Quality-adjusted house price indexes

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

July 19, 2019

Abstract

The constant-quality assumption in repeat-sales house price indexes (HPIs) introduces a significant time-varying attribute bias. The direction, magnitude, and source of the bias varies throughout the market cycle and across metropolitan statistical areas (MSAs). We mitigate the bias using a data-driven textual analysis approach that identifies and includes salient text from real estate agent remarks in the repeat-sales estimation. Absent the text, MSA-level HPIs are biased downwards by as much as 7% during the financial crisis and upwards by as much as 20% after the crisis. The geographic concentration of the bias magnifies its effect on local HPIs.

Key Words: House price index, machine learning, repeat sales, textual analysis

[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

[1] We thank Redfin for providing access to the data employed in this study. The views expressed in this study are those of the authors and do not necessarily reflect the position of Redfin. All errors are the responsibility of the authors.

1 Introduction

Recent empirical research documents the far reaching effects that house price movements have on the economy. House price movements have been linked to, among other things, consumer spending (Mian et al., 2013; Berger et al., 2017), employment (Mian and Sufi, 2014; Adelino et al., 2015), and economic growth (Loutskina and Strahan, 2015). Information about house price movements is disseminated to the public via house price indexes (HPIs). Thus, HPIs play an important role in the decision making process of households, investors, lenders, and, to the extent that aforementioned academic studies inform public policy, government officials. The most prominent HPIs use the repeat-sales methodology proposed in Bailey et al. (1963) and popularized by Case and Shiller (1989).[1]

Repeat-sales HPIs measure house price movements using the price change between sales of the same house under the assumption that the quality of the house does not change over time.[2] The purpose of this study is to show that not only does the constant-quality assumption not hold, but that it also introduces a significant bias in repeat-sales HPIs. By design, the repeat-sales approach controls for time-invariant attributes of the house but does not address the time-varying nature of real estate (e.g. depreciation and renovation). Although this bias is well documented (Palmquist, 1980; Goetzmann and Spiegel, 1995), this is the first study to directly link the bias to specific time-varying attributes of the house and, more importantly, provide a solution that mitigates the bias.[3]

To mitigate the bias, we use machine learning methods to identify salient words and

[1] The Case-Shiller HPI is the most commonly cited HPI in both the popular press and academia. For example, a 2012 *Wall Street Journal* article states that "There are many ways to measure changes in house prices, but the Standard & Poor's/Case-Shiller index has become many economists' favored benchmark in recent years." CoreLogic acquired the rights to the Case-Shiller HPI in 2013.

[2] We use "quality" as an all-encompassing term that refers to property condition, property quality, and neighborhood quality. Property condition refers to a time-varying measure of the house's maintenance and upkeep, property quality refers to a nearly time-invariant measure of the workmanship and materials used in the house's construction, and neighborhood quality refers to a time-varying measure of locational amentities or public service attributes. Although the three measures contain time-varying attributes, constant-quality HPIs assume all attributes are time-invariant.

[3] Recognizing that the constant-quality assumption does not hold, the Case-Shiller HPI attempts to address the bias using a series of filters to control for the first moment of the error term and a robust weighting procedure to control for the second moment of the error term. Although the Case-Shiller index methodology addresses a portion of the time-varying attribute bias, we show that the HPIs are still biased since the methodology does not identify and/or control for all relevant time-varying attributes that contribute to the bias. We discuss the Case-Shiller index methodology in Section 2.2.

phrases (*tokens*) in listing agents' written descriptions (*remarks*) about the house. Then, we include these tokens as controls in the repeat-sales regression. Although the direction, magnitude, and source of the time-varying attribute bias varies throughout the market cycle and across metropolitan statistical areas (MSAs), our flexible data-driven approach identifies and selects a unique set of relevant time-varying tokens for each MSA. This flexibility is critical given the localized nature of real estate and the myriad of underlying information that researchers must control for. We demonstrate the efficacy of our approach by constructing Case-Shiller and quality-adjusted HPIs for eight large MSAs in the United States. We find that the Case-Shiller HPIs are biased downwards by as much as 7% during the recent financial crisis and upwards by as much as 20% after the crisis.

After documenting the bias, we show that it is directly related to homeowners' (dis)incentive to perform maintenance and renovations during housing market booms and busts. We also show that our quality-adjusted HPIs effectively control for the intensity of the improvements made to the house (or lack thereof) between repeat-sales. In other words, our enhanced approach controls for both whether the house was renovated (extensive margin) *and* the degree to which the house was renovated (intensive margin). The distinction is important since our approach controls for not only the fact that an improvement occured, but also the varying intensity (cosmetic vs. structural) in which the improvement occured. Similarly, our approach controls for not only the likelihood that improvements will occur throughout the market cycle, but also the varying intensity in which they occur throughout the market cycle. Moreover, our data-driven approach is extremely flexible in that it does not require the researcher to (i) ex-ante specify a list of controls or (ii) have location-specific knowledge for every MSA of interest. Instead of controlling for a static list of predefined variables, our approach selects a unique set of time-varying controls for each MSA. These unique sets of controls often include location-specific improvements, such as building code updates, that bias some but not all of the MSAs in this study.

Controlling for the time-varying attribute bias takes on added importance when constructing *local* HPIs since the bias is often geographically concentrated within a MSA.[4] This is a concern since recent studies frequently employ local HPIs that are increasingly suscep-

[4] We use 5-digit zip codes and define the HPIs calculated at this smaller geographic level as *local* HPIs.

tible to a time-varying attribute bias. For example, local HPIs have been used to examine the link between subprime lending and foreclosures (Mian and Sufi, 2009) and gentrification (Guerrieri et al., 2013). However, the zip codes of interest in these studies (i.e. subprime and gentrified) are the zip codes whose local HPIs are most susceptible to the time-varying attribute bias we document. For this reason, the coefficient estimates in these studies are likely biased (overstated or attenuated) since the local HPIs do not properly control for the (dis)incentive to renovate in gentrified (subprime) zip codes. These findings take on added importance as studies increasingly examine the effect of house price dispersion not only across MSAs (Van Nieuwerburgh and Weill, 2010; Gyourko et al., 2013), but within MSAs (Landvoigt et al., 2015; Bogin et al., 2018).

2 Methodology

2.1 Repeat-sales HPIs

We assume the transaction price for property n sold in time period t is given by

$$p_{nt} = \delta_t + z_n\beta_z + x_{nt}\beta_x + \mu_n + \phi_{nt} + v_{nt} \tag{1}$$

where p_{nt} is the transaction price, δ_t is the price level at time t, z_n (x_{nt}) is a vector of time-invariant (time-varying) housing attributes observed by the researcher, μ_n (ϕ_{nt}) is a vector of time-invariant (time-varying) housing attributes not observed by the researcher, and v_{nt} is a zero-mean disturbance term uncorrelated with any attributes of the property or the date of sale.

Time-invariant or nearly time-invariant variables, such as dwelling size and location, explain a large portion of the variation in price levels in Equation 1. However, estimating δ_t requires the inclusion of every relevant variable in the estimation. In residential housing markets, this estimation is complicated by the fact that houses are heterogeneous assets with numerous attributes that are difficult to quantify. In the context of Equation 1, it is common for z_n and x_{nt} to only include a few variables; thereby increasing the likelihood of an omitted variable bias.

To overcome this issue, repeat-sales HPIs are frequently employed since they can be constructed without collecting a comprehensive set of physical (beds, bathrooms, etc.) and locational (neighborhood amenities, etc.) attributes. Repeat-sales HPIs are estimated by assuming x_{nt} does not vary within property across time and differencing Equation 1 for same-house sales that transacted at times t and $t' < t$.[5]

$$\Delta p_{nt} = p_{nt} - p_{nt'} = \Delta\delta_t + \Delta\phi_{nt} + \Delta\upsilon_{nt} \tag{2}$$

By design, the price changes in Equation 2 do not reflect time-invariant housing attributes, so the researcher only needs to collect time-varying variables. Likewise, when the time-varying attributes of the property are uncorrelated with the date of sale, omitting these variables from the regression will not bias estimates of δ_t. Thus, when time-varying attributes of the property are either assumed constant or uncorrelated with the date of sale, the repeat-sales approach only needs two fields (transaction price and transaction date) to construct a repeat-sales HPI.

In theory, the repeat-sales approach directly addresses and controls for the heterogeneity of the housing stock which plagues other HPI methodologies.[6] It does so while simultaneously requiring minimal data and offering a straightforward implementation. However, these advantages come at a cost in the form of the "constant quality" assumption. Constant-quality HPIs assume not only that the physical and locational attributes of each house remain constant over time, but also that the condition and quality of those attributes remain constant over time. This assumption is clearly violated by the fact that houses are depreciating assets that change over time due to aging, maintenance, obsolescence, and/or renovations. Thus, constant-quality HPIs that assume the time-varying housing attributes (x_{nt}) in Equation 2 are time-invariant are susceptible to a significant bias.

[5] It is common in repeat-sales estimators to assume x_{nt} does not vary within property across time. All of our results remain when Δx_{nt} is allowed to change.

[6] We focus solely on repeat-sales HPIs given their prevalence in the literature. In unreported results, we find similar improvements when using our methodology to construct hedonic HPIs. See Ghysels et al. (2013) and Bollerslev et al. (2016) for further discussion of alternative HPI methodologies.

2.2 Case-Shiller HPIs

The Case-Shiller index methodology attempts to address the time-varying attribute bias using filters and a robust weighting procedure. The filters remove "sales that occur less than 6 months after a previous sale" and repeat-sales pairs that expericence "changes in the physical characteristics" of the house (CoreLogic, 2018). In addition to the filters, a Robust Interval and Value-Weighted Arithmetic Repeat-Sales (Robust IVWARS) algorithm is used to account for "sale pairs that include anomalous prices or that measure idiosyncratic price changes" (CoreLogic, 2018). The Robust IVWARS procedure assigns a weight of one (no down-weighting) or a weight of less than one but greater than zero for repeat-sale pairs with abnormal price changes.

To facilitate a direct comparison with our quality-adjusted HPIs (see Section 2.4), we first estimate a series of HPIs using Case-Shiller's filters and Robust IVWARS procedure. The exact weights used to create the Case-Shiller HPIs are not publicly available, so we use the weights provided in the descriptive statistics in CoreLogic (2018). Specifically, the weight for each repeat-sales pair that transacted at time periods t' and $t' < t$ is given by

$$w_{t',t} = w^1_{t',t} w^2_{t',t} \tag{3}$$

$$w^1_{t',t} = 1 - 0.5 \times \mathbf{1}(q_{90} < |APA_{t',t}|) - 0.25 \times \mathbf{1}(q_{95} < |APA_{t',t}|) \tag{4}$$

$$w^2_{t',t} = (\hat{a} + \hat{b}(t - t'))^{-0.5} \tag{5}$$

In Equation 3, $w_{t',t}$ assigns a weight based on the absolute value of the annualized price appreciation of the repeat-sales pair, $APA_{t',t}$, and the time between sales, $t - t'$. In Equation 4, repeat-sales pairs with an $APA_{t',t}$ below the 90th percentile are given $w^1_{t',t} = 1$, repeat-sales pairs with an $APA_{t',t}$ between the 90th percentile and the 95th percentile are given $w^1_{t',t} = 0.5$, and repeat-sales pairs with an $APA_{t',t}$ greater than the 95th percentile are given $w^1_{t',t} = 0.25$. Repeat-sales pairs are also weighted based on the time between the first and second sale ($w^2_{t',t}$) under the assumption that houses with longer holding periods are more likely to have underwent physical changes. In Equation 5, $\hat{a}$ and $\hat{b}$ are the least-squares estimates from the regression of the squared residuals from Equation 2 on $t - t'$.

Using the methodology above, we construct Case-Shiller HPIs at both the MSA and local (zip code) level. At the MSA-level, the Robust IVWARS procedure identifies and down-weights extreme price appreciation (e.g. major renovations) based on the statistical distribution of all annualized price changes in the MSA. As HPIs are disaggregated, the procedure's reliance on the statistical distribution of annualized price changes in the local area is a concern since time-varying attribute biases are geographically concentrated. In other words, the weight ($w_{t',t}$) assigned to a repeat-sales pair in Equation 3 may differ at the MSA and local levels because local percentiles may differ from MSA-level percentiles.

To help illustrate the issue with the Robust IVWARS procedure at the local level consider a zip code that is gentrifying. Houses in gentrifying neighborhoods are frequently renovated during the gentrification process. If those renovations are not properly controlled for then the APA associated with the renovations will bias the HPI upwards. Now assume, for simplicity, that only one zip code within the MSA is gentrifying. In this instance, the repeat-sales pairs that were renovated in the gentrifying zip code will likely be down-weighted ($q_{90} < |APA_{t',t}|$) in the MSA-level HPI. However, given the geographic concentration of the renovations within the gentrifying zip code a fraction of the renovations may not be down-weighted in the local HPI since they are no longer abnormal based on the distribution of local APAs. This simple example illustrates that as Case-Shiller HPIs are disaggregated they become increasingly susceptible to a time-varying attribute bias. Thus, it is especially important to identify and mitigate time-varying attribute biases when constructing local HPIs.

2.3 Indicator-adjusted HPIs

Although the Case-Shiller index methodology down-weights repeat-sales pairs with anomalous prices, the HPI is still biased since the anomalous pairs are included. The down-weighting procedure also fails to identify and control for lower intensity improvements since its identification strategy relies on "extreme" pricing anomolies. An alternative approach proposed in the academic literature is to identify renovated properties and include an indicator in Equation 2 as follows

$$\Delta p_{nt} = p_{nt} - p_{nt'} = \Delta\delta_t + \Delta f_{nt}\psi + \Delta\phi_{nt} + \Delta v_{nt} \tag{6}$$

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.[7] 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 sucessive 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.

2.4 Quality-adjusted HPIs

The binary nature of the renovation variable in Equation 6 is a crude control since it offers little insight into the type or intensity of the renovation that occured. We address this limitation using textual analysis. Specifically, we use a new identification strategy that exploits the textual information in an agent's written description of the house when it is listed for sale. Given that recent renovations positively impact both the probability of sale

[7] 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, since 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.

and sales price, the listing agent has a strong incentive to mention recent renovations (if any) in their remarks. Critically, the remarks also offer insight into the type and intensity of the renovation. For example, the remarks can distinguish between cosmetic (new carpet, fresh paint, etc.) and stuctural (additions, remodeled kitchen, etc.) improvements that have differential effects on house prices. In the following subsection we highlight several examples of the relevant time-varying textual information that is provided in the public remarks section of the multiple listing service.

2.4.1 Agent remarks

To more precisely control for the intensity of the renovations performed (if any) between repeat-sales of the same house we incorporate textual information from the listing agent's remarks about the house into the regression. Table 1 displays remarks for three repeat-sales pairs in Washington, D.C. The three pairs were selected to demonstrate the importance of controlling for the intensity of the improvements. A key feature of the repeat-sales pairs is that they provide a remark about the house both before and after the renovation occurred. In other words, the remarks not only provide information about the quality of the house at the time of both transactions, but also allow us to infer if and how much the quality changed between the two transactions.

Take, for example, the first repeat-sales pair in Table 1. The house initially sold for $55,000 in April 2012. Note that the agent's remarks for the first transaction mention that the "house has been condemned." Less than a year later the house sold for $455,000. The quality of the house must have changed substantially to justify the rapid price appreciation. This conjecture is supported by the remarks which note that the subsequent transaction includes "new construction", "custom" finishes, and a "new addition". In theory, the Case-Shiller HPI should discard repeat-sales pair #1 since the index removes pairs in which the house has been physically altered (CoreLogic, 2018). However, the index methodology notes that their data source (deed records) "does not usually describe the physical characteristics of properties (other than the size and alignment of land parcels)", so "it is not possible to identify all of these sales." For this reason, repeat-sales pair #1 would be mistakenly included in the index since the "new addition" mentioned in the remarks is not yet reflected in the house size (sqft) field.

The second and third repeat-sales pairs in Table 1 represent renovations of monotonically decreasing intensity, but increasing frequency within the data. Similar to the first pair, repeat-sales pair #2 includes a major structural renovation that, if not properly controlled for in the repeat-sales estimation, will bias the HPI upwards. The remarks for repeat-sales pair #2 note that the house needed repairs to fix hurricane damage in February 2012. When the house resold seven months later the remarks suggest that those repairs were completed. As expected, the repaired house sold for considerably more. The third repeat-sales pair in Table 1 represents a relatively low intensity renovation. The remarks for repeat-sales pair #3 note that the first transaction was distressed and that cosmetic improvements (e.g. "new carpet", "new paint", "updated appliances") were performed during the nine month holding period. Although the improvements were cosmetic, they still bias the HPI when not properly controlled for.

