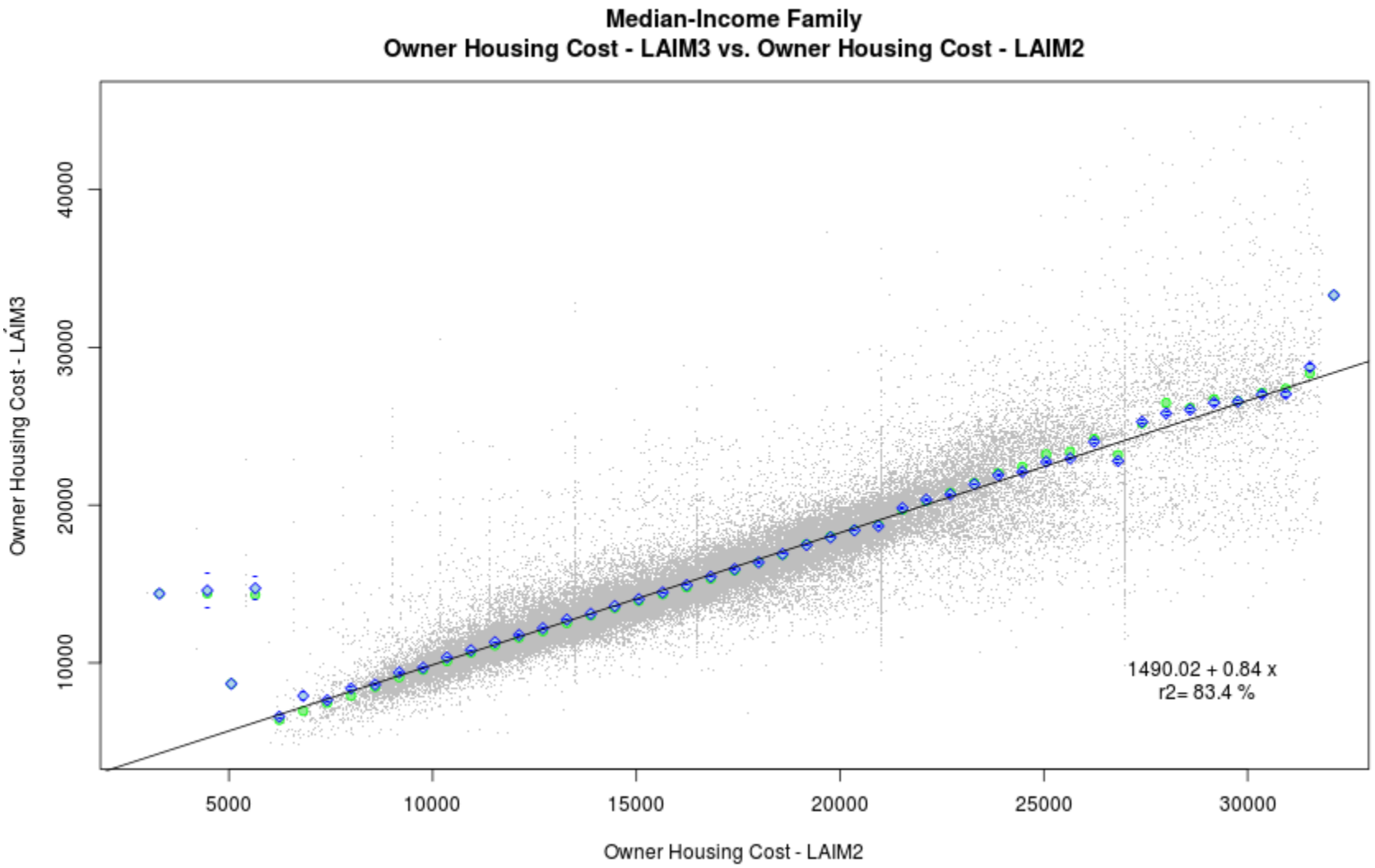
Figure 15: Plots of LAIM3.0 Outputs vs. LAIM2.0 Aggregated Outputs[18]

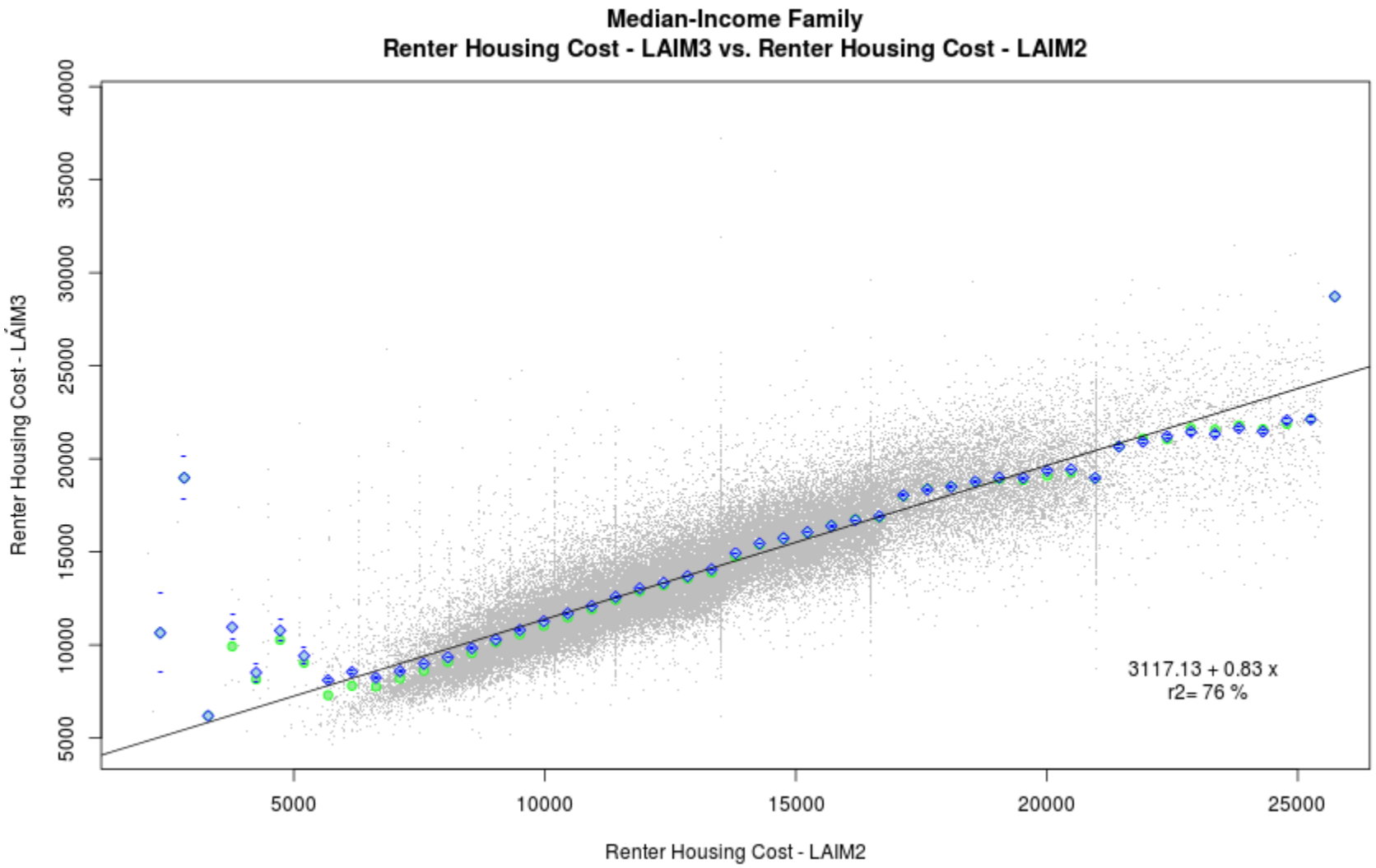


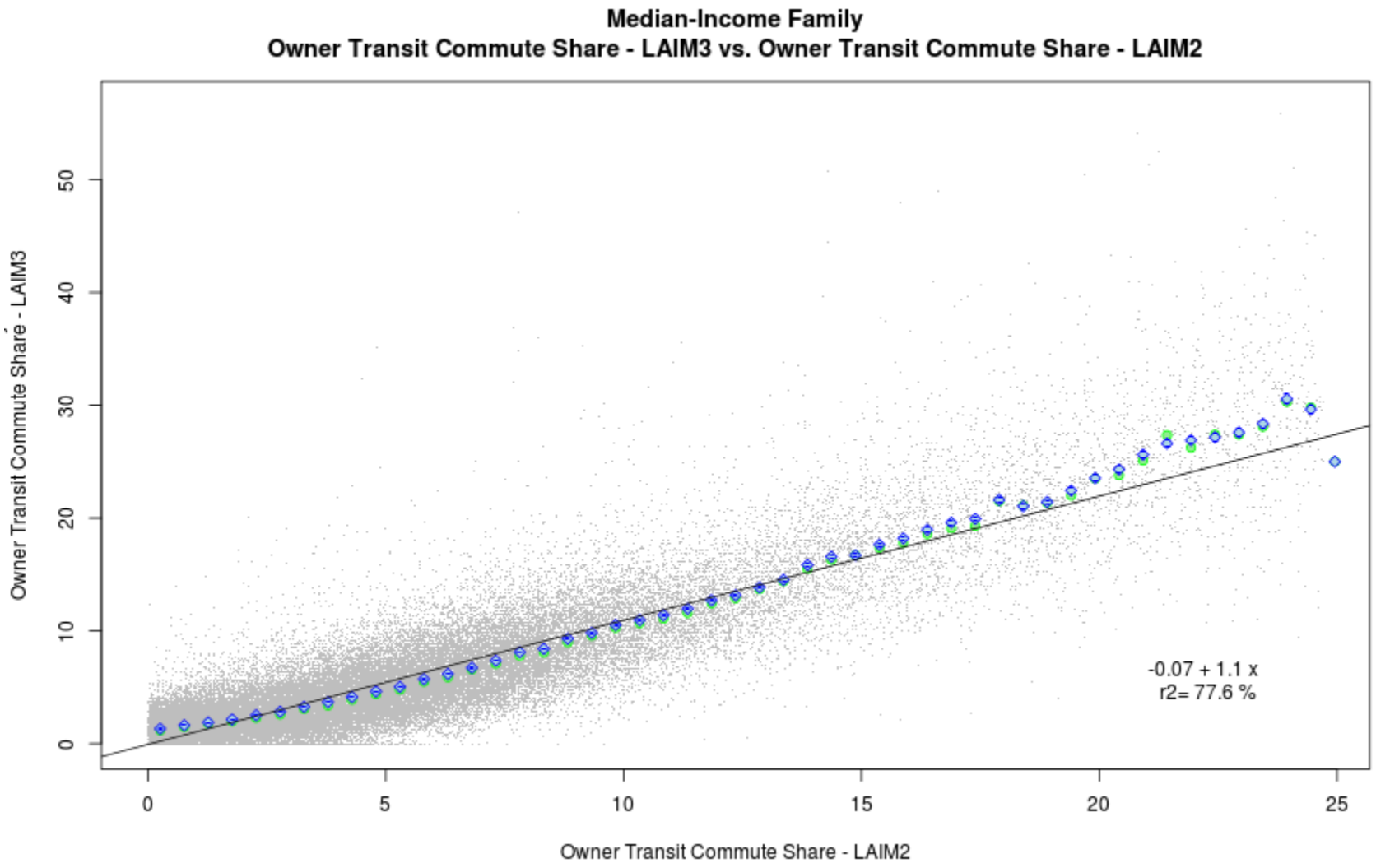
[18] These plots display a few things – the grey dots are the value for every block group, the blue diamond is the mean y-value for 50 bins in x, the green circles

