REVIEW OF LOCATION AFFORDABILITY INDEX HOUSING COST MODEL

FINAL REPORT – MAY 28, 2013

FINAL REPORT SUBMITTED BY:

Econsult Solutions, Inc.
1435 Walnut Street, Suite 300
Philadelphia, PA 19102

Penn Institute for Urban Research
3733 Spruce St., Vance Hall 430
Philadelphia PA 19104

PENN INSTITUTE *for* URBAN RESEARCH

ECONSULT SOLUTIONS
economics | policy | strategy

SUMMARY

The Location Affordability Index (LAI) is designed to provide estimates of the cost-of-living at the local level across the United States. The broad goal of the index is to estimate how expensive it is to live in particular neighborhoods once transportation costs are taken into account. Housing and transportation costs are the two inputs for this index, as both expenditures comprise a significant share of the average household's budget. Further, the index seeks to control for the fact that median housing costs in a neighborhood do not necessarily reflect the costs that a particular household could expect to pay there.

To accomplish this, The Center for Neighborhood Technology (CNT) has assembled a wide variety of datasets that describe neighborhood characteristics including a large proprietary dataset of transportation infrastructure. The data are then used to estimate statistical models of transportation and housing costs that allow customizable estimates of these costs for neighborhoods across the United States.

This report will review the housing component of the LAI and make suggestions for improvements to its data, conceptual framework, and statistical methods. While the index provides a reasonable assessment of housing and transportation affordability, there are significant caveats that should be addressed:

1. The data used for housing are not current and the estimates produced for housing are not quality adjusted. Each of these concerns should be prominently disclosed to users of the LAI so there is no misperception about these important shortcomings.
2. The housing component of the LAI is not available separate from transportation costs on the LAI map. Given the measurement issues with the housing component and the relative uniqueness of the transportation cost measure, these costs should be made available separately.
3. More work needs to be done to demonstrate the value of the twelve household types chosen for the model[1]. There is no aggregate nor summary evidence presented that the model's cost predictions vary significantly by these household types. Illustrative examples could also demonstrate this usefully.
4. There are a variety of econometric concerns that should be explored including geographically non-random errors.
5. For rental housing costs, it appears unambiguous that residuals should be added back in to the estimated housing costs; for owner-occupied housing costs, the balance of the considerations suggest this as well, although there is more ambiguity.
6. A rigorous definition and defense of the goals of the index should be created so that users better understand its purpose, and the extent to which it achieves these goals can be more accurately assessed.
7. The data, model results, and code, including all transformations, should be made available on the website to increase transparency and replicability.

[1] The twelve chosen household types on the Location Affordability Index are: Regional Typical (default), Regional Moderate, Core Typical, Single-Income Family, Dual-Income Family, Low Income, Moderate Income, Single Person Very Low Income, One-Worker Family Very Low Income, Single Professional, Single Worker, and Retirees.

In addition to these suggestions that might improve the basic LAI model, it is useful to consider how the LAI compares to completely different models that could be constructed with other data. This includes the possibility of constructing a user cost measure for owner-occupied housing. While these improvements may be beyond the scope and budget, they should nevertheless be recognized as potentially superior alternatives should sufficient resources be provided.

1.0 PURPOSE AND BASIC TOP-LEVEL CONSTRUCTION

The goal of the LAI is to measure how expensive neighborhoods are for different household types. Neighborhood affordability is calculated as the sum of transportation costs and housing costs for different household types in a given location. While the main goal of the index is to present information on transportation and housing costs by neighborhood, the challenge the index seeks to overcome is that average costs in a neighborhood will be affected by the demographics of the households in that neighborhood. If, for example, a single individual wishes to live on a block that is 70% families of five, then the average housing and transportation costs are not the appropriate measure for him or her. The LAI attempts to control for differences in neighborhood housing and transportation costs that result from neighborhood demographics, and "to focus on the built environment". The result is the ability to better predict how much it would cost to live in a neighborhood for a particular household type. They show these results in two ways. One is by producing an index value for every block group for twelve different kinds of households that were selected in consultation with HUD and other stakeholders. This allows individuals with common household types to quickly compare neighborhoods. The other is with the "My Transportation Cost Calculator," which allows customizable estimates of neighborhood costs based on individual demographics.

The intended audience of the LAI includes a variety of users. This includes households and realtors trying to understand the expected housing and transportation costs in particular neighborhoods. In addition, the index is intended to aid planners, policymakers, and developers in making data-driven decisions for planning and investments. A third possible user group, listed by CNT, includes researchers interested in assessing neighborhood affordability.

The LAI begins with a variety of datasets describing the approximately 210,000 U.S. Census block groups. The two outcomes being measured are transportation and housing costs. Transportation is measured using auto-ownership, auto use, and transit use. Housing is measured using selected monthly owner costs (SMOC) and gross rent (GR) from the U.S. Census American Community Survey (ACS).

Regression models are then utilized to estimate these dependent variables as functions of independent variables. The transportation cost models uses fourteen independent variables that fall under five groups: household characteristics, household density, street connectivity and walkability, transit access, and employment access and diversity. The housing model uses the same variables with the addition of Core Based Statistical Area (CBSA) median SMOC and CBSA median GR.

The regression models allow predictions of what SMOC and GR would be in each block group for a particular type of household, defined by income, household size, and commuters per household. For each

household type, a predicted SMOC and GR for each block group are then combined using the proportion of the block group that owns versus rents for total average housing costs. These housing costs are combined with similar estimates for transportation cost for an overall LAI index number.

1.1 "My Transportation Cost" Calculator

In addition to LAI numbers for each block group and household type, users have the option of using the "My Transportation Cost" calculator. This customizable interface uses the results of the LAI modeling and data exercise to predict housing and transportation costs, enabling users to select particular values for themselves, including household income, number of vehicles, average miles driven, and even housing costs.

2.0 HOUSING MODEL OVERVIEW

2.1 Summary

This section will provide a more detailed overview of how the housing portion of the LAI is calculated. The independent and dependent variables will be discussed, and details will be provided for the regression model and post-regression estimation procedures.

The housing model is used to create conditional expectations of housing costs at the block group level. The conceptual basis for conditional expectations is that housing costs are driven by two basic kinds of factors: household characteristics and the built environment. To predict what a particular household would spend in a particular neighborhood, one should focus on the built environment rather than costs that are driven by household characteristics. The housing regression model seeks to explain housing costs using variables describing household characteristics and the built environment. The regression model creates predicted values that correspond to particular levels of household characteristics. A censoring procedure is then employed to ensure that the predicted housing costs are bound by the actual supply of housing in a neighborhood. Finally, the predicted and censored gross rent (GR) and selected monthly owner costs (SMOC) are averaged by percent of households in a block that are homeowners versus renters, to create an overall average housing cost.

2.2 Dependent Variables Measuring Housing Cost

The dependent variable in the housing cost model includes housing costs measured for two groups: homeowners and renters. SMOC seeks to capture a range of housing expenses for homeowners, and includes condo fees, mortgage payments, real estate taxes, insurances costs, utilities, and other costs for mobile homes. For renters, housing costs are captured using median GR. Both variables are measured as medians at the block group level.

2.3 Independent Variables

The neighborhood independent variables are intended to capture two broad determinants of neighborhood housing costs: household characteristics and built environment characteristics. The following table lists the independent variables, which are summarized briefly below.

Table 1: Neighborhood Independent Variables

Household Characteristics
Median income
Per capita income
Average household size
Commuters per household
Household Density
Residential density
Gross density
Street Connectivity and Walkability
Block density
Intersection density
Transit Access
Transit connectivity index
Transit access shed
Transit access shed frequency of service
Employment Access and Diversity
Employment access index
Job diversity index
Average median commute distance
City-wide variables
CBSA median SMOC
CBSA median GR

Household Characteristics

Measure of household characteristics come directly or indirectly from the American Community Survey's (ACS) five-year estimates. They include median household income, per capita household income, average household size, and commuters per household. Median household income and average household size come directly from the ACS. Per capita income is estimated using median household income divided by

average household size. Commuters per household is estimated from the number of workers over age 16 who do not work at home from the ACS and the total occupied housing units.[2]

Household Density

Four measures of density are used: two related to household density, and two related to street connectivity and walkability. Gross density is simply the number of households divided by the total land area in the block group. Residential density is a slightly more refined measure of population density that only includes sub-areas within the block group (e.g., the blocks) that are residential. It therefore captures population density in residential areas while subtracting from the areas within block groups that have non-residential land uses (i.e. recreational parks).

Street Connectivity and Walkability

The two measures of street connectivity and walkability capture a different kind of density that reflects the pedestrian-friendliness of a neighborhood. Block density uses street maps to define physical blocks (as opposed to Census Block areas), and measures the number of blocks divided by the number of land acres. In other words, it measures the blocks per acre. Intersection density, the second measure of connectivity and walkability, counts the number of street intersections per block group divided by the total land area.

Transit Access

Another important component of the built environment is transit access. These measures are constructed using the Center for Neighborhood Technology's (CNT) proprietary public transportation dataset. This dataset covers all major public transit agencies in areas with populations over 250,000[3], and contains information on stations, stops, and frequencies for all bus, rail, and ferry services. This data is used to construct three measures of transit access: the transit connectivity index (TCI), transit access shed (TAS), and transit access shed frequency of service (TASFOS).