The three repeat-sales pairs in Table 1 would be included in the Case-Shiller HPI since they are arms-length transactions that, on the surface, were not physically altered.[8] In other words, all three repeat-sales pairs would be considered constant quality. However, a cursory review of the remarks shatters the facade of constant quality. That same cursory review also reveals that numeric representations of the text in the remarks are not readily available. In the next subsection, we demonstrate how to incorporate the text in the remarks into a repeat-sales estimation.

2.4.2 Quality-adjusted index methodology

The textual information in agents' remarks has been shown to improve predictive performance (Nowak and Smith, 2017) and mitigate biased coefficient estimates by controlling for property attributes not recorded or easily quantifiable in traditional data sets (Liu et al., 2019). We advance the methodology in these studies to specifically target the time-varying attribute bias inherent in constant-quality HPIs. To the best of our knowledge, this is the first study to provide a solution that mitigates this well known but often overlooked issue. To

[8] Although the repeat-sales pairs would be included in the Case-Shiller HPI their effect may be partially addressed by the down-weighting procedure.

do so, we assume that ϕ_{nt} in Equation 2 can be approximated using a weighted combination of indicator variables for the presence of relevant tokens in the remarks

$$\phi_{nt} = \sum_{s \in \mathcal{S}} \theta_s r_{nts} + e_{nt} \tag{7}$$

In Equation 7, $\mathcal{S}$ is the set of relevant tokens, θ_s is the weight or implicit price of token s, and e_{nt} is an approximation error. When token s is in the remarks for property n sold at time t then $r_{nts} = 1$, otherwise $r_{nts} = 0$. In Equation 7, the presence of token s is expected to change the transaction price by θ_s. For example, we find the tokens *renovated*, *granite*, and *stainless* all have $0 < \theta_s$ which indicates that listings that include these tokens are expected to sell for more than properties without these attributes, ceteris paribus.

The approximation for $\Delta\phi_{nt}$ can be derived by differencing Equation 7

$$\Delta\phi_{nt} = \sum_{s \in \mathcal{S}} \theta_s \Delta r_{nts} + \Delta e_{nt} \tag{8}$$

Using this approximation, we estimate a quality-adjusted repeat-sales HPI that controls for time-varying attributes of the property. Substituting Equation 8 into Equation 2 yields

$$\Delta p_{nt} = \Delta\delta_t + \sum_{s \in \mathcal{S}} \theta_s \Delta r_{nts} + \Delta e_{nt} + \Delta v_{nt} \tag{9}$$

In Equation 9, the set of relevant tokens in the remarks, $\mathcal{S}$, is assumed to be known. In practice, $\mathcal{S}$ is unknown and estimating Equation 9 is infeasible. A feasible estimator of Equation 9 requires the researcher to either specify $\mathcal{S}$ ex-ante based on prior knowledge or estimate $\mathcal{S}$ from the data. As pointed out in King et al. (2017), human beings excel at associating words and phrases with topics but perform poorly when asked to create a list of all relevant words and phrases associated with a topic from scratch. Since the relevant tokens vary across the MSAs considered in this study, we favor a data-driven approach to identify $\hat{\mathcal{S}}$.

Although the relationship in Equation 9 suggests it is possible to use variable selection methods to identify $\hat{\mathcal{S}}$ as the set of the strongest predictors of Δp_{nt}, the high-dimensional nature of text complicates the process. Specifically, the data-driven approach requires we consider a large set of tokens, $\mathcal{K}$, where $\mathcal{S} \subset \mathcal{K}$. We choose $\mathcal{K}$ as the set of the 2,000 most frequent tokens in the remarks since the results are nearly identical when setting $\mathcal{K}$ equal to either 1,000 or 3,000 candidate tokens. When $\mathcal{K}$ is large, using AIC, BIC, or adjusted R^2 to identify $\hat{\mathcal{S}}$ is computationally prohibitive. Noting this, the Least Absolute Shrinkage and Selection Operator (LASSO) and its variants do not explicitly estimate all possible models but rather use an ℓ_1 penalty that allows for variable selection. More importantly, the LASSO problem is convex and can be easily estimated using numerical methods even when the number of variables is large. When there is heteroscedasticity in the error term, Belloni et al. (2012) suggest the following modification to the LASSO

$$\{\hat{d}, \hat{h}\} = \arg\min_{d,h} \sum_{n,t} \left(\Delta p_{nt} - \Delta d_t - \sum_{k \in \mathcal{K}} h_k \Delta r_{ntk} \right)^2 + \lambda \sum_{k \in \mathcal{K}} |h_k| \upsilon_k \tag{10}$$

In Equation, 10, $0 < \lambda$ is an overall penalty parameter and υ_k is a token-specific penalty term. The shape of the penalty on h_k, $|h_k|$ yields a $\hat{h}$ with some entries equal to zero. A token with a coefficient equal to zero is not included in the final model; conversely, the $\hat{\mathcal{Q}}$ tokens with non-zero coefficients constitute $\hat{\mathcal{S}}$.

It is important to note that $\hat{\mathcal{S}}$ is the set of tokens that are the strongest predictors of price changes to the same house. The set of tokens that are the strongest predictors of price levels can be obtained using a similar variable selection approach and the hedonic model in Equation 2. In other words, $\hat{\mathcal{S}}$ includes only tokens that proxy for time-varying information associated with ϕ_{nt}, but does not include tokens that proxy for time-invariant information associated with μ_n. Thus, $\hat{\mathcal{S}}$ may not be appropriate for use in hedonic models. Calculating quality-adjusted hedonic HPIs is possible but is beyond the scope of this study.

Given $\hat{\mathcal{S}}$, Equation 9 can be estimated using least-squares. Define $\hat{\delta}_t^*$ as the least-squares estimate of δ_t based on Equation 9 using $\hat{\mathcal{S}}$ and $\hat{\delta}_t$ as the least-squares estimate of δ_t based on Equation 2. The Theorem below provides necessary conditions for token s to adjust the HPI. In the appendix, we provide an additional proof to demonstrate that $\hat{\delta}_t^*$ is equal to a HPI estimated without tokens, $\hat{\delta}_t$, plus a correction factor.

Theorem (HPI Adjustment Theorem)**.** *The least-squares estimates of the repeat-sales HPI* *with tokens, $\hat{\delta}_t$, and without tokens, $\hat{\delta}_t^*$, are related by*

$$\hat{\delta}_t = \hat{\delta}_t^* - \sum_{s \in \hat{\mathcal{S}}} \hat{\pi}_{st} \hat{\theta}_s \tag{11}$$

where $\hat{\pi}_{st}$ is the coefficient from a regression of differenced r_{nts} on differenced indicators for the time periods of sale for each repeat-sales transaction pair.

The Theorem indicates $\hat{\delta}_t^*$ is equal to $\hat{\delta}_t$ plus a weighted sum of the implicit prices for the tokens, $\hat{\theta}_s$. The $\hat{\pi}_{st}$ are the least-squares coefficients from a differenced linear probability model

$$\Delta r_{nts} = \Delta \pi_{st} + \Delta \epsilon_{nts} \tag{12}$$

Similar to the repeat-sales estimators in Equation 12, the differenced indicator variables for the tokens are regressed on the differenced indicators for the time period. When $0 < \pi_{st}$ ($\pi_{st} < 0$), token s is more (less) likely to appear in the agents' remarks in time period t.

When $\hat{\pi}_{st}\hat{\theta}_s = 0$ in Equation 11, token s has no impact on the HPI. Two conditions must be satisfied for token s to have an impact on the HPI. First, when $\hat{\theta}_s = 0$, token s has no impact on price and cannot impact the HPI. Second, when $\hat{\pi}_{st} = 0$, the token does not appear in remarks more frequently in time period t than in other time periods. Therefore, the token is not correlated with the indicator variable for sales in time period t, d_t, and cannot impact the HPI. When either of these conditionas are met, $\hat{\pi}_{qt}\hat{\theta}_s = 0$ and the omission of information related to token s does not bias the HPI up or down in any time period.

3 Data overview

The multiple listing service (MLS) data used in this study was provided by Redfin. The MLS data includes transaction level information for eight large MSAs that were selected based on historical data availability.[9] Using MLS data ensures the transactions are arms-

[9] Redfin is a publicly traded real estate brokerage firm that uses modern technology to help people buy and sell houses. Redfin launched its home-buying and selling (i.e. brokerage) services in Seattle in 2006. Over time Redfin has expanded to over 80 markets across the United States. In some markets the local MLS does not provide historical data to new brokerage firms.

length (similar to the Case-Shiller HPI) and offers the added benefit of panoptic coverage.[10] The transaction level data includes physical and locational housing attributes as well as information about the transaction itself, such as the transaction price and date, that we use to create the repeat-sales HPIs.

The MLS data also includes the agents' public remarks. The remarks are unstructured textual descriptions of the house that are entered by the listing agent. Listing agents, who have presumably toured both the inside and outside of the house, use the remarks to highlight important information about the house that may not be available in other areas of the MLS listing. In other words, the unstructured format of the remarks allows the agent to comment on features of the house that are difficult to quantify (e.g. quality). Since the remarks are entered by the listing agent when the house is listed for sale on the MLS we are able to identify and control for changes to the quality of the house (if any) that occured since the previous listing. Examples of the remarks are displayed in Table 1.

Prior to creating the repeat-sales HPIs we apply several filters to clean the transaction data. We also preprocess the text in the remarks. In unreported results, we find that the data filters and text preprocessing have a modest effect on $\hat{\mathcal{S}}$ but a negligible effect on $\hat{\delta}_t^*$. Because our primary focus is on the HPI, we provide additional insight into the data filters, descriptive statistics at the MSA-level, and text preprocessing procedure in an internet appendix.

4 Results

4.1 MSA HPIs

Figure 1 plots two repeat-sales HPIs for Miami, Phoenix, San Francisco, and Washington D.C. in Panels A to D, respectively. The first HPI in each panel is a Case-Shiller HPI that includes a 95% confidence interval. The second HPI in each panel is a quality-adjusted HPI that controls for the time-varying attributes of real estate using $\hat{\mathcal{S}}$. The Case-Shiller and quality-adjusted HPIs are estimated using the methodology in Sections 2.2 and 2.4,

[10] Real estate agents were involved in approximately 88% of all housing transactions in 2016 NAR (2016).

respectively.

The difference between the HPIs in each MSA demonstrates the importance of controlling for time-varying housing attributes when constructing repeat-sales HPIs. The results also highlight homeowners' (dis)incentive to perform maintenance and renovations throughout the market cycle.[11] For example, the Washington D.C. HPIs in Panel D show that the Case-Shiller index is biased upwards by as much as 7% leading up to the financial crisis (2002-2007), biased downwards by as much as 7% during the financial crisis (2008-2012), and biased upwards by as much as 6% soon after the financial crisis (2013-2017). Given that homeowners have more (less) incentive to maintain and renovate their house during a rising (falling) market, these results suggest the bias is directly related to time-varying attributes (i.e. changes to the quality) of the house.

Although a similar pattern is found in Miami (pre-crisis: 9%, crisis: -5%, post-crisis: 9%) and Phoenix (pre-crisis: 2%, crisis: -4%, post-crisis: 7%), the magnitude and timing of the bias varies across the MSAs. For example, the San Francisco Case-Shiller HPI is biased upwards by as much as 21% after the financial crisis, but does not experience a downward bias during the financial crisis. In an internet appendix we provide similar plots for four additional MSAs (Baltimore, Boston, Los Angeles, and Portland) to further highlight the relationship between the time-varying attribute bias and market cycle.

Figure 2 provides additional support for our conjecture that the bias inherent in constant-quality HPIs is directly related to time-varying housing attributes. The figure displays the contribution of each token to the difference between the HPIs in Figure 1 over time. We only plot tokens that adjust the constant-quality HPI more than 0.005 or less than -0.005 in at least one time period. As emphasized by Equations 11 and 12, the only tokens that adjust the HPI up or down are those with non-zero implicit prices whose frequency varies over time. This necessary condition precludes certain tokens from adjusting the HPI despite having a large implicit price.

Several interesting insights emerge from Figure 2. First, the tokens in Figure 2 represent time-varying attributes that relate to either a sales condition associated with the transac-

[11] We plot the difference between the two HPIs in an internet appendix to more clearly highlight the relationship between the time-varying attribute bias and the market cycle.

tion or the quality of the property. For example, tokens identifying distressed transactions (e.g. *shortsale*, *homepath*, bank *owned*, *hud*) are present in all four panels. Similarly, tokens identifying property improvements (e.g. *new*, *renovated*, *updated*) are present in each panel. Unsurprisingly, the distressed tokens' contribution are negative and the property improvement tokens' contribution are positive.

Second, the contribution of the tokens varies throughout the market cycle. As expected, the distressed tokens are not only more likely to adjust the Case-Shiller HPI during the crisis period, but the contribution of the distressed tokens is also the largest during this period. Take, for example, the *homepath* token in Miami (Panel A) and Washington D.C. (Panel D). The contribution of the token, which identifies foreclosed houses owned by Fannie Mae, is considerably larger in 2009-2011 relative to 2014-2017. In contrast, the contribution of the tokens that identify property improvements are considerably larger during the post-crisis period relative to the crisis period.

Third, the MSA-specific tokens highlight the efficacy of the data-driven variable selection process we employ. For example, the *impact* token is unique to Miami in Panel A. The *impact* token identifies "impact resistant" hurricane windows that are required for all new construction in southern Florida per the 2002 change to Florida's building code. Given the high cost (approximately $500 to $600 for each single hung window) and discounts offered by insurance companies, it is no surprise that houses that install hurricane impact resistant windows sell for a considerable premium in Miami. If not properly controlled for in a repeat-sales HPI the installation of these windows between repeat-sales will bias the HPI upwards.

4.2 Local HPIs

Over the past decade researchers have become increasingly interested in price dynamics within MSAs. The surge of interest coincides with increased data availability as both CoreLogic and the Federal Housing Financing Administration (FHFA) now produce annualized local HPIs at the 5-digit zip code level. Using the newly available local HPIs, recent studies demonstrate that the underlying assumption of homogeneous price dynamics in MSA-level HPIs (i.e. Figure 1) is tenuous by documenting significant heterogeneity in the price dynamics of local HPIs (Landvoigt et al., 2015; Bogin et al., 2018). However, the HPIs employed in these studies do not control for the time-varying attribute bias that we document in the previous section.

In this section, we show that the time-varying attribute bias is geographically concentrated within MSAs and investigate the extent to which the geographic concentration of the bias affects local HPIs. We document the geographic concentration of the time-varying attribute bias by testing for localization of the set of down-weighted properties in the Case-Shiller HPI, $\mathcal{L}$, using the non-parametric method described in Duranton and Overman (2005). More specifically, the empirical density of pairwise distances between all elements (locations) in $\mathcal{L}$, $f(\mathcal{L})$, is used to test for localization. Significance is determined using pointwise confidence intervals based on repeated sampling from the complement of $\mathcal{L}$, $\mathcal{L}^c$. When there is extra mass near zero in $f(\mathcal{L})$ compared to the randomly drawn sets from $\mathcal{L}^c$, this is evidence that $\mathcal{L}$ is significantly localized.

Figure 3 displays $f(\mathcal{L})$ and pointwise confidence intervals when the pairwise distances are placed into buckets 0.5 miles wide. Figure 3 indicates there is significant localization of down-weighted repeat-sale pairs in Phoenix, San Francisco, and Washington D.C. as a significant portion of $f(\mathcal{L})$ lies above the confidence intervals for distances near zero. In an internet appendix we also find evidence of significant localization of the down-weighted repeat-sales pairs with anomolous prices in Baltimore, Boston, Los Angeles, and Portland. The only exception is Miami where there appears to be significantly less localization of the down-weighted repeat-sale pairs.

