Online Appendix:

Quality-adjusted house price indexes

Adam D. Nowak [a,c] Patrick S. Smith [b,c]

[a]West Virginia University, College of Business & Economics; Email: adam.nowak@mail.wvu.com
[b]San Diego State University, Fowler College of Business; Email: patrick.smith@sdsu.edu
[c]Redfin Corporation, Consultant

Appendices

Contents

A Adjustment Theorem Proof (Internet Appendix)

Proof. D is an $N \times T-1$ matrix and R is an $N \times \mathcal{Q}$ matrix. Each row $n = 1, ..., N$ in both D and R corresponds to a repeat-sales pair. Excluding one time period in order to avoid perfect multicollinearity, each column $t = 1, ..., T-1$ in D corresponds to a unique time period. Each column $q = 1, ..., \mathcal{Q}$ in R corresponds to a unique token in $\hat{\mathcal{S}} = \{\hat{\mathcal{S}}_1, ..., \hat{\mathcal{S}}_{\mathcal{Q}}\}$. Because of the one-to-one relationship between the columns in R and the set $\hat{\mathcal{S}}$, the phrase *token* q is understood to mean the token in $\hat{\mathcal{S}}$ associated with column q in R. Similarly, the phrase *time period* t is understood to mean the time period associated with column t in D.

$D_{nt} = 1$ if the second sale in the repeat-sales pair occurs in time period t, $D_{nt} = -1$ if the first sale in the repeat-sales pair occurs in time period t, and $D_{nt} = 0$ if neither sale or both sales in the repeat-sales pair occurs in time period t. Similarly, $R_{nq} = 1$ if only the second sale in the repeat-sales pair contains token q, $R_{nq} = -1$ if only the first sale in the repeat-sales pair contains token q, and $R_{nq} = 0$ if neither sale or both sales in the repeat-sales pair contain token q. Finally, define y as the $N \times 1$ vector of differenced log transaction prices.

By definition, the least-squares price index when not controlling for tokens is given by

$$\hat{\delta} = [D'D]^{-1}D'y$$

The normal equations for the least-squares price index when controlling for tokens, $\hat{\delta}^*$, and the least-squares implicit prices for the $\mathcal{Q}$ tokens, $\hat{\theta}$, are given by

$$\begin{bmatrix} D'D & D'R \\ R'D & R'R \end{bmatrix} \begin{bmatrix} \hat{\delta}^* \\ \hat{\theta} \end{bmatrix} = \begin{bmatrix} D'y \\ R'y \end{bmatrix}$$

The first $T-1$ equations solve

$$D'D\hat{\delta}^* + D'R\hat{\theta} = D'y$$

Rearranging and premultiplying by $[D'D]^{-1}$

$$\begin{aligned}\hat{\delta}^* &= [D'D]^{-1}D'y - [D'D]^{-1}D'R\hat{\theta} \\ &= \hat{\delta} - [D'D]^{-1}D'R\hat{\theta}\end{aligned}$$

The product $[D'D]^{-1}D'R$ can be written as

$$[D'D]^{-1}D'R = [[D'D]^{-1}D'R_{\bullet 1}, [D'D]^{-1}D'R_{\bullet 2}, \ldots, [D'D]^{-1}D'R_{\bullet \mathcal{Q}}] = [\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{\mathcal{Q}}]$$

Where $\hat{\pi}_q = (\hat{\pi}_{q1}, \hat{\pi}_{q2}, \ldots, \hat{\pi}_{qT-1})'$. This implies

$$\hat{\delta}^* = \hat{\delta} - \sum_{q=1}^{\mathcal{Q}} \hat{\pi}_q \hat{\theta}_q$$

□

B Data Overview (Internet Appendix)

B.1 Data filters

The MLS data used to construct the repeat-sales HPIs was provided by Redfin (Redfin, 2017). Prior to constructing the MSA-level and local HPIs we apply several filters to the single-family detached residential transaction data. We list the filters below and provide a detailed overview of the number of transactions that are dropped in each MSA in Table B1. Records are dropped that do not meet the following criteria:

1. zip code and tract are both available
2. sale date $\leq$ 2017
3. \$50,000 $\leq$ sale price $\leq$ \$3,000,000
4. 500 $\leq$ square feet of living area $\leq$ 6,000
5. 1 $\leq$ bedrooms $\leq$ 6
6. 1 $\leq$ bathrooms $\leq$ 6
7. 0 $\leq$ age $\leq$ 200
8. lot size $\leq$ 5 acres
9. 0 $\leq$ time-on-market $\leq$ 730
10. length(remark) $\geq$ 10 characters
11. unique remark
12. num sales in year $\geq$ 1,000
13. num sales in zip code each year $\geq$ 25
14. unique listing id
15. house sold more than once (i.e., repeat-sale)

Table B1: Filtered transaction data by MSA

	MSA								
Filter	ATL	BAL	BOS	LA	MIA	PDX	PHX	SF	DC
none	930,855	276,040	831,670	1,403,686	453,749	366,006	1,179,603	693,640	702,496
zip and tract avail	930,855	276,040	831,670	1,403,686	453,749	366,006	1,179,603	693,640	702,496
sale date $\leq$ 2017	930,787	276,040	831,638	1,403,572	453,733	366,004	1,179,593	693,639	702,496
\$50K $\leq$ price $\leq$ \$3M	864,622	265,700	826,556	1,383,112	440,032	365,766	1,149,497	686,109	671,211
500 $\leq$ sfla $\leq$ 6,000	845,899	262,241	819,135	1,375,271	436,177	364,685	1,145,463	683,926	658,792
1 $\leq$ beds $\leq$ 6	841,465	261,009	816,166	1,372,428	435,245	363,937	1,143,841	682,544	655,096
1 $\leq$ baths $\leq$ 6	837,779	260,447	815,013	957,868	433,461	363,427	1,142,918	682,057	652,793
0 $\leq$ age $\leq$ 200	837,613	259,859	804,553	957,220	433,391	363,371	1,142,753	679,677	651,797
lot size $\leq$ 5 acres	835,760	257,066	795,094	954,463	433,210	357,207	1,142,091	678,435	643,610
0 $\leq$ tom $\leq$ 730	834,618	253,165	794,140	943,921	431,374	355,432	1,141,524	677,360	625,209
remark $\leq$ 10 char	831,568	176,118	785,964	931,638	421,273	344,389	1,131,977	655,682	413,637
unique remark	663,754	173,096	776,717	909,520	409,954	337,164	1,111,025	585,564	404,037
1K+ sales in year	663,651	172,916	776,491	909,518	409,291	336,480	1,110,722	584,980	403,535
25+ sales in zip each year	662,686	172,733	773,195	904,080	408,828	336,197	1,110,145	584,208	403,353
unique listing id	661,198	172,708	770,574	900,938	408,634	335,136	1,109,420	573,307	403,322
repeat-sales	152,268	61,438	303,749	235,850	128,704	125,399	478,580	202,255	172,989

Notes: Table B1 tabulates the number of records that are dropped for each filter across the nine MSAs examined in this study. The final row for each column identifies the number of repeat-sales transactions that are included in the MSA-level HPIs.

B.2 Descriptive statistics by MSA

The following table provides descriptive statistics for select housing characteristics for each of the nine MSAs examined in this study. The descriptive statistics are provided for the repeat-sales transaction data highlighted in Table B1. In addition to the descriptive statistics, we also list the time period of the data used in the construction of the MSA-level HPIs and the counties represented in the MSA-level HPIs. We construct the quality-adjusted and Case-Shiller HPIs using the exact same repeat-sales data to ensure we provide an apples-to-apples comparison.

Some, but not all, of the represented counties overlap with the represented counties used to construct the "official" Case-Shiller HPIs. For example, the three counties represented in our Miami repeat-sales data are identical to the counties represented in the "official" Case-Shiller "Miami-Fort Lauderdale-Pompano Beach, FL" HPI. In contrast, the four counties represented in our Portland repeat-sales data are a subset of the seven counties represented in the "official" Case-Shiller "Portland-Vancouver-Beaverton, OR-WA" HPI. Our HPIs only include four of the seven counties due to data restrictions (i.e., the other three counties are not available in our data set). That said, the four counties included in our data set represent the core of the Portland MSA.

Table B2: Descriptive statistics for repeat-sales by MSA

	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Atlanta (N = 152,268)						
Price (000s)	50.00	131.50	228.82	179.90	274.30	3,000.00
Age	0.00	7.00	23.40	16.00	33.00	199.00
Sfla (000s)	0.51	1.60	2.26	2.09	2.74	6.00
Bedrooms	1.00	3.00	3.67	4.00	4.00	6.00
Bathrooms	1.00	2.00	2.60	2.50	3.00	6.00
Years:	2000-2017					
Counties:	Cobb GA, DeKalb GA, Fulton GA, Gwinnett GA					
Baltimore (N = 61,438)						
Price (000s)	50.00	207.00	347.97	300.00	435.00	3,000.00
Age	0.00	23.00	45.82	47.00	63.00	198.00
Sfla (000s)	0.50	1.20	1.80	1.58	2.15	5.99
Bedrooms	1.00	3.00	3.60	4.00	4.00	6.00
Bathrooms	1.00	2.00	2.35	2.50	3.00	6.00
Years:	2002-2017					
Counties:	Anne Arundel MD, Baltimore County MD, Howard MD, Baltimore City MD					
Boston (N = 303,749)						
Price (000s)	50.00	215.00	374.05	311.00	440.00	3,000.00
Age	0.00	26.00	55.87	51.00	80.00	200.00
Sfla (000s)	0.50	1.30	1.88	1.68	2.25	6.00
Bedrooms	1.00	3.00	3.31	3.00	4.00	6.00
Bathrooms	1.00	1.00	1.90	2.00	2.50	6.00
Years:	2000-2017					
Counties:	Bristol MA, Essex MA, Middlesex MA, Norfolk MA, Plymouth MA, Suffolk MA, Worcester MA					
Los Angeles (N = 235,850)						
Price (000s)	50.00	255.50	611.27	463.00	786.00	3,000.00
Age	0.00	25.00	48.43	52.00	66.00	145.00
Sfla (000s)	0.50	1.27	1.87	1.65	2.26	6.00
Bedrooms	1.00	3.00	3.25	3.00	4.00	6.00
Bathrooms	1.00	2.00	2.26	2.00	3.00	6.00
Years:	2000-2017					
Counties:	Los Angeles CA, Orange CA					

Table B2: Descriptive statistics for repeat-sales by MSA (cont.)

	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Miami (N = 128,704)						
Price (000s)	50.00	179.90	361.07	280.00	420.00	3,000.00
Age	0.00	12.00	27.87	22.00	42.00	151.00
Sfla (000s)	0.50	1.50	2.11	1.92	2.53	6.00
Bedrooms	1.00	3.00	3.39	3.00	4.00	6.00
Bathrooms	1.00	2.00	2.44	2.00	3.00	6.00
Years:	2000-2017					
Counties:	Broward FL, Miami-Dade FL, Palm Beach FL					
Portland (N = 125,399)						
Price (000s)	50.00	215.00	324.37	280.00	384.90	3,000.00
Age	0.00	11.00	38.24	31.00	60.00	165.00
Sfla (000s)	0.50	1.26	1.80	1.62	2.18	5.98
Bedrooms	1.00	3.00	3.29	3.00	4.00	6.00
Bathrooms	1.00	2.00	2.15	2.00	2.50	6.00
Years:	2003-2017					
Counties:	Clackamas OR, Clark WA, Multnomah OR, Washington OR					
Phoenix (N = 478,580)						
Price (000s)	50.00	135.00	237.68	194.00	278.00	3,000.00
Age	0.00	7.00	19.20	14.00	28.00	167.00
Sfla (000s)	0.52	1.47	1.98	1.80	2.29	5.99
Bedrooms	1.00	3.00	3.40	3.00	4.00	6.00
Bathrooms	1.00	2.00	2.29	2.00	2.50	6.00
Years:	2000-2017					
Counties:	Maricopa AZ, Pinal AZ					
San Francisco (N = 202,255)						
Price (000s)	50.00	320.00	594.65	500.00	745.00	3,000.00
Age	0.00	19.00	43.82	44.00	62.00	199.00
Sfla (000s)	0.50	1.22	1.78	1.60	2.13	5.99
Bedrooms	1.00	3.00	3.27	3.00	4.00	6.00
Bathrooms	1.00	2.00	2.11	2.00	2.50	6.00
Years:	2000-2017					
Counties:	Alameda CA, Contra Costa CA, Marin CA, San Francisco CA, San Mateo CA					

Table B2: Descriptive statistics for repeat-sales by MSA (cont.)