The TCI is a spatially weighted average of frequency of service for transit stops within 1.5 miles of the block group. The measure begins by drawing twelve concentric 1/8th mile circles around each transit stop. Then a weighted average of service frequency within the block group is estimated using the following: the frequency of service at the transit stops whose 1/8th mile circles overlap with the block group, the percent of the block group's total land area encompassed by the concentric circles, and regression based weights for each of the twelve concentric circles. The end result is the TCI, which is a single weighted measure of the frequency of transit service for a block group.

The TAS measures the total accessible area within thirty minutes of public transit from each block group. This is estimated by making assumptions about travel time and walking time to determine which stops are within thirty minutes of a block group. Then the total area within a quarter mile of each of those stops is

[2] An adjustment is made for the percent of the population who are in group housing, and are therefore not in occupied housing units.
[3] There are three exceptions: Dayton, OH; Roanoke, VA; and York-Hanover, PA.

summed. This measure captures how far one can quickly go using public transit, and the accessibility of nearby areas from a given block group.

The TASFOS measure is estimated by summing the average rides per week at all of the same public transit stops within 30 minutes of a block group. This measure is similar to the TAS but captures an important dimension of the convenience of nearby accessible areas.

Employment Access and Diversity

An important component of the value of a neighborhood is the availability of nearby employment opportunities. The ability to find a job closer to home (or a home closer to their job, if the latter is chosen first) is a valuable amenity as it allows for shorter commute distance. Therefore, the number and variety of nearby jobs is something that can increase the value and cost of housing. Three measures are used here: the Employment Access Index (EAI), the Job Diversity Index, and median commute distance. Importantly these measures not only capture nearby job availability, but also serve as a proxy for nearby economic activity, which may also increase the value of living in a given neighborhood.

The EAI is a gravity model that attempts to measure the availability of jobs near a given neighborhood. This measure uses the Longitudinal Employer-Household Dynamics Origin Destination Employment Statistics (LODES) dataset from the U.S. Census, which provides a block group level estimate of total employment. This index for a particular block group is calculated as:

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

Where E is the index value, i indexes all n block groups in the U.S., p_i is the employment in block group i, and r_i is the distance between the given block group and block group i in miles.

While the EAI measures the total availability of nearby employment, it does not account for differences in employment opportunities. The Job Diversity Index is an attempt to measure this, and to serve as a "proxy for the mix of economic activity". This measure uses the same basic data as the EAI but first groups total employment by job categories. The LODES data provides twenty detailed job types, and for each of these an EAI is calculated for each block group. Each job type EAI is used separately as the independent variable in a regression on block group level autos per household from the ACS. The statistical significance and coefficients of each job type EAI is used to group them into seven bigger categories. Table 2 below from CNT shows the twenty detailed job types and the seven bigger job categories they correspond to:

Table 2: Major and Detailed Job Categories

Industry Code From LODES	NAIC two Digit Sectors and Descriptions
12	54 (Professional, Scientific, and Technical Services)
13, 19	55 (Management of Companies and Enterprises), 53 (Real Estate and Rental and Leasing), 81 (Other Services [except Public Administration])
9, 20	51 (Information), 92 (Public Administration)
1, 2, 3, 11, 15, 16, 17, 18	11 (Agriculture, Forestry, Fishing and Hunting), 21 (Mining, Quarrying, and Oil and Gas Extraction), 22 (Utilities), 53 (Real Estate and Rental and Leasing), 61 (Educational Services), 62 (Health Care and Social Assistance), 71 (Arts, Entertainment, and Recreation), 72 (Accommodation and Food Services)
5, 8, 10	31-33 (Manufacturing), 48-49 (Transportation and Warehousing), 52 (Finance and Insurance)
7	44-45 (Retail Trade)
4, 6, 14	23 (Construction), 42 (Wholesale Trade), 56 (Administrative and Support and Waste Management and Remediation Services)

Source: CNT

Next, EAIs are estimated for each of these seven groups. The final step is using the seven EAIs into a single Herfindahl-Hirschman index measure of diversity.

The final measure of employment access is average median commute distance. This also uses the LODES data, which not only reports where people work, but where those people live. This measure calculates the median straight-line distance to work for each census block. The average of these distances is then estimated at the census block group level.

It is unclear why the coefficients in a regression with autos per household are the ideal way to group jobs into larger categories. A more detailed explanation for the theory underlying this methodology would be useful to assess whether some alternative might be preferable.

City-wide Variables

Two city-wide variables are used to control for regional differences in housing costs: regional median SMOC and regional median GR.

2.4 Regression Framework

There are a total of four regressions used in the housing cost model. Two are needed so that a separate model can be run for GR and SMOC. In addition, the transit access variables are not available for all geographic areas[4], therefore two sets of regressions are run for each dependent variable: one with all the data and the smaller set of independent variables that excludes transit, and one with less observations but the full set of independent variables that includes transit.

The goal of the regression is to estimate $SMOC_i^j$ and GR_i^j, housing cost in block group i, conditional on being household type j. To do this the following models are estimated for SMOCi and GRi, the selected monthly owner costs and gross rents in block group i:

$$SMOC_i = f(X_i, W_i) + \epsilon_i$$
$$GR_i = g(X_i, W_i) + \varepsilon_i$$

This models housing costs for renters and owners as functions of X_i and W_i, where X_i is a vector containing the four household characteristic variables, and W_i is a vector containing the rest of the twelve independent variables. The regression results in estimates of the functions f() and g() that can be used to generate predicted values of $SMOC_i$ and GR_i. Importantly, these coefficients allow predicted values for what housing costs would be if household characteristics were a particular value. In other words, predictions of $SMOC_i$ and GR_i conditional on X_i equal to particular values x_i. The twelve household types the LAI index is produced for represent twelve different values of x_i.

One challenge in regression analysis is to determine the most appropriate functional form of f() and g(). The CNT chose a flexible second order functional form that allows for the possibility of interacting all independent variables. This model is as follows:

$$\boldsymbol{H(Z)} = \boldsymbol{a_0} + \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \left(\alpha_n \times \boldsymbol{f(z_n)} + \sum_{\boldsymbol{m=n}}^{\boldsymbol{N}} \boldsymbol{\beta_{nm}} \times \boldsymbol{f(z_n)} \times \boldsymbol{f(z_m)} \right) + \boldsymbol{\delta}$$

Where Z is defined as the matrix containing all of the vectors X_i and W_i for all i, and z_n is the nth of the N independent variables, where N is 13 for the non-transit regressions and 16 for the full regressions.

$$\boldsymbol{Z} = \begin{pmatrix} X_1 & W_1 \\ \vdots & \vdots \\ X_i & W_i \end{pmatrix}$$

In addition, due to non-normality of the independent variables, and/or the non-linearity of the relationships between independent variables, several possible transformations were explored. Table 3 below shows the

[4] For the two housing variables, around 60% of block groups have the requisite transit data available.

transformations investigated, and the frequency with which they were used across the four regressions. The exponential transformation was the most common. The CNT reports that the transformations were chosen via an iterative process to maximize r-squared and statistical significance. The dependent variables were transformed by dividing by mean values.

Table 3: Functional Forms of LAI by Time Used

Function	Percent of Time Used	Functional Form Calculation
Exponential	29%	$\exp(x/ x_{max})/\exp(x_{avg} / x_{max})$
Natural Log	21%	$\ln(x+1- x_{min})/\ln(x_{avg} +1- x_{min})$
Inverse	14%	$(x_{avg} +1- x_{min})/(x+1- x_{min})$
Square Root	10%	$\text{sqrt}(x- x_{min})/\text{sqrt}(x_{avg} - x_{min})$
Inverse Square Root	10%	$(\text{sqrt}(x_{avg} - x_{min})+1)/(\text{sqrt}(x- x_{min})+1)$
Inverse Natural Log	9%	$(\log(x_{avg} +1- x_{min})+1)/(\log(x+1- x_{min})+1)$
Linear	7%	x/x_{avg}
Inverse Exponential	0%	$(\exp(x_{avg} / x_{max})+1-\exp(x_{min} / x_{max}))/(\exp(x/ x_{max})+1-\exp(x_{min} / x_{max}))$

Variables are selected by running the full model with all interactions and then removing the insignificant variables until all included coefficients have p-values less than or equal to .05.

2.5 Top and Bottom Censoring to Control for Available Housing Stock

One possible shortcoming with using predictions from a regression model to estimate conditional housing costs for a neighborhood is that the supply of housing is fixed in the short-run and predictions may fall outside the available universe of housing in that neighborhood. For example, if one of the twelve household types had an income of $10 million a year, it is likely that the predicted housing costs would be extremely high. While model predictions would likely represent a reasonable estimate of the housing costs for households with $10 million in income, it is also the case that in most neighborhoods houses of the predicted price and quality would not actually exist, and that all of the existing homes would be far less expensive. Likewise if one of the twelve household types had zero income this would generate predicted house values outside of what is available in many neighborhoods.

While these are extreme examples, on the margin it is possibly for more common income levels to generate predictions outside of what is available in some neighborhoods. Therefore, a censoring is performed using the 10^{th} and 90^{th} percentile of SMOC and GR for each block group, which is data available from the ACS. If predicted values are below the 10^{th} percentile then the 10^{th} percentile is used, if they are above the 90^{th} percentile, then that value is used.

This is an important improvement over a simple regression method using only predicted values. Given the choice of ACS data and the overall methodology, this is a reasonable approach to controlling for the

housing supply constraint in a neighborhood. There does not appear to be any obvious improvement that could be recommended for this methodology given the overall data and modeling choices.