Next we examine whether the geographic concentration of the time-varying attribute bias affects local HPIs. To do so, we create and compare local Case-Shiller HPIs to our local quality-adjusted HPIs. We construct the local Case-Shiller and quality-adjusted HPIs using the methodology outlined in Sections 2.2 and 2.4. Although the methodology is similar, the local quality-adjusted HPIs differ from their MSA-level counterparts in several ways. First, they are constructed at the zip code level, instead of MSA. Second, they are constructed on an annual basis, instead of quarterly. Third, the set of tokens used as controls in the estimation of local HPIs is zip code specific.

Following the FHFA criteria detailed in Bogin et al. (2018), we only construct local HPIs for zip codes with at least 100 repeat-sales transactions and filter out transaction pairs with a holding period less than or equal to a year and/or an annual average appreciation rate greater (less) than 30% (-30%). When constructing the Case-Shiller HPI (sans tokens) we also apply the Robust IVWARS procedure. In Figure 4, the difference between our quality-adjusted local HPIs and the Case-Shiller local HPIs is set to 0 in 2001 (Panels A to C) or the second year the data is available (2003 in Panel D). For each year and MSA, we then calculate several intervals of interest to highlight the distribution, magnitude, and cyclicality of the time-varying attribute bias within the MSA. The intervals include the entire range, 5th and 95th percentiles, and the interquartile range (25th and 75th percentiles).

As shown in the HPI Adjustment Theorem, when the relative frequency of token s does not vary across time within the zip code, token s does not impact the HPI. If no bias exists, then no tokens are included in the estimation of the local quality-adjusted HPIs and the range in Figure 4 will be exactly zero in every time period. However, the results in Figure 4 document the presence of a significant bias, thereby highlighting the acute need to control for time-varying housing attributes when constructing local HPIs.

Several additional stylized facts emerge from the results in Figure 4. First, the magnitude of the time-varying attribute bias varies considerably by zip code within each MSA. This finding, along with the localization plots in Figure 3, supports our conjecture that the time-varying attribute bias is geographically concentrated within the MSA. Second, the magnitude of the range varies throughout the market cycle. For example, the magnitude of the range in San Francisco (Panel C) is much larger after the financial crisis (2015-2017) relative to the pre-crisis period. Similarly, the magnitude of the bias at the extremes is much larger than the bias at the MSA level in all four panels. Third, the direction of the bias varies throughout the market cycle. For example, the range in Miami and Phoenix is clearly skewed downwards during the financial crisis and upwards after the financial crisis. These findings further support the conjecture that the type and intensity of housing improvements differ systematically throughout the market cycle.

4.3 Robustness checks

The results in Figures 1 to 4 indicate that the tokens control for time-varying attributes of the property. In an internet appendix we investigate, among other things, the extent to which conventional controls for distressed transactions, flips, and renovations obviate the need for the tokens. To do so, we either include an indicator variable for these transactions in the repeat-sales estimation or filter out these transactions. The indicator for a distressed sale is equal to 1 if the transaction is either real estate owned (REO) or a short sale. The indicator variable for a flip is equal to 1 if the holding period was less than 12 months. The indicator variable for a renovation is equal to 1 if the property was renovated in the past 12 months. We then construct and compare the indicator-adjusted HPIs (without tokens) to the quality-adjusted HPIs (with tokens). If the information in the tokens is redundant after controlling for the three transaction types, the quality-adjusted HPI should not differ from the indicator-adjusted HPI. The results of these robustness checks clearly indicate that the tokens control for information above and beyond that which is conveyed by the indicator variables. Furthermore, the time-varying attribute bias persists when distressed transactions, flips, and renovations are filtered out of the repeat-sales sample. These findings indicate that our improved quality-adjusted HPIs are not simply a byproduct of including heterogeneous transaction types in the repeat-sales estimation.

5 Conclusion

We find evidence of a significant time-varying attribute bias in repeat-sales HPIs that use the Case-Shiller index methodology (Case and Shiller, 1989; CoreLogic, 2018). The direction, magnitude, and source of the bias varies throughout the market cycle and across MSAs, thereby complicating its resolution. We also find that the bias is geographically concentrated within MSAs, further complicating the construction and interpretation of local HPIs. Harnessing recent advances in machine learning and computing power, we provide a data-driven textual analysis approach that mitigates the time-varying attribute bias. Our approach identifies the relevant text in agents' written remarks about the house and includes the textual information as controls in the repeat-sales estimation. The use of repeat-sales allows us to not only identify the quality of the house during successive transactions, but also infer and control for changes to the quality of the house when estimating the HPI. In doing so, we disentangle pure price changes from those associated with changes in quality - thereby creating a quality-adjusted HPI.

References

Adelino, M., Schoar, A., and Severino, F. (2015). House prices, collateral, and self-employment. *Journal of Financial Economics*, 117(2):288–306.

Bailey, M. J., Muth, R. F., and Nourse, H. O. (1963). A regression method for real estate price index construction. *Journal of the American Statistical Association*, 58(304):933–942.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

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.

Berger, D., Guerrieri, V., Lorenzoni, G., and Vavra, J. (2017). House prices and consumer spending. *The Review of Economic Studies*, 85(3):1502–1542.

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.

Bogin, A., Doerner, W., and Larson, W. (2018). Local house price dynamics: New indices and stylized facts. *Real Estate Economics*, Forthcoming.

Bollerslev, T., Patton, A. J., and Wang, W. (2016). Daily house price indices: Construction, modeling, and longer-run predictions. *Journal of Applied Econometrics*, 31(6):1005–1025.

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

Case, K. E. and Shiller, R. J. (1989). The efficiency of the market for single-family homes. *The American Economic Review*, 79(1):125–137.

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

CoreLogic (2018). S&p corelogic case-shiller home price indices methodology. *S&P Dow Jones Indices, Version: April 2018*, pages 1–34.

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

Ghysels, E., Plazzi, A., Valkanov, R., and Torous, W. (2013). Forecasting real estate prices. In *Handbook of economic forecasting*, volume 2, pages 509–580. Elsevier.

Goetzmann, W. N. and Spiegel, M. (1995). Non-temporal components of residential real estate appreciation. *The Review of Economics and Statistics*, 77(1):199–206.

Guerrieri, V., Hartley, D., and Hurst, E. (2013). Endogenous gentrification and housing price dynamics. *Journal of Public Economics*, 100:45–60.

Gyourko, J., Mayer, C., and Sinai, T. (2013). Superstar cities. *American Economic Journal: Economic Policy*, 5(4):167–99.

King, G., Lam, P., and Roberts, M. E. (2017). Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*.

Landvoigt, T., Piazzesi, M., and Schneider, M. (2015). The housing market (s) of san diego. *American Economic Review*, 105(4):1371–1407.

Liu, C., Nowak, A., and Smith, P. (2019). Asymmetric or incomplete information about asset values? *The Review of Financial Studies*, forthcoming.

Loutskina, E. and Strahan, P. E. (2015). Financial integration, housing, and economic volatility. *Journal of Financial Economics*, 115(1):25–41.

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

Mian, A., Rao, K., and Sufi, A. (2013). Household balance sheets, consumption, and the economic slump. *The Quarterly Journal of Economics*, 128(4):1687–1726.

Mian, A. and Sufi, A. (2009). The consequences of mortgage credit expansion: Evidence from the us mortgage default crisis. *The Quarterly Journal of Economics*, 124(4):1449–1496.

Mian, A. and Sufi, A. (2014). What explains the 2007–2009 drop in employment? *Econometrica*, 82(6):2197–2223.

NAR (2016). National association of realtors 2016 profile of home buyers and sellers.

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

Palmquist, R. B. (1980). Alternative techniques for developing real estate price indexes. *The Review of Economics and Statistics*, 73(1):442–448.

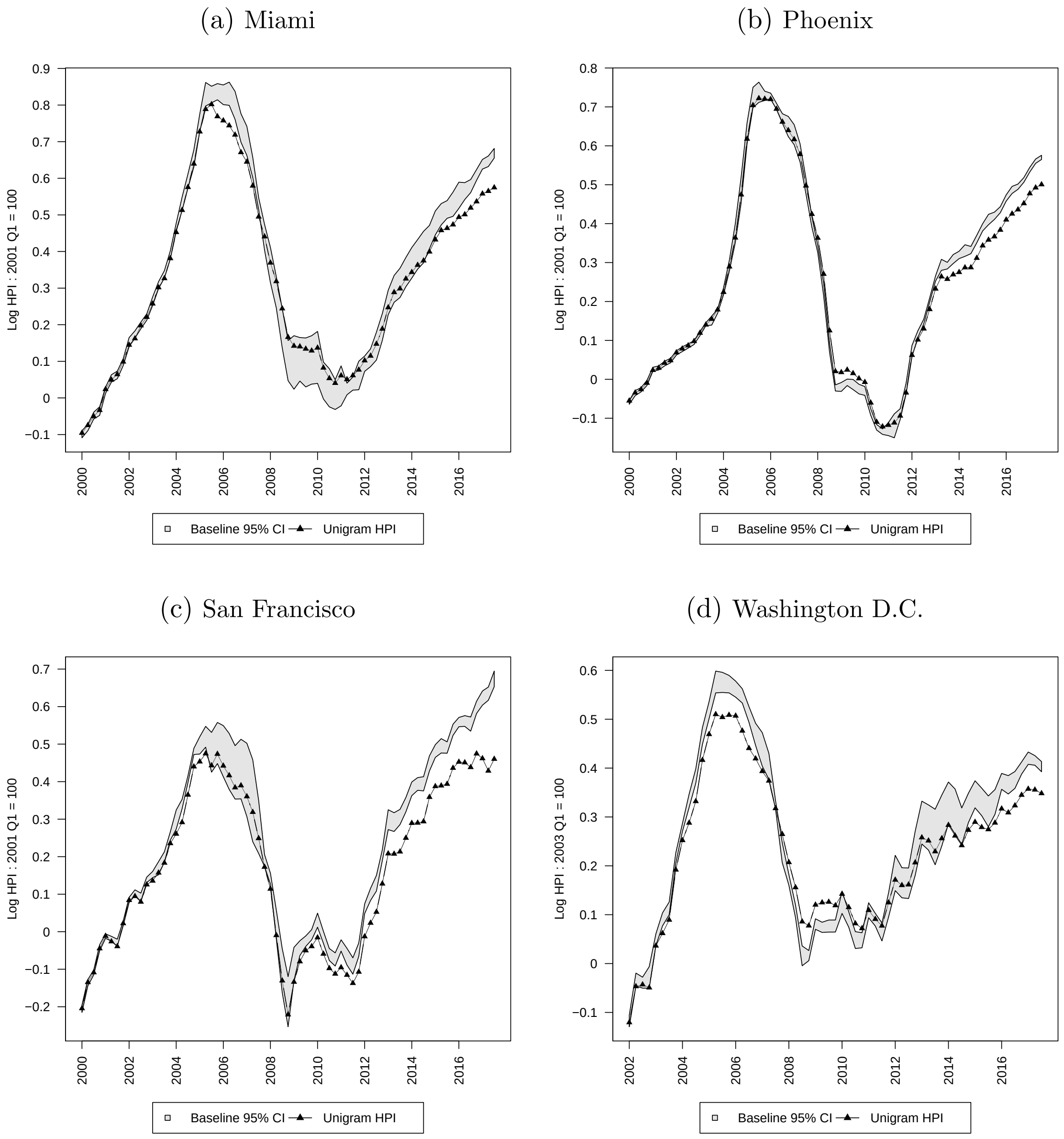
Van Nieuwerburgh, S. and Weill, P. O. (2010). Why has house price dispersion gone up? *The Review of Economic Studies*, 77(4):1567–1606.

Table 1: Remarks on the varying intensity of renovations

<table>
<tr><td colspan="6">Repeat-sales pair #1</td></tr>
<tr><th>MSA</th><th>Zip code</th><th>Sale date</th><th>Price</th><th>Sqft</th><th>Renovated</th></tr>
<tr><td>DC</td><td>20112</td><td>04/12/2012</td><td>$55,000</td><td>1,769</td><td>No</td></tr>
<tr><td colspan="6">MLS Remark: Location! Wooded 1 acre lot with well and septic. House has been condemned, please check with Health Dept. for well and septic status. Do not enter the home on property. Email lister with any questions. Sign on property. Show anytime.</td></tr>
<tr><td>—"—</td><td>—"—</td><td>02/27/2013</td><td>$455,000</td><td>—"—</td><td>Yes</td></tr>
<tr><td colspan="6">MLS Remark: 90% New Construction on main house plus new addition. Green home w custom everything. Don't settle for builder grade. Maintenance free exterior with vinyl siding, trex deck. Wired in speaker system, Cat 5, RG6. Hardwood floors. Upgraded carpet. Custom cabinets, tile work, fire place, built ins, cathedral ceilings, recessed lighting, up lighting, granite, high end appliances. Too much to list!</td></tr>
<tr><td colspan="6">Repeat-sales pair #2</td></tr>
<tr><th>MSA</th><th>Zip code</th><th>Sale date</th><th>Price</th><th>Sqft</th><th>Renovated</th></tr>
<tr><td>DC</td><td>22066</td><td>02/11/2012</td><td>$426,500</td><td>1,440</td><td>No</td></tr>
<tr><td colspan="6">MLS Remark: Great opportunity to buy in close-in Great Falls and repair/renovate to suit after hurricane damage; wonderful quiet street, walk to Riverbend Park at end of street; Property sold strictly "as is".</td></tr>
<tr><td>—"—</td><td>—"—</td><td>09/17/2012</td><td>$595,000</td><td>—"—</td><td>Yes</td></tr>
<tr><td colspan="6">MLS Remark: Completely renovated home in close-in Great Falls, quiet neighborhood setting. New kitchen and baths, hardwoods throughout the main, spacious lower level with 2 bonus rooms for in-law or au pair. Large rec room with wet bar. Ample closet and storage space. Best value in Langley school district. At end of Weant Drive, enjoy nature trails and parkland along the Potomac River.</td></tr>
<tr><td colspan="6">Repeat-sales pair #3</td></tr>
<tr><th>MSA</th><th>Zip code</th><th>Sale date</th><th>Price</th><th>Sqft</th><th>Renovated</th></tr>
<tr><td>DC</td><td>20109</td><td>08/08/2008</td><td>$134,000</td><td>1,066</td><td>No</td></tr>
<tr><td colspan="6">MLS Remark: Excellent bank owned property in great location! 4 bdrms, 2.5 baths., deck and fenced rear yard. Countrywide pre-qual required for offer submission. Free appraisal & credit report if financed thru Countrywide. Seller to choose settlement agent.</td></tr>
<tr><td>—"—</td><td>—"—</td><td>05/09/2009</td><td>$232,000</td><td>—"—</td><td>Yes</td></tr>
<tr><td colspan="6">MLS Remark: Great 4 bedroom detached home at end of street. Walk to school from move-in ready brick front raised rambler with new carpet and new paint throughout, laminated, hardwood floors and updated kitchen appliances and updated bathrooms vanities. Level, partially fenced backyard. Free one year home warranty.</td></tr>
</table>

Notes: Table 1 displays three repeat-sales transaction pairs in which the property was purchased, rehabbed, and sold within one year. The repeat-sales pairs were selected to provide examples of renovations of varying intensity. The intensity of the renovation monotonically decreases from repeat-sales pair #1 to #3.