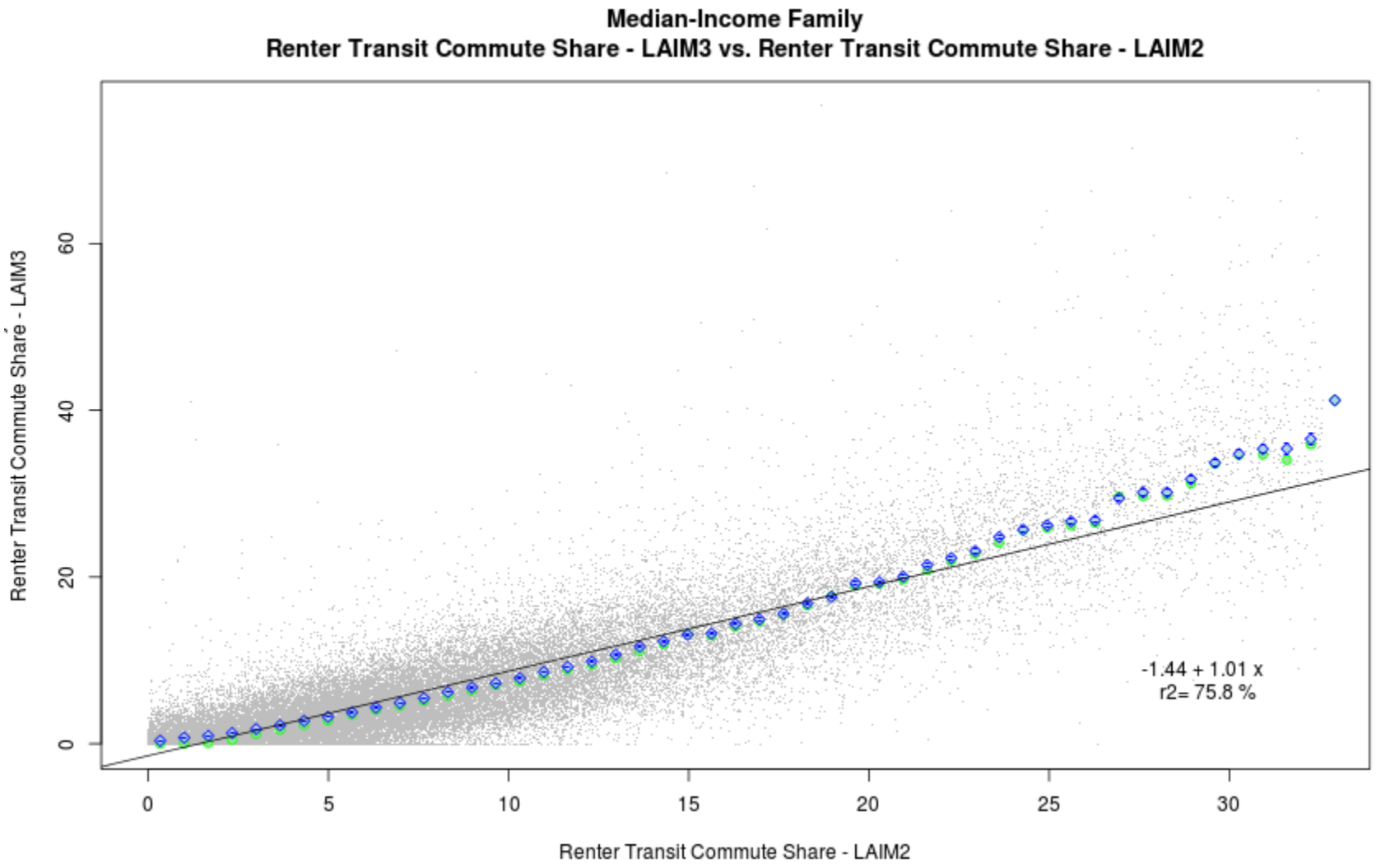


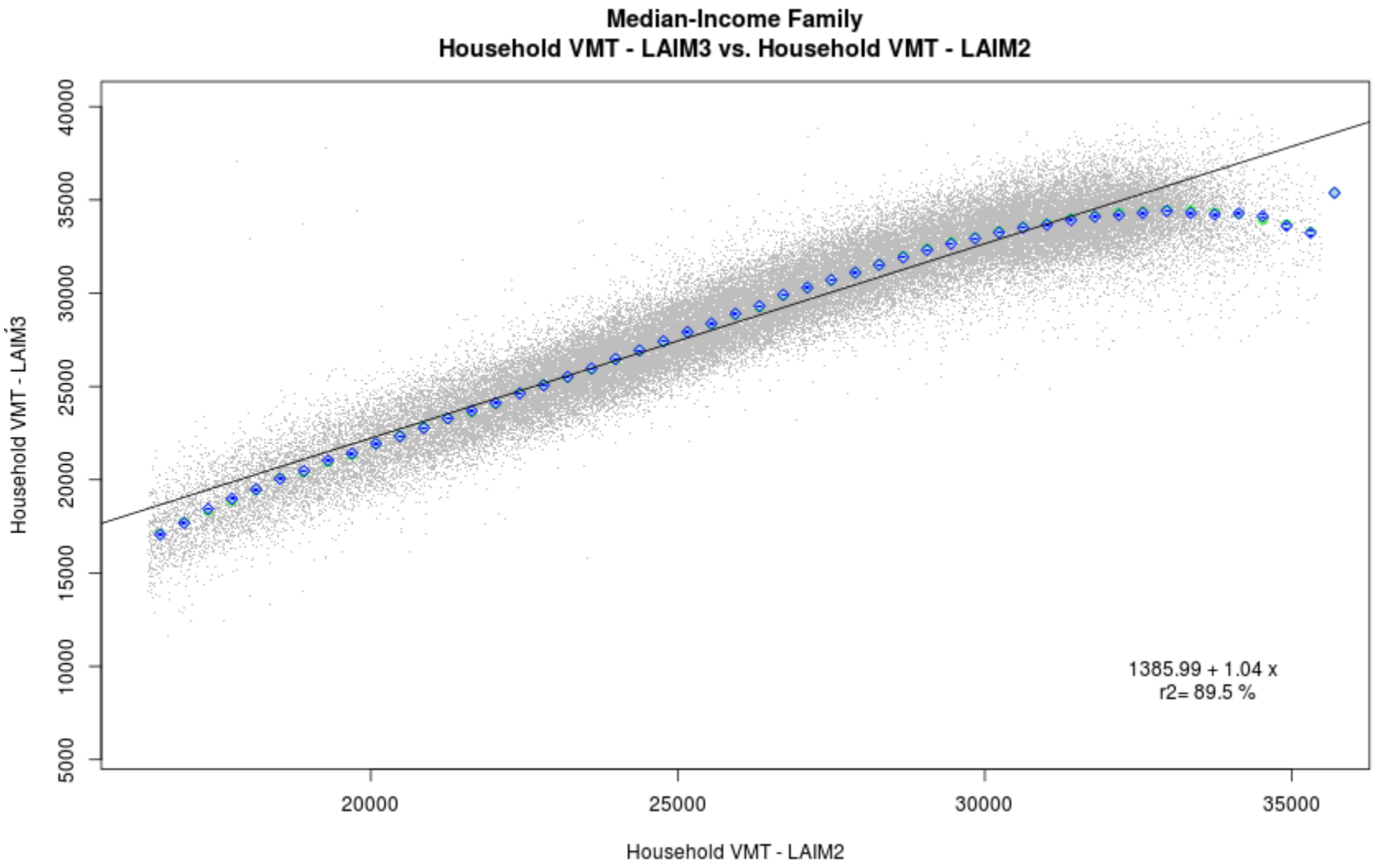
are the median, and the line is the linear fit to the grey dots with fit stats in the lower right.


Median-Income Family
Owner Housing Cost - LAIM3 vs. Owner Housing Cost - LAIM2
Owner Housing Cost - LAIM3
10000
20000
30000
40000
5000
10000
15000
20000
25000
30000
Owner Housing Cost - LAIM2
1490.02 + 0.84 x
r2= 83.4 %