2.6 Overall Housing Cost Index

The final step, once block group level predicted SMOC and GR costs are estimated, is to create a weighted average for overall housing costs in a block. The weighting is done using the percent of the block group that is owner-occupied housing versus the percent that is renter-occupied.

3.0 COMPARISON TO H+T INDEX

3.1 Data Similarities/Differences

The Housing and Transportation Affordability (H+T) and Location Affordability (LAI) indices use identical dependent variable measures of housing costs. In addition, they compute a weighted average of rental and owner-occupied housing costs using the same weighting technique. The H+T index did not use a regression framework, and so there are no independent variables to compare. However, the independent variables for the LAI can be compared to the independent variables for the H+T auto estimates.

For the most part, these variables are identical. However, there are some small differences that represent improvements in the LAI. For example, the Transit Connectivity Index uses a more refined measure with 1/8th mile concentric circles rather than ½ mile and 1/4th mile previously used in the H + T. In addition, average block size has been replaced with block density in the LAI model, which is a small improvement that nevertheless provides more accurate data.

There are also three additional independent variables used that were not included in the H+T models: transit access shed frequency of service, job diversity, and average median commute distance. There is good reason to include all three variables and they capture aspects of the neighborhood not fully captured by the previously used eleven variables. The inclusion of these in the LAI is thus an improvement over the H+T methodology.

3.2 Regression vs. Non-Regression

The most fundamental difference between the methods involves the use of a regression framework. This was a change suggested in previous critiques of the H+T method, and is a substantial improvement. A key fact in understanding housing costs is that the median may not be relevant to particular households. While there are improvements that can and should be made to the model used - and these will be discussed later in the report - the regression-based approach taken by the Center for Neighborhood Technology (CNT) in the LAI is an important improvement.

3.3 Linear Regression Improvement

One important improvement to the regression model used for LAI and the modeling approach used for the H+T index is the use of a linear model rather than a rational functional form. Following recommendations from the previous Econsult/Penn IUR report, CNT adopted a linear regression with interaction terms and data transformations to account for non-linearities. While there are potential improvements still to be made here, this represents a significant improvement that allows for a simpler yet flexible model with easier to interpret coefficients. The ability to use OLS also means the model contains the desirable statistical properties of OLS such as unbiasedness and efficiency.

3.4 Allowance of Multiple Household Types

A final improvement over the broad H+T methodology is that the LAI allows for a variety of household types rather than just using Core Based Statistical Area (CBSA) median household characteristics. The addition of the “My Transportation Cost” calculator adds further flexibility and customization. These improvements are along the lines previously suggested by Econsult/Penn IUR and should make the index more useful to all types of users.

4.0 DATA ISSUES

4.1 Use of Old Housing Cost Data

The most immediately apparent weakness with the Location Affordability Index (LAI) is that it is based on 2006-2010 data. House prices within and between cities can change drastically over a few years, and using data that is, on average, four years old creates a significant risk that the estimates are not accurate.

This significant variability in local housing prices over time can be seen in data from Philadelphia. Table 4 below shows quality-controlled changes in house prices for seventeen neighborhood submarkets from Econsult Solutions. The indices suggest a wide range of changes, with some neighborhoods increasing in price while others drastically decrease. For example, the West/Southwest submarket saw prices fall 37% from 2006 to the most recent quarter, while the Lower North is up 26% over the same period. Even over the relatively shorter time period from Q1 2009 to Q1 2013 there is a large variance in changes, ranging from a decrease of 30% to an increase of 6%. Importantly these submarkets divide the city into 17 areas and are therefore vastly larger than block groups, of which there are over 1,200 used in the LAI index for Philadelphia. This means there will be even greater variance between block groups than what is shown between submarkets.

Table 4: Changes in House Prices in Philadelphia Submarkets

Submarket	Q1 2006 to Q1 2013	Q1 2009 to Q1 2013
Central	5%	2%
Central Northeast	-21%	-11%
Combined South	10%	3%
Lower Far Northeast	-10%	-10%
Lower North	26%	6%
Lower Northeast	-29%	-22%
Lower Northwest	-6%	-10%
Lower Southwest	2%	-6%
North	-24%	-24%
North Delaware	-34%	-22%
RiverWards	-7%	-9%
University	-3%	-18%
Upper Far Northeast	-15%	-10%
Upper North	-5%	-14%
West Park	9%	-10%
Upper Northwest	-12%	-13%
West/Southwest	-37%	-30%

Source: Econsult Solutions (2013)

A delay of four years is a significant lag when it comes to the cost of housing. Thus, this remains an important shortcoming of the LAI.

4.2 Inclusion of Older Mortgages

The lack of timeliness due to using the 2006-2010 data is compounded by yet another important issue with using selected monthly owner costs (SMOC) from the ACS: the housing costs include older mortgages. This measure includes mortgage payments which are a significant driver of housing costs for homeowners. However, by taking the median in a block group, the sampling method includes all mortgages, including some which are very old. Table 5 below from the American Housing Survey shows that in 2011, the median year of origination for homeowners with mortgages is 2006. Nearly 40% are for mortgages prior to 2005, and 18% are prior to 2000.

Table 5: Year of Mortgage Origination (2011)

Year Primary Mortgage Originated	Percent	Cumulative
2010 to 2014	20%	100%
2005 to 2009	41%	80%
2000 to 2004	22%	40%
1995 to 1999	9%	18%
1990 to 1994	4%	10%
1985 to 1989	2%	6%
1980 to 1984	1%	3%
1975 to 1979	1%	2%
1970 to 1974	1%	1%
Median (year)	**2006**	

Source: American Housing Survey (2012)

The age of the mortgage originations is consequential for the LAI. Mortgages reflect both market interest rates and house prices at the time of origination. The median mortgage costs in a neighborhood therefore reflect housing costs across time based on average origination dates. For a person comparing the costs of living in two neighborhoods, the housing costs attainable for people who moved into the neighborhood ten or even two years ago are not relevant since those prices are no longer available. What matters are current prices, and old mortgages do not necessarily reflect that.

The result is that housing costs will be biased in neighborhoods with older mortgages. For neighborhoods where prices have fallen over time this will be an upward bias, in those where they have increased it will be a downward bias.

4.3 Inclusion of Older Rental Leases

The gross rent (GR) measure from the ACS suffers a similar problem as the SMOC. Rather than mortgage terms set in the past, the gross rent measure has rents set in the past, both due to the ACS lag and to long-term leases. While the majority of leases tend to be for a year or less, there is still nominal rigidity in rents even when new leases are signed. In addition, landlords often offer continuing tenants discounts, which further downward biases average rents relative to market rents (Genesove, 2003). Given that rents move relatively slowly compared to house prices, this is likely to be a relatively minor problem issue, and considerably less so than the SMOC measurement issues.

4.4 Use of ACS Block Group Data

The use of block group level ACS data means estimates have a large degree of uncertainty. For every block group in the U.S., the average margin of error for block group level SMOC is 37% of the level. This means that on average, it can be said with 90% certainty that the true cost falls within plus or minus 37% of

the reported value. The cost of this uncertainty should be weighed against the benefits of looking at a precise geography like block group, rather than a larger geography like census tract.

4.5 Non-Separation of Housing and Transportation Costs

While individual block group level estimates are available with housing and transportation costs separated, the data displayed in the maps for the LAI do not separate these estimates. Maps should be available for housing and transportation cost estimates separately. The housing cost component has much more uncertainty than the transportation cost component. In addition, the transportation costs are a unique measure not provided elsewhere, whereas as measures of housing costs are available from other sources. Users should have the option of combining the transportation estimates with alternative measures of housing costs, and examining the transportation costs alone on the map.

5.0 ECONOMETRIC ISSUES

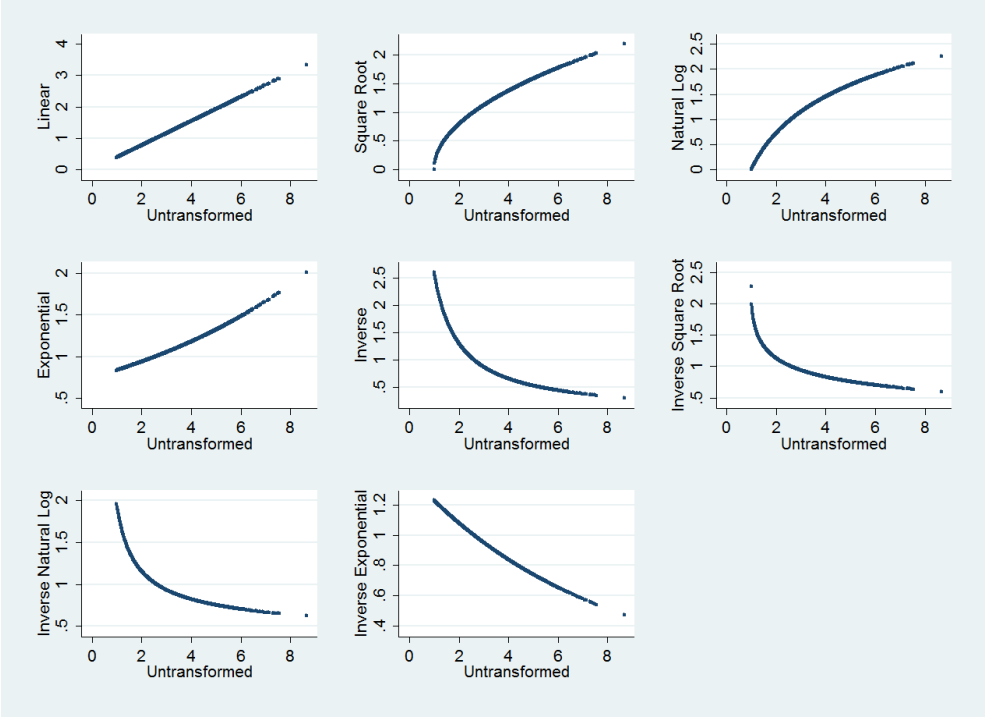
Adopting a regression-based approach to housing cost means that a variety of modeling and econometric decisions must be made. While the approach taken by the Center for Neighborhood Technology (CNT) represents a good start, there are many areas for improvement. An important consideration is improving the model fit, including attaining a high r-squared measure. However, it is also important to not overfit the data or excessively data mine. These concerns are often at odds and should both be considered in model selection.