	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Washington D.C. (N = 172,989)						
Price (000s)	50.00	310.00	493.86	435.50	605.00	3,000.00
Age	0.00	19.00	38.67	39.00	55.00	200.00
Sfla (000s)	0.50	1.25	2.05	1.79	2.59	6.00
Bedrooms	1.00	3.00	3.92	4.00	4.00	6.00
Bathrooms	1.00	2.00	2.75	2.50	3.50	6.00
Years:	2002-2017					
Counties:	Alexandria VA, Arlington VA, District of Columbia DC, Fairfax County VA, Loudoun VA, Montgomery MD, Prince George's MD, Prince William VA					

Notes: Descriptive statistics are displayed for select housing characteristics by MSA. The descriptive statistics only include the repeat-sales of single-family detached houses that are used to construct the MSA-level HPIs.

B.3 Text preprocessing

Prior to tokenizing the remarks, we perform a minimal amount of preprocessing. The primary goal of the preprocessing procedure is to clean and standardize the remarks. The remarks are preprocessed as follows:

1. Convert to lower case.
2. Replace commas (,), periods (.), ampersands (&), and the word *and* with a space.
3. Replace all special characters with a space.
4. Remove apostrophes.
5. Remove all remaining single letters.
6. Replace all numbers with a space. Numbers can be in either numeric or character form.
7. Remove duplicate empty spaces.
8. Depluralize.
9. Trim empty spaces at the beginning and end of the remark.

In unreported results we find that additional preprocessing, such as stemming the remarks, has a negligible effect on the results we report. See Section D.2 for additional discussion.

C Robustness Checks (Internet Appendix)

C.1 Additional MSAs

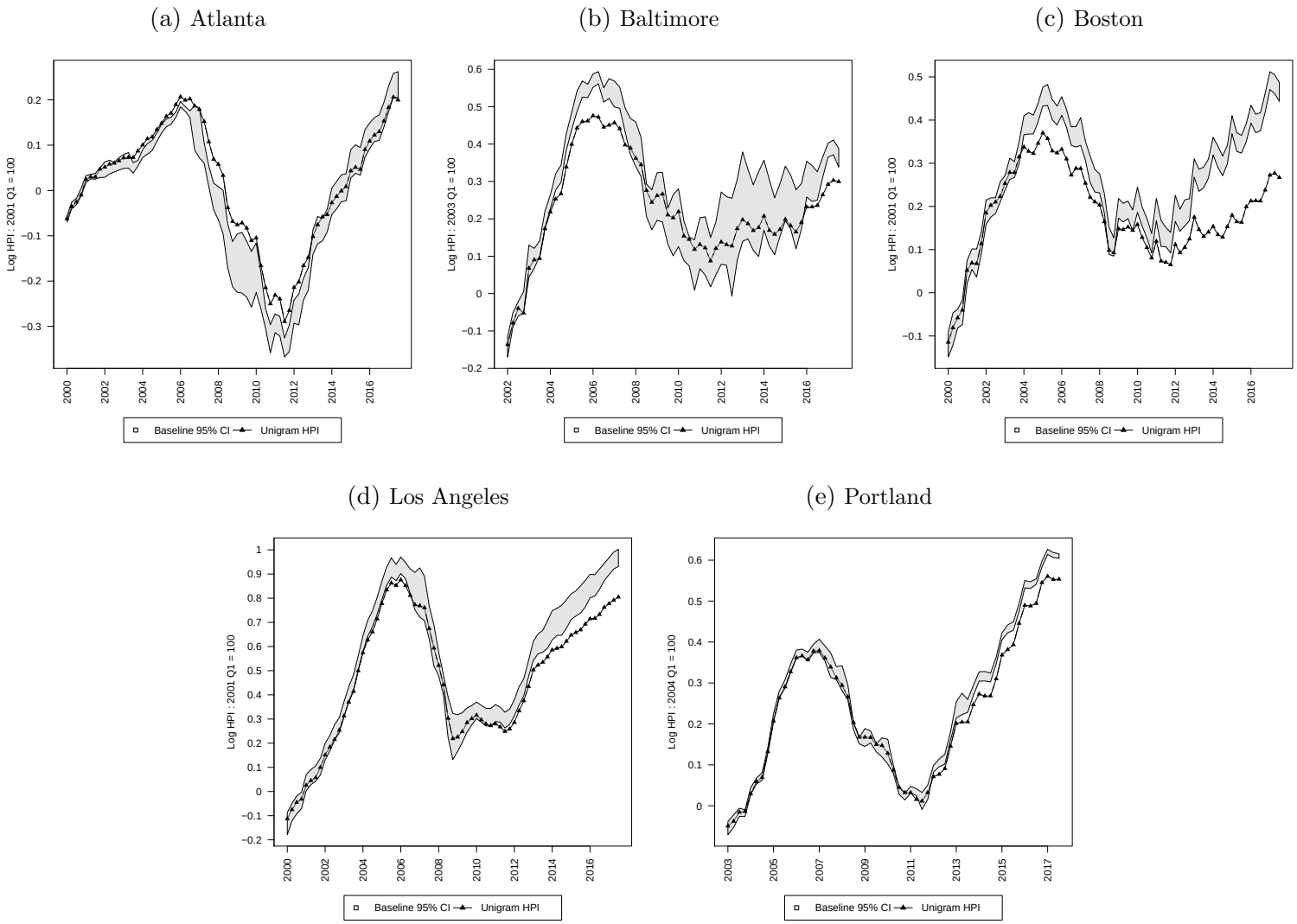
The body of the paper provides quality-adjusted HPIs for four MSAs: Miami, Phoenix, San Francisco, and Washington D.C. In this section, we provide quality-adjusted HPIs for five additional MSAs: Atlanta, Baltimore, Boston, Los Angeles, and Portland. Figure C1 displays the MSA-level Case-Shiller and quality-adjusted HPIs, Figure C2 displays the Duranton and Overman (2005) localization plots, and Figure C3 displays the dispersion of differences for the local HPIs within the five additional MSAs.

Redfin provided MLS data for seven additional MSAs that are not included in this study (Redfin, 2017). The data was provided to examine a different research question in a concurrent study. We do not include the seven MSAs here because some local MLSs do not provide historical data to new brokerage firms. For this reason, the seven additional MSAs do not include the historical data necessary to capture the entire market cycle discussed in this paper. The seven MSAs and the corresponding time period of the data provided by the local MLSs are as follows:

1. Austin, TX (2007 - 2017)
2. Chicago, IL (2005 - 2017)
3. Dallas, TX (2007 - 2017)
4. Denver, CO (2007 - 2017)
5. Houston, TX (2008 - 2017)
6. San Diego, CA (2007 - 2017)
7. Seattle, WA (2007 - 2017)

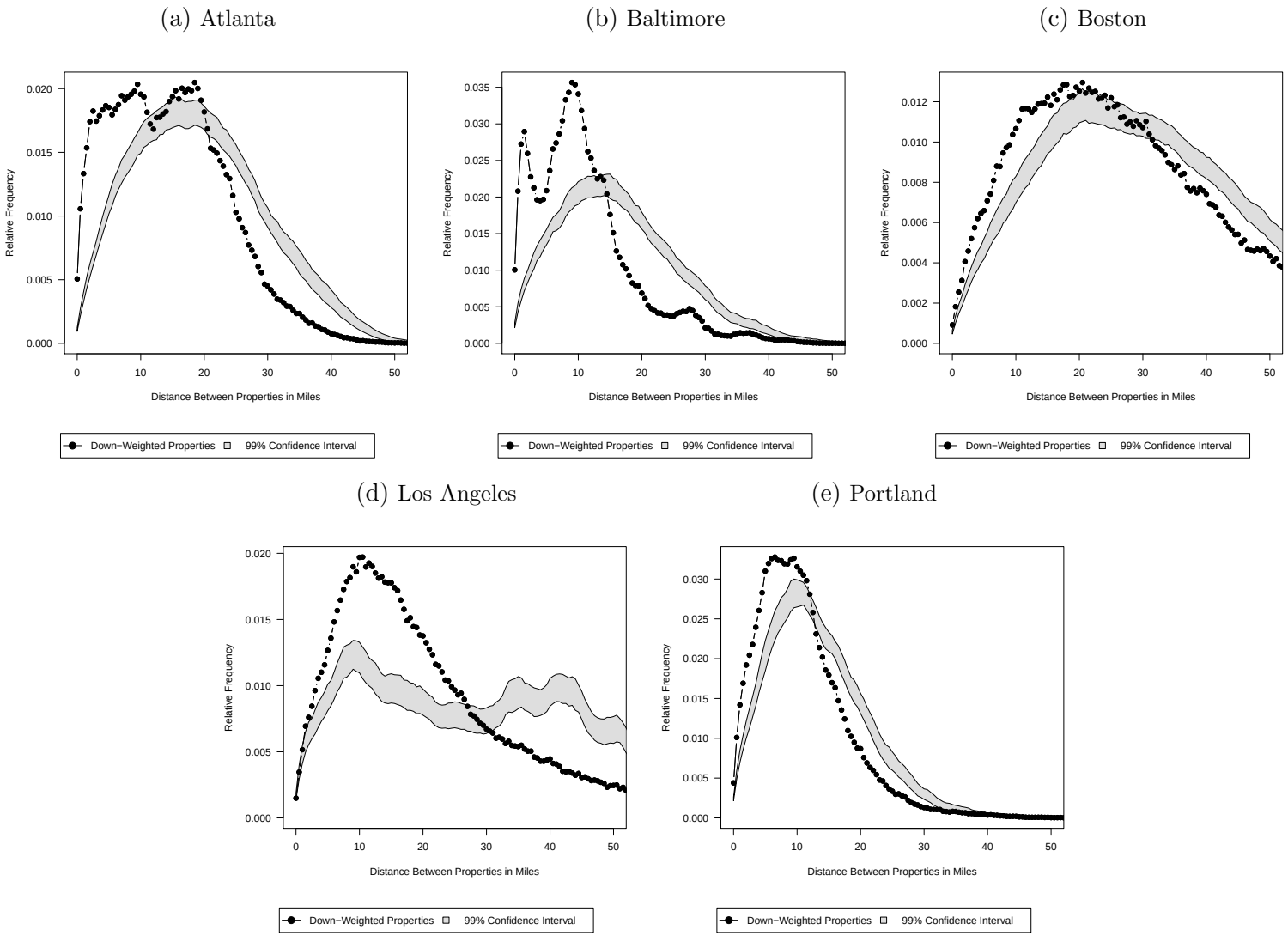
The quality-adjusted HPIs for the seven MSAs not included in this study display similar biases to those in the paper. The HPIs for Austin (3%), Chicago (12%), Dallas (3%), Denver (13%), Houston (4%), San Diego (8%), and Seattle (4%) are all biased upwards during the post-crisis period. The HPIs are available by request.