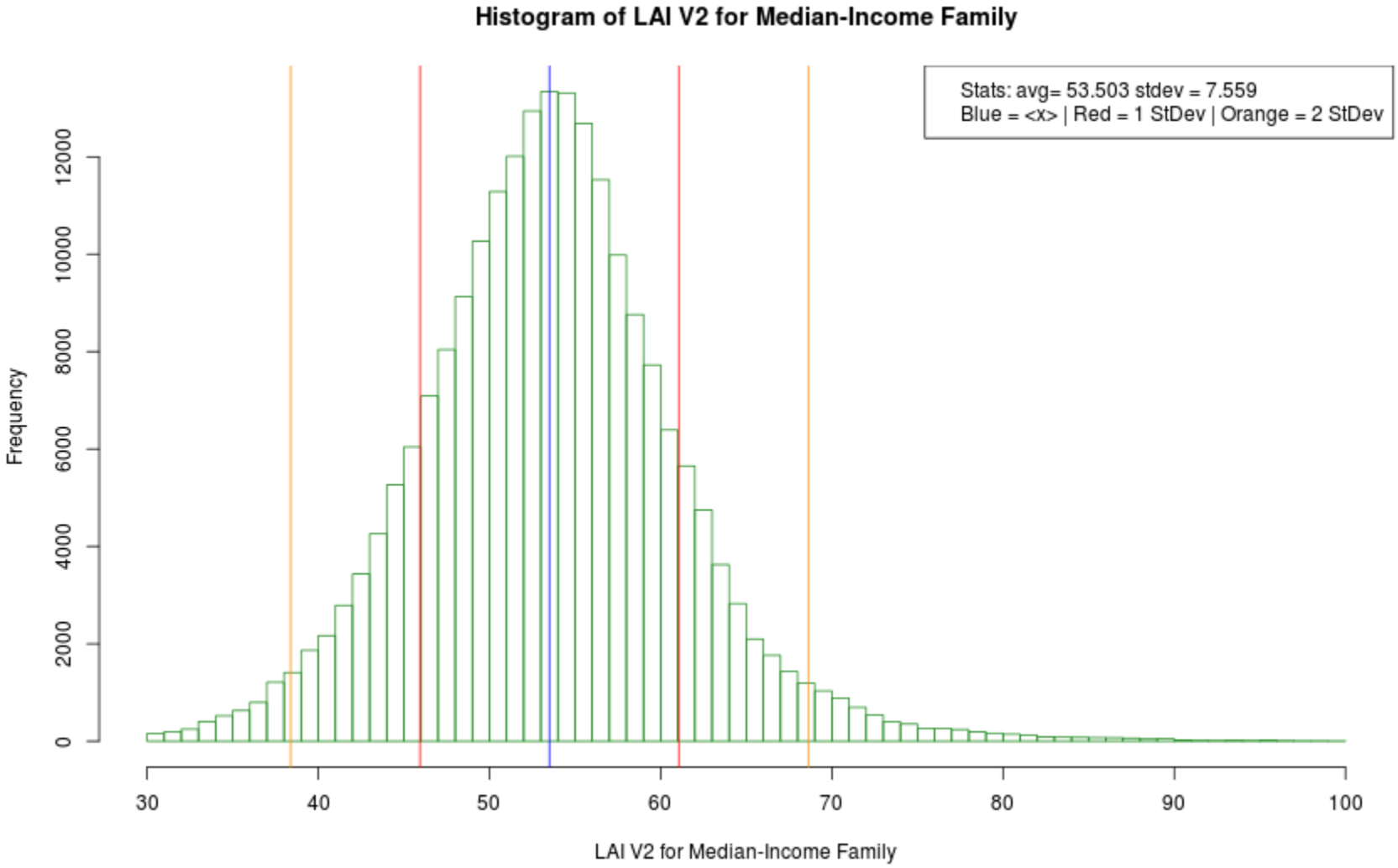

Median-Income Family
Renter Housing Cost - LAIM3 vs. Renter Housing Cost - LAIM2
Renter Housing Cost - LAIM3
Renter Housing Cost - LAIM2
5000
10000
15000
20000
25000
30000
35000
40000
3117.13 + 0.83 x
r2= 76 %

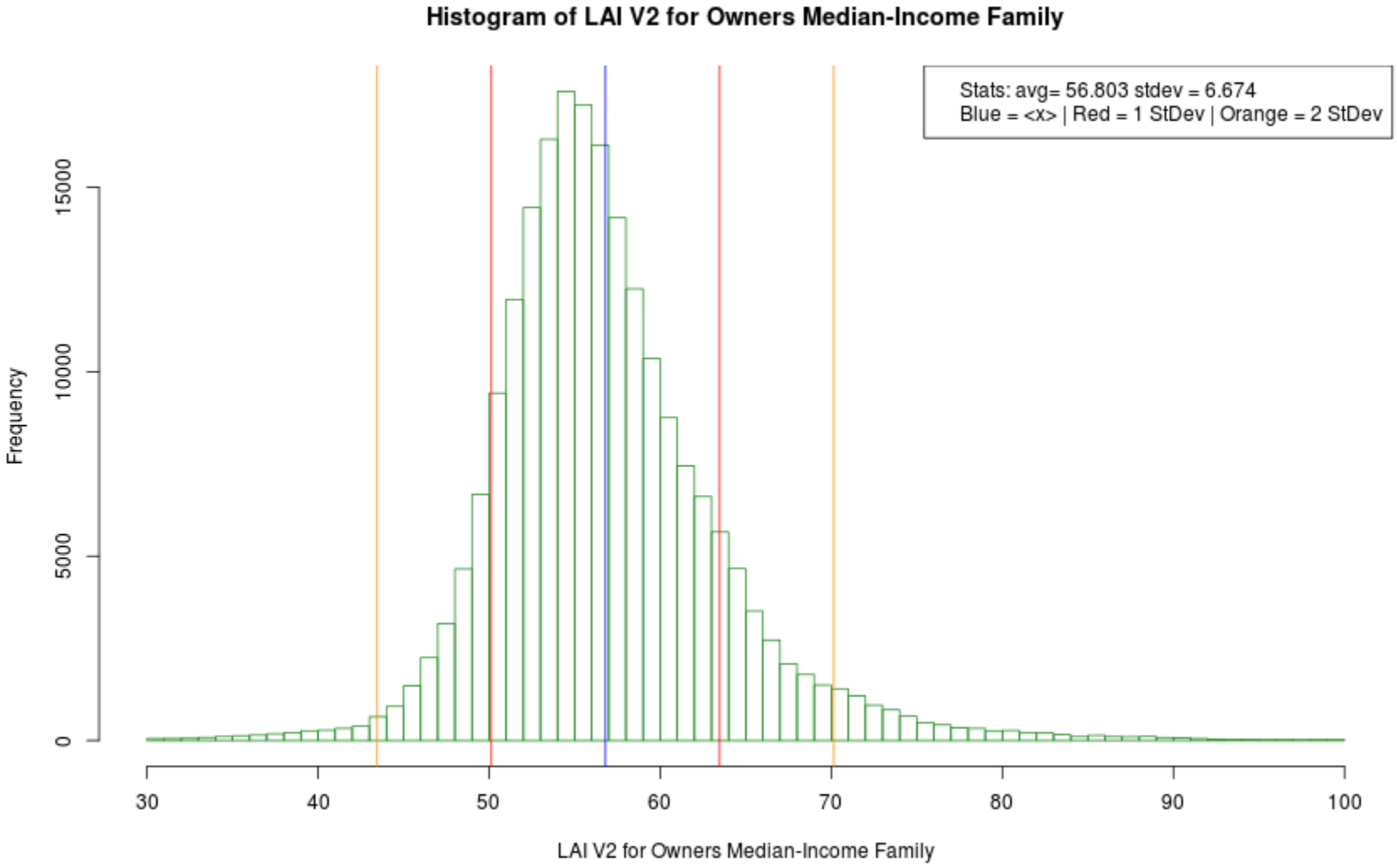

Median-Income Family
Owner Transit Commute Share - LAIM3 vs. Owner Transit Commute Share - LAIM2
Owner Transit Commute Share - LAIM3
Owner Transit Commute Share - LAIM2
0
10
20
30
40
50
0
5
10
15
20
25
-0.07 + 1.1 x
r2= 77.6 %

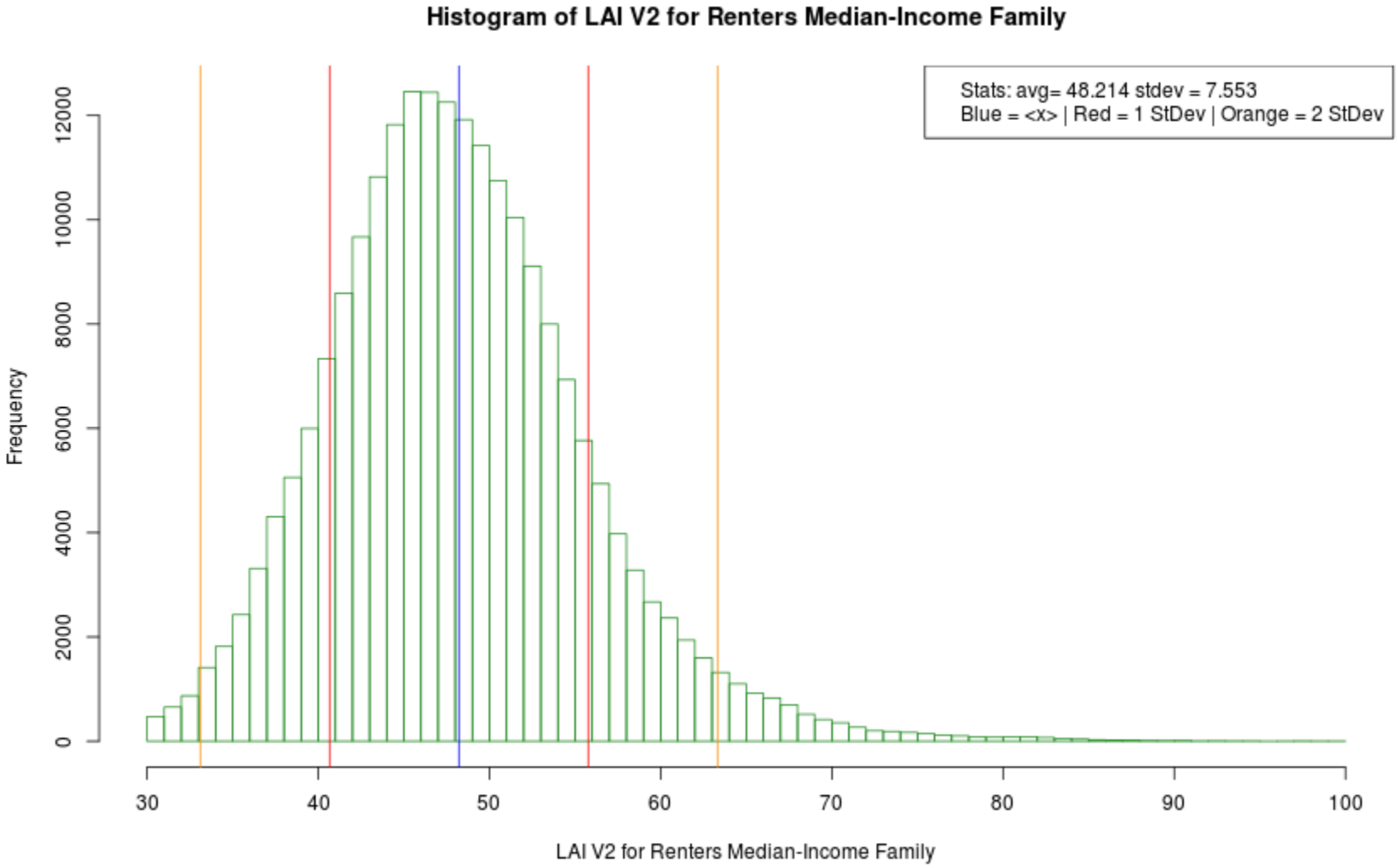

Median-Income Family
Renter Transit Commute Share - LAIM3 vs. Renter Transit Commute Share - LAIM2
Renter Transit Commute Share - LAIM3
Renter Transit Commute Share - LAIM2
0
20
40
60
0
5
10
15
20
25
30
-1.44 + 1.01 x
r2= 75.8 %