Figure 1: Quality-adjusted repeat-sales HPIs



(a) Miami

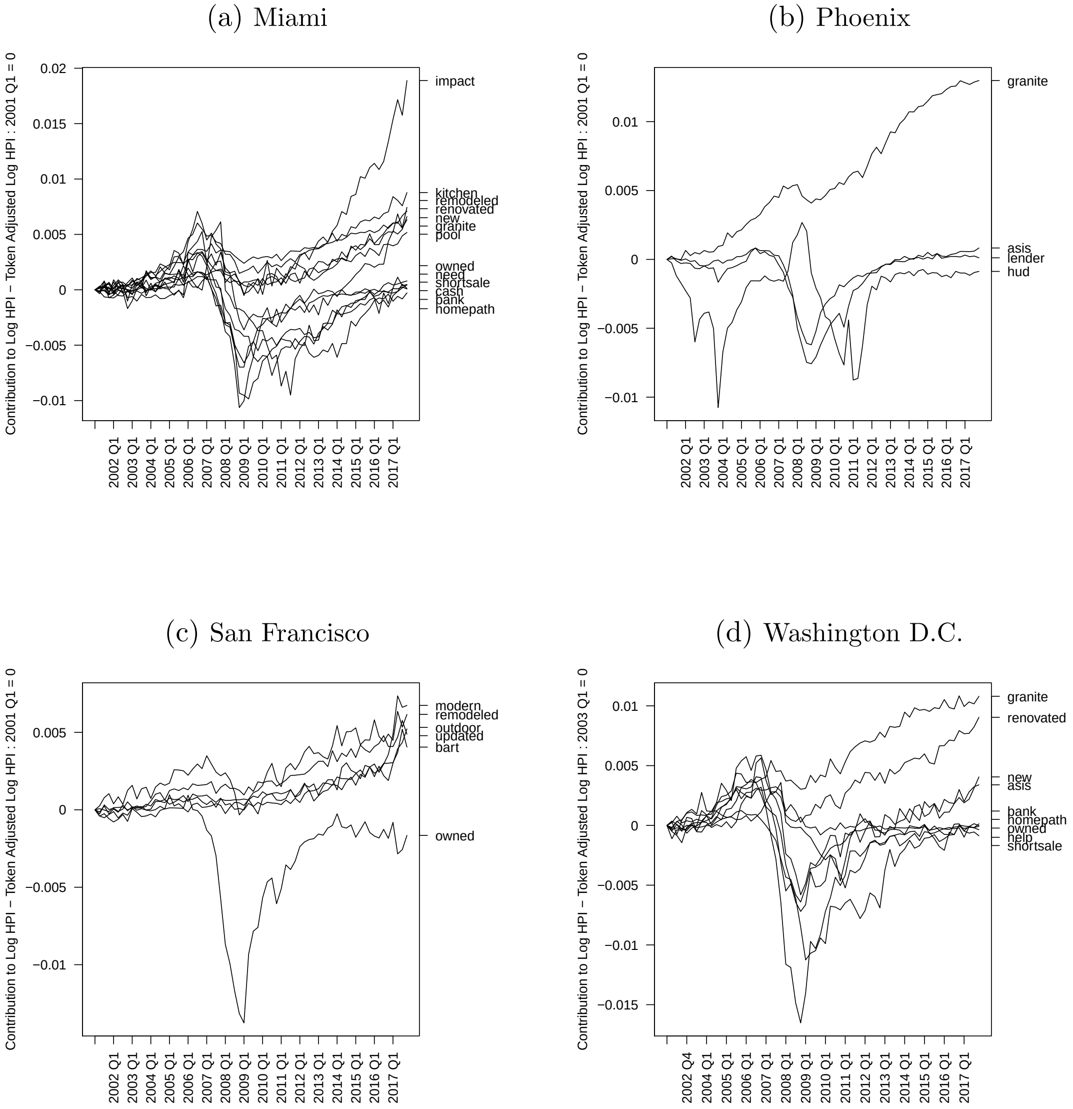
(b) Phoenix

(c) San Francisco

(d) Washington D.C.

Notes: Figure 1 displays the Log HPI from a repeat-sales estimation incorporating tokens as controls for time-varying attributes for Miami, Phoenix, San Francisco, and Washington D.C 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 2.4.

Figure 2: Contribution of the tokens to the difference in Log HPI



(a) Miami

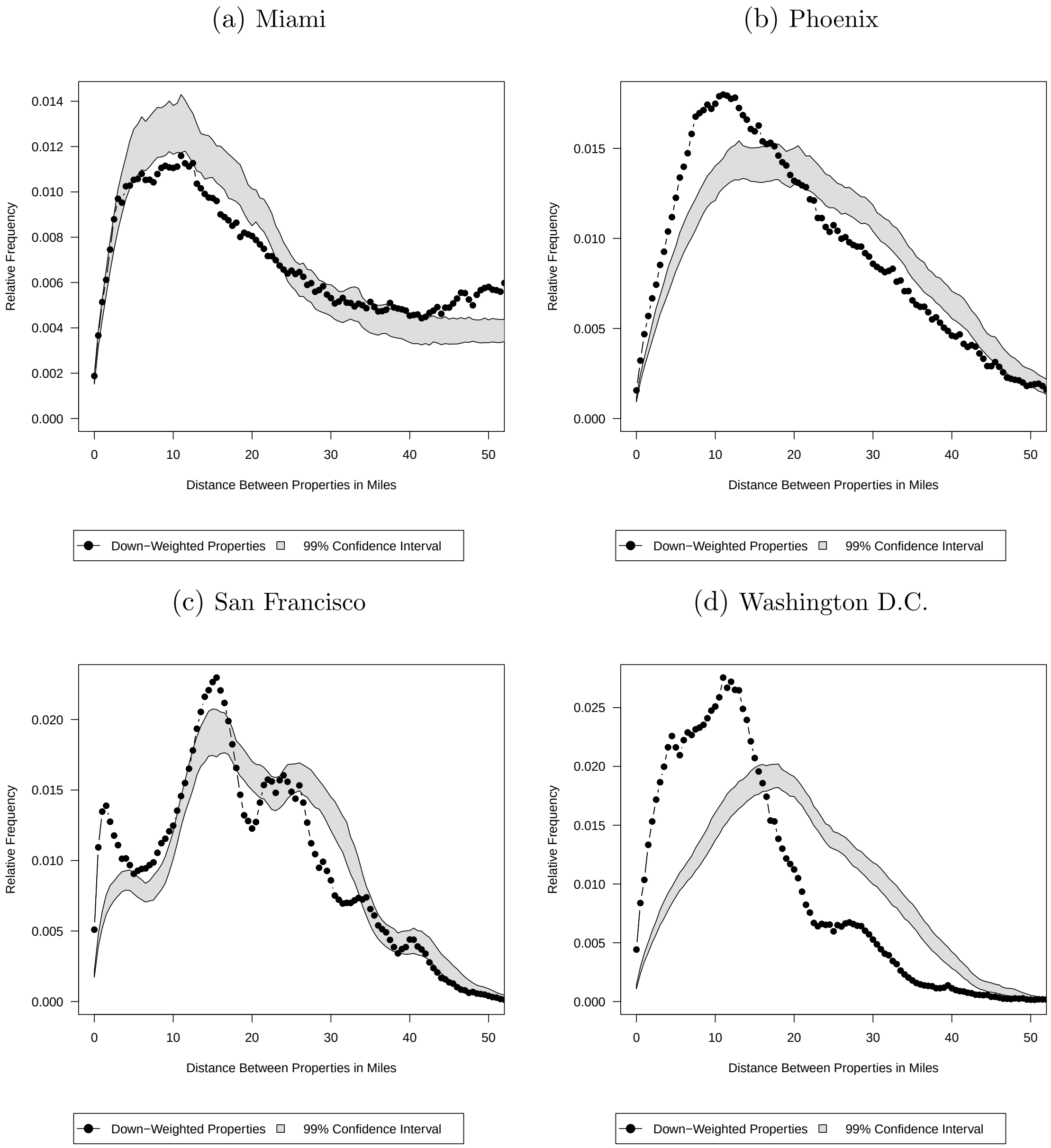
(b) Phoenix

(c) San Francisco

(d) Washington D.C.

Notes: Figure 2 displays the contribution of tokens to the difference between the two HPIs in Figure 1. Only tokens that increase the Log HPI more than 0.005 or less than −0.005 in any time period are included. 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 labels at the right-hand side are associated with the contribution of the token to the difference between the two Log HPIs in the final time period. The labels are adjusted up or down for presentation purposes.

Figure 3: Geographic concentration of time-varying attribute bias


(a) Miami

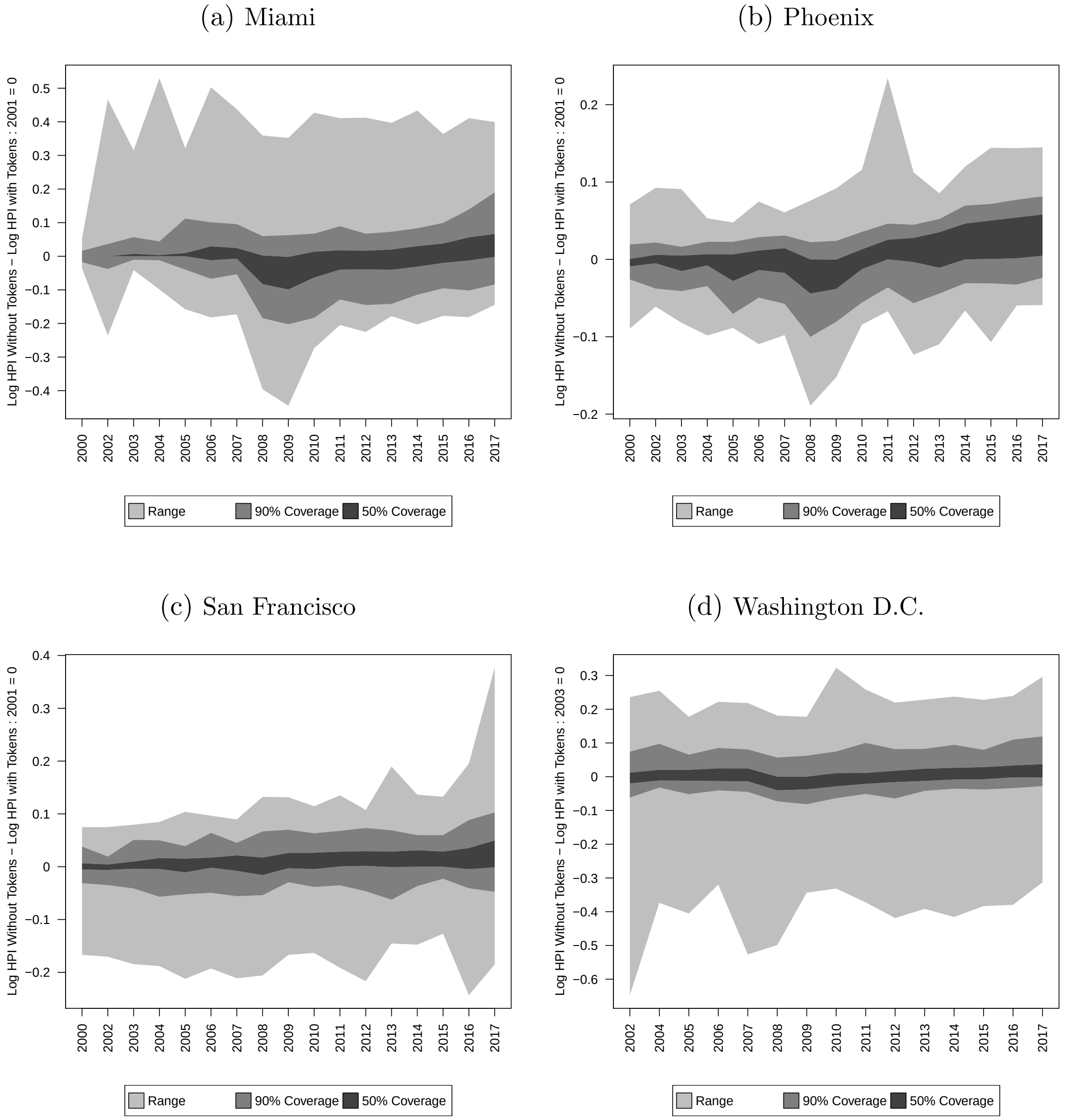
(b) Phoenix

(c) San Francisco

(d) Washington D.C.

Notes: Figure 3 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 4: Dispersion of differences between zip code HPIs with and without tokens


(a) Miami

(b) Phoenix

(c) San Francisco

(d) Washington D.C.

Notes: Figure 4 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 132, 145, 129, and 156 unique zip codes in Miami, Phoenix, San Francisco, and Washington D.C., 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.

Appendices

Contents

A Proof for HPI Adjustment Theorem (Internet Appendix)

Proof. Define N as the number of repeat-sales and T as the number of time periods. Excluding the first time period in order to avoid perfect multicollinearity, define D as the $N \times T-1$ matrix where $D_{nt} = 1$ 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 in the repeat-sales pair occurs in time period t. For $\mathcal{Q} = |\hat{\mathcal{S}}|$, define R as the $N \times Q$ matrix of differenced indicator variables where $R_{nr} = 1$ if the second sale in the repeat-sales pair contains token q, $R_{nr} = -1$ if the first sale in the repeat-sales pair contains token s, and $R_{ns} = 0$ if neither sale or both sales in the repeat-sales pair contain token s. 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

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

where y is the vector of transaction prices. The normal equations for the least-squares price index when controlling for tokens, and the least-squares implicit prices for the $\mathcal{Q}$ tokens are given by

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

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

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

Where $\hat{\pi}_s = (\hat{\pi}_{s2}, \hat{\pi}_{s3}, \ldots, \hat{\pi}_{sT})'$. This implies

$$\hat{\delta} = \hat{\delta}^* - \sum_{s \in \hat{\mathcal{S}}} \hat{\pi}_s \hat{\theta}_s$$

□

B Data Overview (Internet Appendix)

B.1 Data filters

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

Some, but not all, of the represented counties overlap with the represented counties used to construct the Case-Shiller HPIs. For example, the three counties represented in our Miami repeat-sales data are identical to the counties represented in Case-Shiller's "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 Case-Shiller's "Portland-Vancouver-Beaverton, OR-WA" HPI. That said, the four counties included in our repeat-sales data represent the core of the Portland MSA.

Table B2: Descriptive statistics for repeat-sales by MSA

	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
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					
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					

	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
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					
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 processed in the following order:

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.

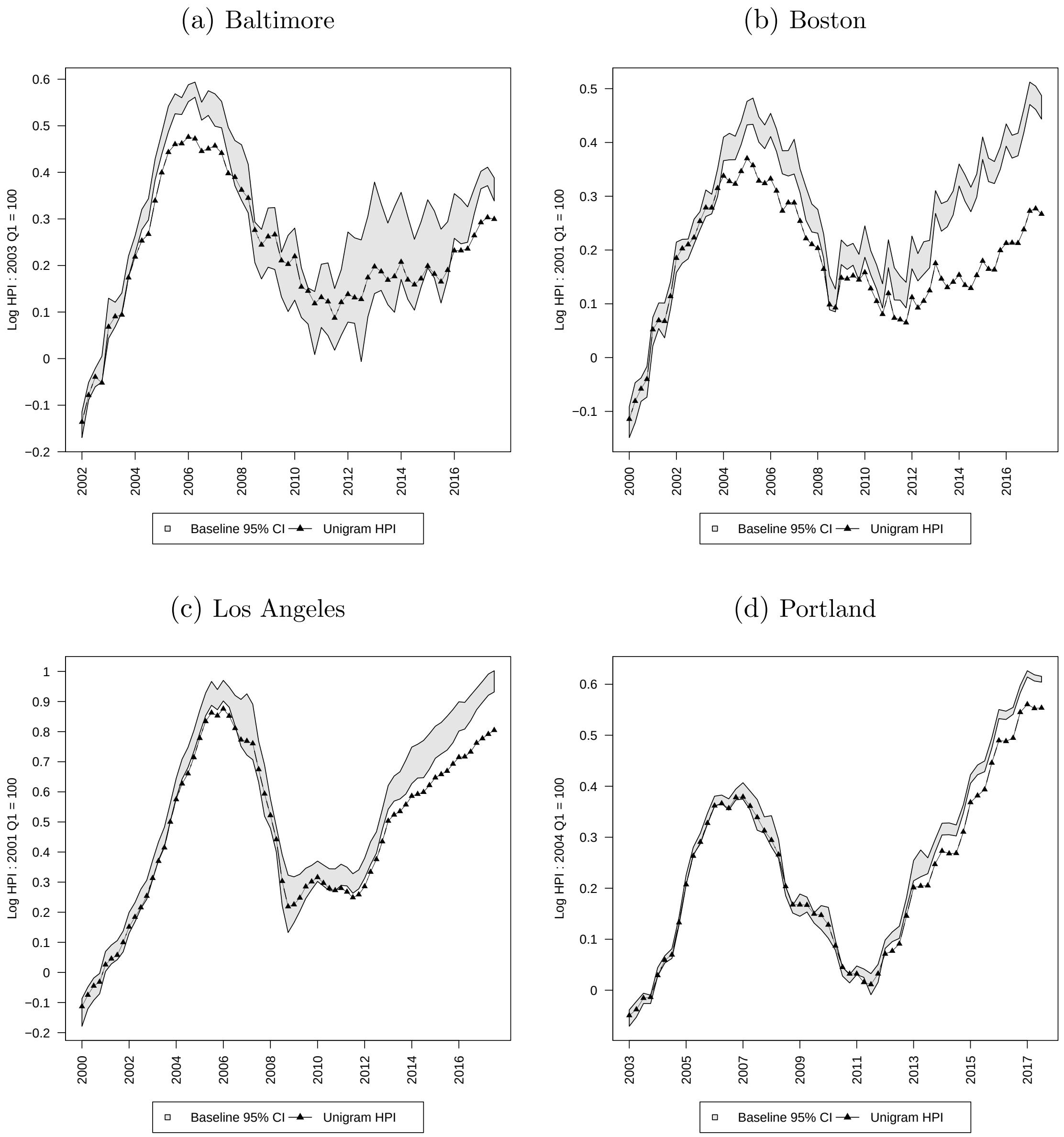
In unreported results we find that additional preprocessing, such as stemming the remarks, has a neglible effect on the results we report. See Section D.2 for additional discussion.