5.1 Data Transformation

The first issue is the transformation of the independent and dependent variables. Considering seven possible transformations iteratively means there are 7^{16} possible combinations of transformations that could be used. The large number of possible combinations and no systematic selection criteria creates the possibility that the final model is a result of excessive data mining and generates an overfitting of the data. It is important to note that the second order model used already incorporates some non-linearity in it due to the own-interaction term.

Additionally, for certain variables the tested transformations include a significant amount of redundancy given that the transformations will often be nearly indistinguishable. Consider, for example the average household size variable. Figure A below reproduces the scattergrams with the seven transformations on the y-axis, and the untransformed variable on the x-axis. Visual inspection suggests several of the transformations would be collinear. The linear, exponential, and inverse exponential, for example, all appear to be linear products of the untransformed variable.

Figure A: Average Household Size Variable Transformations



A correlation matrix verifies the high degree of correlation between many of the transformations. Excluding the untransformed variable, twenty-two out of the total of twenty-eight correlations are .95 or greater, and ten are 0.99 or greater. To the extent that these transformations can be closely approximated by a linear scaling of the untransformed variable or each other, the extra transformations add marginal value while the large possible number of possible combinations runs the risk of overfitting the data.

Table 6: Average Household Size Variable Transformation Correlations

	None	Linear	Square Root	Natural Log	Exp.	Inverse	Inverse Square Root	Inverse Natural Log	Inverse Exp.
Untransformed	1.00								
Linear	1.00	1.00							
Square Root	0.99	0.99	1.00						
Natural Log	0.98	0.98	1.00	1.00					
Exponential	1.00	1.00	0.97	0.97	1.00				
Inverse	(0.93)	(0.93)	(0.98)	(0.98)	(0.90)	1.00			
Inverse Square Root	(0.93)	(0.93)	(0.98)	(0.98)	(0.91)	1.00	1.00		
Inverse Natural Log	(0.92)	(0.92)	(0.97)	(0.98)	(0.90)	1.00	1.00	1.00	
Inverse Exponential	(1.00)	(1.00)	(0.99)	(0.99)	(0.99)	0.95	0.95	0.95	1.00

An additional issue with such transformations is the risk of creating extreme outliers. While the untransformed measure has an observation that is a large 15.7 standard deviations from the mean, the inverse distribution exacerbates this significantly with the largest observation becoming over 300 standard deviations from the mean. With so many transformation iterations possible, there is a risk that large outliers of transformed independent variables will coincide randomly with residual outliers and thus spuriously generate a high regression fit.

Given the large number of potential transformations, and absent a prior theoretical or empirical reason for a particular transformation, a more standard approach would be to transform each variable so that it is normally distributed by subtracting the mean and dividing by the standard deviation. Alternatively, a simple and flexible measure with an easy interpretation could be used, for instance, the natural logarithm or square root. This runs less risk of data mining and overfitting, as well as exacerbating outliers. While improving the r-squared is an important goal of model selection, it is possible to rely too highly on these criteria and the issues in this section highlight some possible reasons why.

5.2 Functional Form

The regression models utilize a flexible functional form with a large number of interactions. Similar to the issue of the transformations in section 5.1, it is unclear whether these interactions and the selection process are overfitting the data. Given this risk it would be desirable to start with interactions that are most economically defensible and for which theoretical support exists. From this baseline, insignificant interactions can be removed, and reduced numbers of interactions investigated. Focusing on a smaller number of plausible interactions will allow for the coefficients to be explored in more detail to see whether they represent plausible values of marginal changes in housing costs for the corresponding independent variables, which will help mitigate against overfitting the data.

5.3 Geographically Non-Random Errors

An important consideration when using geographically distributed data is the potential non-randomness of errors. The usual desirable properties of ordinary least squares (OLS) regression require assumptions about the randomness of the errors, and geographically distributed errors run the risk that these assumptions will be violated. Consider a simplified representation of the LAI model:

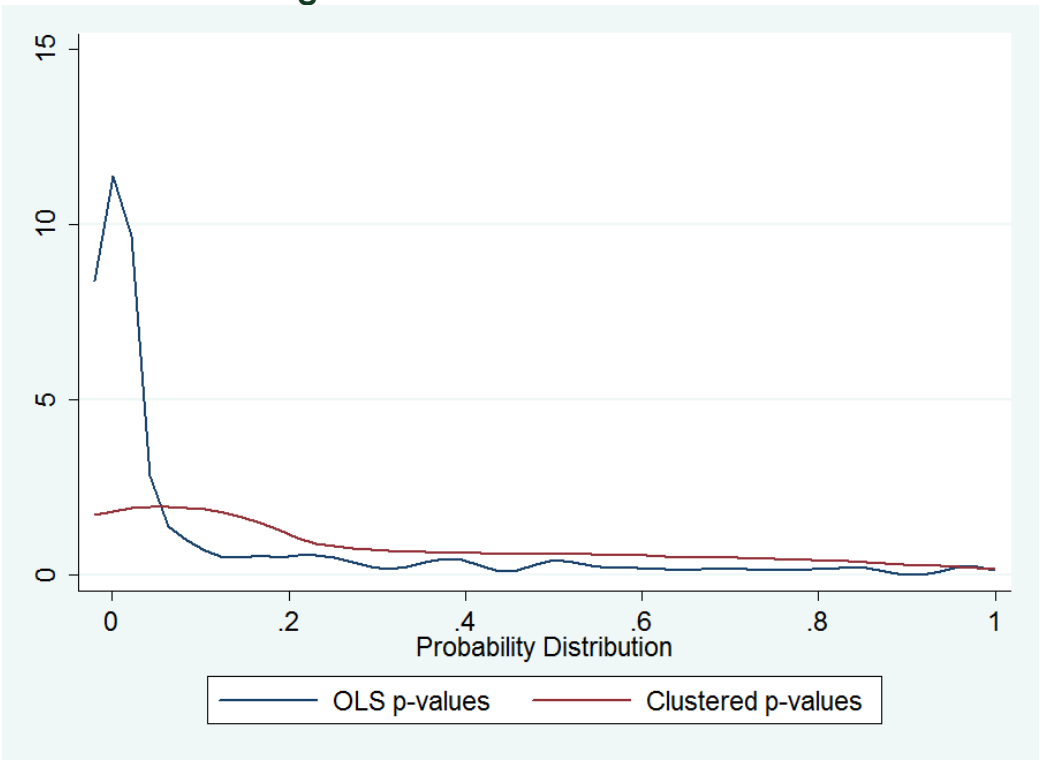
$$Y_{ij} = \beta X_{ij} + c_j + \epsilon_{ij}$$

Where Y_{ij} are the housing costs (either selected monthly owner costs or gross rent) for block group i in CBSA j, $X_{ij}$is a vector of independent variables, $\epsilon_{ij}$is the random error for block group i in CBSA j, and c_j is the unobserved CBSA specific random error for CBSA j. There are several potential problems that can arise in these circumstances. The most benign case is geographic-level heteroskedasticity. In this case, the conditional expectation of c_j given $X_{ij}$is zero so that OLS is consistent and unbiased. However, the model leaves c_j in the error term, so that expected variance varies by group:

$$Var(c_j + \epsilon_{ij}) = \sigma^2 + \gamma_i$$

The result is that the normal standard errors estimated with OLS will be incorrect, and any inference based on these will be inaccurate. This is potentially consequential for the Location Affordability Index (LAI) since model selection is done by iteratively removing variables with insignificant p-values. In order to illustrate the potential importance of clustering, a regression was run with SMOC as the dependent variable and all interaction terms included. Then the same regression was run with a clustered variance estimate based on county level grouping. Figure B below shows the distribution of p-values for the two estimates. Overall the clustering results in significant decrease in the percent of coefficients that are statistically significant.

Figure B: Distribution of P-Values



An even more consequential problem results if the omitted group level error c_j is correlated with the regressors X_{ij}. In this case the coefficients β are biased and inconsistent. A common approach is to use fixed-effects regression. Even if the group level heterogeneity c_j is uncorrelated with the regressors OLS, c_j while unbiased and consistent is still inefficient relative to random effects regression. While the inclusion of CBSA level median SMOC and GR will likely help to control for some CBSA level unmeasured differences, the model should test for the use of fixed-effects or random effects.

As an illustrative example, using state-level fixed effects for the SMOC regression with transit variables increases the adjusted r-squared to 0.6804 from 0.6730. Utilizing county-level fixed effects increased it further to 0.6949. This is the effect on the adjusted r-squared of only including the fixed effects, which would likely be further improved by removing newly statistically insignificant variables. Using county fixed effects and only the sixteen independent variables directly with no interactions the adjusted r-squared is 0.6737. This means that fixed effects can account for more variation than all of the interaction terms combined.