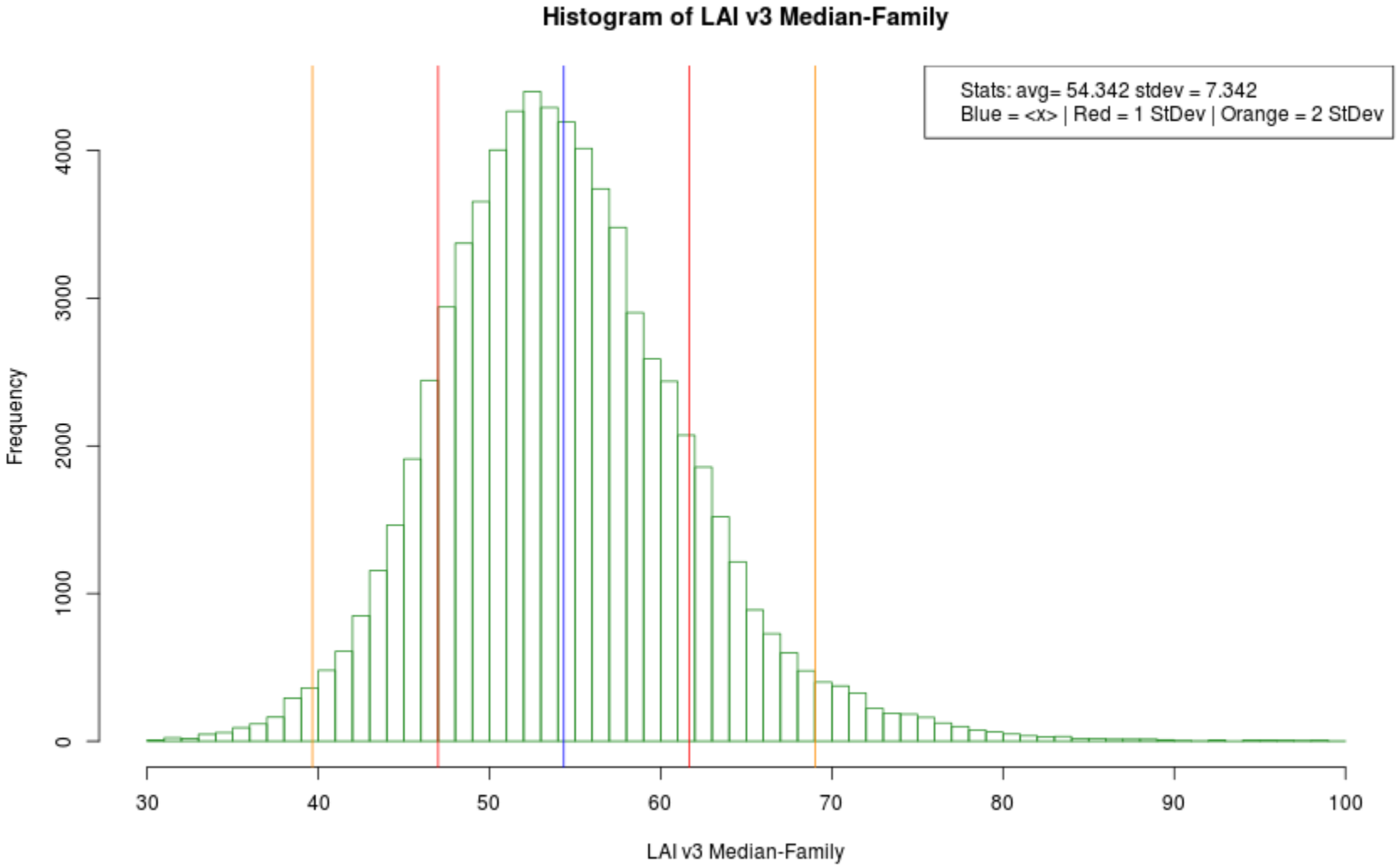

Median-Income Family
Household VMT - LAIM3 vs. Household VMT - LAIM2
Household VMT - LAIM3
Household VMT - LAIM2
5000
10000
15000
20000
25000
30000
35000
40000
20000
25000
30000
35000
1385.99 + 1.04 x
r2= 89.5 %

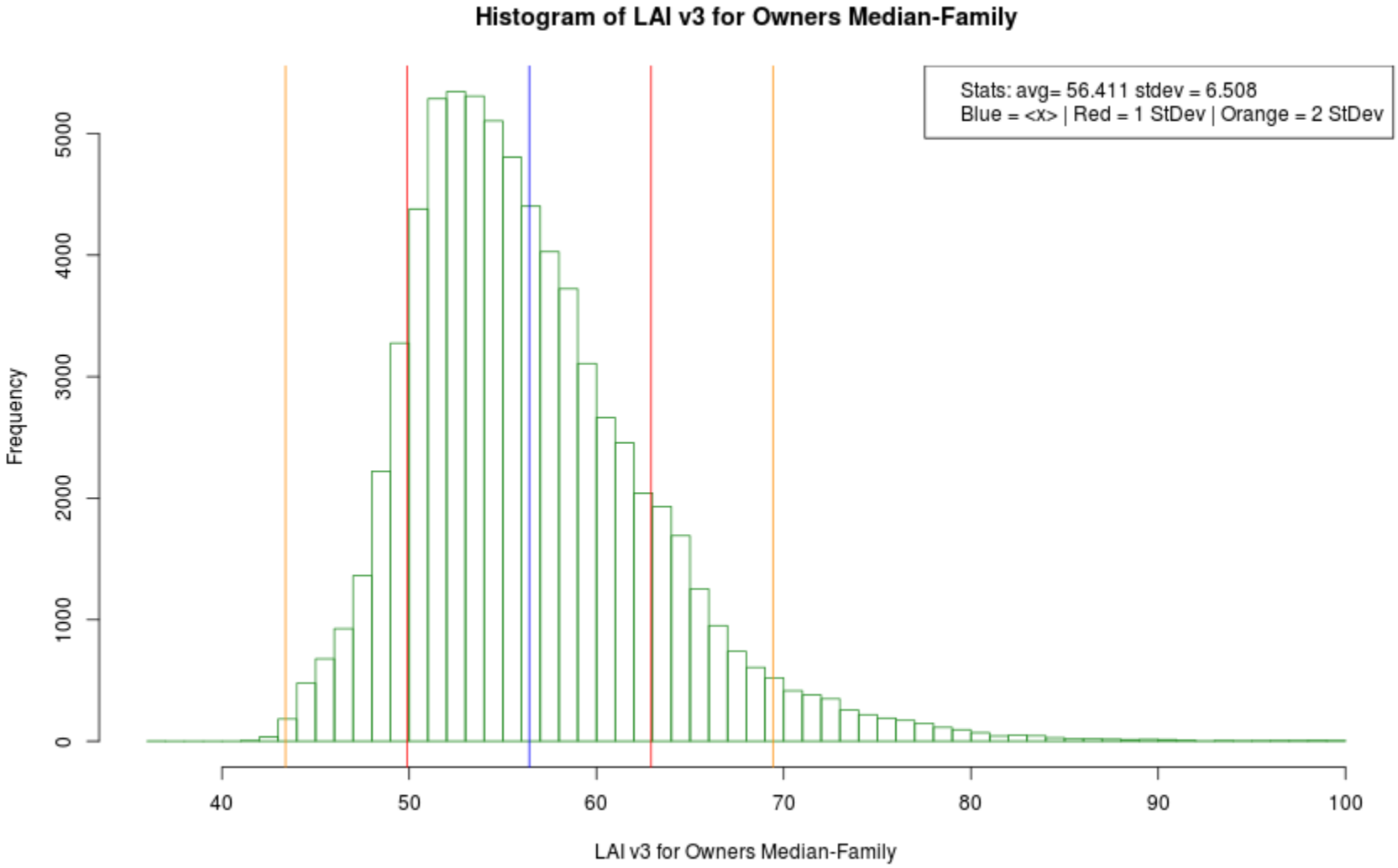
The following two histograms show the distribution for the overall LAI (Housing Cost + Transportation Cost)/Income for the tenure weighted combination of owners and renters, for LAIM2.0 and LAIM3.0, as well as the error-bar plot showing the correlation for the housing cost over income, the transportation cost over income and the combined ratio.

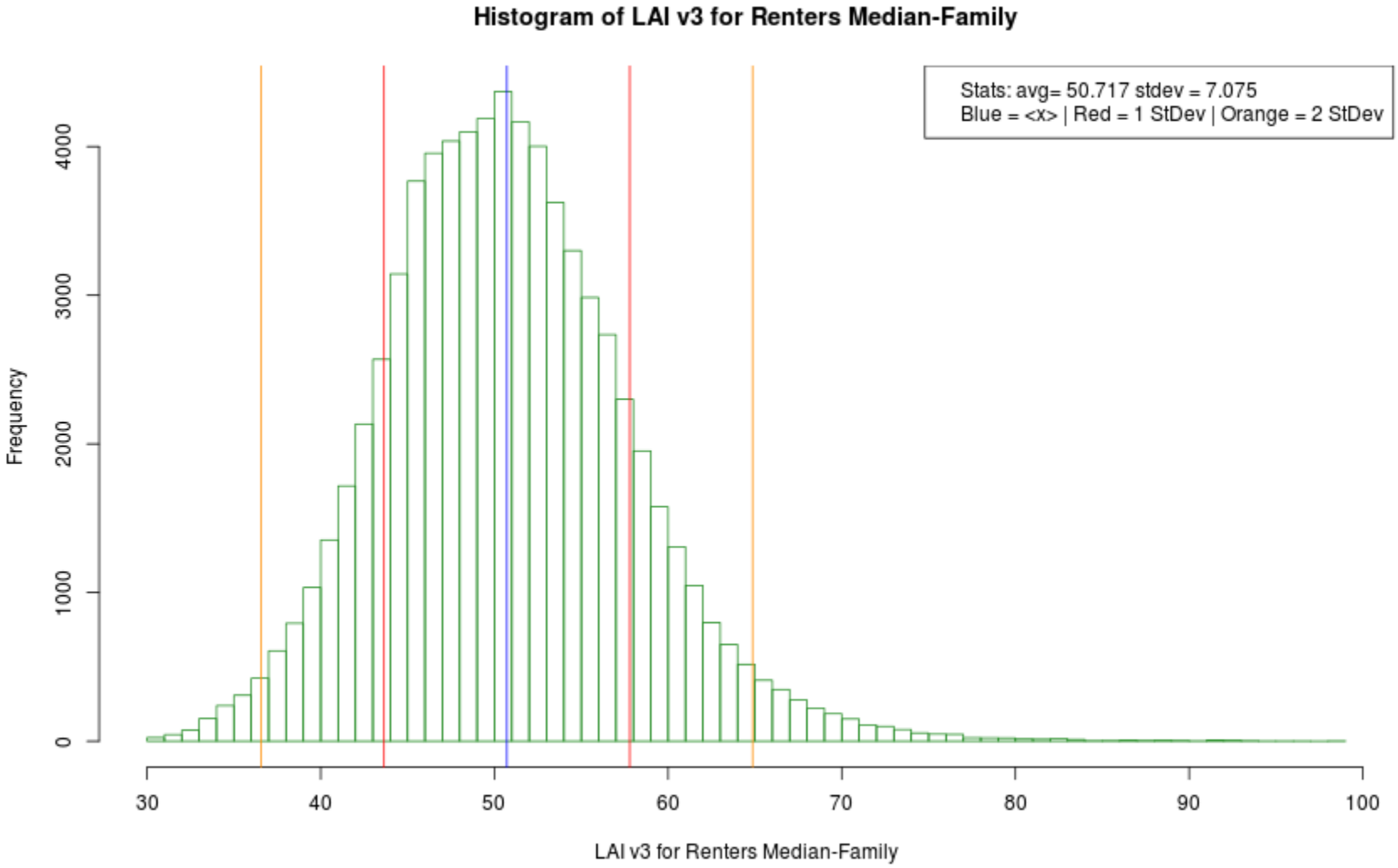

Histogram of LAI V2 for Median-Income Family
Stats: avg= 53.503 stdev = 7.559
Blue = <x> | Red = 1 StDev | Orange = 2 StDev
Frequency
0
2000
4000
6000
8000
10000
12000
30
40
50
60
70
80
90
100
LAI V2 for Median-Income Family