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 four additional MSAs: Baltimore, Boston, Los Angeles, and Portland. The MSAs were selected primarily based on historical data availability, but also to provide variation in terms of location-specific attributes such as price, supply constraints, and the underlying housing stock. 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 four additional MSAs.

Figure C1: Additional MSA-level quality-adjusted HPIs



(a) Baltimore

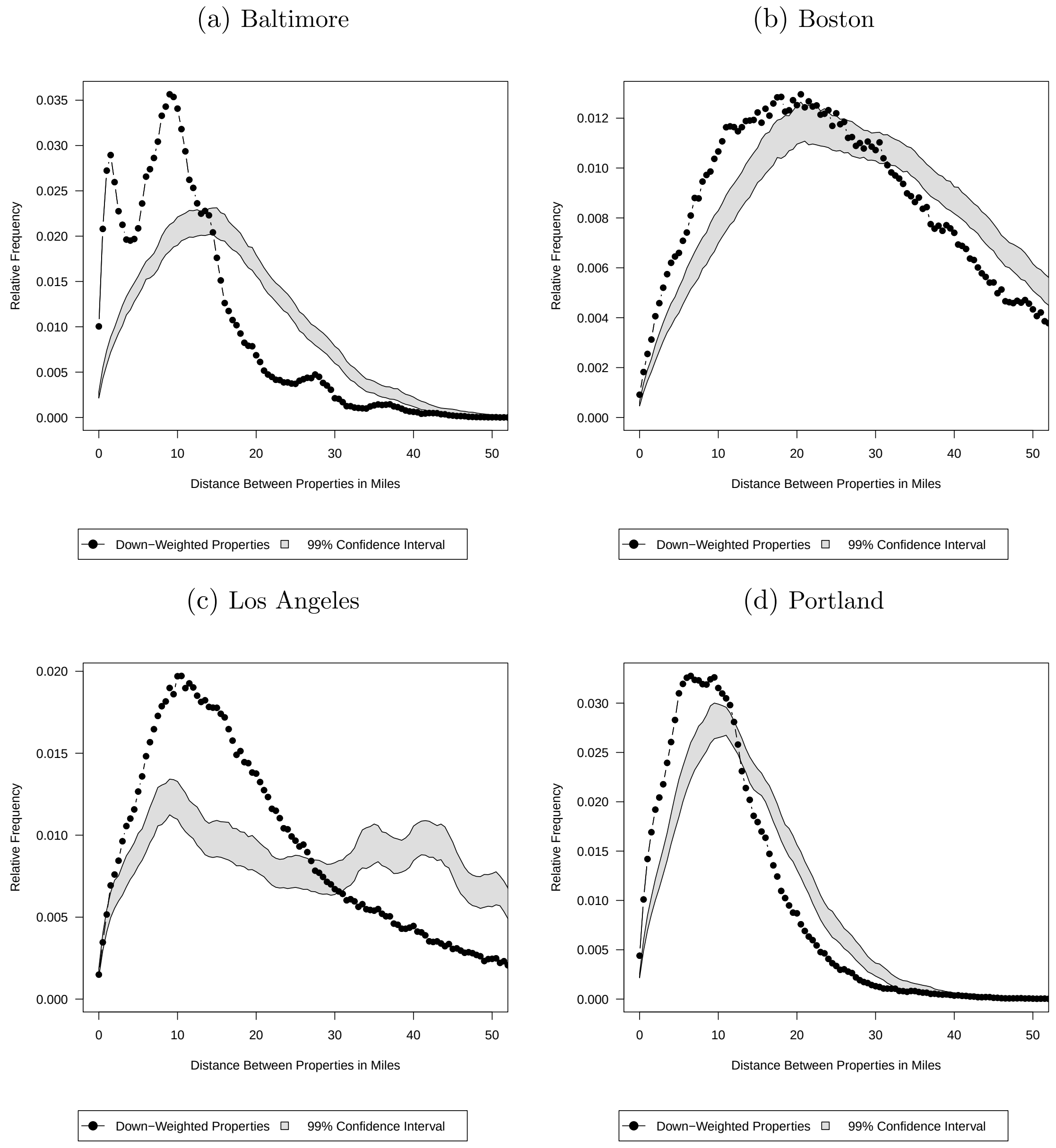
(b) Boston

(c) Los Angeles

(d) Portland

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 2.4.

Figure C2: Geographic concentration of bias for additional MSAs


(a) Baltimore

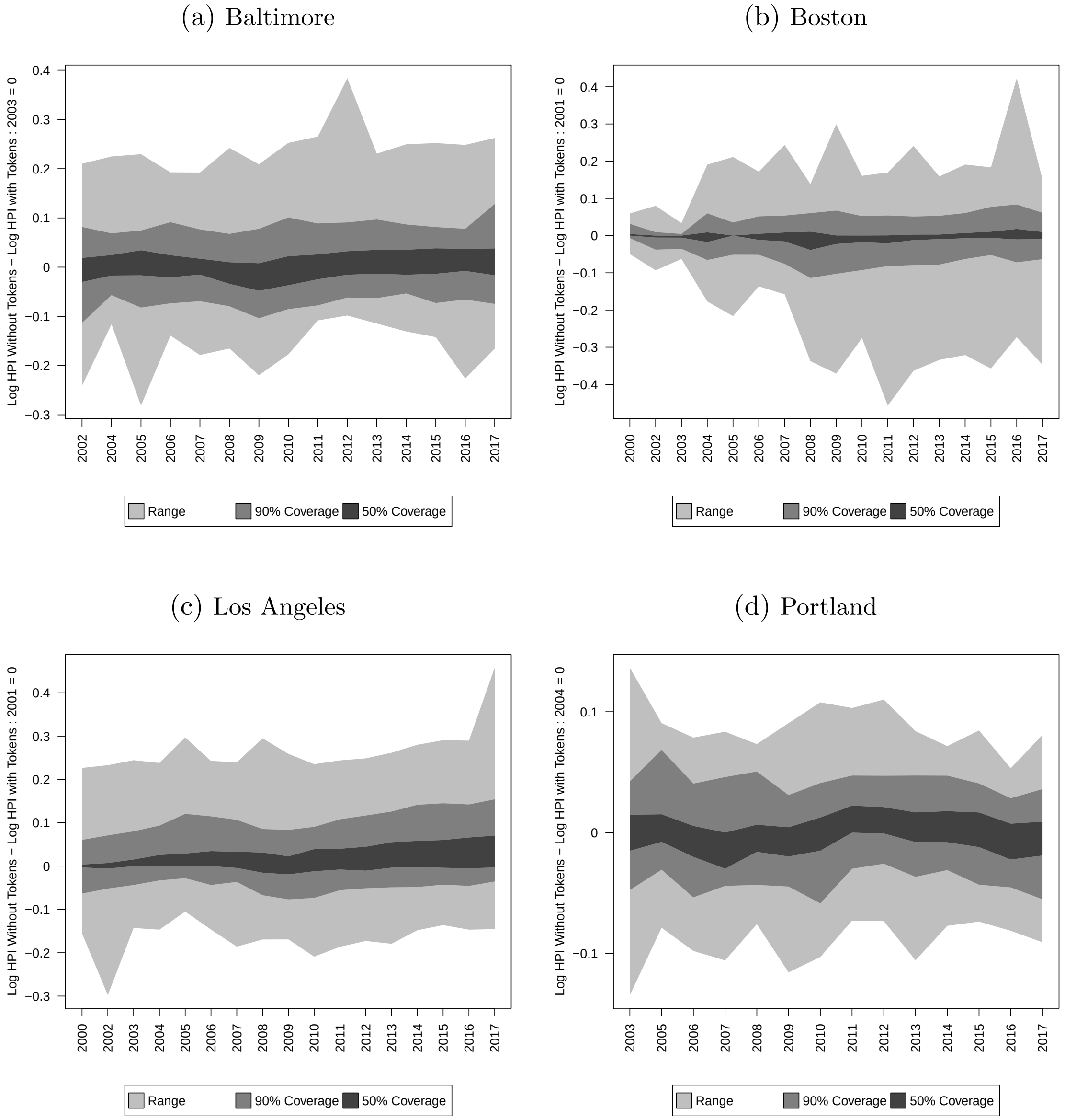
(b) Boston

(c) Los Angeles

(d) Portland

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


(a) Baltimore

(b) Boston

(c) Los Angeles

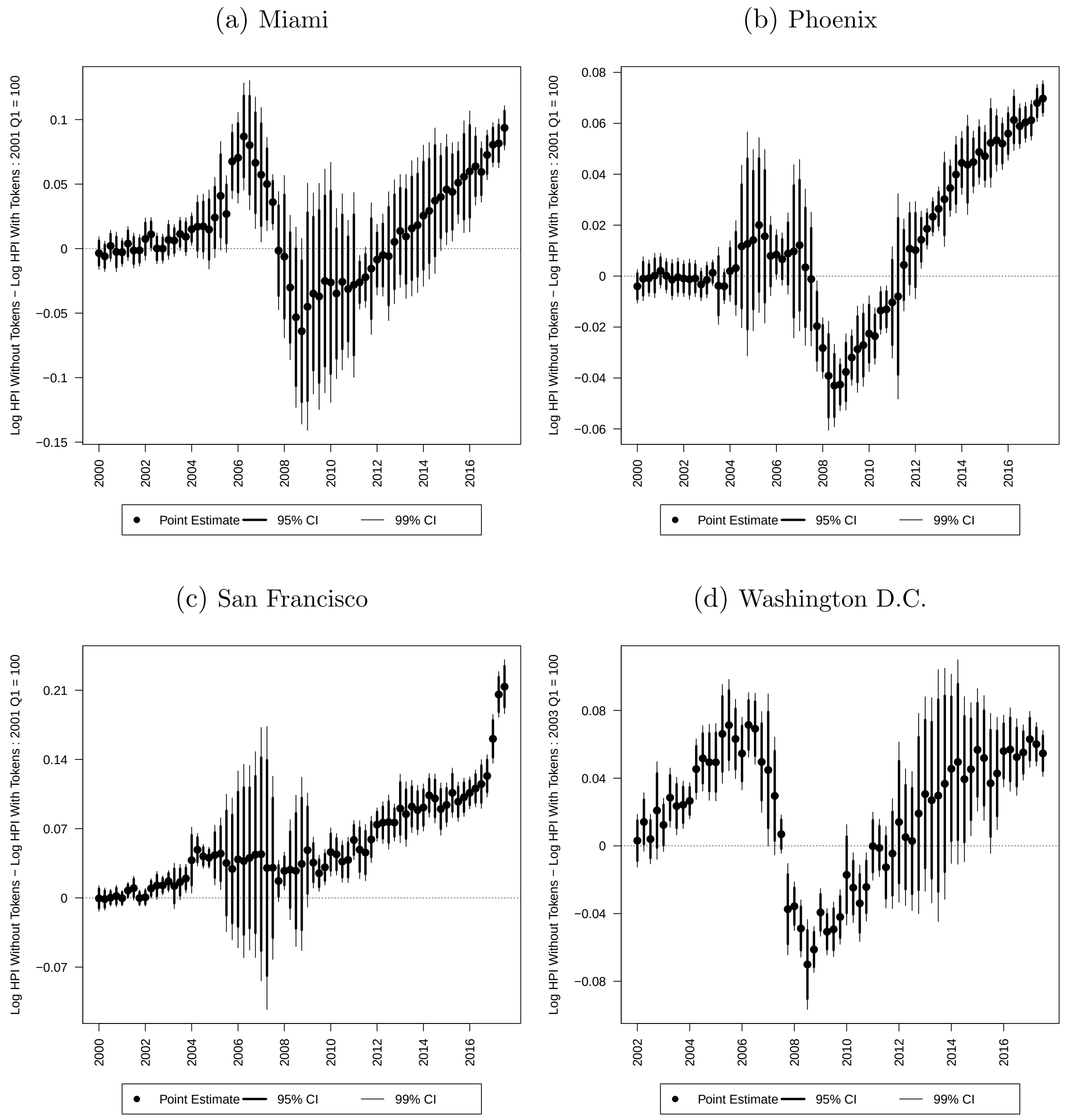
(d) Portland

Notes: Figure 4 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 74, 384, 285, and 81 unique zip codes in Baltimore, Boston, Los Angelese, 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 four additional MSAs (Baltimore, Boston, Los Angeles, and Portland) in Figure C1 of this internet appendix.

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


(a) Miami

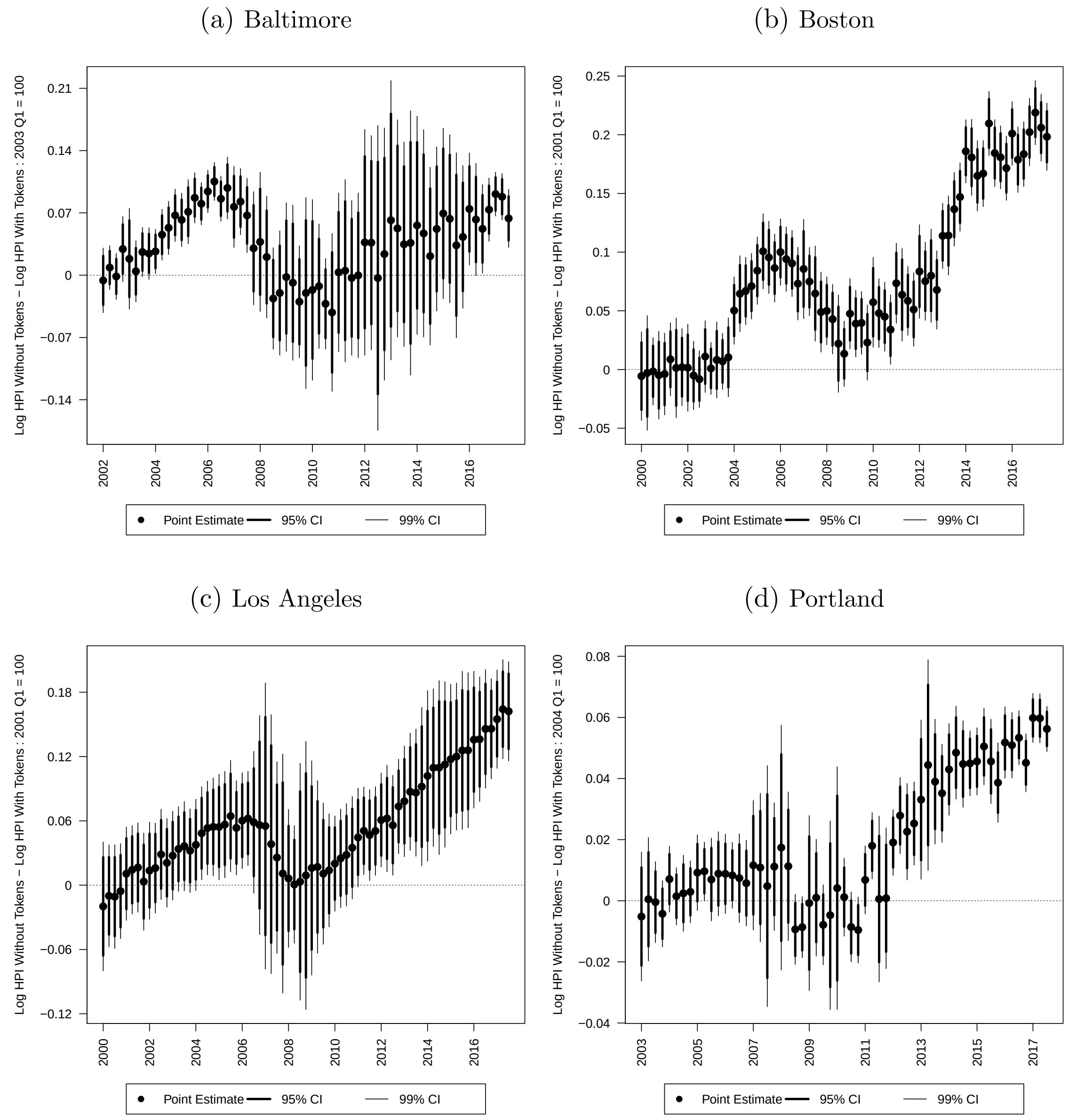
(b) Phoenix

(c) San Francisco

(d) Washington D.C.