Given that the model won't be used to generate predictions for CBSAs outside the sample, there is little lost by including fixed effects. Overall, given that supply restrictions at the city level can be an important determinant of between city housing costs, it is conceptually more plausible to think of intra-city differences in housing costs, rather than inter-city differences, as determined by the demographic and structural factors measured by the LAI. The inclusion of CBSA level median SMOC and GR suggests the CNT recognizes this, however the use of fixed or random effects is a potentially preferable way to control for this.

A related form of non-random errors is spatial autocorrelation. This occurs if the unexplained variation in block group level housing costs is likely to be correlated across space. Given that local amenities are important determinants in house prices, and the included independent variables cannot hope to capture all local amenities, there is strong prior reason to suspect spatial autocorrelation will exist. Importantly, many amenities occur at smaller geographies than fixed effects, random effects, or the inclusion of city-wide SMOC and GR can control for.

One common way to model spatial autocorrelation is:

$$\epsilon_{ij} = \lambda \sum_{h=1}^{n} W_{ih}\epsilon_{hj} + \eta_j$$

Where λ is the spatial autocorrelation coefficient, and W_{ih} is a spatial weight matrix that gives weights based to errors for other observations based on spatial criteria. Examples of weighting include: a 1 for neighbors and a 0 for all others, inverse distance, inverse distance squared, x nearest neighbors, and others. Tests for spatial autocorrelation like Moran's I can be used to determine whether this is necessary.

The alternative approaches to modeling group level heterogeneity should be explored and it is likely that clustering, fixed effects, random effects, and/or spatial autocorrelation will be desirable. Implementing clustering, random effects, or fixed effects is simple using modern statistical software. Controlling for spatial autocorrelation is more complicated and often requires the use of specific GIS software in addition to the normal statistical software.

5.4 Geographically Varying Coefficients

An implicit assumption in the model is that the marginal contribution of the independent variables to housing cost does not vary throughout the country. Intuitively, it is easy to think of reasons why this would not be case. To take one example, areas where it is very warm or very cold, for instance, the marginal valuation of transportation access may be higher as the disamenity of travel time is higher when the weather is less pleasant.

The implication of such geographic differences in marginal impacts is that the coefficients in the model should be allowed to vary by geography. This can easily be accommodated using interaction variables. For simplicity consider a linear model whereX_{ij} is a vector with k variables and β is a vector of k corresponding coefficient:

$$Y_{ij} = \beta X_{ij} + \epsilon_{ij}$$

Dividing the sample into m = 1,2,…M regions, let D_m be a dummy variable equal to 1 if the block group is in region m, and 0 otherwise. Then the coefficient matrix β_m is the impact of the k coefficients in geography m and can be estimated using the model:

$$Y_{ij} = \sum_{m=1}^{M} D_m \beta_m X_{ij} + \epsilon_{ij}$$

As an illustrative example, allowing the coefficients to vary for the Northeast region of the country and for California increased the adjusted r-squared for the SMOC regression with transit variables from 0.6730 to 0.6917. While this is a relatively small improvements in r-squared this represents only one simple geography measure selected without experimentation. In addition, newly significant variables were not removed. A further refinement of geographically varying coefficients should yield greater improvements in adjusted r-squared.

It is not difficult to implement geographically varying coefficients using modern statistical software packages. For example, in Stata this can be done by defining a categorical variable that indicated the larger geographical area that the block group belongs to, for instance state or region, and utilizing the xi regression option that allows interactions to be easily and intuitively specified. While the number of possible geographies that could be used here is large and would therefore be time consuming to examine exhaustively, testing a few obvious geographies for varying coefficients would not. Utilizing small level geographies like county or CBSA would result in a large number of coefficients and would therefore be more complicated and would cost far more degrees of freedom than using larger geographies.

5.5 What do Residuals Reflect and Should They Be Included or Not?

There is another important econometric issue that is conceptual as well: should unexplained variations in housing costs be included or excluded from the index?

The regression model turns housing costs in each block group into two components: explained costs and unexplained costs. Consider the simplified regression model:

$$Y_i = \hat{f}(X_i, W_i) + \hat{\epsilon}_i$$

Here the $\hat{f}$ and $\hat{\epsilon}_i$ represent the estimations resulting from the regression, X_i represents demographic variables, and W_i represents all of the other variables that describe the built environment. The regression coefficients in $\hat{f}(X_i, W_i)$ represent the part of housing costs that can be explained by the independent variables, while $\hat{\epsilon}_i$ represents the unexplained portion.

The LAI approach is to model block group costs for particular values of X_i corresponding to m different household types. Letting the superscript m index the values of X_i takes to describe the various household types, the LAI generates block group housing costs by replacing X_i with $X_i^m$and creating predicted values:

$$\widehat{Y_i^m} = \hat{f}(X_i^m, W_i)$$

The portion of housing costs that are unexplained are not included in the prediction. However, this raises a conceptual issue of how best to treat the residuals $\hat{\epsilon}_i$. From the perspective of a potential index user, is it relevant information that housing costs in a particular neighborhood are higher than would be expected given the demographics and built environment of that neighborhood?

Partly this depends on the ultimate source of these unexplained costs. If they correspond to unmeasured differences in neighborhood demographics, then whether including them or excluding them is more accurate depends on their effect on housing costs and the level of the variable in a given block group relative to the level for the chosen household type. There is no reason to presume a priori that excluding them is more accurate.

If the residuals represent unmeasured differences in built environment costs then they should be included in the predictions as are the costs attributable to the measured built environment variables in W_i. This could include unmeasured built environment differences like good views, or higher quality homes.

A third possibility is that residuals may represent costs of housing due to positive (or negative) amenities that are not part of the "built environment" in the sense of being part of the neighborhood infrastructure, but nevertheless affect the desirability of living in a particular neighborhood. This could include quality of nearby schools, the variety of retail choices nearby, or crime levels. Because these amenities will be included in market costs of living in the neighborhood, it suggests they should be added into the predictions of neighborhood costs. In addition, if it is possible to empirically measure these amenities, then they should be considered for inclusion in the model as independent variables.

To a certain extent, the desirability of including amenities depends on what the index intends to measure. A fully quality controlled measure of housing costs would compare the costs of a particular level of housing services and would control for amenity differences. For households considering moving into particular neighborhoods the cost of the amenities represents a real out-of-pocket expense, even if that expense is tied to greater amenity values. An optimal index then would both display a constant quality cost of housing services, but also inform users of the cost and level of bundled amenities. Since this index is not attempting that optimal measurement goal, it is difficult to judge how unmeasured amenities should be treated without a rigorously stated goal of the index. Despite uncertainty that arises from the lack of a rigorous definition, the basic stated goal of measuring expensiveness does suggest the residuals should be included.

Alternatively, residuals may represent lower or higher housing costs due to older mortgages with lower housing costs, or neighborhoods with higher or lower average down payments. Ideally in such cases, for the purposes of index users looking to examine the costs of locating in the neighborhood, these residuals should not be included in the housing costs. Importantly, this will not be an issue for GR.

Overall, there are multiple potential sources of unexplained variation in housing costs. It would appear most consistent with the treatment of demographic and built-environment independent variables to include the residuals in the predicted housing costs. In addition, this would be most consistent with the basic stated goal of measuring pure expensiveness. However, absent a theoretically rigorous statement of index purpose it cannot be said with certainty whether excluding or including them is more conceptually correct.

If the residuals are to be included, the procedure for doing this is straightforward. First, the regression model is run so that the coefficients in $\hat{f}()$ are estimated. Next, predicted values $\widehat{Y_i}$are generated using the actual values of the variables in X_i, which are in turn used to generate predicted residuals using the formula:

$$\hat{\epsilon}_i = \widehat{Y_i} - \hat{f}(X_i, W_i)$$

Then m predicted values are generated using the values of X_i^mcorresponding to the m household types:

$$\hat{Y}_i^m = \hat{f}(X_i^m, W_i)$$

Lastly, $\tilde{Y}_i^m$, the final predicted housing costs for household type m in block group i, are generated by adding the unexplained variation in housing costs to the predicted housing costs based on household type:

$$\tilde{Y}_i^m = \hat{Y}_i^m + \hat{\epsilon}_i$$

This approach would be more conceptually relevant to the users of the LAI, as it would provide a more accurate measure of the cost of moving into a neighborhood.

5.6 Household Types

The LAI allows the estimation of costs for twelve household types. Yet, there is no evidence presented for why these twelve types are optimal. In addition, there is no indication given of the extent to which the differences in household types drive the differences in predicted housing costs. It would be useful to see summary statistics illustrating the extent to which predicted housing costs for each household type varied on average. In addition, case studies that show how unadjusted housing costs compare to predicted housing costs for each of the twelve household types in particular block groups would be useful.

Importantly, this result would be much clearer from the model results if variables were all standardized using z-scores and a simple model with no interactions was estimated. In this case each coefficient has the same interpretation that would indicate its importance in the regression: the effect of a one-standard deviation change in the independent variable. Even if a more complex model with interactions was the final result, the extent to which predictions varied using this simple model and the coefficients on z-score transformed independent variables would both be indicative of the impact of household demographics and the twelve household types on estimated housing costs.

6.0 CONCEPTUAL ISSUES

While measurement and statistical issues are an important consideration, the construction of any cost of living index raises a variety of conceptual issues. This section will review two important issues and discuss how they relate to the usefulness of the approach to the potential audiences of the LAI.