Figure C1: Additional MSA-level quality-adjusted HPIs



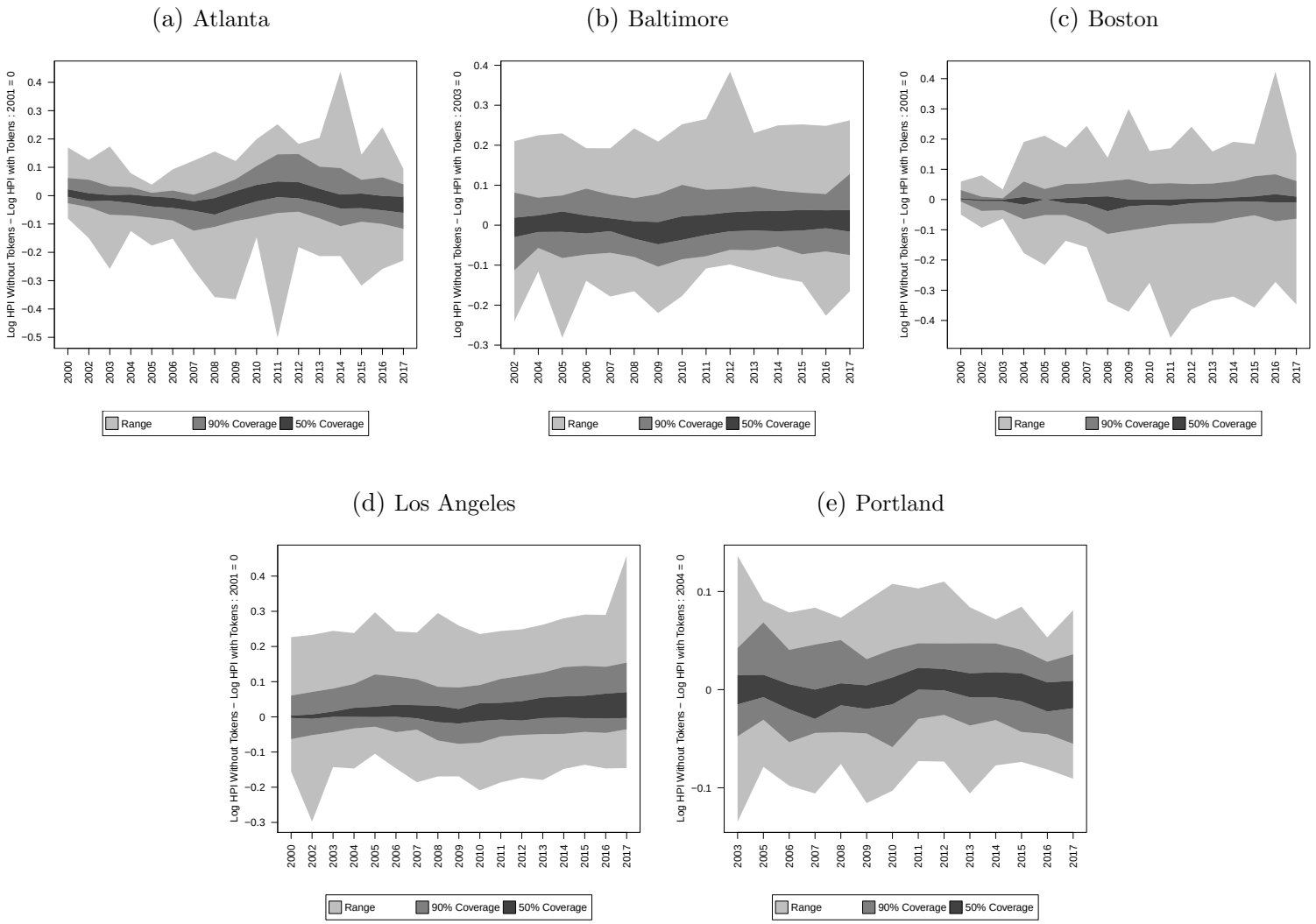
Notes: Figure C1 displays the Log HPI from a repeat-sales estimation incorporating tokens as controls for time-varying attributes and the 95% confidence interval for the Log HPI from a repeat-sales estimation without tokens where standard errors are clustered at the property level. The difference between the two Log HPIs is set to 0 in 2001 Q1 where possible else the second earliest Q1 available in the data. The HPI without tokens uses the Case-Shiller methodology in Section 2.2 and the HPI with unigram tokens uses the quality-adjusted methodology in Section 2.3.

Figure C2: Geographic concentration of bias for additional MSAs



Notes: Figure C2 displays the density of pairwise distances between properties that are down-weighted in the Case-Shiller HPI and the pointwise confidence intervals based on repeat sampling of properties that are not down-weighted. This measure of localization uses the methodology in Duranton and Overman (2005).

Figure C3: Dispersion of differences between local HPIs for additional MSAs

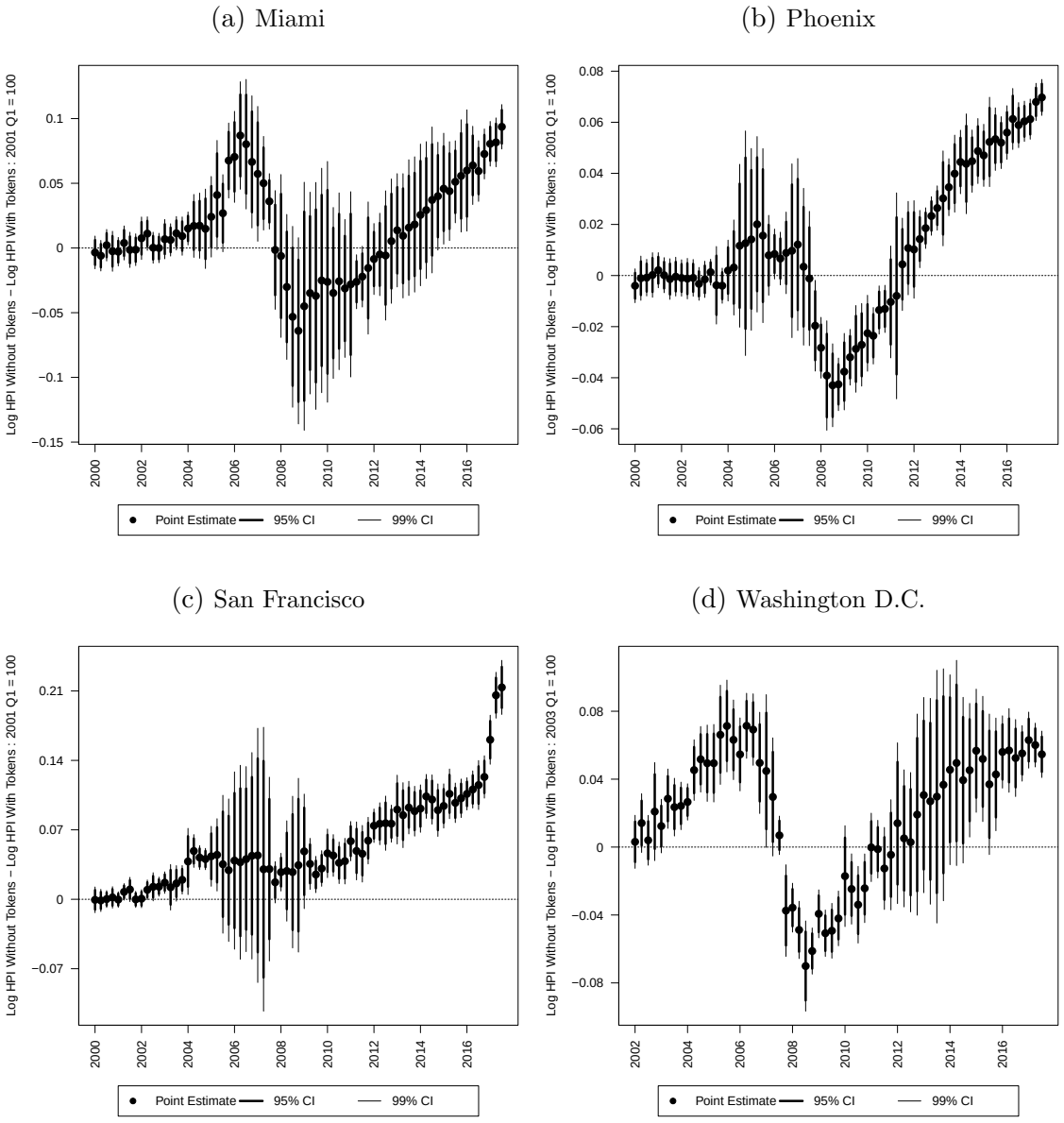


Notes: Figure C3 displays the difference between the Case-Shiller and quality-adjusted local HPIs at the zip code level. Only zip codes with at least 100 repeat-sales transactions are included. This requirement yields 94, 74, 384, 285, and 81 unique zip codes in Atlanta, Baltimore, Boston, Los Angeles, and Portland, respectively. Each panel indicates the range, 5th-95th percentiles, and 25th-75th percentiles. The difference between the HPIs is set to 0 in 2001 where possible else the second earliest year available in the data.

C.2 Difference in repeat-sales HPIs

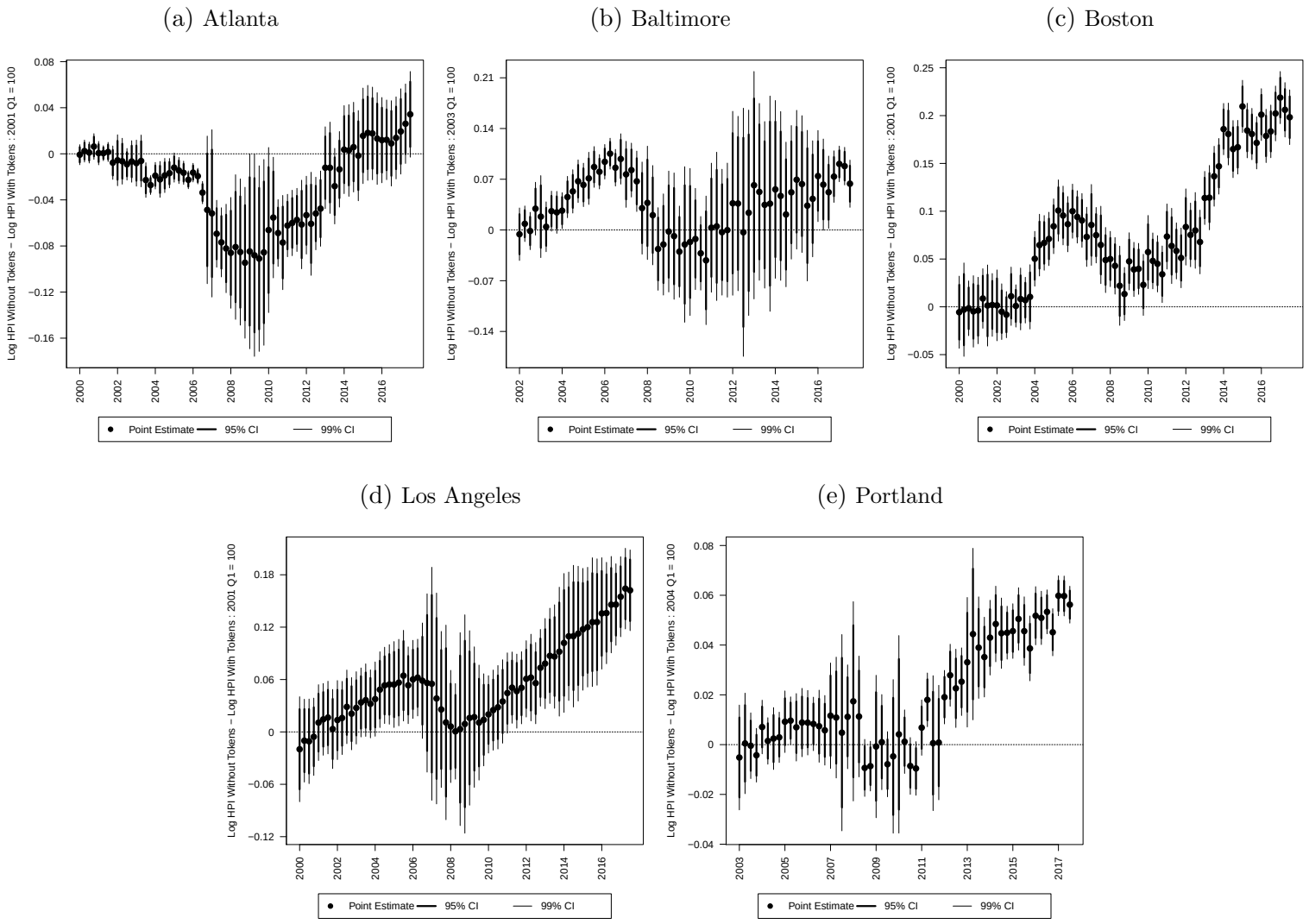
In this section, we plot the difference between the Case-Shiller and quality-adjusted HPIs for each MSA. In doing so, we demonstrate that the size, magnitude, and direction of the time-varying attribute bias fluctuates throughout the market cycle and across MSAs. Figure C4 plots the difference for the four MSAs (Miami, Phoenix, San Francisco, and Washington D.C.) in Figure 1 of the body of the paper. Figure C5 plots the difference for the five additional MSAs (Atlanta, Baltimore, Boston, Los Angeles, and Portland) in Figure C1 of this internet appendix.