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

Figure C5: Additional difference in repeat-sales HPIs


(a) Baltimore

(b) Boston

(c) Los Angeles

(d) Portland

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 is provided for each point estimate.

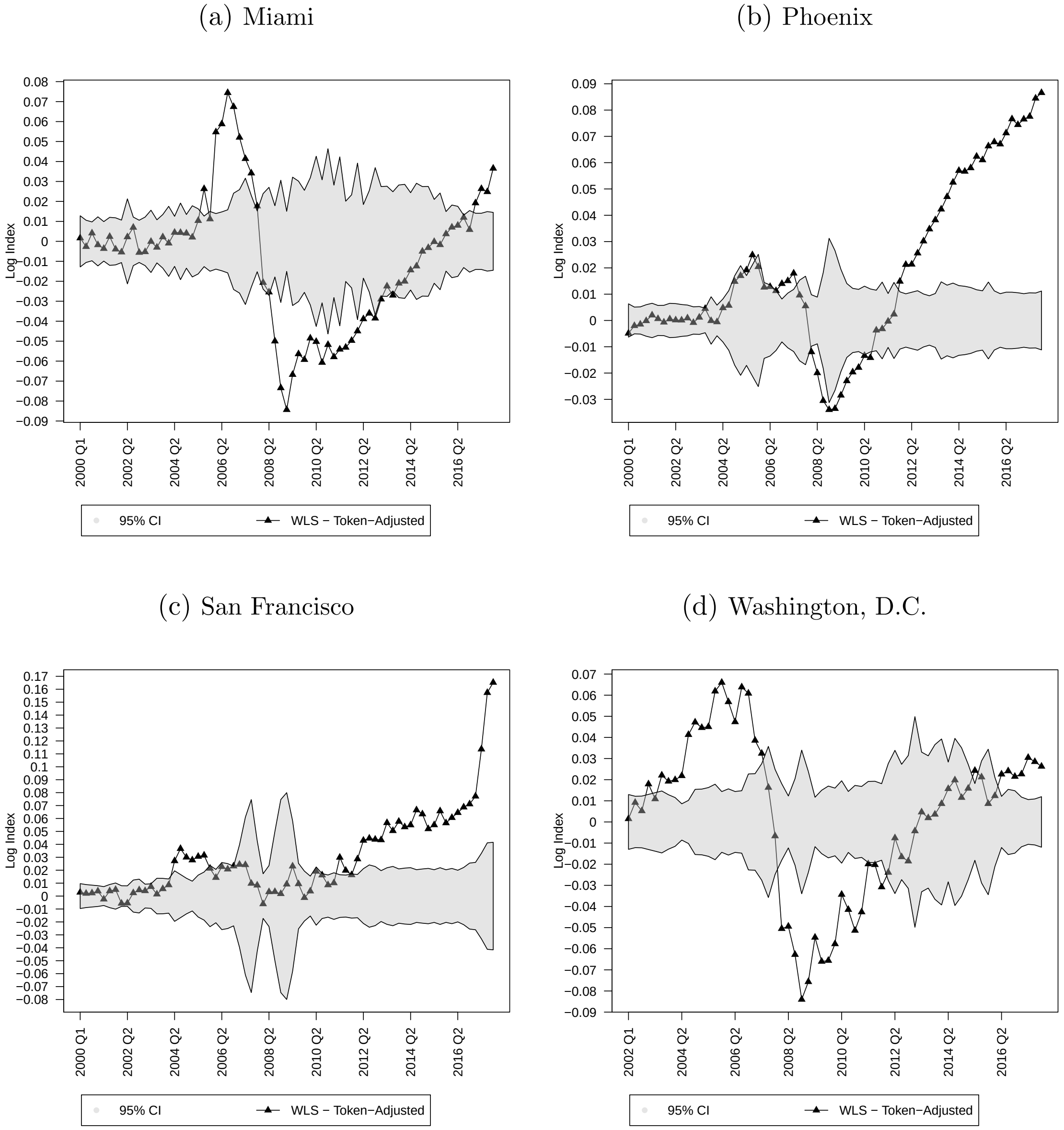
C.3 Indicator-adjusted HPIs

The results in the body of the paper indicate that the text in agents' remarks (*tokens*) control for time-varying attributes of the property. This section investigates the extent to which conventional controls for flips, distressed sales, and renovations obviate the need for tokens. 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 is equal to 1 if the transaction was a real estate owned (REO) or short sale transaction. The indicator variable for a renovation is equal to 1 if the property was renovated in the past 12 months.

We then construct and compare indicator-adjusted HPIs that do not include the tokens (see methodology in Section 2.3) to our quality-adjusted HPIs that include the tokens (see methodology in Section 2.4). 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 variable(s) 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 Baltimore, Boston, Los Angeles, and Portland in Figure C10.