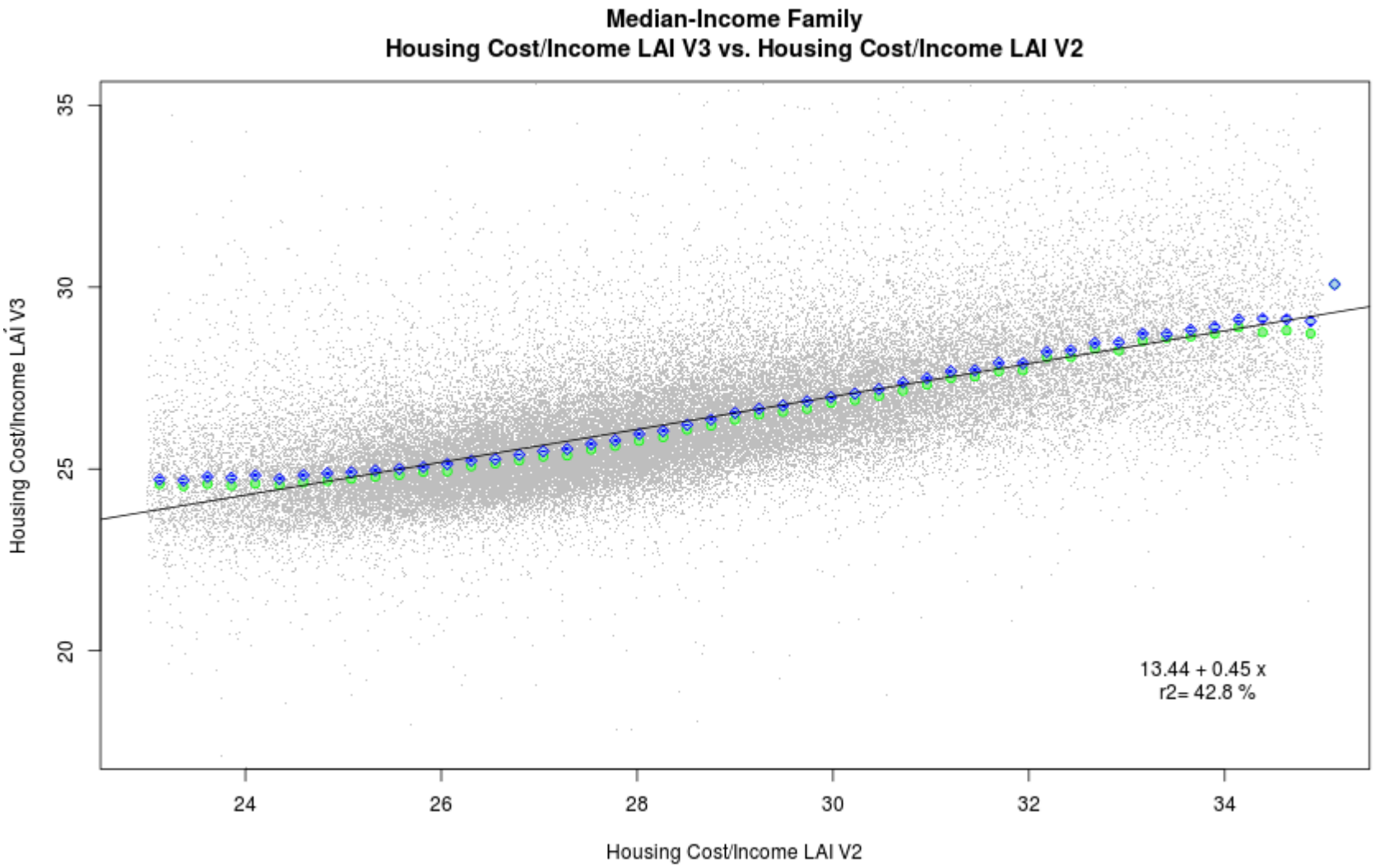

Histogram of LAI V2 for Owners Median-Income Family
Stats: avg= 56.803 stdev = 6.674
Blue = <x> | Red = 1 StDev | Orange = 2 StDev
Frequency
0
5000
10000
15000
30
40
50
60
70
80
90
100
LAI V2 for Owners Median-Income Family

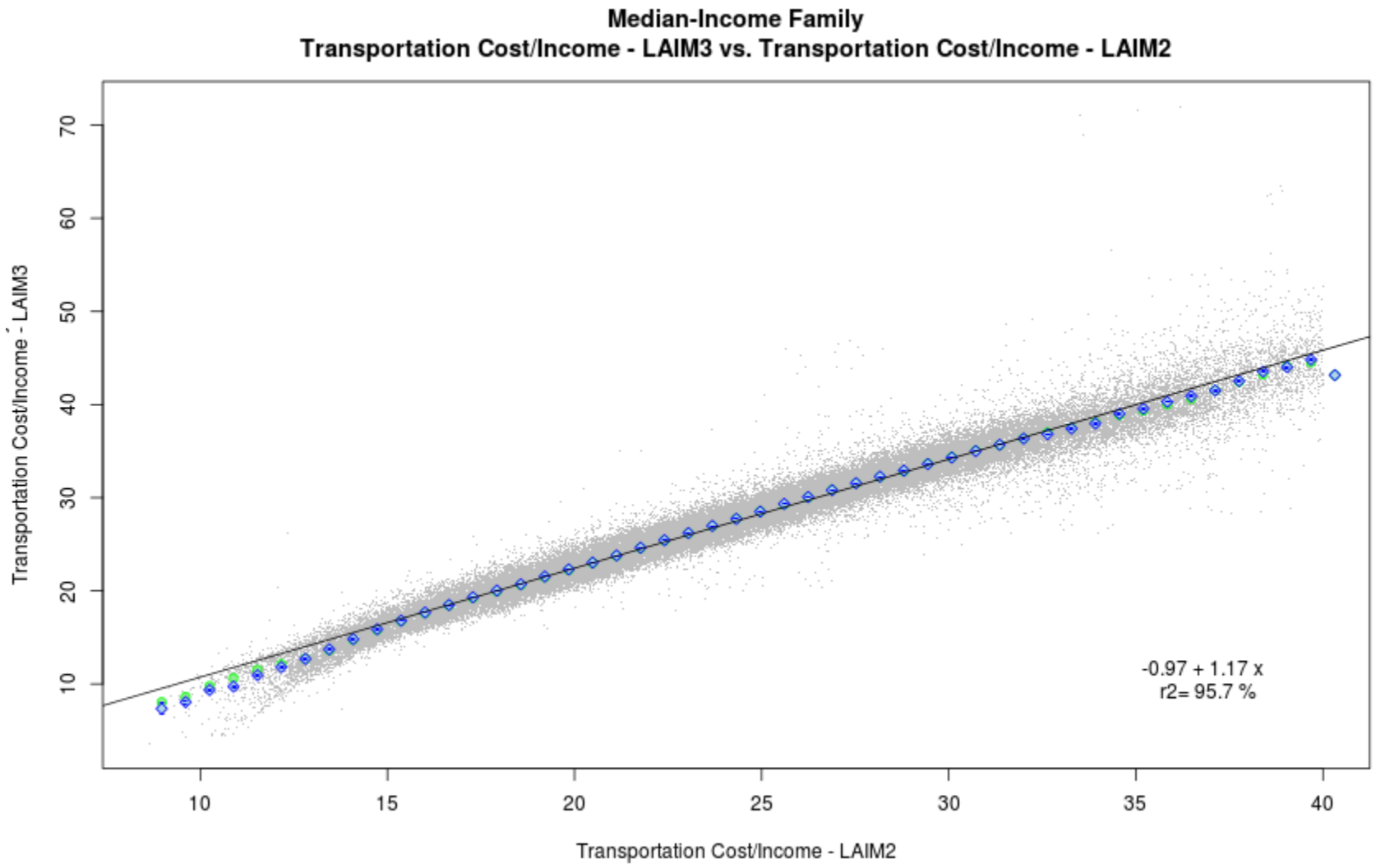

Histogram of LAI V2 for Renters Median-Income Family
Stats: avg= 48.214 stdev = 7.553
Blue = <x> | Red = 1 StDev | Orange = 2 StDev
Frequency
0
2000
4000
6000
8000
10000
12000
30
40
50
60
70
80
90
100
LAI V2 for Renters Median-Income Family