6.1 Payments Approach vs. User Cost/Owners' Equivalent Rent

The literature on including housing in cost-of-living indexes focuses on the difficulty of measuring the cost of owner-occupied housing. In large part this is due to the fact that housing is a long-lived asset, and the money spent on housing in a particular period does not necessarily correspond to the value or true cost of the flow of services derived from it. This is in contrast to most goods where spending closely matches consumption in terms of both timing and amount. The money a household spends on apples in a given month, for example, will closely match their total consumption of apples in that month. While many goods contain some asset nature, for instance clothing, out-of-pocket spending remains a good approximation of the true cost unless the asset life is significantly long (Diewart, 2003). Housing represents a unique challenge in this regard.

There are several approaches that can be taken to measuring housing costs. The approach taken in LAI, and used in the CPI from 1950 to 1983, is the "payments approach". Housing cost is measured as the sum of out-of-pocket spending on the various costs of owning a home. This includes mortgage payments, insurance, property taxes, and maintenance. However, this approach remains unpopular among governments and economists for measuring the cost of housing for several reasons.

Importantly, the payments approach creates a measure of housing cost that is sensitive to the portfolio choices of households. For example, if a homeowner decides to pay off mortgage, their mortgage expenses will go up significantly in the period the payment is made, and will then go to zero thereafter. Equivalently, households may differ in the amount of down payment they make which would affect the cost as measured by a payments approach. However, because these are purely financial decisions and do not alter the value of the housing services produced, they do not alter the real cost of living in a given neighborhood, yet the payments approach results in a large change in cost. In addition, the payments approach fails to account for the fact that if house prices go up, then the real cost of owning a home in a given neighborhood has relatively fallen for current homeowners.

Given the shortcomings of the payments approach, the more commonly used measures of housing costs are the rental equivalence approach and the user cost approach. The former uses the price that owner-occupied housing would rent for as the opportunity cost of owning it.[5]The latter estimates the cost of housing by looking at the cost of buying a house in one period and selling it in the next.

[5] Traditionally rental equivalence was argued for on the grounds that in equilibrium the cost of owning a house should equal rent. However, more recent literature has emphasized that rents are an appropriate measure not because of an equilibrium condition, but

However, some of the main weaknesses of the payments approach do not apply to the LAI given its specific purposes, and are not relevant to its likely users. First, while a cost of living index is intended to capture a measure of the costs for everyone, the LAI index is meant to address what the costs would be for someone newly moving into a neighborhood. While portfolio choices over time can lead to volatility in a payments approach for current owners, if measuring costs for marginal owners then this is not a relevant concern.

However, households do make different choices about down payments and other mortgage options, which presents a complication. Given that a cost-of-living index is meant to be an aggregate that measures costs for everyone, it would be conceptually difficult and necessarily incomplete to select a particular set of mortgage parameters and assume they hold for all households counted in such an index. However, the LAI is not a cost-of-living index concerned with all households. Through allowing the selection of different household types, the LAI index is concerned with measuring relative housing costs across geographies while holding certain things fixed, for instance household size and income. Therefore it would be entirely consistent and coherent for the LAI to measure housing payments by holding mortgage terms constant. This could be either by specifying a transparent set of mortgage parameters or by letting users specify them. For users to have the most accurate expectations of costs it would be useful for the index to generate expected mortgage terms for new homebuyers based on current market conditions. This could be done by predicting mortgage terms for households based on demographic information similar to the predictions of housing and transportation costs for the current LAI.

Importantly, however, this is not the current approach taken by the LAI. As discussed in the data limitations section of this report, the LAI index simply assumes that the average mortgage costs in a block group reflect the marginal costs of living there. While a payments approach that measured costs for newly moving households would be relevant and appropriate for those looking to understand relative differences in the local costs, the current payments approach is not.

A final weakness of the payment approach is the absence of possibility for price appreciation to reduce the costs of living in a neighborhood. To accommodate this payments approach could include a neighborhood-based measure of expected appreciation. However, while house prices may be somewhat forecastable in the short-run, if households are examining costs with the assumption of living in the neighborhoods for more than a few years then they there are looking at costs over a horizon where prices are not forecastable.[6] In addition, for this to be consequential forecasted changes would have to differ across neighborhoods, which would be an extremely difficult and highly uncertain task. While a measure of neighborhood level expected appreciation may be a useful if complicated and difficult addition, it remains useful to have a measure of housing costs that makes no assumption *ex ante* about house price appreciation. In addition, while not the preferred measure of economists and price statisticians, for most users excluding expected appreciation will also coincide with their expectations of what such an index would measure.

---

because owners have the option of renting a home and therefore rents are the opportunity cost of occupying an owned house. See Poole, Ptacek, and Verbrugge (2005) and Diewert, Nakamura, and Nakamura (2009).

[6] See: Quigley and Raphael (2004)

A counterpoint would be to argue that rental equivalence has such expected appreciation "baked in" to the estimates and does not entail the difficult task of forecasting. This simplicity is an important motivation behind the use of rental equivalence by many governments.

A common criticism of rental equivalence is that the average quality of rental stock in an area may differ greatly from the average quality of owner-occupied stock. Because the CPI measures changes over time this is not a problem so long as the two stocks experience similar rates of inflation. However, the LAI is concerned with relative price levels, not price changes. These differences are consequential for the LAI and would suggest strongly against adopting a rental equivalence approach that does not measure the cost of owner-occupied housing based on the cost to the household consuming the housing services.

The other option to evaluate housing choices is the user costs of homeownership that addresses the questions of the frictions in the housing markets introduced by transaction costs, and differential tax treatment that limit the rental equivalence method. Where and when user costs are high relative to rental costs, according to this approach, the marginal home seeker is expected to choose to be a renter (Diaz and Luengo-Prado, 2008). A strand of this literature expands on the concept of user cost to include public amenities such as jobs, public school, public safety to point to where a homebuyer (or a renter) gets the most for their money (Fisher et al, 2009). From this perspective households first choose a community and then a quantity of housing in order to maximize their utility over a bundle of goods and services such as public goods, commutes and housing prices (Banzhaf and Farooque, 2012). Under that approach, the price of housing services for homeowners is the current marginal cost of purchasing the unit minus the net present value of the unit (expected sale price net of depreciation, transaction cost and taxes). The variable used to model the user cost of owning are current house price, interest rate, down payment, tax payment, mortgage interest deduction, holding period and expected appreciation rate. An implementation of a full user cost model that appropriately captures expected appreciation and holding period might not be practical, but a basic model based on purchase cost, amortization period and local property tax might be useful to identify monthly housing costs for a household looking to move to a neighborhood and provide a reasonable proxy of the ownership cost burden based on income.

There are also a range of possible improvements that could be made to the existing method. An appropriately measured payments approach for newly moving households would be an important improvement to the LAI that nevertheless left the overall methodology of a payment approach unchanged. A more conceptually rigorous but challenging change would be to adopt a user cost measure. More specific details of alternative methods and improvements will be furthered discussed in section 7.

6.2 Quality Control

An important limitation of the LAI methodology is that the cost of housing is not quality controlled. In a basic cost of living framework the good being compared in two periods or places should be of constant quality. However, two neighborhoods with different housing costs can also reflect vastly different levels of housing services (housing quality, amenities). For households and policymakers a quality-adjusted measure of housing affordability would be an important consideration in order to determine if two locations truly differ from an affordability perspective. For example MIT's Housing Affordability Initiative (HAI) takes into account

locational amenities in its pricing model to measure the ratio of affordable units to the number of units in a town, which represents the town's affordability index (Fisher et al., 2009). This approach explicitly defines the goal of an affordability measure as capturing not only affordable structures but also units that provide access to jobs, safe communities, open spaces and good schools.

The methodology developed to calculate this index takes into consideration the user cost for owner occupied housing based on estimated current house values, interest rates and tax liabilities. Two affordability indexes are measured, an unadjusted one based only on housing costs for homeowners and for renter, an adjusted one that takes into account the level of amenities available in the town. The multiple dimensions that enter into the calculation of this affordability index (housing costs, access to jobs, safety and school quality) are presented in the results. The purpose of these dimensions is to make the index useful to policy-makers as they decide where public investment should be made and what type of investment is needed depending on the place. The primary underlying criteria is the price/income ratio adjusted for job accessibility, quality of school and safety. Thus this measure is aimed at an affordability concept that is user-cost affordability. The strong conceptual basis to the user cost approach allows a robust comparative analysis of housing costs across locations. This method is a conceptually well-identified analysis of the affordability question in a user-costs framework that adjusts for the level of amenities.

7.0 ALTERNATIVE MEASURES OF HOUSING COSTS

7.1 New Measurement Approaches

The appropriate measure of housing cost to be used for the Location Affordability Index (LAI) depends on its intended use. As reviewed in the preceding section, there are three main methods to capture housing costs: the payments approach, the owner's equivalent rent, and the user cost. Each method presents strengths and weaknesses to fulfill three different objectives:

- Indicator of where households spend more than a given share of their income on housing and transportation
- Indicator of where housing and transportation costs combined are cost efficient from an individual perspective
- Indicator of where public investment should be put in place to improve social welfare and encourage transportation efficient development

The more simple aim of the LAI is to provide a measure of the expensiveness of housing (rental or owner occupied) and transportation across neighborhoods. However, the index's attempt to control for demographics and the particular focus on the built environment suggests there other objectives than mere expensiveness. Therefore, to understand which approach is most suitable it would be useful to have a more rigorously stated definition of the index goals. However, regardless of theoretical desirability, the practical ability to implement each of the three main approaches to estimating local housing costs and their resulting strengths and weaknesses can still be discussed.