Overall we find that 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

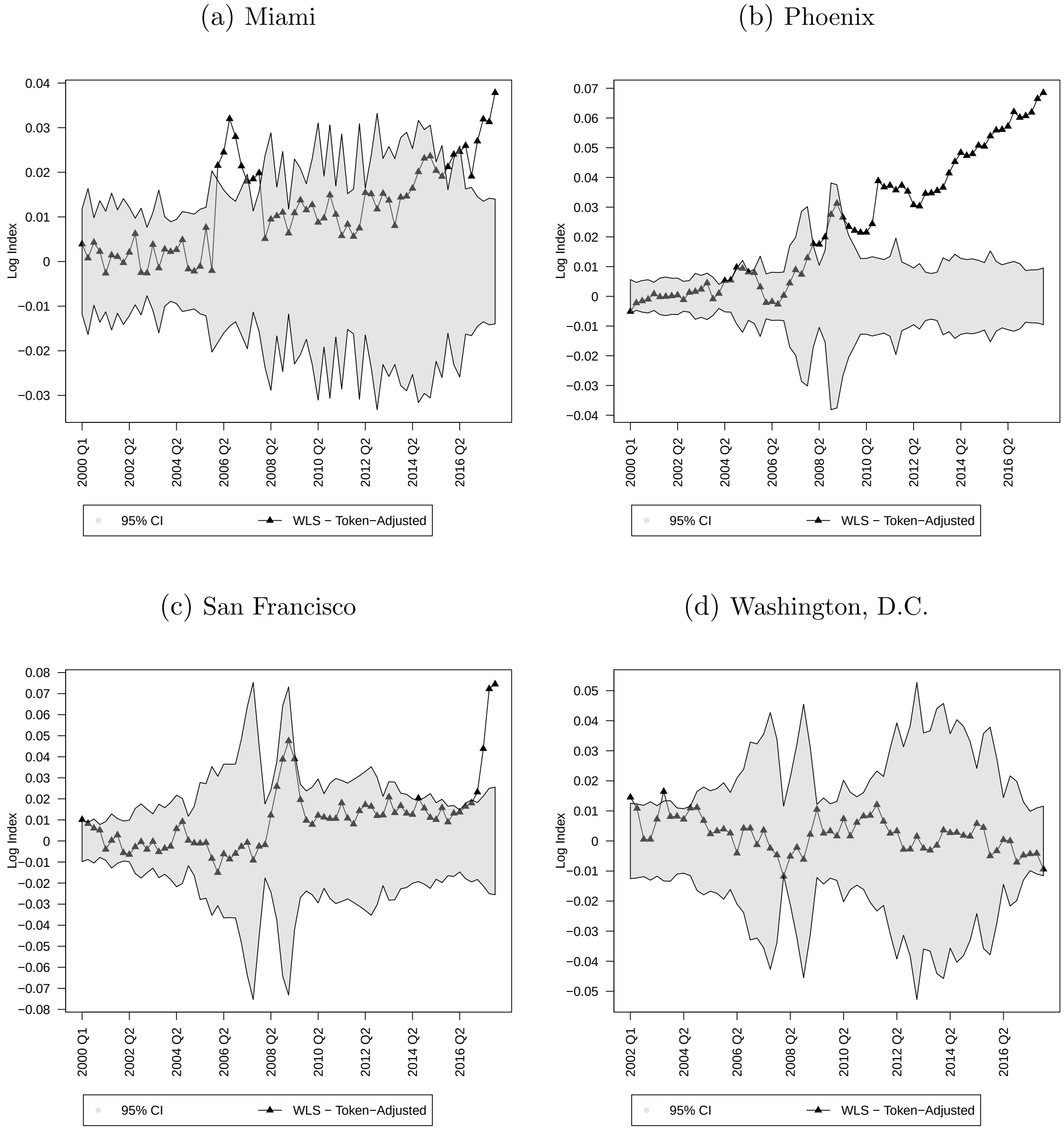
(b) Phoenix

(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

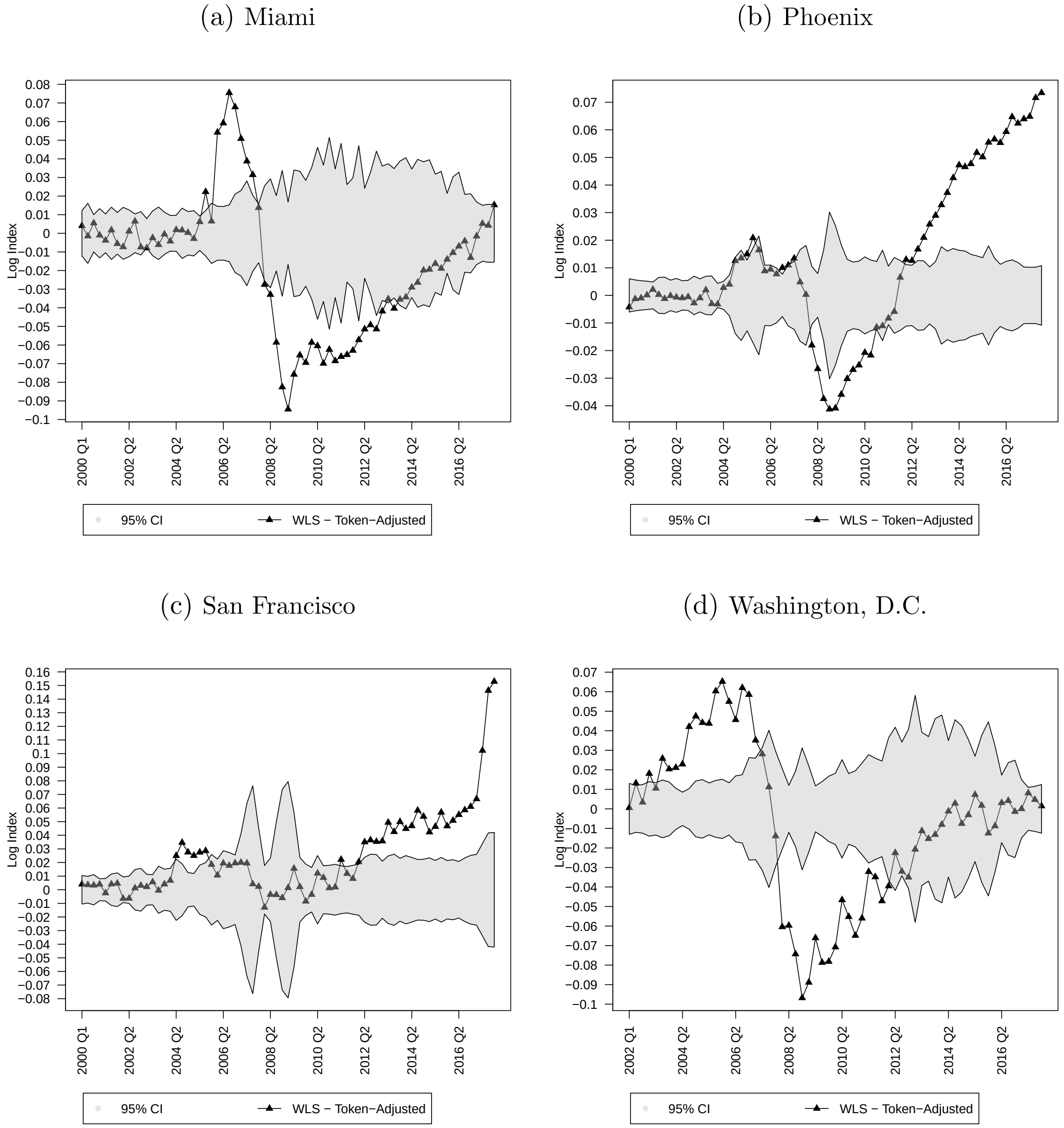
(b) Phoenix

(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

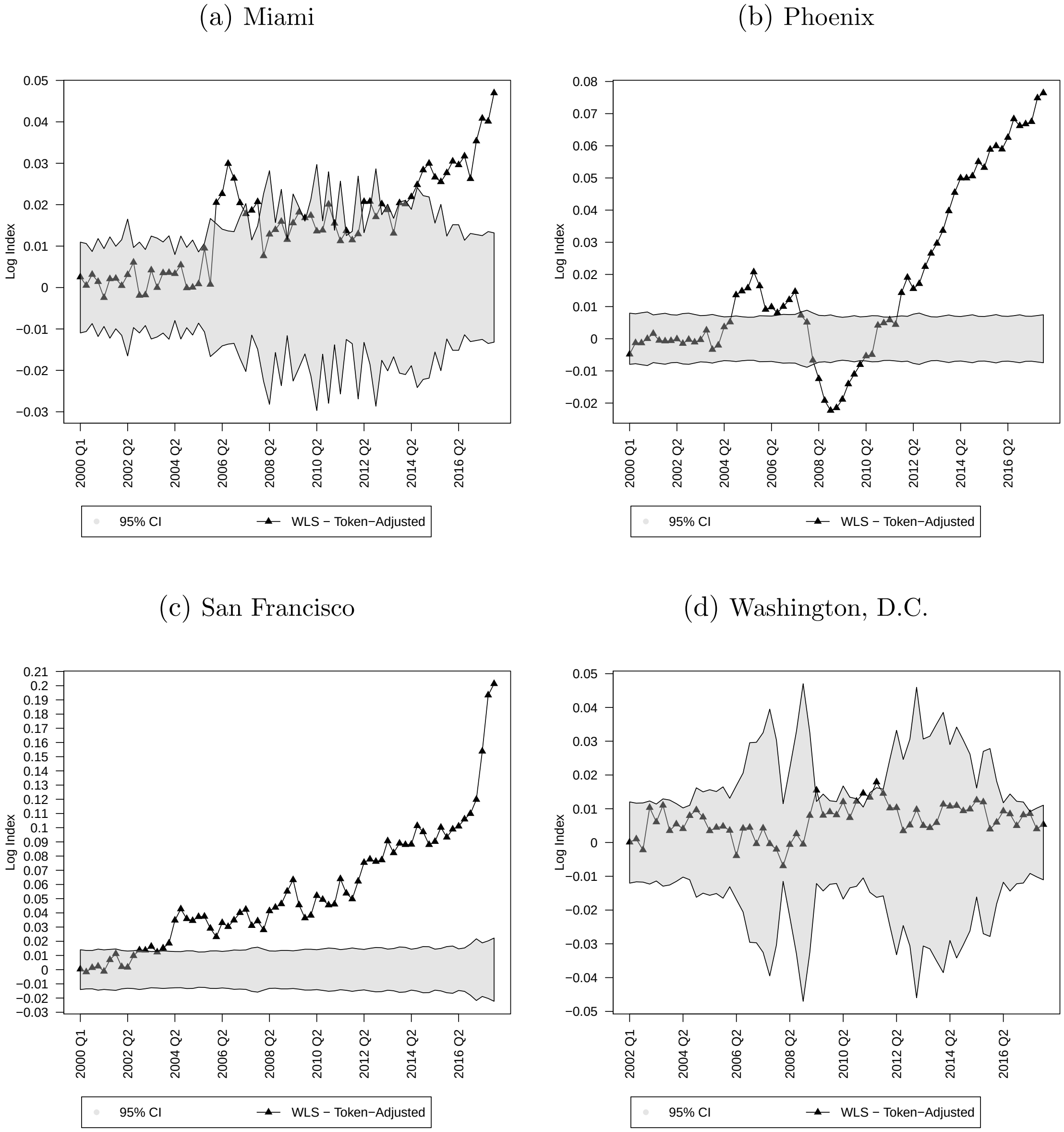
(b) Phoenix

(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


(a) Miami

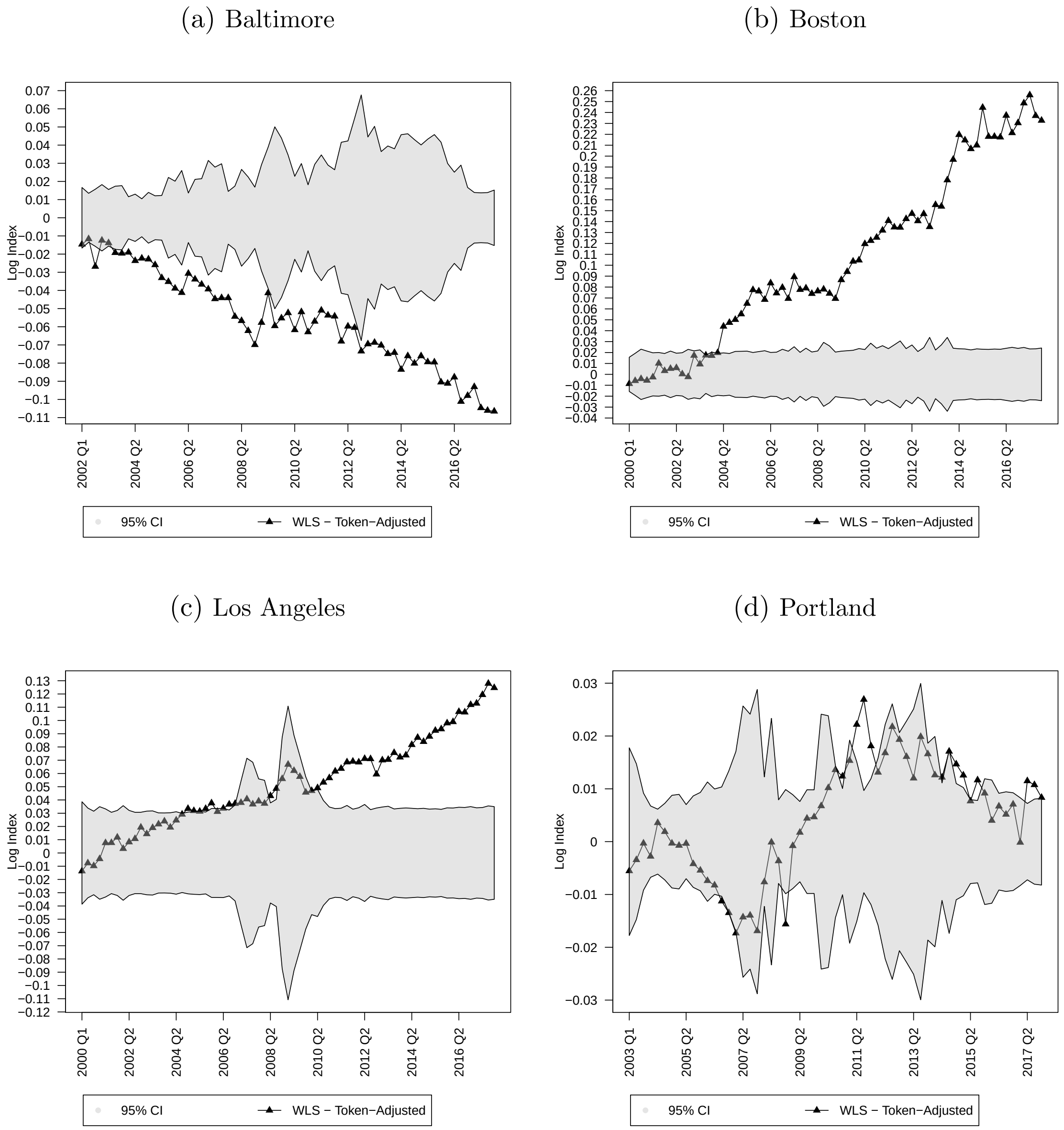
(b) Phoenix

(c) San Francisco

(d) Washington, D.C.

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



(a) Baltimore

(b) Boston

(c) Los Angeles

(d) Portland

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).

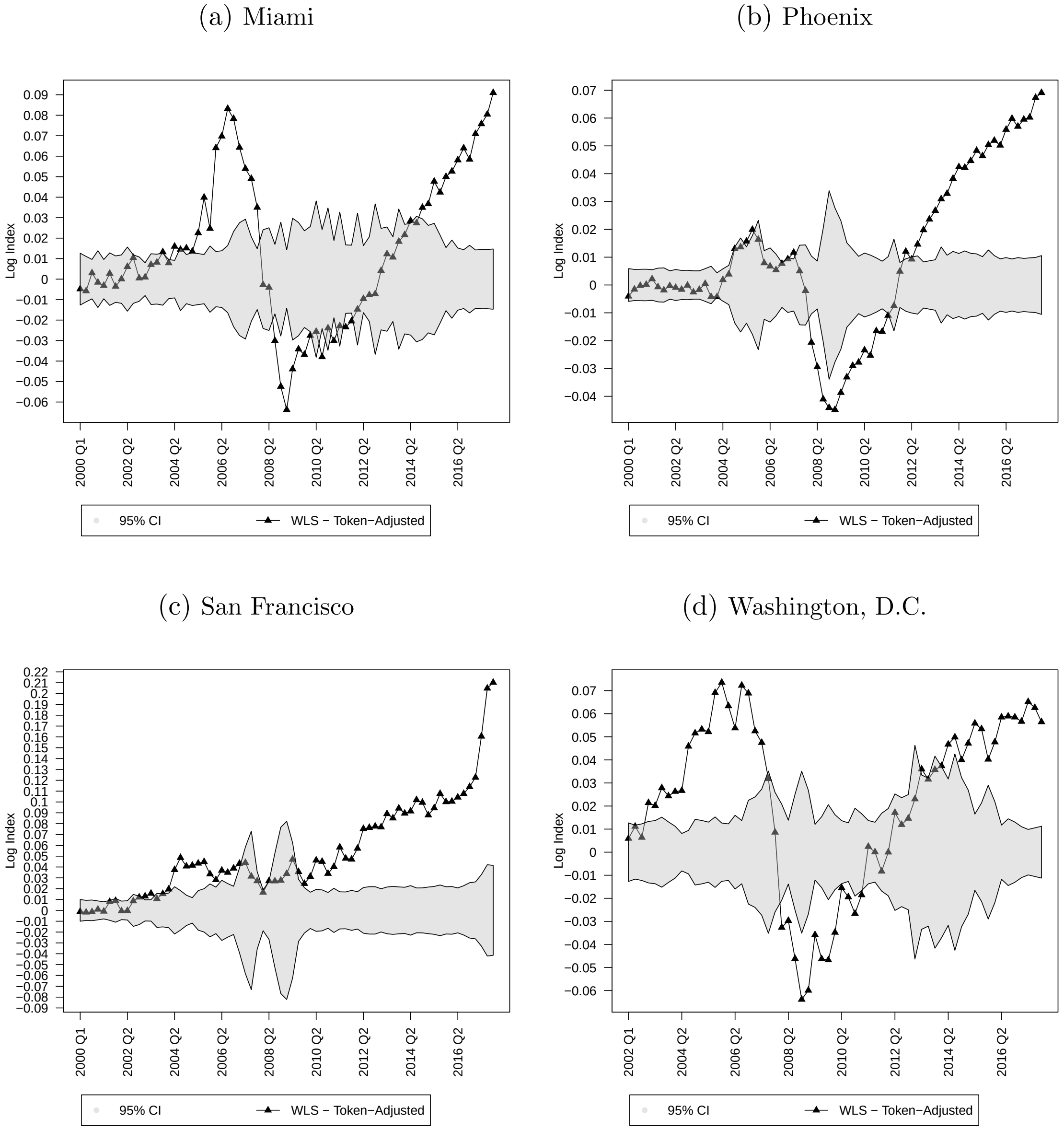
C.4 HPIs without flips, distressed sales, and renovations

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 type(s) 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 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