Figure C4: Difference in repeat-sales HPIs with and without tokens



Notes: Figure C4 displays the difference between the two repeat-sales HPIs displayed in Figure 1 of the body of the paper. The point estimate represents the difference between the Case-Shiller HPI (without tokens) and our quality-adjusted HPI. A 95% and 99% confidence interval are provided for each point estimate.

Figure C5: Additional difference in repeat-sales HPIs



Notes: Figure C5 displays the difference between the two repeat-sales HPIs displayed in Figure C1. The point estimate represents the difference between the Case-Shiller HPI (without tokens) and our quality-adjusted HPI. A 95% and 99% confidence interval are provided for each point estimate.

C.3 Indicator-adjusted HPIs

An alternative approach to mitigate the time-varying attribute bias that has been proposed in the academic literature is to identify renovated properties and include an indicator in Equation 2 of the body of the paper as follows

$$\Delta p_{nt} = p_{nt} - p_{nt'} = \Delta\delta_t + \Delta f_{nt}\psi + \Delta\phi_{nt} + \Delta v_{nt} \tag{1}$$

where $\Delta f_{nt} = f_{nt} - f_{nt'}$ and f_{nt} represents an indicator variable equal to 1 for a house that was recently renovated and 0 otherwise.[1] The literature identifies renovated properties based on either the length of time between the repeat-sales (Clapp and Giaccotto, 1999; Bourassa et al., 2013), building permits (McMillen and Thorsnes, 2006; Billings, 2015), or changes to physical attributes across successive transactions (Bogin and Doerner, 2018).

Although the three identification strategies differ, they are similar in that they do not identify every renovation or control for the varying intensity of renovations. For example, Bourassa et al. (2013) consider any house that sold more than once in a year a flip (i.e., a house that was renovated and sold within a short period). This identification strategy likely underestimates the number of renovations since it only identifies successful flips that sold within an arbitrary one year holding period. One obvious concern is that high intensity renovations that introduce the largest bias may take longer than a year to complete - especially if they require permits. However, relying solely on building permits (McMillen and Thorsnes, 2006) or changes to physical attributes (Bogin and Doerner, 2018) will also underestimate renovations since structural changes to the house are not a necessary condition of a renovation.

The results in the body of the paper indicate that the text in agents' remarks (*tokens*) control for time-varying attributes of the property. Here, we investigate the extent to which conventional controls for flips, distressed sales, and renovations obviate the need for tokens.

[1] A similar approach is employed to control for distressed transactions (e.g., REOs). Distressed properties introduce a time-varying attribute bias, albeit in the opposite direction, because homeowners have no incentive to maintain the house leading up to the foreclosure, homeowners may damage the house when moving out, and/or the house may be damaged after the homeowners move out.

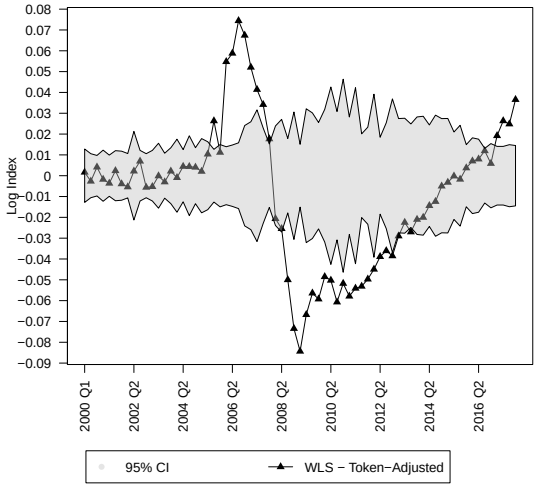
We do this by including an indicator variable for the three transactions types in the repeat-sales estimation. The indicator variable for a flip equals 1 if the holding period was less than 12 months. The indicator for a distressed sale equals 1 if the transaction was a real estate owned (REO) or short sale transaction. The indicator variable for a renovation equals 1 if the property was renovated during the 12 months prior to the sales transaction.

We then construct and compare indicator-adjusted HPIs that do not include the tokens (see Equation 1) to our quality-adjusted HPIs that include the tokens (see methodology in Section 2.3). If the information in the tokens is redundant after including the indicator variables for the three transaction types, then the quality-adjusted HPI should not be statistically different from the indicator-adjusted HPI. Figures C6 to C9 display the results for Miami, Phoenix, San Francisco, and Washington D.C. The four figures differ only in terms of which indicator variables are included in the construction of the indicator-adjusted HPI. Figure C6 includes the flip indicator, Figure C7 includes the distressed sale indicator, Figure C8 includes the renovation indicator, and Figure C9 includes all three indicators. We also plot indicator-adjusted HPIs with all three indicators for Atlanta, Baltimore, Boston, Los Angeles, and Portland in Figure C10.

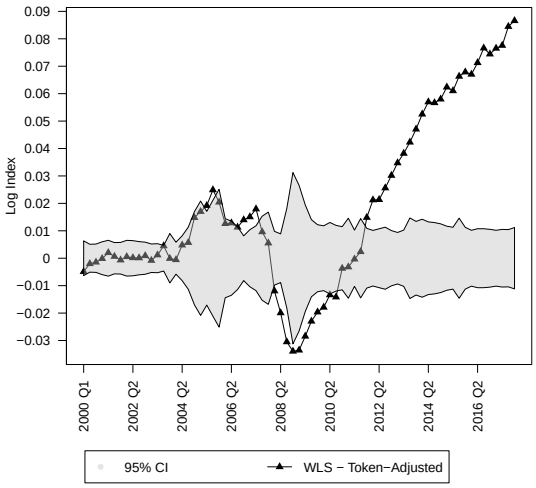
Overall we find the quality-adjusted HPIs are statistically different than the indicator-adjusted HPIs in all but one MSA (Washington D.C.). This finding highlights the fact that the indicators control for renovations, flips, distressed sales conditions, *and* additional information beyond that contained in the indicator variables. Additional corroborating evidence is provided in the next section where we drop the three transaction types (flips, distressed sales, and renovations) from the repeat-sales sample.

Figure C6: HPIs controlling for flips

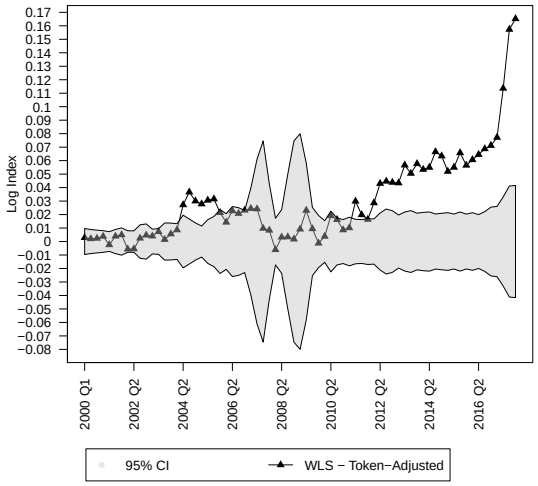
(a) Miami



(b) Phoenix



(c) San Francisco

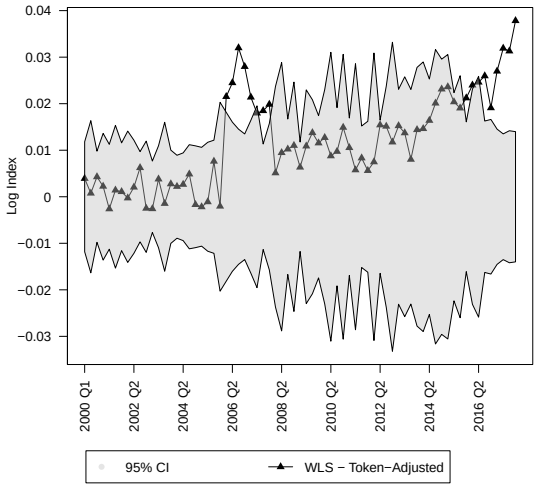


(d) Washington, D.C.

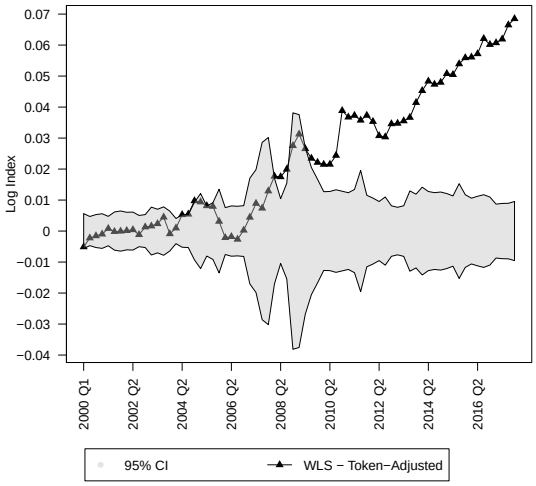
Notes: Figure C6 displays the difference between the Log HPI with and without tokens when including an indicator for houses that were flipped (holding period less than or equal to 12 months).

Figure C7: HPIs controlling for distressed transactions

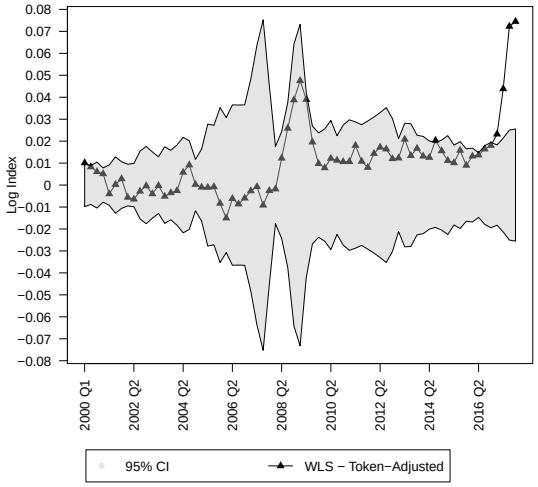
(a) Miami



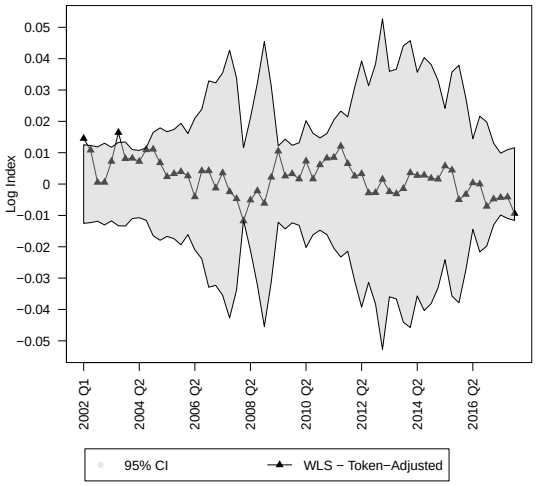
(b) Phoenix



(c) San Francisco



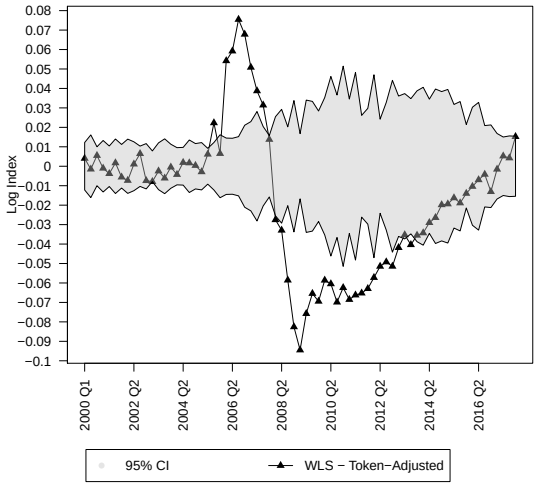
(d) Washington, D.C.