First is the method currently used in the LAI, which most closely resembles a payments approach. A payments approach requires detailed survey information at the local level that gathers household-level information about mortgage payments, real estate taxes, home insurances, utilities, and condo or homeowner association fees. Despite its limitations, the primary being the timeliness issue mentioned earlier, there appears to be no substitutable dataset to the five-year American Community Survey (ACS) to create such a measure.

As mentioned earlier, it would be desirable to limit the estimation to recent movers, but that would require a special run of the census since such variables are not part of the public data. In any case, it is important to warn the user that the data on housing costs are lagged and therefore do not represent current market conditions. In addition, wherever possible users should be provided with both housing and transportation cost indexes separately so that users can take advantage of the transportation cost information combined with more current measures of housing costs.

The rental equivalent approach appears impractical to implement at such a local scale. The main potential source of data for this approach would be the data collected by the Bureau of Labor Statistics to compute the Consumer Price Index Owners' Equivalent Rent, however the sample is too limited to be able to estimate such a variable at the census block group level.

The user cost approach might be the one approach for which more current alternative data sources could be available to measure the housing costs of homeowners. This could be coupled with gross rent from the either the five-year ACS or from Zillow to obtain more current housing costs at the block group level.

Creating housing cost indicators based on the user cost approach requires information about sale prices and property taxes at the local level on an annual basis. It would address the limited ability of the payments approach to measure affordability constraints for potential residents. Examples of such housing affordability indices have been developed by a number of professional organizations using a limited number of variables and relatively simple calculations. These indices are designed to evaluate affordability from a potential homebuyer's perspective (presently, no such national index for renters or for combined renters and owners seems to exist). Calculators have been developed allowing users to identify the houses for which his or her household can qualify on a mortgage, based on industry standards and a household's characteristics.

Among the established indices are the Housing Affordability Index (HAI) developed by the National Association of Realtors (NAR) that measures if the median-priced home is affordable to a median-income household, and the Housing Opportunity Index (HOI) developed by the National Association of Home Builders (NAHB) that calculates the share of homes in an MSA affordable to a median-income family. These indices are not built to measure the share of current residents who are burdened by their housing payments (be it mortgage payments or rents) but rather if a median household can purchase a median home in that area based on current mortgage underwriting. Developed at the neighborhood level and combined with rent data, a similar approach could be used for the LAI. Another potential source of data is Zillow.com, which provides rental and ownership costs calculated at the property, Zip Code, and city level, and is constantly updated. Zillow estimates are calculated using their extensive database of sale transactions, property assessment information and Multiple Listing Service (MLS) listings, as well as user-

submitted data. Zillow.com is able to estimate property values based on physical attributes (e.g., location, lot size, square footage, number of bedrooms and bathrooms), tax assessments (e.g., property tax information, actual property taxes paid, exceptions to tax assessments) and prior and current transactions (e.g., actual sale prices over time of the home itself and comparable recent sales of nearby homes).

The main limitation of these three approaches is that NAR, NAHB and Zillow are using private data (NAR collects its own and NAHB uses CoreLogic) which might not be an option for public entities. However, for MPOs and other local governments, getting access to the sale records in their individual jurisdictions is likely feasible. Using these data to calculate estimated monthly costs for buyers based on mortgage payment, taxes, down payment opportunity costs, and insurance and for renters would provide up-to-date estimates at not only the household but also the neighborhood and city levels that could be useful both for individuals and local governments. In the long run, HUD should consider establishing a national public repository of house price transactions that could be of great use to researchers, as well as to local organizations.

Another limitation is that these data are not well suited to create a model similar to the LAI in which property and household characteristics are linked to provide housing costs estimates for different household types. However, property level estimates from Zillow are potentially usable as the default setting for households looking at a particular housing unit. Coupled with the transportation cost estimated by CNT for different household type, the use of a measure similar to Zillow estimates could significantly improve the usability of the index by providing default value at the address level, while still providing the possibility for a household to change these parameters to better fit its situation. Alternatively, the data could be used to create estimated housing costs per square foot, which combined with a square footage level specified by the user could generate block group level estimates of expected housing costs. In addition, models estimating block group level square footage as a function of household demographics could be explored as a way to generate predicted expenditures for different household types and users with particular demographic profiles.

In contrast to the housing measure used in the H+T Indicator, the LAI uses regression analyses to develop different measures of housing cost for different household types. This approach likely requires using census data in one way or another. The LAI regression approach to housing costs enables estimated housing costs to be provided not only for a “typical household” defined in a relatively abstract way, but for 12 different household types based on the household size, gross annual income and number of commuter. It might be useful to add the tenure difference to these types. However, if the goal is to provide address-level indicators for households, then calculators available from private-sources provide much more precise estimates of the housing costs a household would face for different houses based on his tenure choice than LAI does. In addition, a more accurate measure of housing affordability could potentially be created using price per square foot estimates from house price transaction data and either allowing users specify square footage or with a model predicting square footage from demographics data.

7.2 Improvements to the Existing Approach

In addition to the modeling and econometric approaches suggested above, it is worth examining whether there are possible ways to improve the index by utilizing additional data sources or otherwise augmenting the model.

One possibility is that a local price index could be used to make the local housing costs more current. For example, the local indexes produced by Zillow could be used. However, this could only adjust for the lag due to ACS data being based on SMOC measured from 2006-2010. In order to correct for the more serious lags due to older mortgages, it would require knowing the distribution of mortgage ages. Given that the housing costs are based on a distribution of mortgage ages that is unknown, it is unclear what baseline year a local price index would use to adjust prices to current levels.

There are other technical challenges to using Zillow or other local house price indexes to update prices. For example, their indexes are at the neighborhood and zip code level, but not block group. In order to construct block group measures it is likely that transactions data would be required, in which case it would be preferable to construct an entirely new measure altogether as specified in section 7.1.

Finally, attempting make the data current runs the risk giving users the false impression that the timeliness problem has been addressed when it would still remain a significant issue. Overall, given the above considerations, it is preferable to clearly disclose that the index measures lagged average costs rather than attempt to update costs using other sources of house prices. Any attempt to update the costs would likely be highly uncertain, fall far short of making the data current, and would risk giving a false impression of accuracy.

8.0 CASE STUDY: PHILADELPHIA

More can be learned about the LAI estimates by looking at a detailed comparison of SMOC and predictions from the model with other housing cost data sources for a particular metro. Philadelphia has recently computed property values for every building in the city, which provides a current market-based assessment of house prices derived from home sale transactions.

Market value of homes should be related to both SMOC and the predicted SMOC using the regression model. Using only residential properties, the 2013 assessed values and assessed values per square foot were aggregated to the block group level using medians. Figure C below shows that predicted owner costs are generally positively related to market values, with a correlation of 0.75. Figure D below that indicates the predicted owner costs are also positively related to market value per square foot, also with a correlation of 0.75.

Figure C: Block Group Level Comparison of Predicted SMOC and Philadelphia Assessment Data