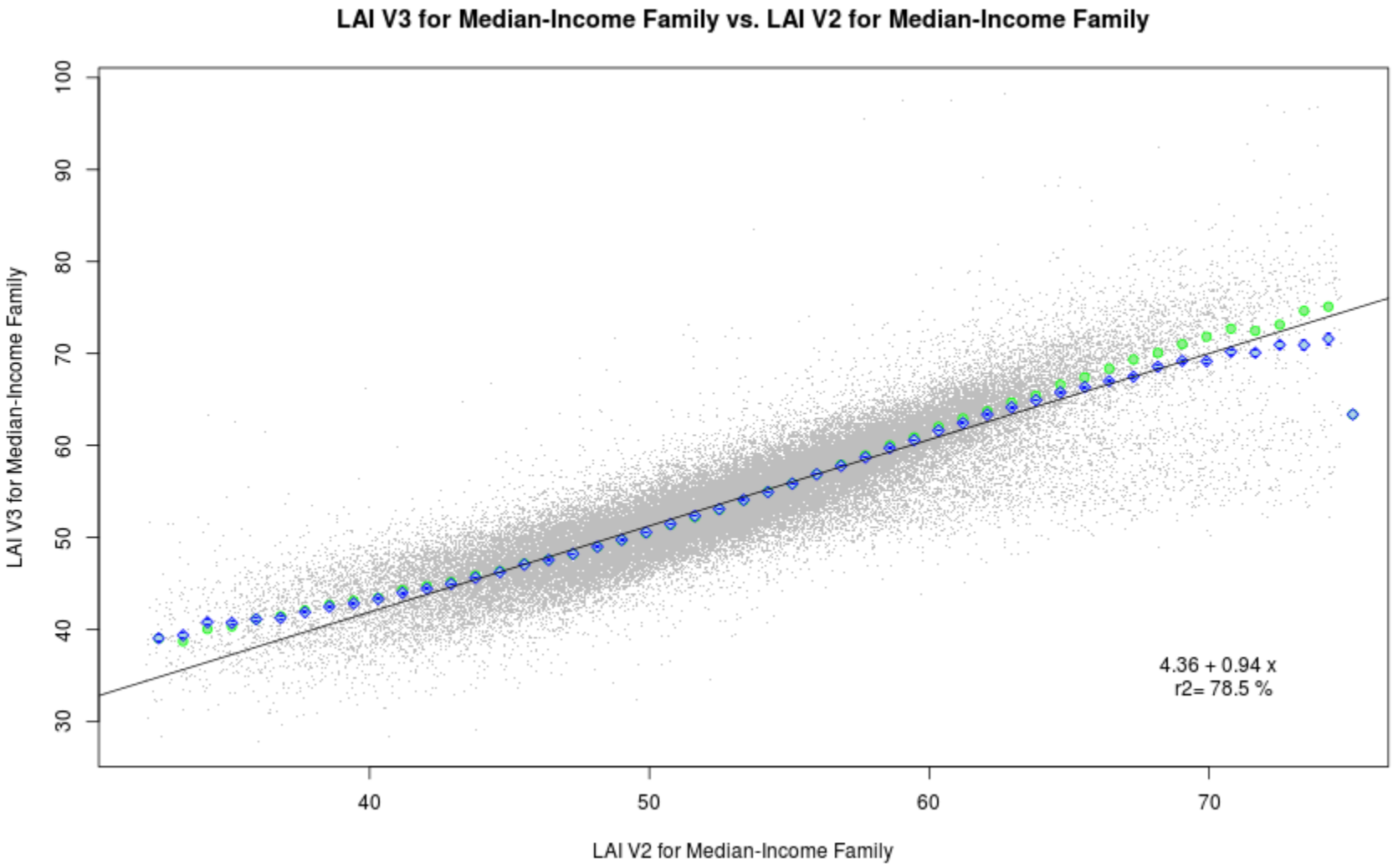

Histogram of LAI v3 Median-Family
Stats: avg= 54.342 stdev = 7.342
Blue = <x> | Red = 1 StDev | Orange = 2 StDev
Frequency
0
1000
2000
3000
4000
30
40
50
60
70
80
90
100
LAI v3 Median-Family


Histogram of LAI v3 for Owners Median-Family
Stats: avg= 56.411 stdev = 6.508
Blue = <x> | Red = 1 StDev | Orange = 2 StDev
Frequency
0
1000
2000
3000
4000
5000
40
50
60
70
80
90
100
LAI v3 for Owners Median-Family


Histogram of LAI v3 for Renters Median-Family
Stats: avg= 50.717 stdev = 7.075
Blue = <x> | Red = 1 StDev | Orange = 2 StDev
Frequency
0
1000
2000
3000
4000
30
40
50
60
70
80
90
100
LAI v3 for Renters Median-Family


Median-Income Family
Housing Cost/Income LAI V3 vs. Housing Cost/Income LAI V2
Housing Cost/Income LAÍ V3
Housing Cost/Income LAI V2
13.44 + 0.45 x
r2= 42.8 %
20
25
30
35
24
26
28
30
32
34


Median-Income Family
Transportation Cost/Income - LAIM3 vs. Transportation Cost/Income - LAIM2
Transportation Cost/Income - LAIM3
Transportation Cost/Income - LAIM2
-0.97 + 1.17 x
r2= 95.7 %
10
20
30
40
50
60
70
10
15
20
25
30
35
40


LAI V3 for Median-Income Family vs. LAI V2 for Median-Income Family
LAI V3 for Median-Income Family
LAI V2 for Median-Income Family
30
40
50
60
70
80
90
100
40
50
60
70
4.36 + 0.94 x
r2= 78.5 %

Conclusions

The LAIM3.0 is a robust model, that gives values that are consistent with LAIM2.0. Even when the values differ significantly between the two models, most of the difference can be explained by the 4 years' worth of change that happened in the US economy, and Table 32 captures this when comparing the mean in 2012 ACS and 2016 ACS values. The over index looks to have changes on average by one percent, increasing at for the Median-Income Family from a mean of 53.3% in 2012 to a mean of 54.3% in 2016.

Appendix G: Household Percentile Index

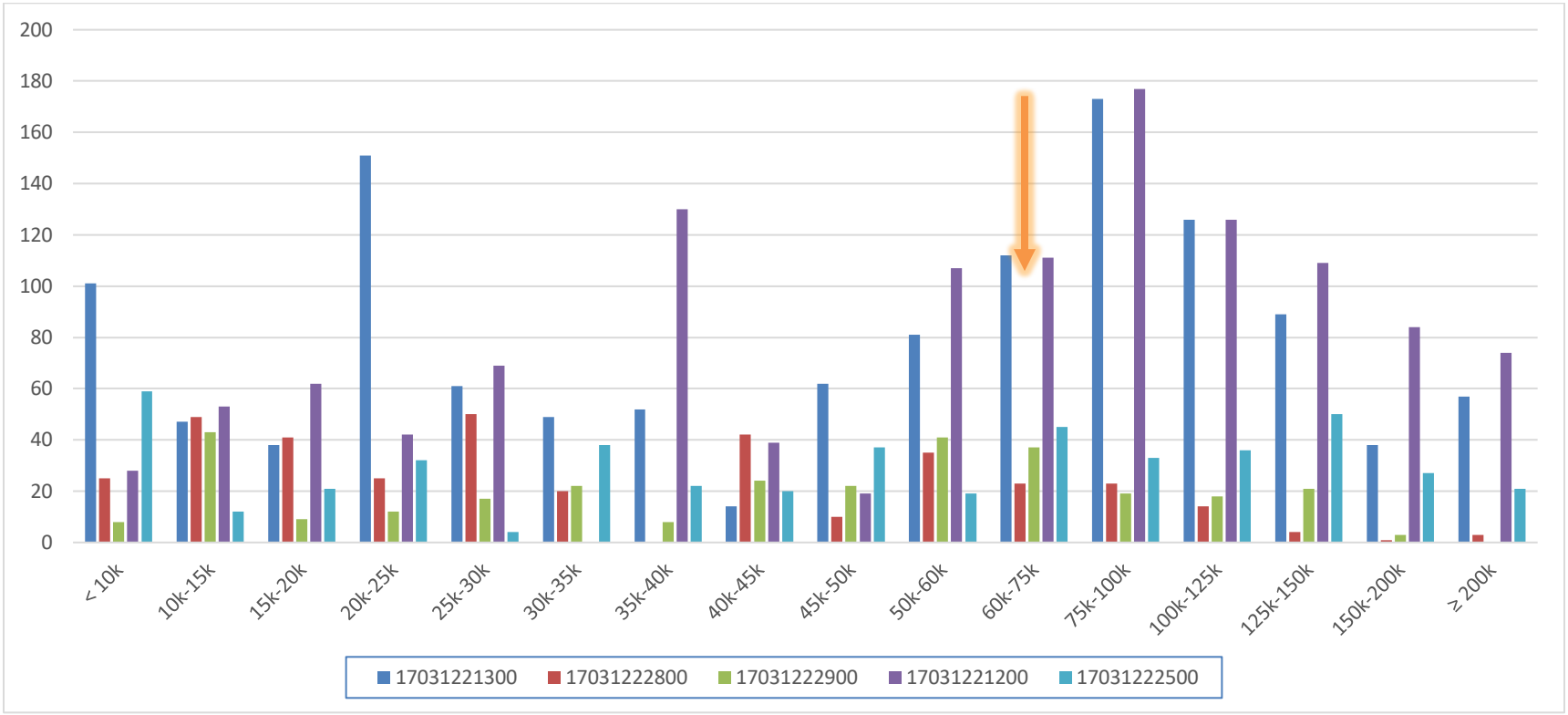
Household Percentile Index

HUD wanted to run a new index to find for each tract how the income of the control household would compare to the income within the tract. In other words, find the percent of all the households within the tract that have a lower income than that of the control household. The census gives the number of households within bins by household income, and that can use this to estimate this percentile. A spread sheet of 5 tracts is included in a neighborhood in Chicago. Note that the control household income is the same (the "Median-Income Family" control household for this example). The formula in sql would look be:

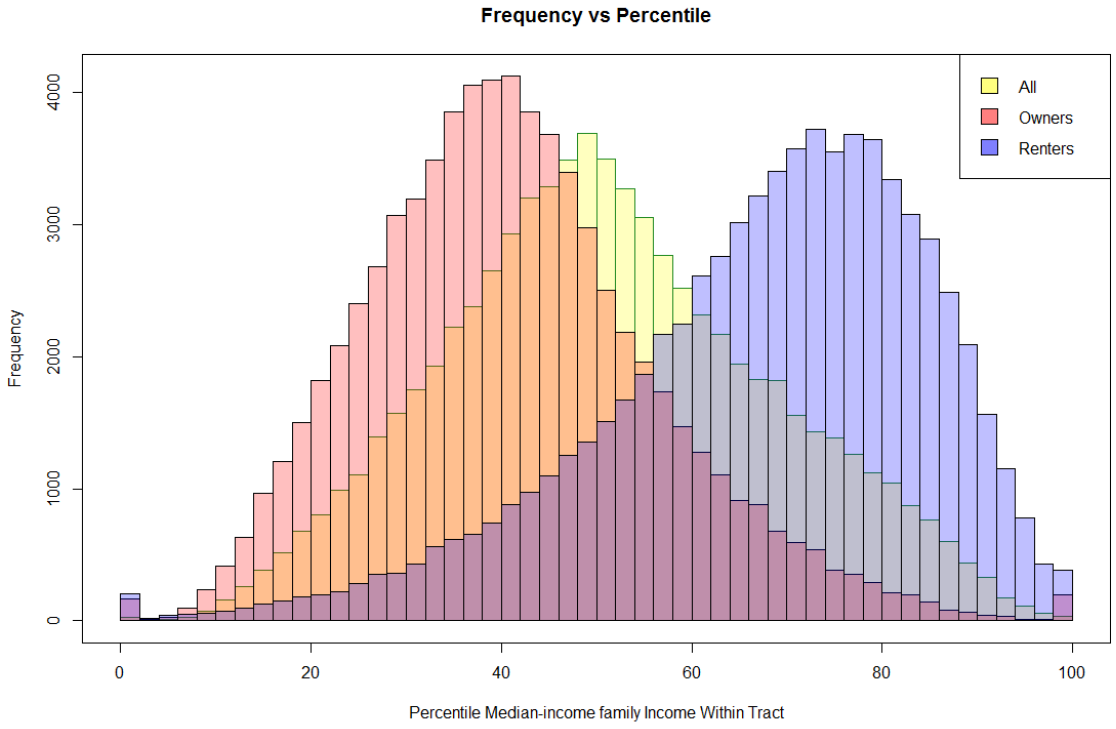
```
 when control_hh_income < 75000

then 100*(hh_income_lt_10k+hh_income_10k_15k+hh_income_15k_20k+

                hh_income_20k_25k+hh_income_25k_30k+hh_income_30k_35k+hh_income_35k_40k+

                hh_income_40k_45k+hh_income_45k_50k+hh_income_50k_60k+

                hh_income_60k_75k*

                ((control_hh_income-60000)/(75000-60000)))/hh_income_universe )
```