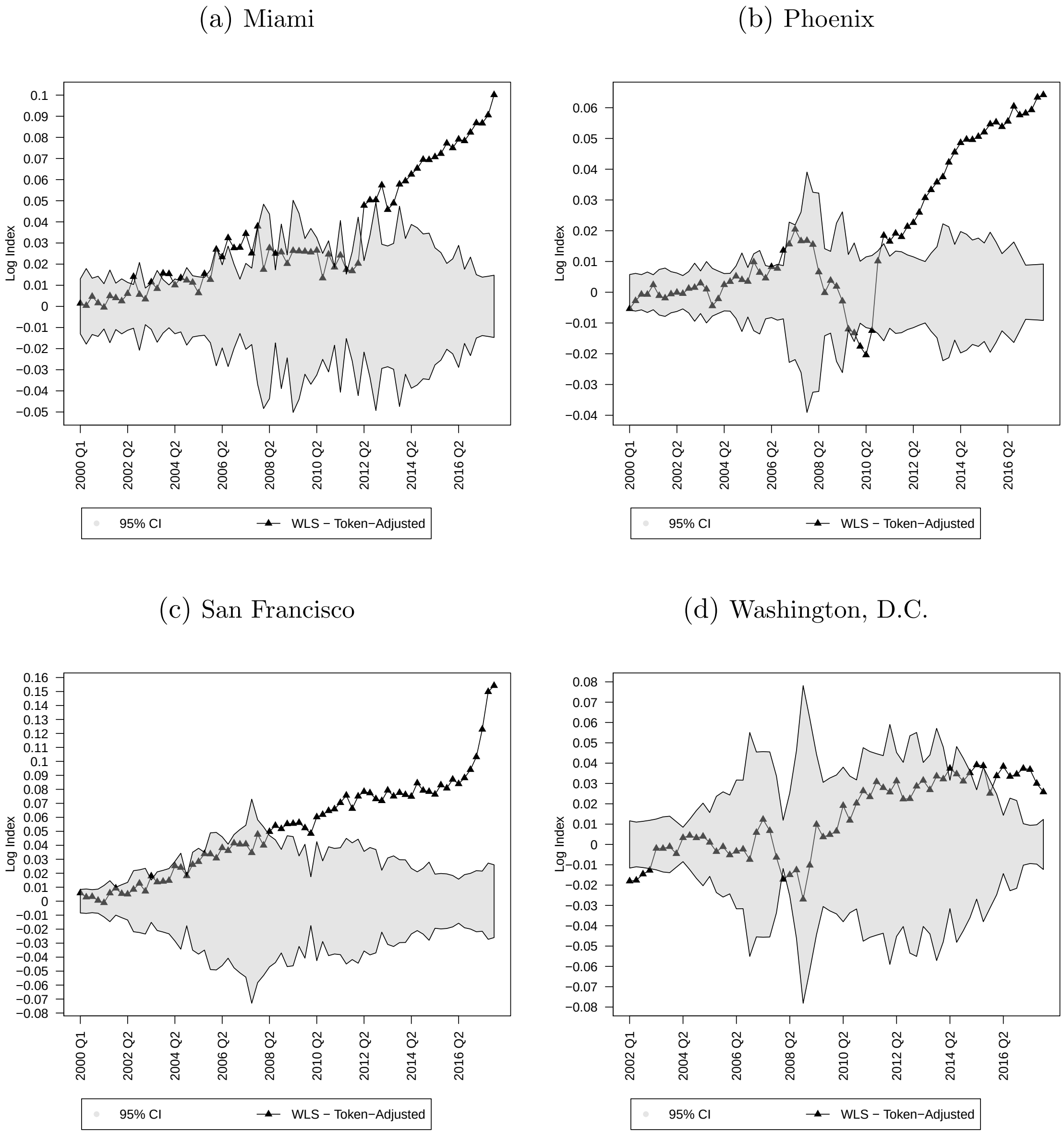
(b) Phoenix

(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

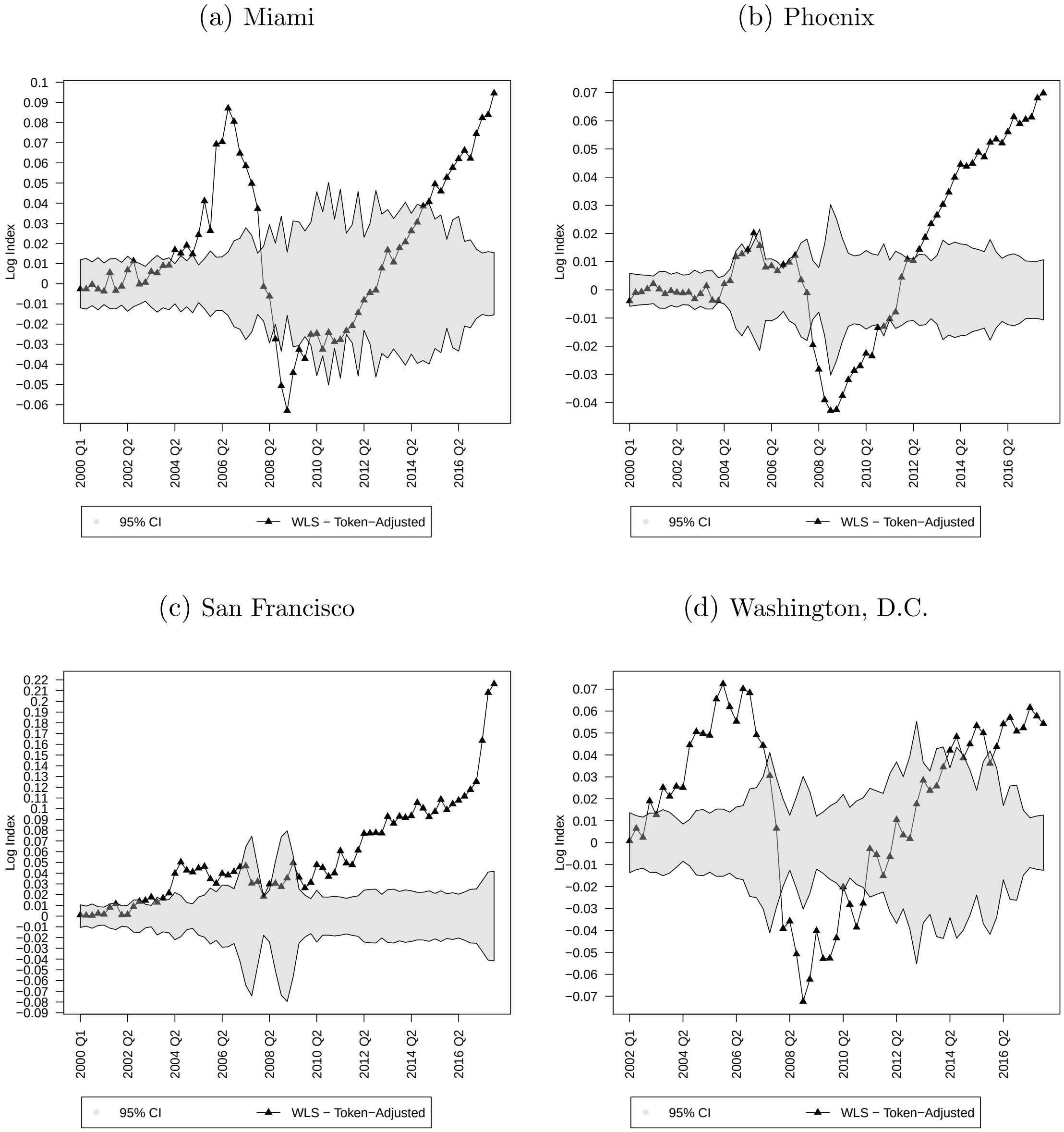
(b) Phoenix

(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

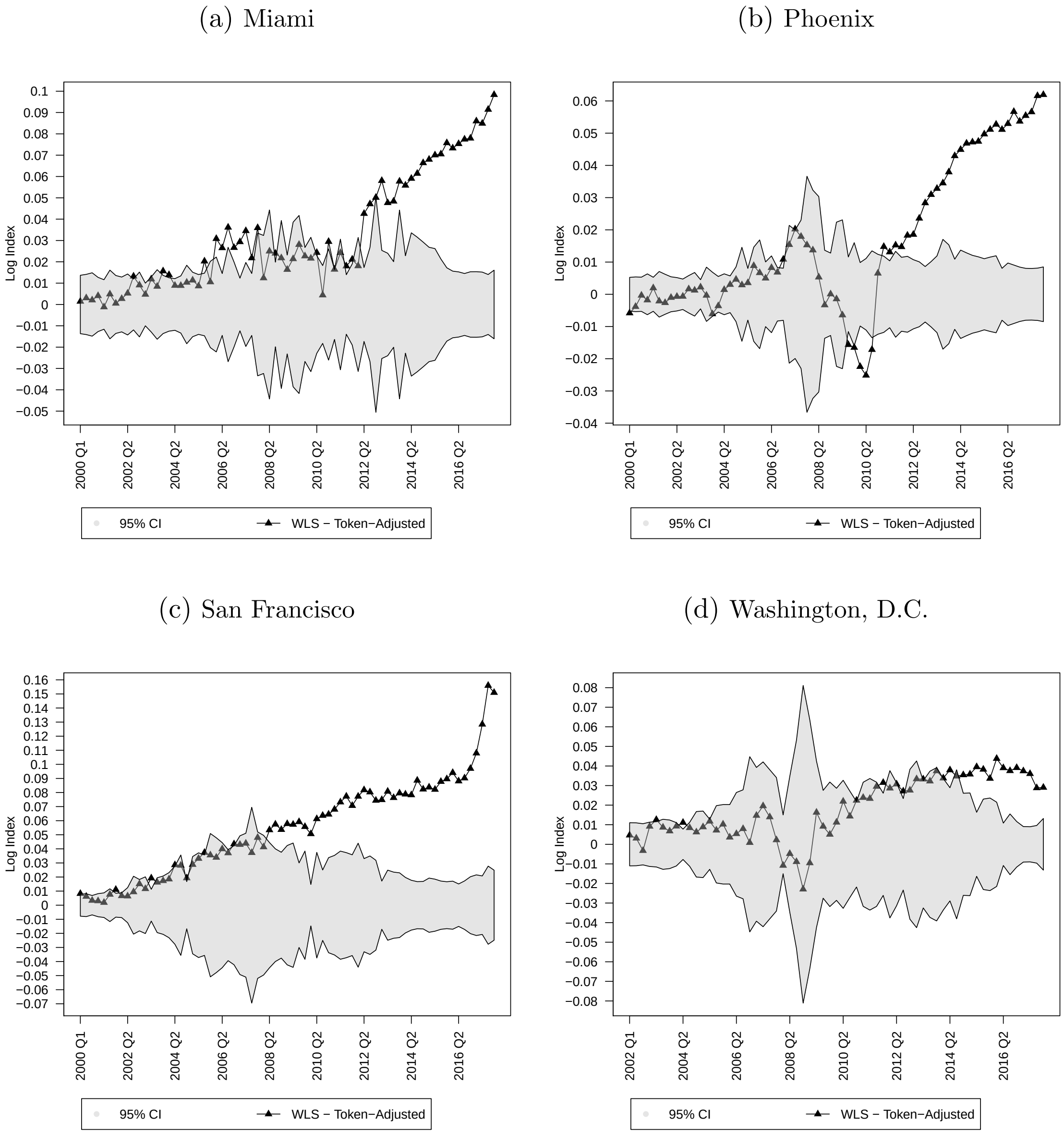
(b) Phoenix

(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

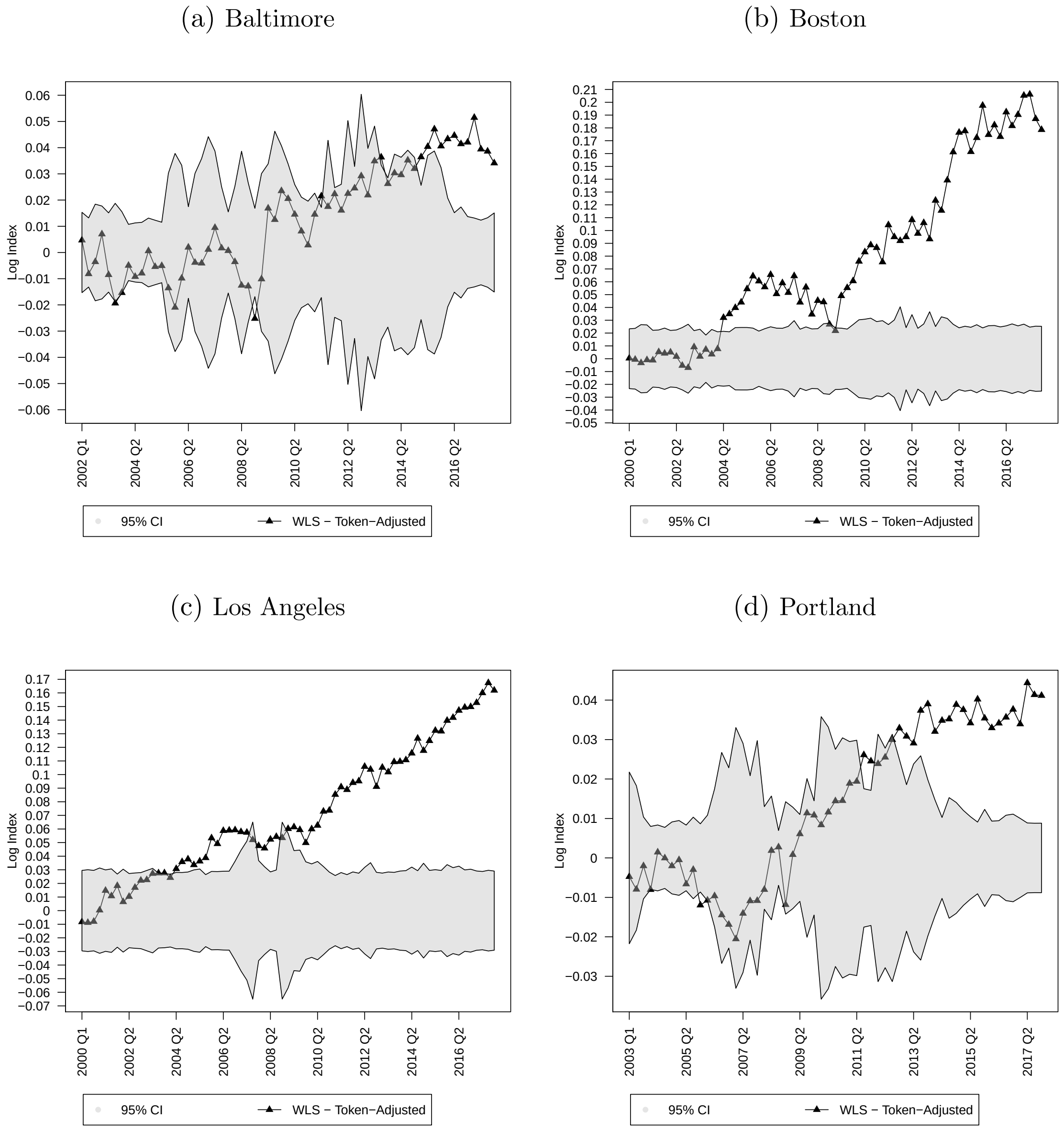
(b) Phoenix

(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 exlude flips, distressed sales, and renovations



(a) Baltimore

(b) Boston

(c) Los Angeles

(d) 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 implict 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 that the implicit prices of the tokens do not vary over time. We recognize that 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 bigger 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 biases 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 by including an interaction between the tokens and annual indicators. 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 that the static implicit price assumption does not introduce a significant bias in the eight MSAs we examine.

Table D1: Time-varying implicit prices HPI

MSA	Min	Mean	Max
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

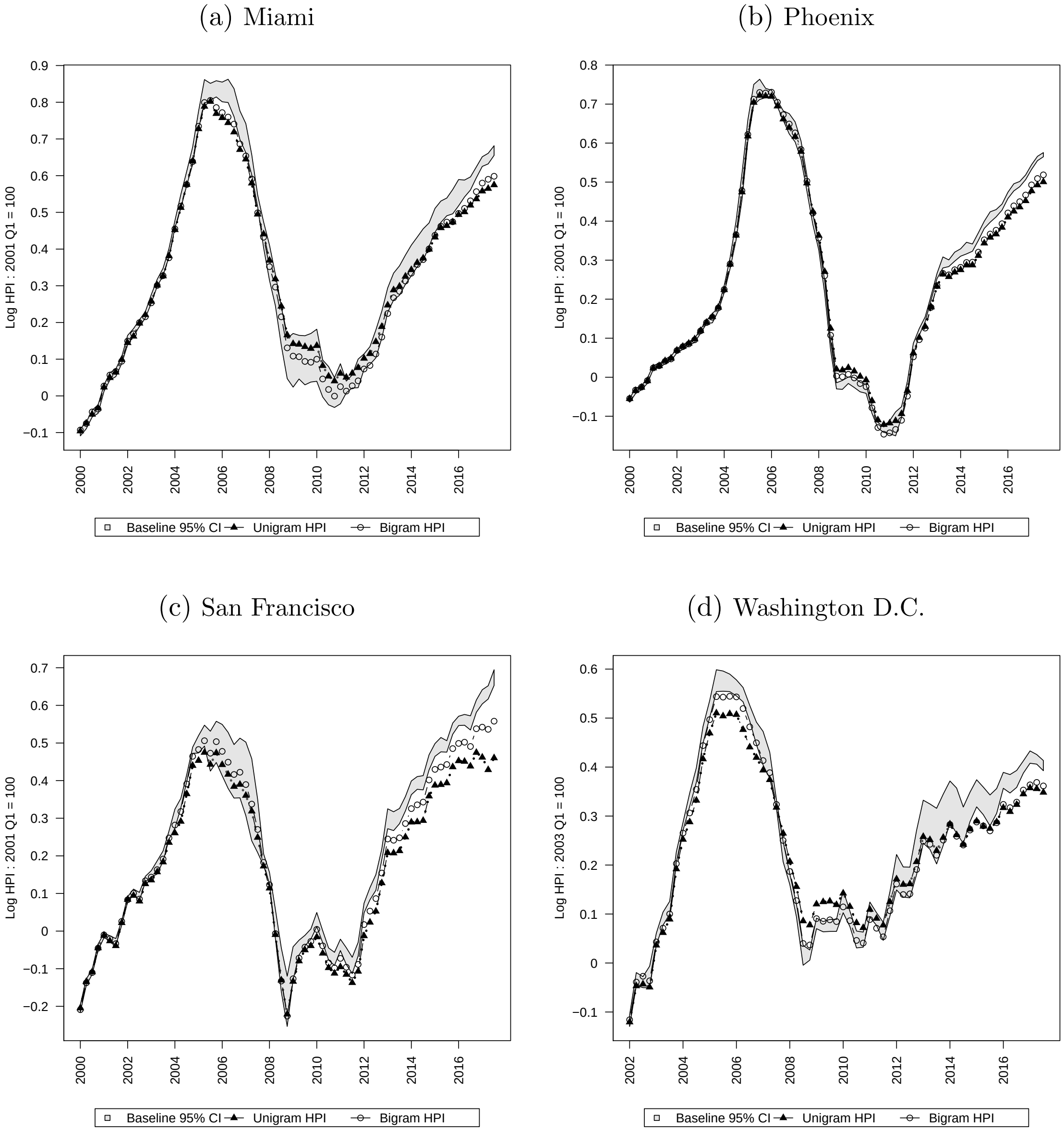
Note: Table D1 displays summary statistics for the difference between HPIs calculated assuming the implicit price for each token is constant and HPIs that allow the implicit price for each token to vary annually.

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 eight MSAs. Note, however, that 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


(a) Miami

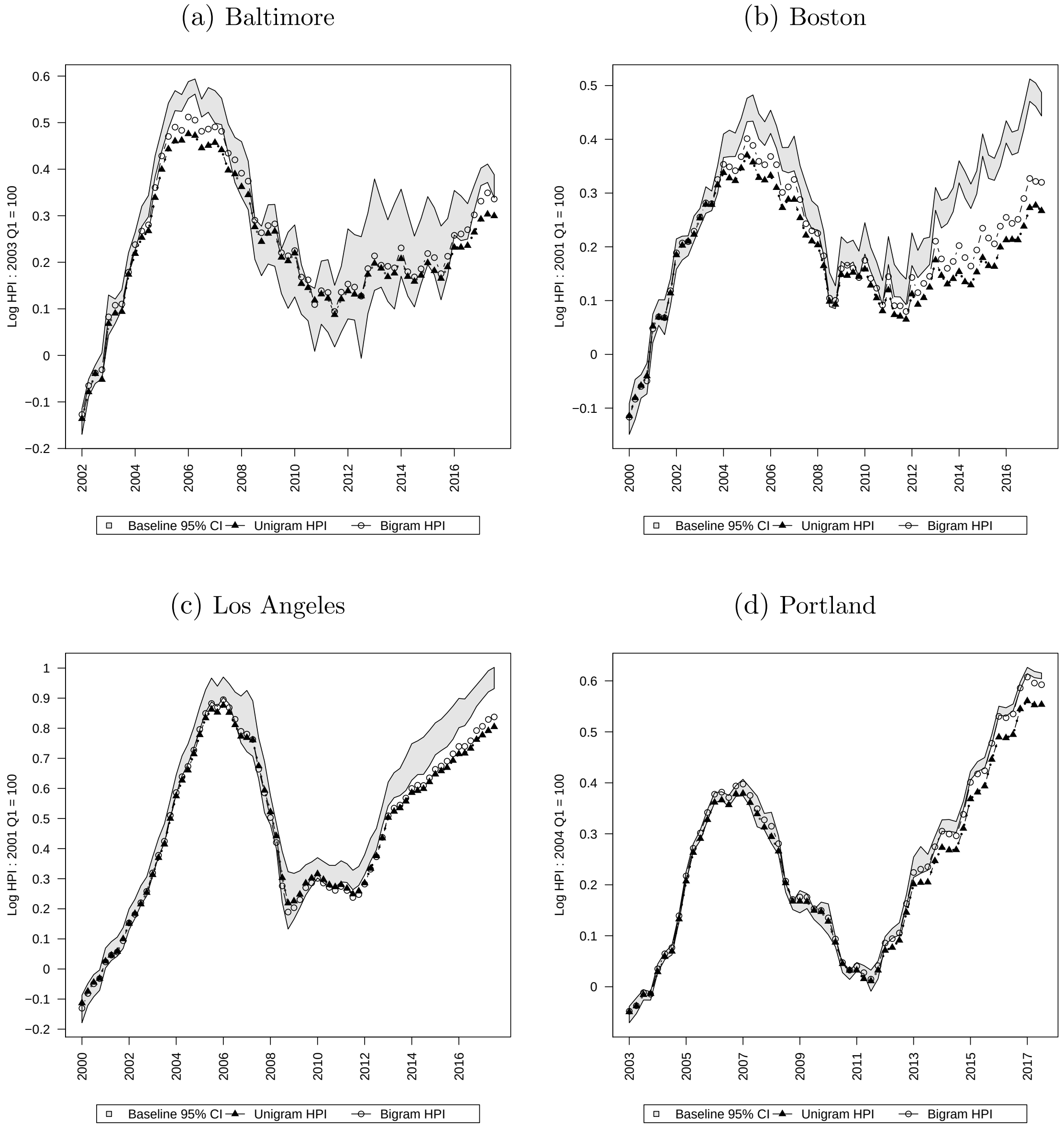
(b) Phoenix

(c) San Francisco

(d) Washington D.C.

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


(a) Baltimore

(b) Boston

(c) Los Angeles

(d) Portland

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 also 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) that identifies and includes 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 that the single-selection procedure that we employ does not introduce a significant bias for the eight 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 HPI

MSA	Min	Mean	Max	$\hat{\mathcal{Q}}$	$\hat{\mathcal{Q}}_{ds}$
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 (Belloni et al., 2014).