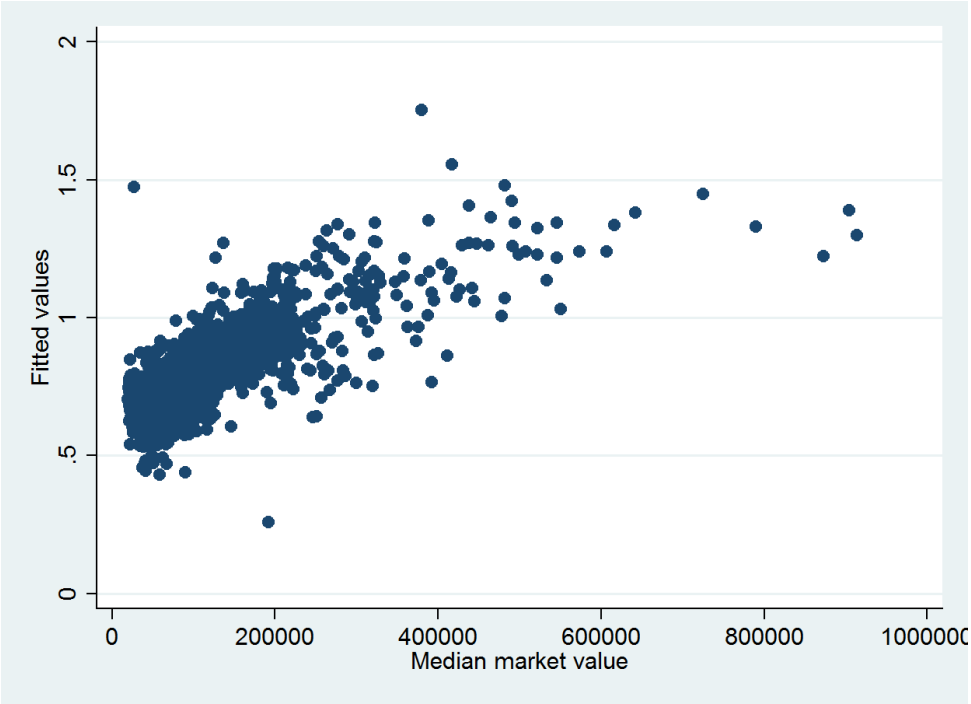
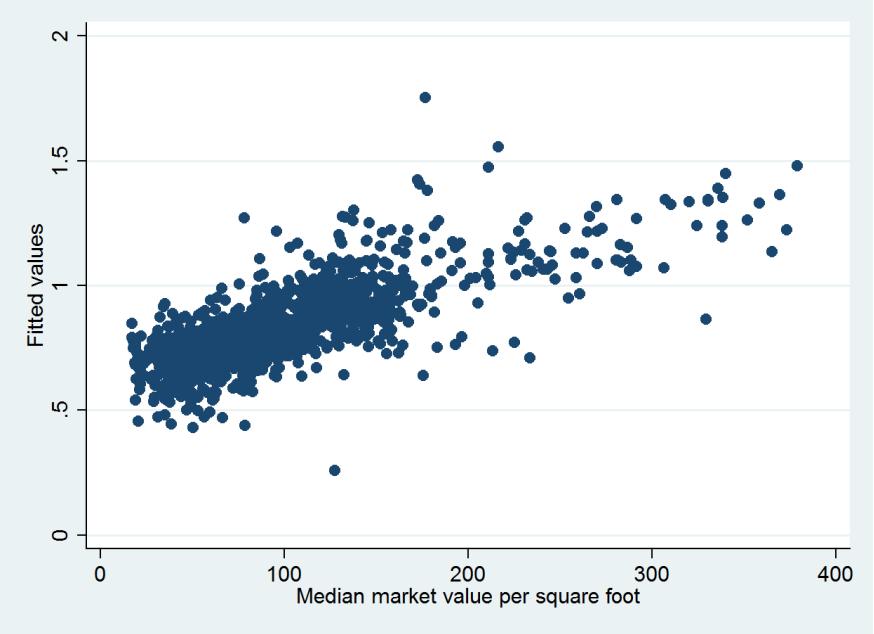
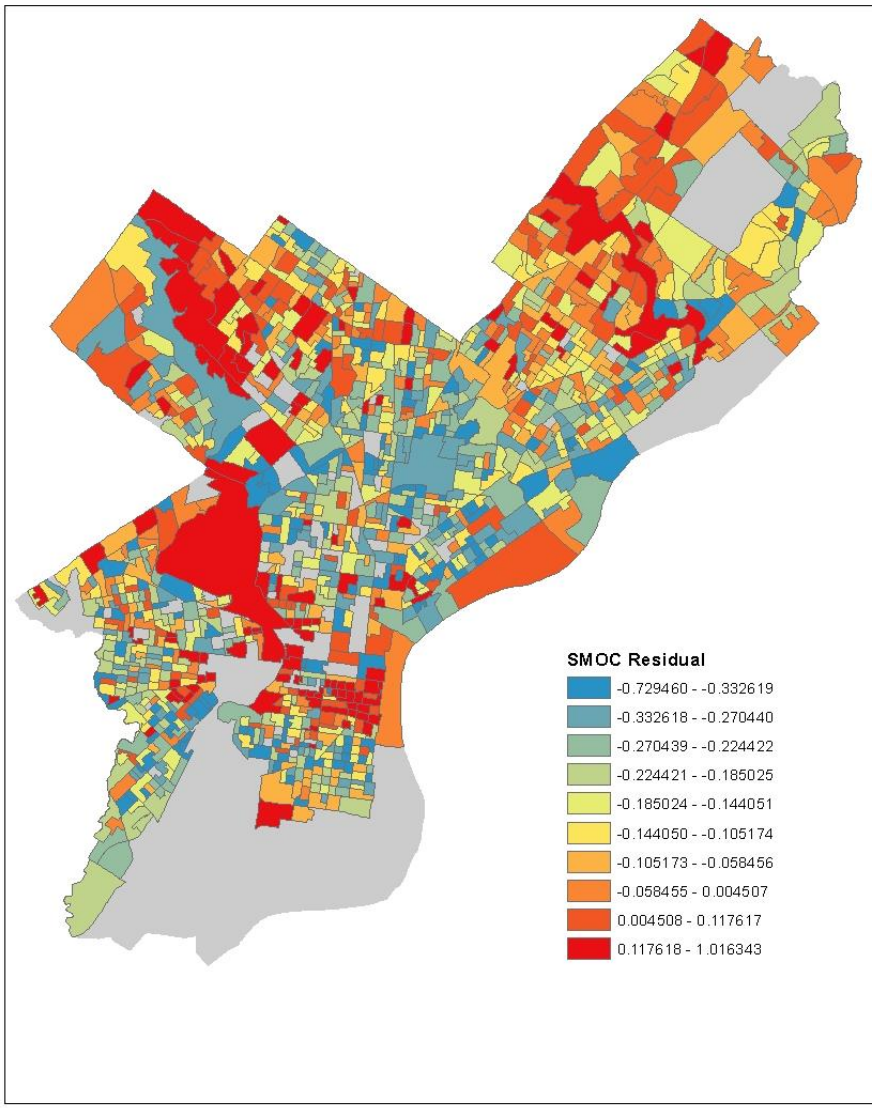


Figure D: Block Group Level Comparison of SMOC and Philadelphia Assessment Data, Price Per Square Foot



In addition, Philadelphia data can be used to examine the geography of residuals. Figure E below presents a map of residuals for the SMOC transit regression. Visual inspection reveals that block groups are clustered near other block groups with similar residual values, which suggests that there is spatial correlation. A Moran's I value of 0.18 with a z-score of 10.71 and p-value of 0.00 provides statistical evidence that spatial correlation exists in the data. This supports the suggestion of section 5.3 that econometric methods to account for geographically non-random errors should be explored.

Figure E: Map of SMOC Residuals for Philadelphia



9.0 DATA MANAGEMENT

Given the proposed use of the index by researchers it is important that the data and modeling process be clear and accurate so that the results can be replicated. Therefore not only the data but the code used to create the models and data transformations should be made available on the LAI website. Resulting model coefficients and statistical results should also be published and easily available. Making statistical program code and data available to researchers common practice among academics and research institutions and should be done for the LAI index.

10.0 RECOMMENDATIONS

The housing cost component of Location Affordability Index (LAI) represents an attempt to measure housing affordability at the block group level. The review presented above does not represent a discrete endorsement of the use of ACS data as the basis for measuring relative housing costs. While the information contained in the index is useful, the extent of the necessary caveats that should accompany the index are significant enough that an endorsement of the ACS data as the source would be too optimistic about the usefulness. In large part the acceptability of the data depends on the user and their purposes. For a family considering where to live in a city it would be of limited use, given the extent to which it fails to capture current costs. Additionally, it would be of limited use to developers and researchers interested in knowing current housing costs for marginal residents.

While a discrete endorsement of the ACS data is inappropriate, given the current data limitations and constraints the overall approach and use of ACS data may be the best option available. Nevertheless, there are several important changes that should be made before the index is presented to users.

1) The data is significantly lagged, which given house price volatility is an important limitation. The fact that these costs are not current, especially for owner-occupied housing, should be prominently disclosed to users up-front.
2) The estimates are not quality adjusted. The availability of housing within a given price range in a neighborhood does not mean that acceptable quality of housing is available, or that affordability is not a problem for that neighborhood. This caveat should also be prominently displayed to users.
3) Users should have the option of displaying predicted housing costs alone on the LAI website mapping engine rather than just the weighted average of the housing and transportation costs.
4) Evidence needs to be given for the desirability of the twelve household types, and the extent to which the choice of these affects housing costs.
5) Several econometric issues merit further exploration, including the use of fixed effects and other issues of geographically non-random errors, a narrower set of transformations, geographically varying coefficients, and theoretically justifiable set of independent variable interactions.

6) A rigorous theoretical definition of what the index seeks to measure should be established and justified to help guide index measurement decisions.
7) Contingent on the answer to the previous recommendation, residuals should potentially be added back to predicted values.
8) The data, model results, and code, including all transformations, should be made available on the website to increase transparency and replicability.

In addition to these improvements in the existing approach, there should be further exploration of different approaches to measuring housing costs. In particular, a user cost approach based on sales transactions data and mortgage payment calculators represents a more conceptually justifiable approach to measuring housing costs. Absent a user cost that fully accounts for expected appreciation, an improved marginal payment approach should be adopted. This approach would also require transactions level home sales data, which further emphasizes that the availability of such data is central to improving upon the LAI approach.

REFERENCES

Banzhaf, Spencer and Omar Farooque. (2012) “Interjurisdictional Housing Prices and Spatial Amenities: Which Measures of Housing Prices Reflect Local Public Goods?” NBER Working Paper 17809, http://www.nber.org/papers/w17809

Diaz, Antonia & Maria Jose Luengo Prado (2008). "On the User Cost and Homeownership," *Review of Economic Dynamics*, 11(3): 584-613.

Diewert, Erwin (2003). “The Treatment of Owner-Occupied Housing and Other Durables in a Consumer Price Index”. Center for Applied Economics Research Working Paper.

Diewert, W. Erwin, Alice O. Nakamura, and Leonard I. Nakamura. "The housing bubble and a new approach to accounting for housing in a CPI." *Journal of Housing Economics* 18.3 (2009): 156-171.

Fisher, Lynn, Henry Pollakowski and Jefferey Zabel. (2009) “Amenity-Based Housing Affordability Indexes,” Real Estate Economics, 34(4), 705-746.

Quigley, John M. and Steven Raphael (2004). “Is Housing Unaffordable? Why Isn’t It More Affordable?”*Journal of Economic Perspectives*, 18(1): 191–214.

Genesove, David. "The nominal rigidity of apartment rents." *Review of Economics and Statistics* 85.4 (2003): 844-853.

Poole, Robert, Frank Ptacek, and Randal Verbrugge. "Treatment of owner-occupied housing in the cpi." *Federal Economic Statistics Advisory Committee (FESAC) on December* 9 (2005): 2005.