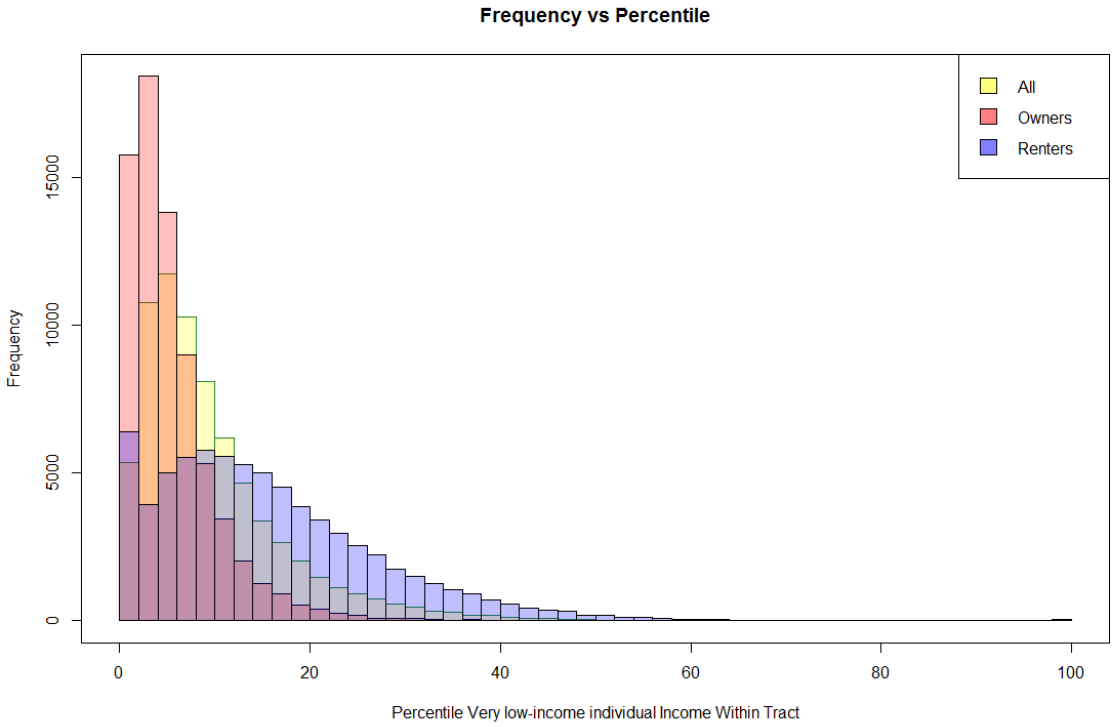
In this case control_hh_income is always $63,327 – the median household income for the Chicago MSA (note that I extrapolate the value for the bin that contains the control household income). On the graphs there is a "glowing" orange arrow at approximately $63m327, showing visually for each tract the percentile. Note that there is a different percentile depending on the income distribution of the households in each tract. While it is "agnostic as to the likelihood of a particular household residing in a given tract" it does depend on the distribution of households in the tract.

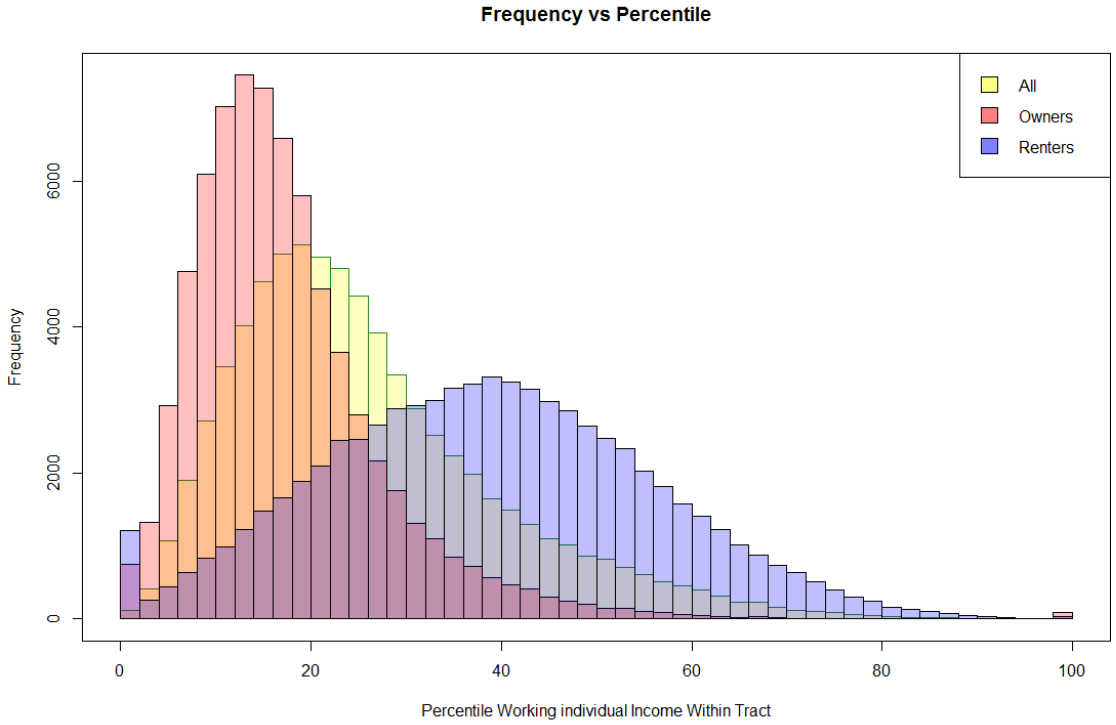
Tract	Percentile	HH Universe	< 10k	10k-15k	15k-20k	20k-25k	25k-30k	30k-35k	35k-40k	40k-45k	45k-50k	50k-60k	60k-75k	75k-100k	100k-125k	125k-150k	150k-200k	≥ 200k
17031221300	54.42	1251	101	47	38	151	61	49	52	14	62	81	112	173	126	89	38	57
17031222800	82.77	365	25	49	41	25	50	20	0	42	10	35	23	23	14	4	1	3
17031222900	70.46	304	8	43	9	12	17	22	8	24	22	41	37	19	18	21	3	0
17031221200	46.64	1230	28	53	62	42	69	0	130	39	19	107	111	177	126	109	84	74
17031222500	57.56	476	59	12	21	32	4	38	22	20	37	19	45	33	36	50	27	21