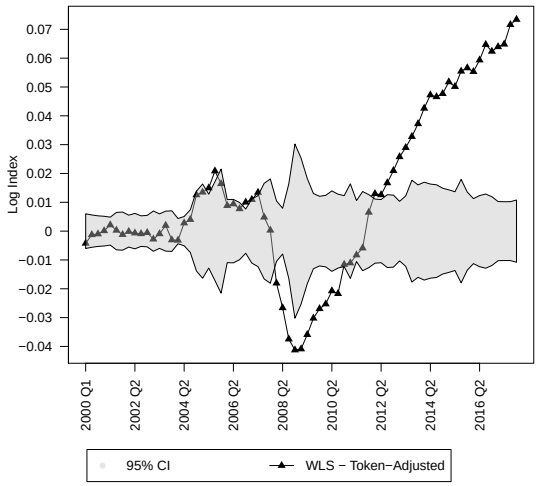
Notes: Figure C7 displays the difference between the Log HPI with and without tokens when including an indicator for houses that were involved in a distressed transaction (REO or short sale).

Figure C8: HPIs controlling for renovations

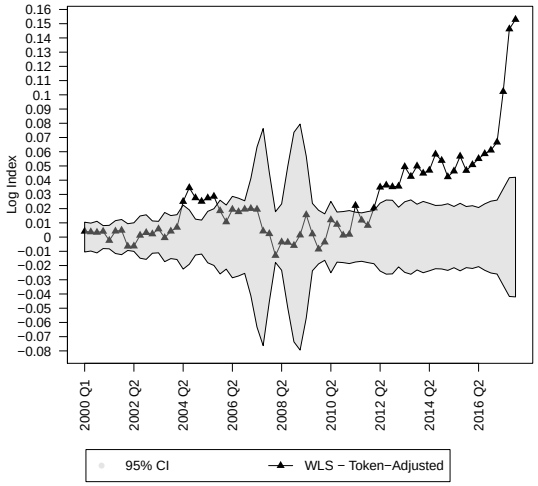
(a) Miami



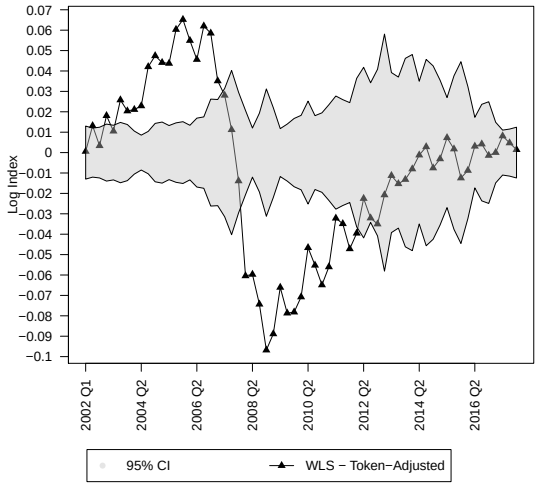
(b) Phoenix



(c) San Francisco

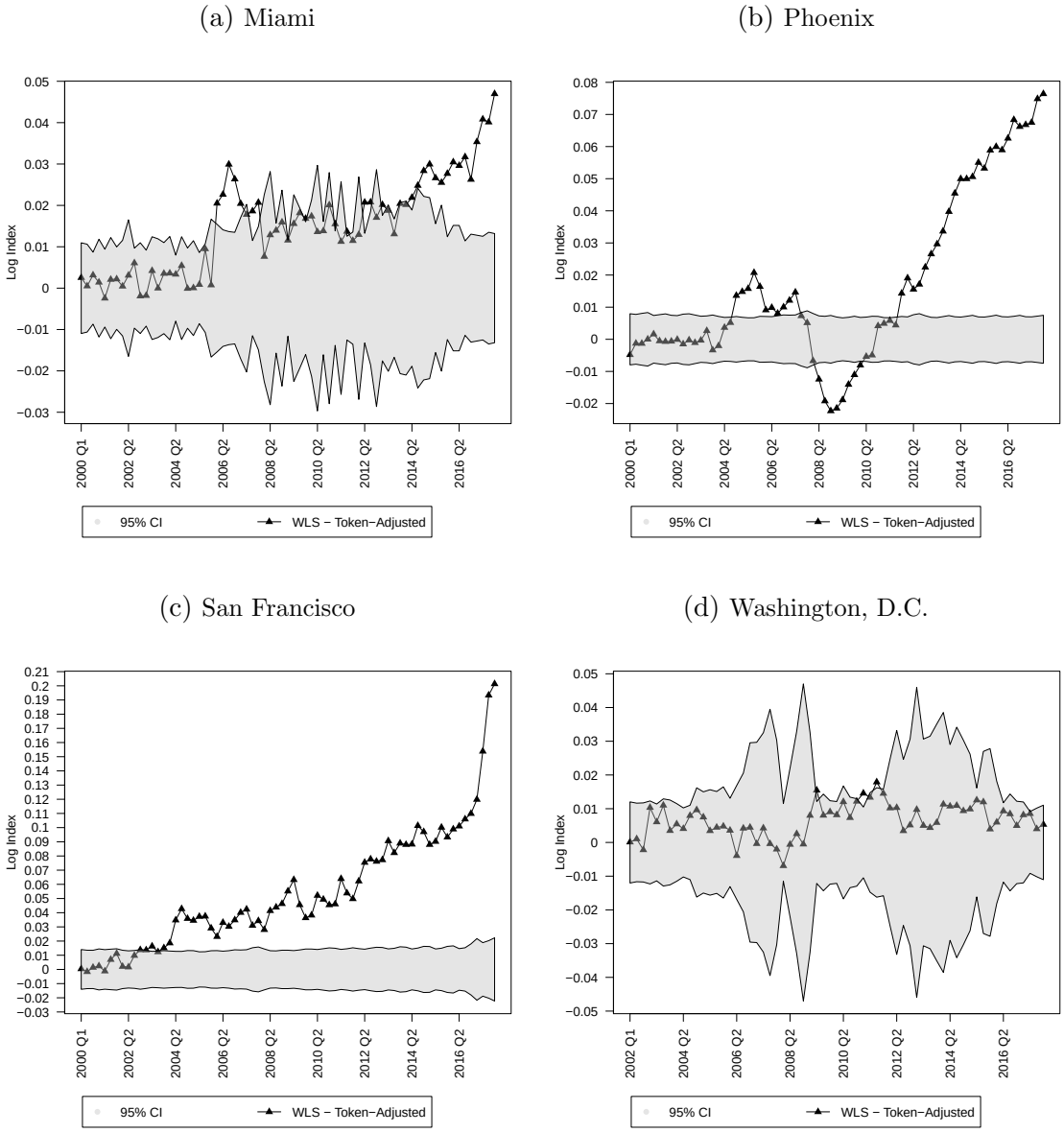


(d) Washington, D.C.



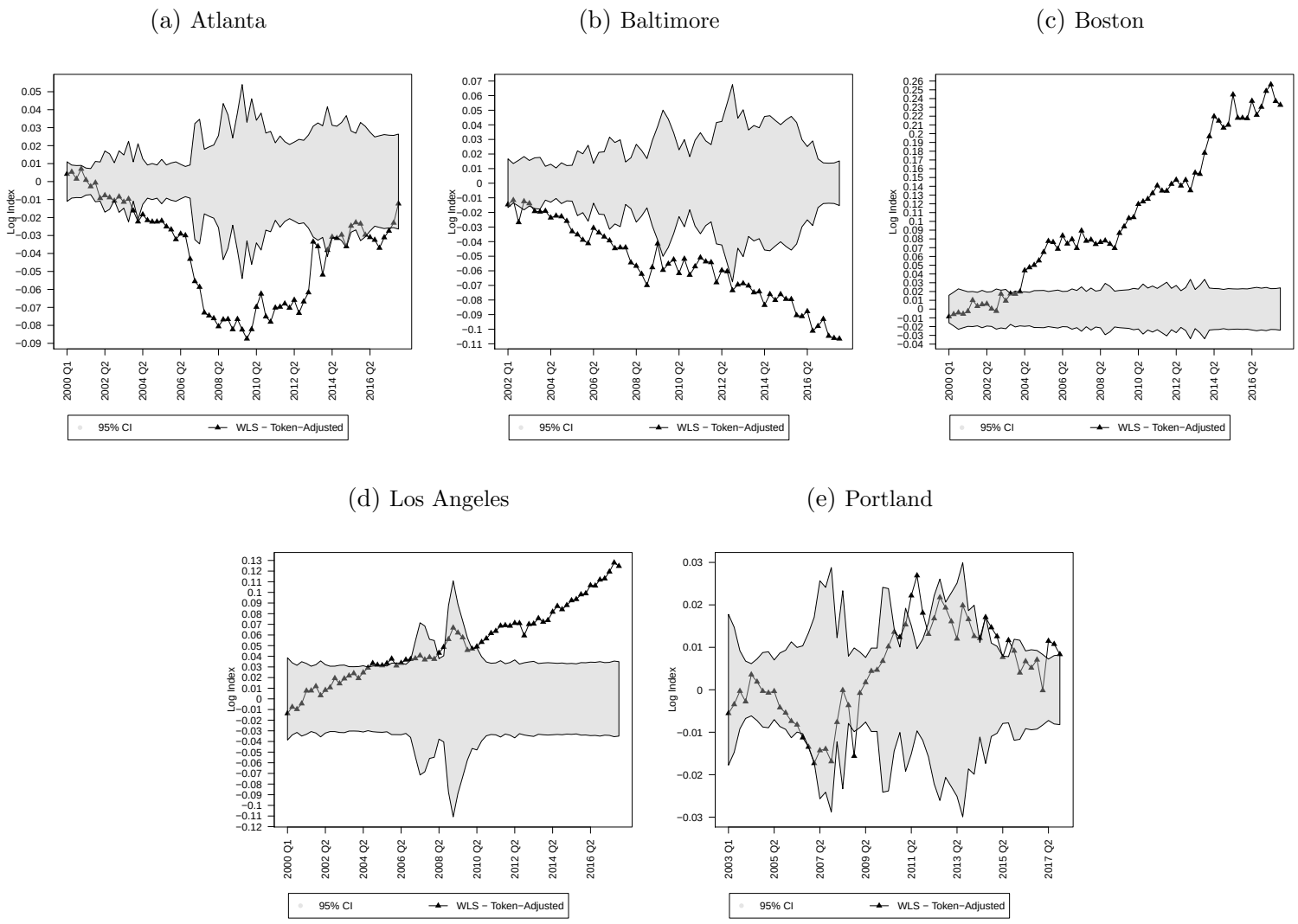
Notes: Figure C8 displays the difference between the Log HPI with and without tokens when including an indicator for houses that were renovated within the past year.

Figure C9: HPIs controlling for flips, distressed transactions, and renovations



Notes: Figure C9 displays the difference between the Log HPI with and without tokens when including indicators for flips (holding period less than or equal to 12 months), distressed sales (REO and short sales), and renovations (recently renovated within past 12 months of transactions).

Figure C10: Additional HPIs that control for flips, distressed transactions, and renovations



Notes: Figure C10 displays the difference between the Log HPI with and without tokens when including indicators for flips (holding period less than or equal to 12 months), distressed sales (REO and short sales), and renovations (recently renovated within past 12 months of transactions).

C.4 HPIs without flips, distressed sales, and renovations

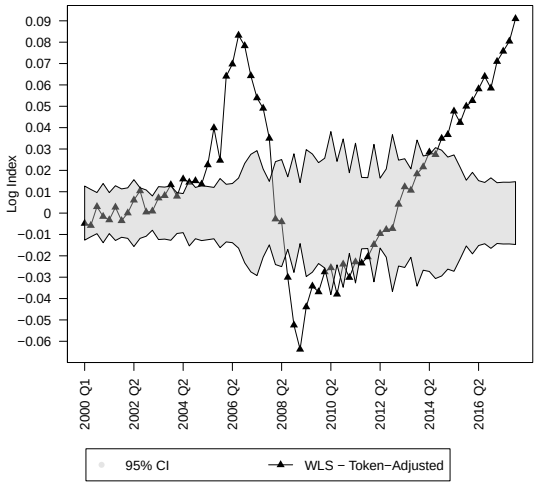
In this section we further examine the degree to which the textual information in agents' remarks control for a time-varying attribute bias. However, instead of including indicator variables for flips, distressed sales, and renovations, we drop all transactions for properties that were involved in at least one of the three transaction types during the study period. After excluding these properties from the sample, we estimate Case-Shiller HPIs (without tokens) and compare them to our quality-adjusted HPIs (with tokens).

Figures C11 to C14 display the results for Miami, Phoenix, San Francisco, and Washington D.C. The four figures differ only in terms of which transaction types are dropped when constructing the two HPIs. Figure C11 removes all transactions for properties that were flipped at least once during the study period, Figure C12 removes all transactions for properties that were sold as a short sale or REO at least once during the study period, Figure C13 removes all transactions for properties that were sold shortly after being renovated at least once during the study period, and Figure C14 removes all transactions for properties that were either flipped, sold under distressed sales conditions, or recently renovated at least once during the study period. Figure C15 corresponds with Figure C14 except that it plots HPIs for Atlanta, Baltimore, Boston, Los Angeles, and Portland.

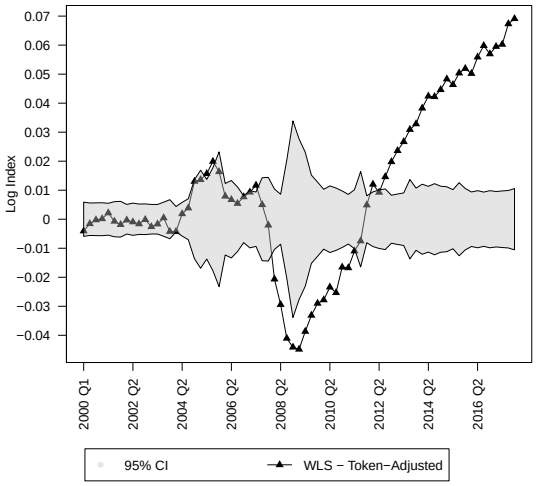
Overall the results highlight the fact that the quality-adjusted HPIs are statistically different than the Case-Shiller HPIs even after dropping properties that were involved in at least one of the three transaction types. This finding highlights the fact that the time-varying attribute bias that our approach identifies and mitigates is not simply the byproduct of including heterogeneous transaction types in the repeat-sales estimation.

Figure C11: HPIs that exclude flips

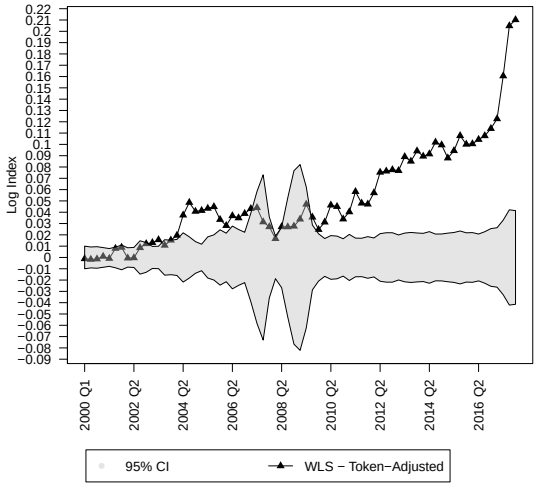
(a) Miami



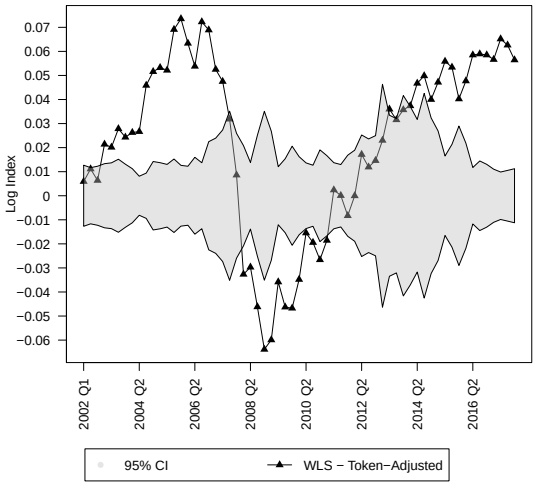
(b) Phoenix



(c) San Francisco



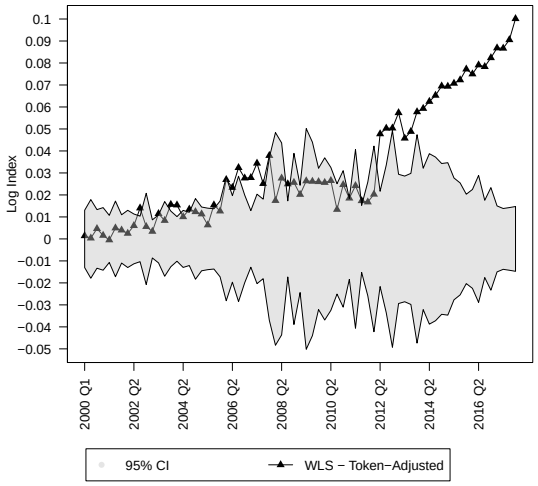
(d) Washington, D.C.



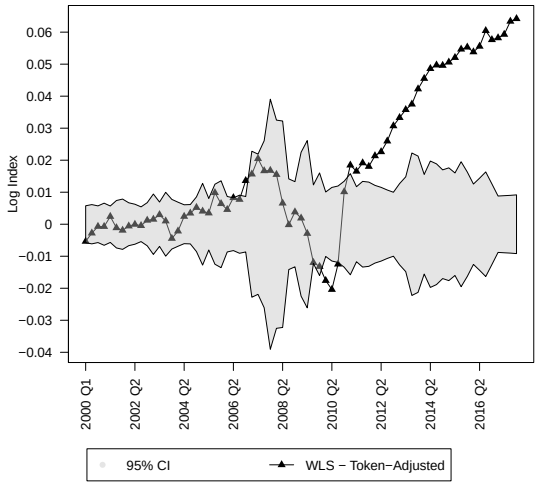
Notes: Figure C11 displays the difference between the Log HPI with and without tokens when dropping houses that were flipped (holding period less than or equal to 12 months).

Figure C12: HPIs that exclude distressed transactions

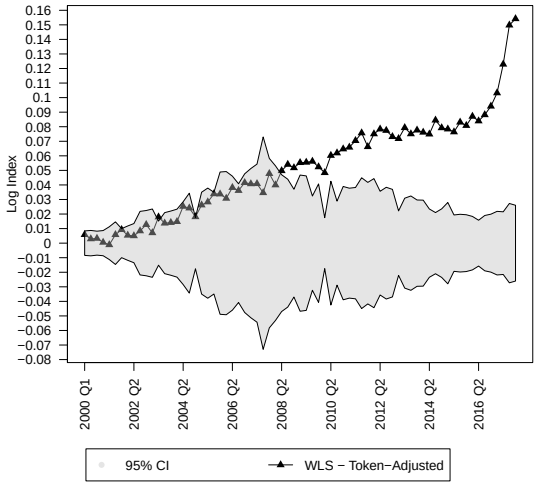
(a) Miami



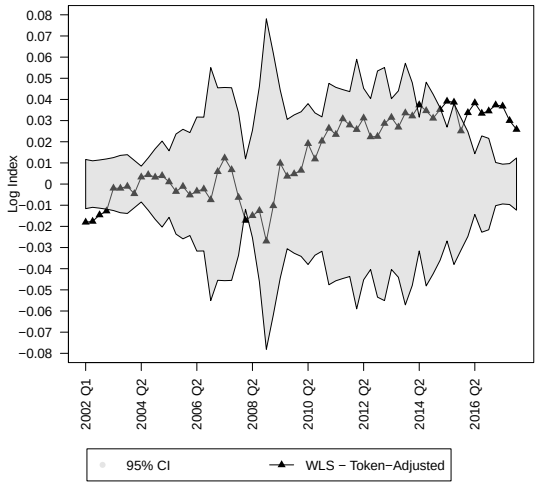
(b) Phoenix



(c) San Francisco



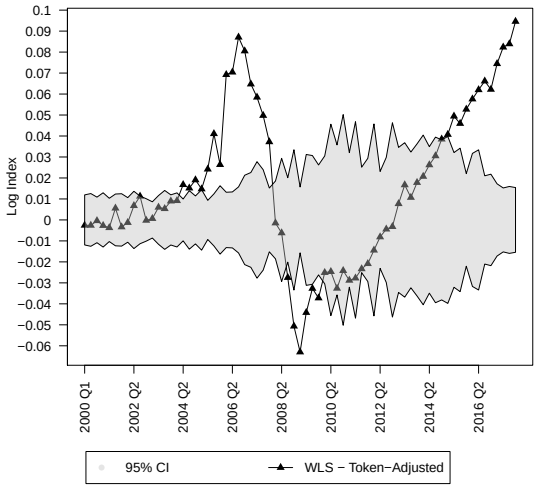
(d) Washington, D.C.



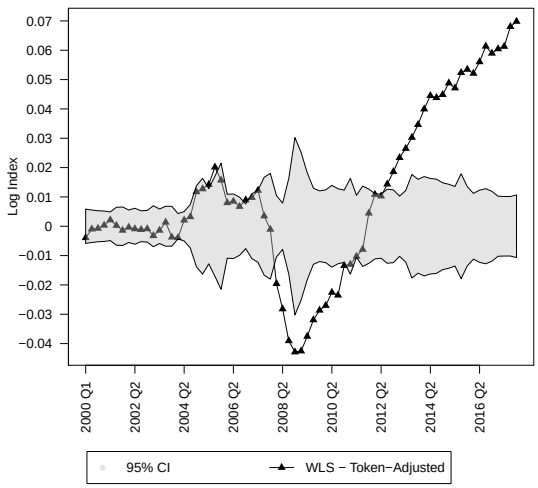
Notes: Figure C12 displays the difference between the Log HPI with and without tokens when dropping houses that were involved in at least one distressed sale (short sale or REO) during the study period.

Figure C13: HPIs that exclude renovations

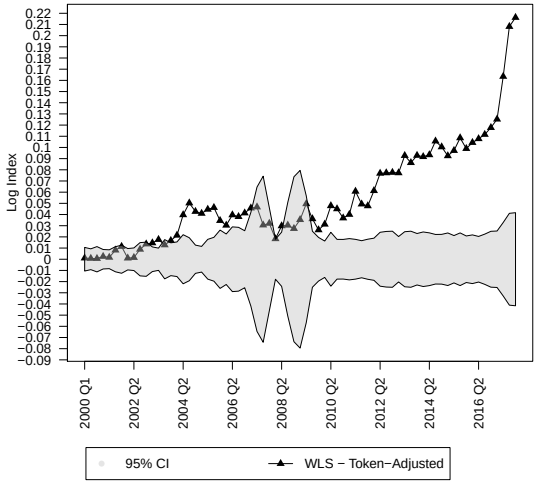
(a) Miami



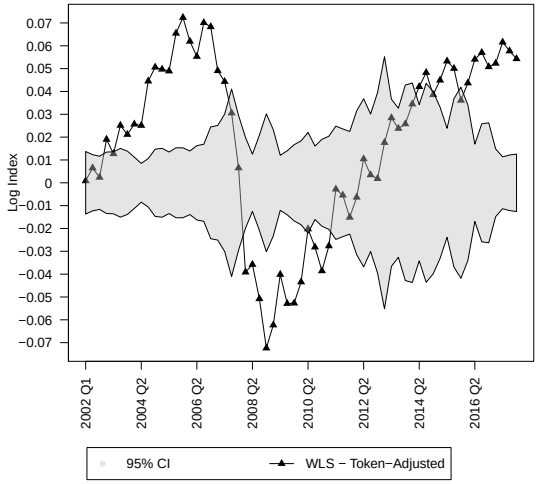
(b) Phoenix



(c) San Francisco



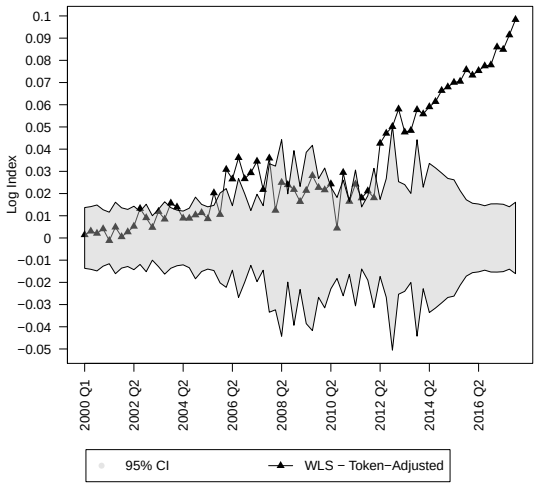
(d) Washington, D.C.