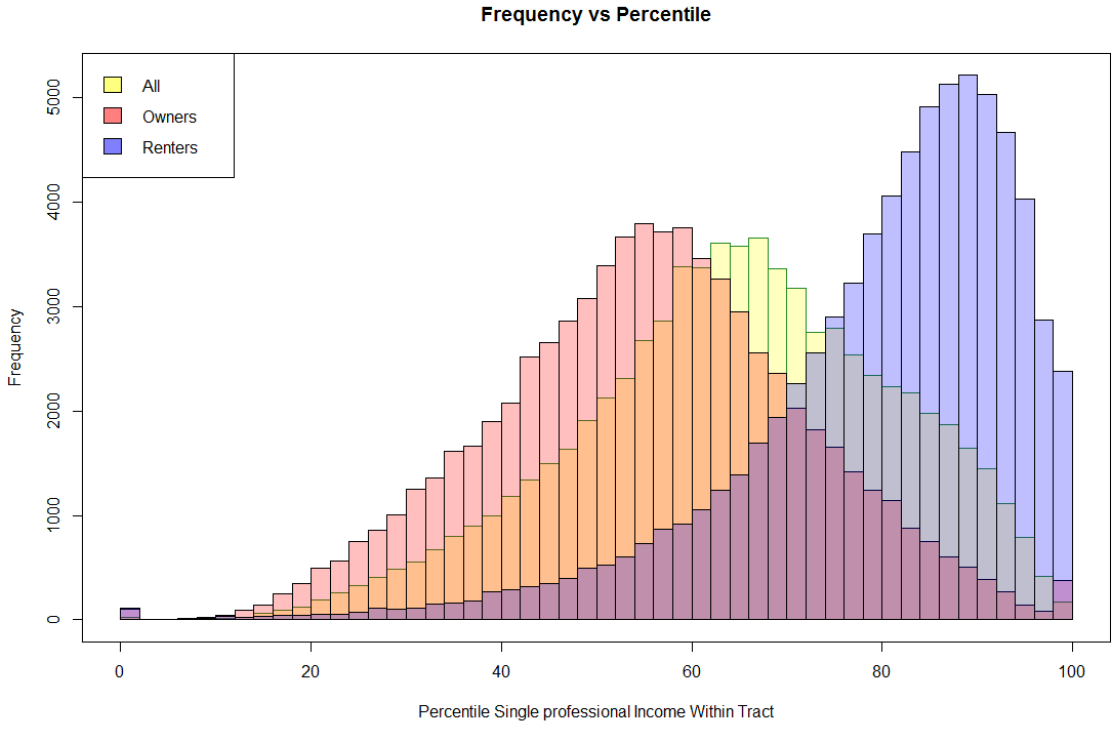


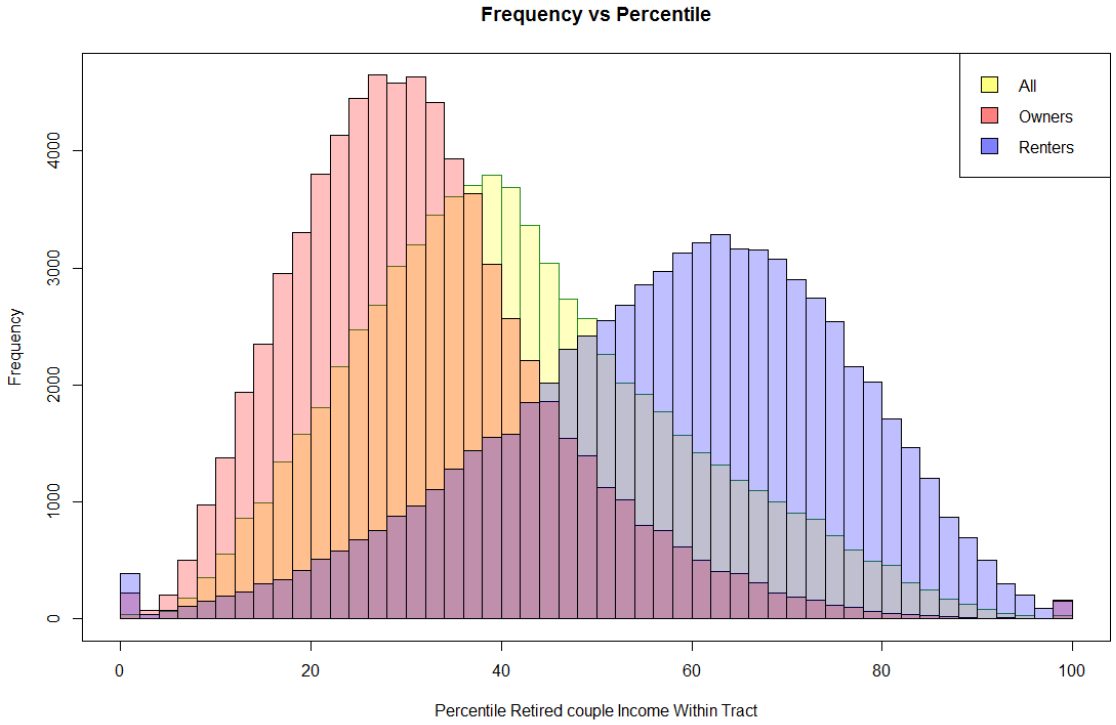
The following histograms show the percentile for all census tracts, for the eight control households.

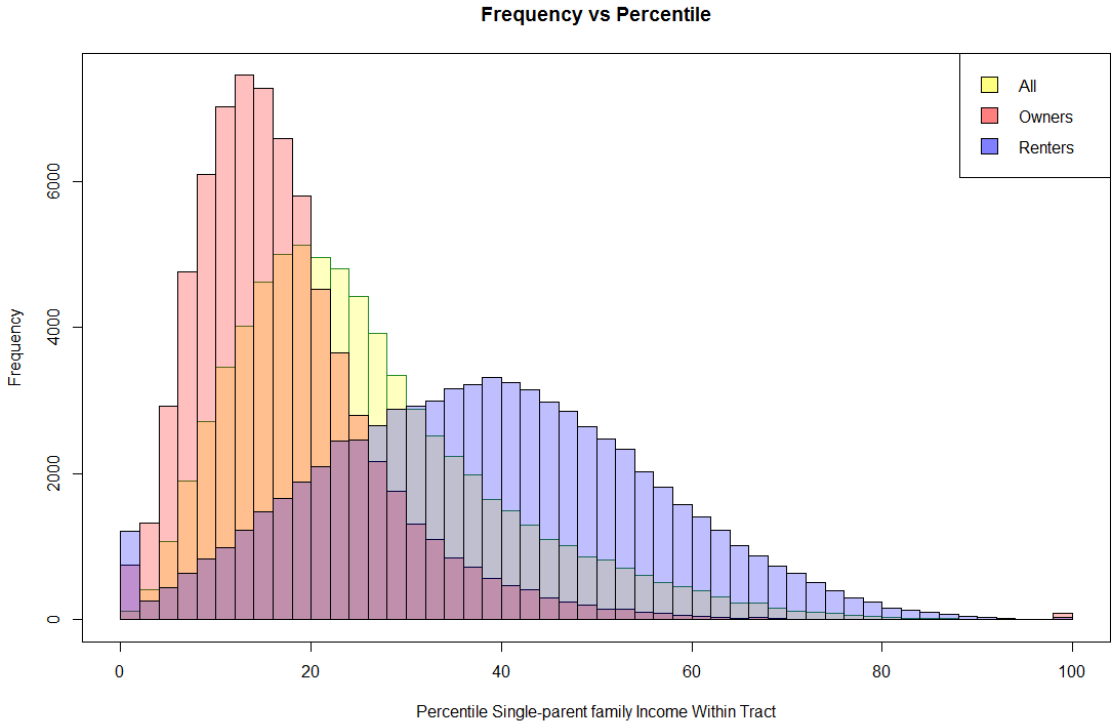


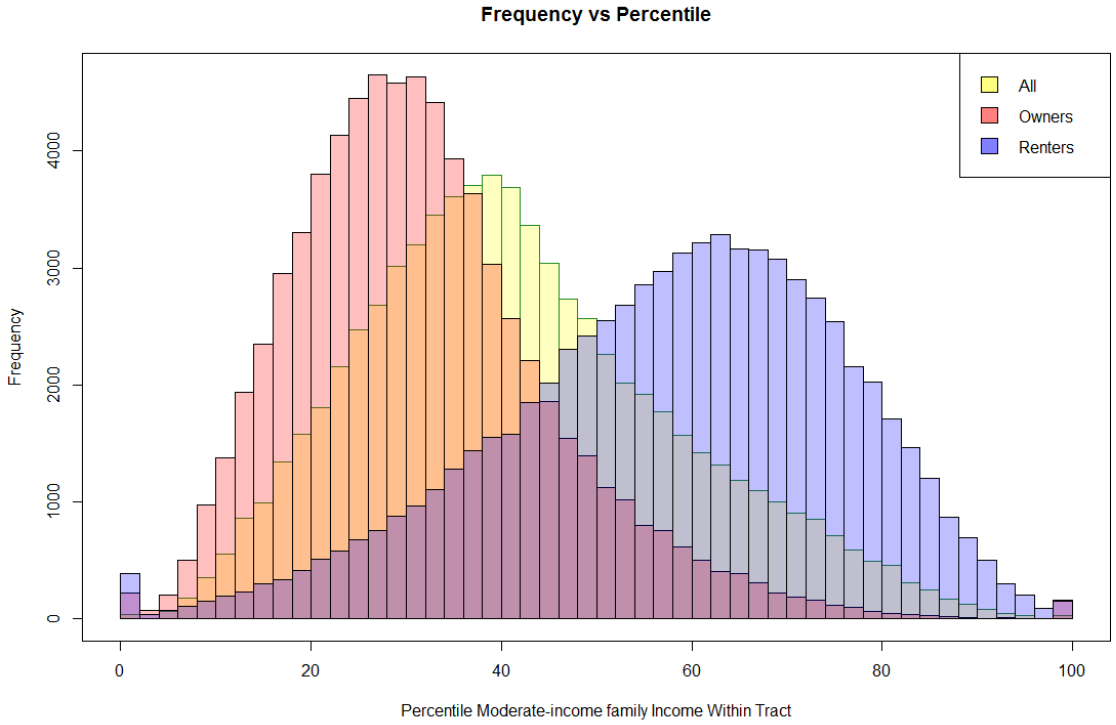


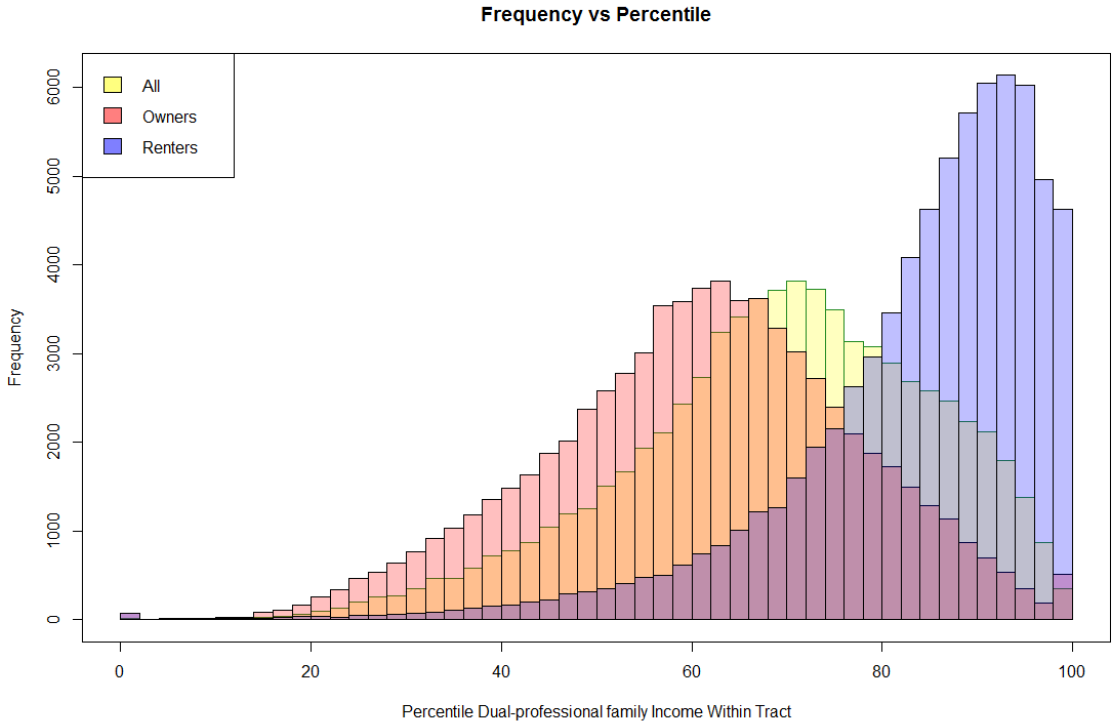

Frequency vs Percentile
All
Owners
Renters
Frequency
0
2000
4000
6000
0
20
40
60
80
100
Percentile Working individual Income Within Tract


Frequency vs Percentile
All
Owners
Renters
Frequency
0
1000
2000
3000
4000
5000
0
20
40
60
80
100
Percentile Single professional Income Within Tract


Frequency vs Percentile
All
Owners
Renters
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
Percentile Retired couple Income Within Tract


Frequency vs Percentile
All
Owners
Renters
Frequency
0
2000
4000
6000
0
20
40
60
80
100
Percentile Single-parent family Income Within Tract


Frequency vs Percentile
All
Owners
Renters
Frequency
0
1000
2000
3000
4000
0
20
40
60
80
100
Percentile Moderate-income family Income Within Tract


Frequency vs Percentile
All
Owners
Renters
Frequency
0
1000
2000
3000
4000
5000
6000
0
20
40
60
80
100
Percentile Dual-professional family Income Within Tract