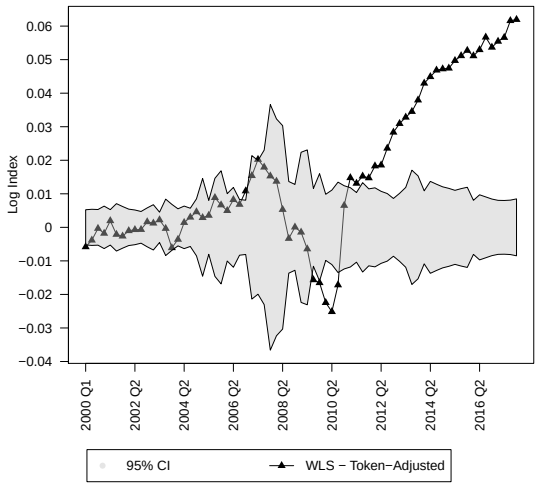
Notes: Figure C13 displays the difference between the Log HPI with and without tokens when dropping houses that underwent a recent renovation at least once during the study period.

Figure C14: HPIs that exclude flips, distressed transactions, and renovations

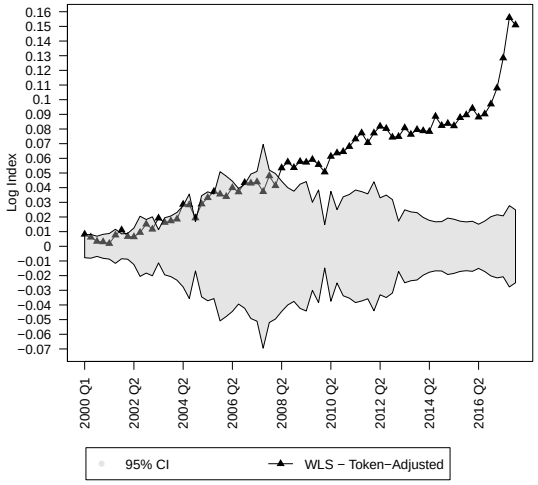
(a) Miami



(b) Phoenix



(c) San Francisco

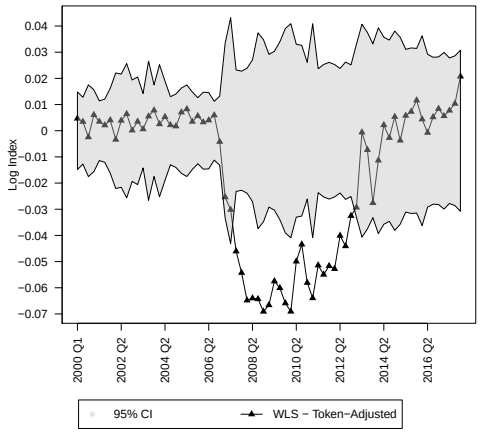


(d) Washington, D.C.

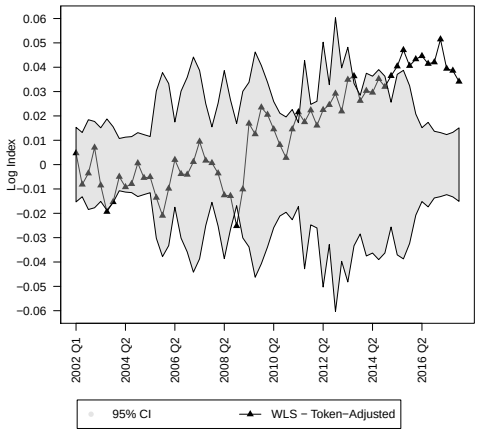
Notes: Figure C14 displays the difference between the Log HPI with and without tokens after dropping all transactions for properties that were involved in at least one flip (holding period less than or equal to 12 months), distressed sale (REO and short sales), or renovation (any renovation in the past 12 months) from the sample.

Figure C15: Additional HPIs that exclude flips, distressed sales, and renovations

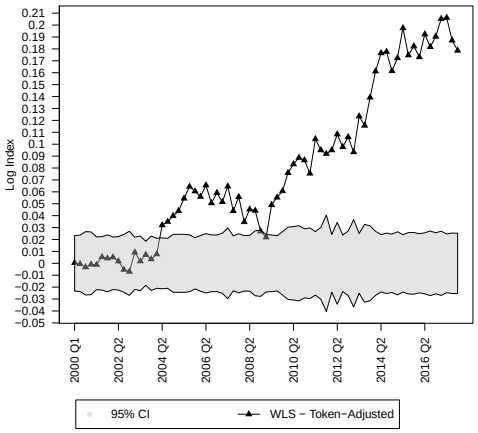
(a) Atlanta



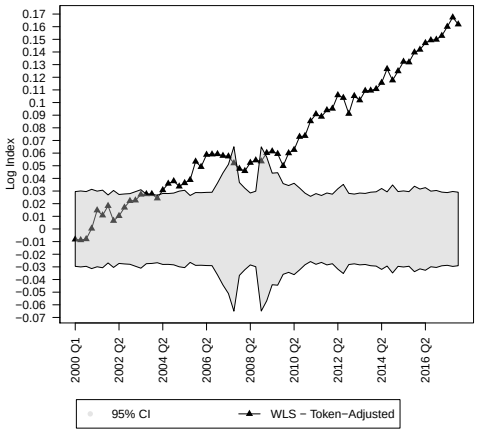
(b) Baltimore



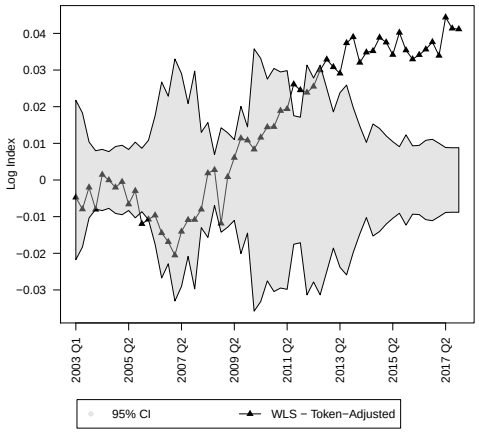
(c) Boston



(d) Los Angeles



(e) Portland



Notes: Figure C15 displays the difference between the Log HPI with and without tokens after dropping all transactions for properties that were involved in at least one flip (holding period less than or equal to 12 months), distressed sale (REO and short sales), or renovation (any renovation in the past 12 months) from the sample.

D Additional Considerations (Internet Appendix)

D.1 Time-varying implicit prices in quality-adjusted HPI

The quality-adjusted HPIs in the body of the paper select and include a set of time-varying tokens in the repeat-sales estimation under the assumption the implicit prices of the tokens do not vary over time. We recognize this assumption may not hold since the implicit prices of the tokens likely vary throughout the market cycle. For example, the magnitude of the implicit price for the *hud* token, which identifies REOs sold by the U.S. Department of Housing and Urban Development (HUD), is likely larger during (2008-2012) than after (2013-2017) the financial crisis. In contrast, the magnitude of the implicit price for the *renovated* token is likely smaller during (2008-2012) than after (2013-2017) the financial crisis since the type and intensity of the renovations being performed differ.

Here, we examine whether holding the implicit prices of the tokens constant impacts the quality-adjusted HPIs reported in the body of the paper. To do so, we allow the implicit prices of the tokens in the quality-adjusted HPI to vary over time at an annual frequency. For an MSA with Y years of data, this increases the total number of implicit prices we must estimate from $|\mathcal{K}|$ to $|\mathcal{K}|Y$ where $|\mathcal{K}|$ is the number of tokens in $\mathcal{K}$. For a MSA with $Y = 15$ years of data and choosing $\mathcal{K}$ as the 2,000 most frequent tokens, this requires estimating $|\mathcal{K}|Y = 30{,}000$ implicit prices.

Define $y(t) = 1$ if time period t is in year y and $y(t) = 0$ otherwise. Define $r_{nyk} = r_{ntk}y(t)$ as the product of the remark indicator for token k and the indicator for year y. This implies $r_{nyk} = 1$ if token k appears in the remarks for a property sold in year y and $r_{nyk} = 0$ otherwise. Define $y = 1, ..., Y$ as the year a property sold where $y = 1$ corresponds to the first year in the data. We estimate the quality-adjusted HPI with annually-varying implicit prices by solving

$$\{\hat{d}, \hat{h}\} = \arg\min_{d,h} \sum \left(\Delta p_{nt} - \Delta d_t - \sum_{k \in \mathcal{K}} \sum_{t=1}^{T} \sum_{t'=1}^{T} (h_{ky(t)} r_{ny(t)k} - h_{ky(t')} r_{ny(t')k}) \right)^2 + \lambda \sum_{k \in \mathcal{K}} \sum_{y=1}^{Y} |h_{ky}| \upsilon_{ky} \tag{2}$$

In Equation 2, $\hat{h}$ is a $|\mathcal{K}|Y \times 1$ vector with a separate implicit price, $\hat{h}_{ky}$, for every token $k \in \mathcal{K}$ in every year $y = 1, ..., Y$. Equation 2 is similar to Equation 9 in the body of the paper but allows for annual variation in the implicit price of the tokens indicated by h_{ky}. We also experimented with a time-varying set of the most frequent tokens in each year, $\mathcal{K}_y$ but found $\mathcal{K}_y$ was nearly identical across years, and as a result the quality-adjusted HPI estimates were not sensitive to annually-varying sets of candidate tokens.

Table D1 displays summary statistics for the difference between MSA-level quality-adjusted HPIs that either (i) assume the implicit prices of the tokens are constant or (ii) allow the implicit prices of the tokens to vary annually. The results indicate the static implicit price assumption does not introduce a significant bias in the nine MSAs we examine.

Table D1: Time-varying implicit prices HPI

MSA	Min	Mean	Max
atl	-0.002	0.002	0.008
bal	-0.004	-0.000	0.004
bos	-0.002	0.005	0.016
dc	-0.007	-0.002	0.004
la	-0.005	-0.001	0.003
mia	-0.002	0.001	0.005
pdx	-0.003	0.001	0.005
phx	-0.005	-0.000	0.004
sf	-0.003	0.002	0.004

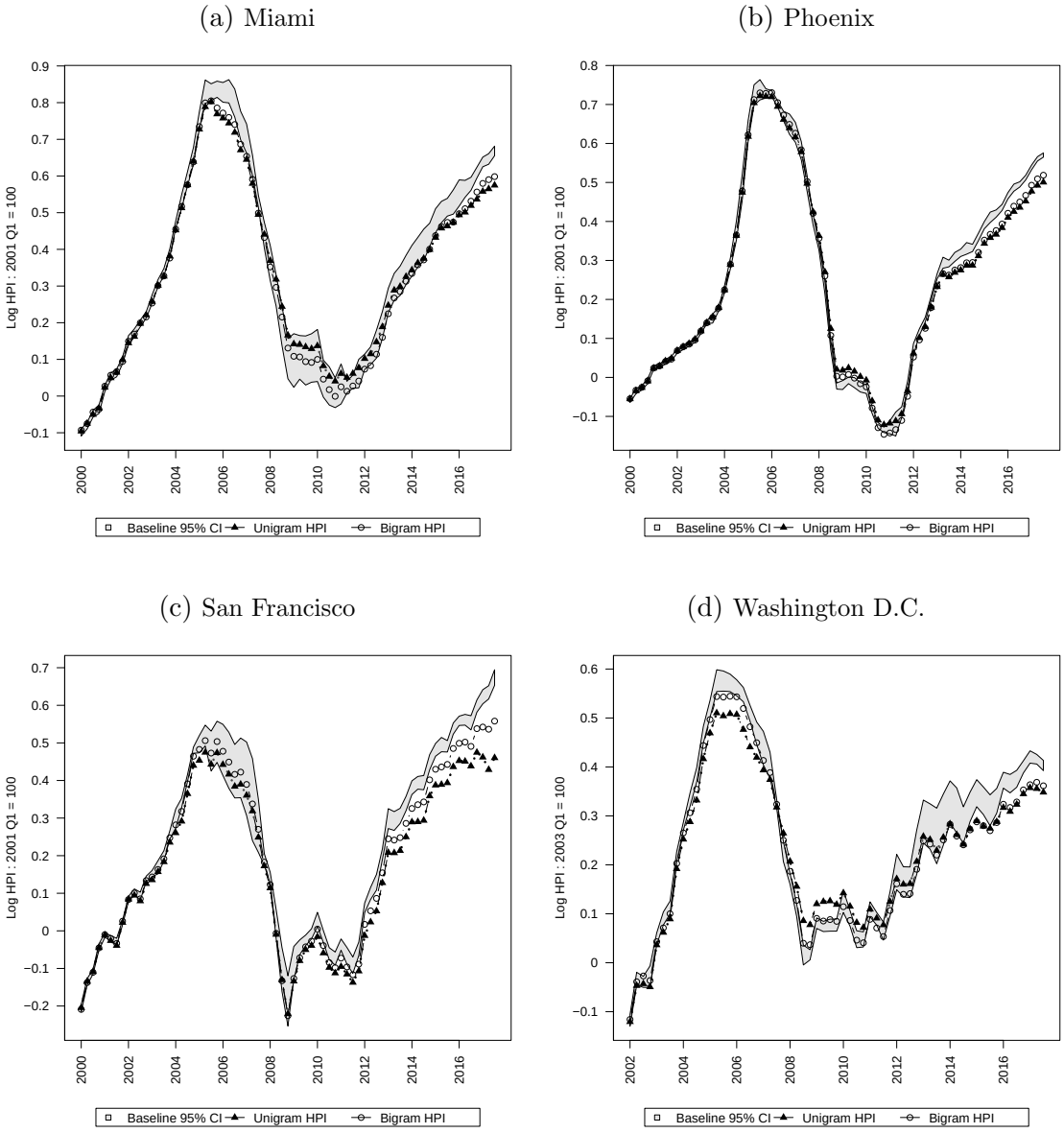
Note: Table D1 displays summary statistics for the difference between HPIs calculated assuming the implicit price for each token is constant and HPIs that allow the implicit price for each token to vary annually.

D.2 Alternative tokenization procedures

For the sake of brevity, we only examine unigram tokens and limit the number of candidate tokens to 2,000 in the body of the paper. Although unreported, we thoroughly examine whether our tokenization procedures bias our findings. In short, we find that increasing/decreasing the number of candidate tokens, using bigrams (two word phrases) or trigrams (three word phrases) instead of unigrams (one word), and/or employing alternative tokenization procedures (stemming, including plurals, etc.) does not have a material impact on the results reported in the body of the paper.

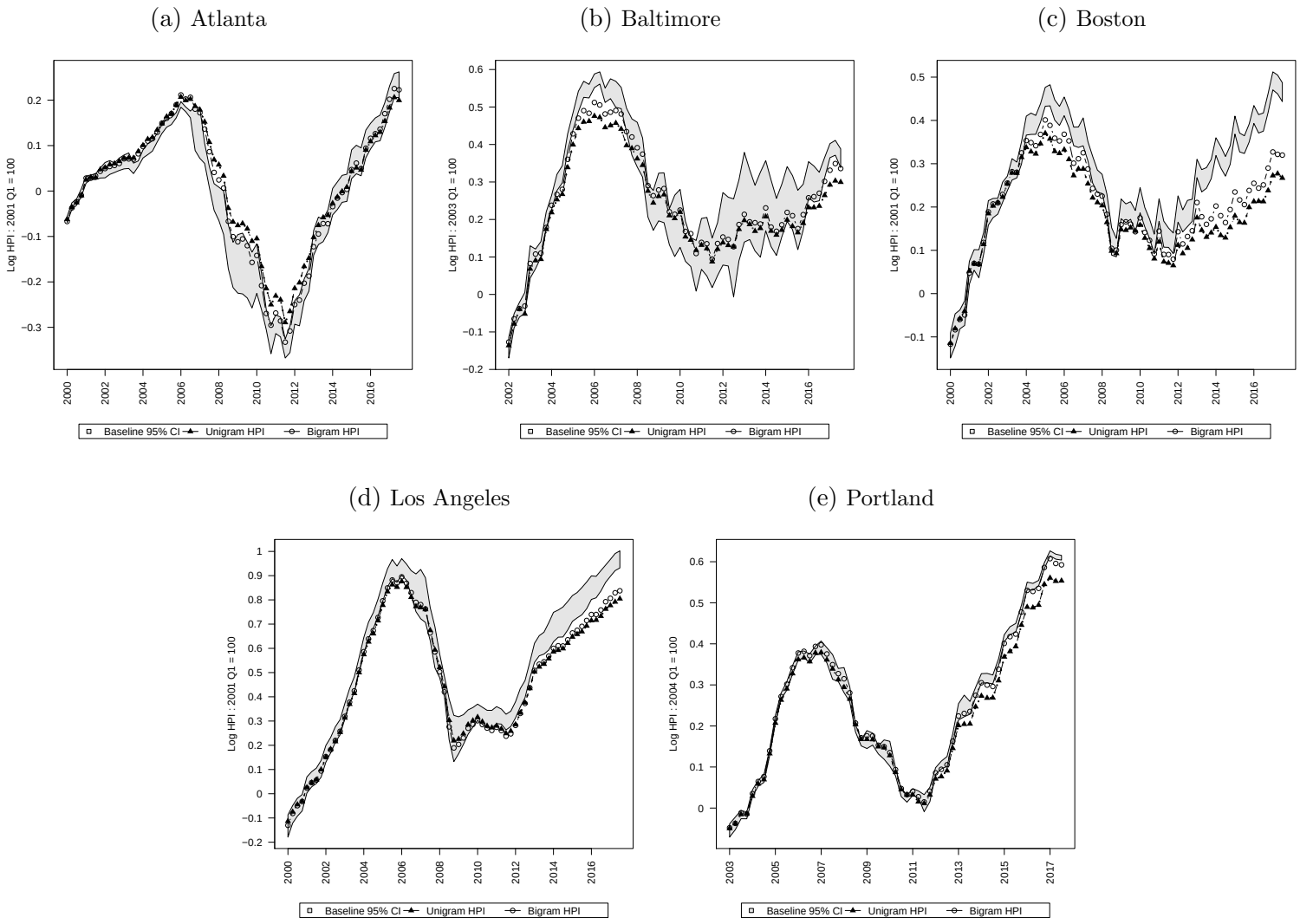
See, for example, the comparison of unigram and bigram quality-adjusted HPIs in Figures D1 and D2 where the bigram token-adjusted HPI tracks the unigram token-adjusted HPI fairly closely across all nine MSAs. Note, however, the bigram token-adjusted HPI does not adjust upwards (downwards) as much during the financial crisis (post-crisis) period. This finding is not surprising given that Nowak and Smith (2017) find that unigrams outperform bigrams for both in-sample and out-of-sample price prediction.

Figure D1: Unigram and bigram quality-adjusted HPIs



Notes: Figure D1 compares the Case-Shiller HPIs to quality-adjusted HPIs that incorporate unigram or bigram tokens.

Figure D2: Additional unigram and bigram quality-adjusted HPIs



Notes: Figure D2 compares the Case-Shiller HPIs to quality-adjusted HPIs that incorporate unigram or bigram tokens.

D.3 Alternative variable selection procedures

For the sake of brevity, we only use one high-dimensional variable selection methodology in the body of the paper. Although unreported, we examine whether the single-selection LASSO procedure we employ biases our findings. One possible concern with the single-selection LASSO procedure is that it only selects tokens that are the strongest predictors of price changes. Modest predictors of price changes that are significantly correlated with d_t may be omitted from $\hat{\mathcal{S}}$. When this is true, $\hat{\mathcal{S}}$ may not be adequate to correct the HPI. To address this concern, we run the double-selection LASSO procedure described in Belloni et al. (2014) to identify and include the strongest predictors of price changes *and* the differenced indicators for quarter of sale in the repeat-sales estimation. The additional tokens are chosen based on their ability to predict the date of sale using a linear probability model.

Table D2 displays summary statistics for the difference between the single-selection HPI and the double-selection HPI log index. The results indicate the single-selection procedure that we employ does not introduce a significant bias for the nine MSAs we examine. By construction, $\hat{\mathcal{S}} \subset \hat{\mathcal{S}}_{ds}$ where $\hat{\mathcal{S}}_{ds}$ is the set of tokens selected by the double selection procedure. Although $\hat{\mathcal{S}}_{ds}$ is larger, Table D2 indicates the additional tokens do not significantly alter the resulting HPI.

Table D2: Double-selection HPIs

MSA	Min	Mean	Max	$\hat{\mathcal{Q}}$	$\hat{\mathcal{Q}}_{ds}$
atl	-0.001	0.003	0.014	166	458
bal	-0.002	0.001	0.003	157	199
bos	-0.007	-0.003	0.001	433	760
dc	-0.001	0.000	0.002	244	340
la	-0.011	-0.002	0.005	314	667
mia	-0.011	-0.004	0.002	159	315
pdx	-0.005	-0.002	0.001	182	251
phx	-0.009	-0.004	0.008	320	853
sf	-0.004	-0.001	0.006	268	444

Note: Table D2 displays summary statistics for the difference between the single-selection HPI and the double-selection HPI log index. The double-selection estimator includes an additional set of tokens as controls. This additional set of tokens is the set of the strongest predictors of the differenced indicators for quarter of sale. $\hat{\mathcal{Q}}$ indicates the number of tokens in $\hat{\mathcal{S}}$ and $\hat{\mathcal{Q}}_{ds}$ indicates the number of tokens selected using the double-selection procedure in Belloni et al. (2014).

References

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Billings, S. B. (2015). Hedonic amenity valuation and housing renovations. *Real Estate Economics*, 43(3):652–682.

Bogin, A. and Doerner, W. (2018). Property renovations and their impact on house price index construction. *Journal of Real Estate Research*, Forthcoming.

Bourassa, S. C., Cantoni, E., and Hoesli, M. (2013). Robust repeat sales indexes. *Real Estate Economics*, 41(3):517–541.

Clapp, J. M. and Giaccotto, C. (1999). Revisions in repeat-sales price indexes: Here today, gone tomorrow? *Real Estate Economics*, 27(1):79–104.

Duranton, G. and Overman, H. G. (2005). Testing for localization using micro-geographic data. *The Review of Economic Studies*, 72(4):1077–1106.

McMillen, D. P. and Thorsnes, P. (2006). Housing renovations and the quantile repeat-sales price index. *Real Estate Economics*, 34(4):567–584.

Nowak, A. and Smith, P. (2017). Textual analysis in real estate. *Journal of Applied Econometrics*, 32(4):896–918.

Redfin (2000-2017). Data: Multiple listing service (MLS) transaction data. *Redfin Corporation (Accessed: February 2018)*.