Measuring Neighborhood Change: The Issue of Ex Post Borders
Appendix Glossary

Appendix Section A: Proofs of Propositions

Appendix Section B: Tables and Figures

Appendix Section C: Coverage Maps Over Time—National and Select CBSAs

Appendix Section D: Missing Data Reweighting and TIGER/Line Tract Intersections

Appendix Section E: Regression Results for LTDB-1970

Appendix Section F: Regression Results for Figures 12, 13, and 15 with All CBSAs

Appendix Section G: Regression Results Weighting by the Number of Tract Intersections

Appendix Section H: Comparing the 1970-LTDB and LTDB for Missing Data Pass-Through and Outliers

Appendix Section A: Proofs of Propositions

Proof of Proposition 1: $\beta_P^* = \frac{Cov(\rho - P_{t,1}, P_{t,1})}{Var(P_{t,1})} = \frac{Cov(\rho, P_{t,1}) - Var(P_{t,1})}{Var(P_{t,1})} = \frac{-\sigma_{P,1}^2}{\sigma_{P,1}^2} = -1$, which follows because since ρ is a constant, $Cov(\rho, P_{t,1}) = 0$.

Proof of Proposition 2a: $\beta_X^* = \frac{Cov(\rho - P_{t,1}, X_{t,1})}{Var(X_{t,1})} = \frac{Cov(\rho, X_{t,1}) - Cov(P_{t,1}, X_{t,1})}{Var(X_{t,1})} = \frac{-Cov(P_{t,1}, X_{t,1})}{Var(X_{t,1})} = \frac{-\gamma\sigma_{X,1}^2}{\sigma_{X,1}^2} = -\gamma$, which follows because since ρ is a constant, $Cov(\rho, X_{t,1}) = 0$.

Proof of Proposition 2b: (i) We know that $P_{t,1} = \rho - \sum_{m\epsilon t}(\beta X_{m,1} + \varepsilon_m)$, and if $X_{t,1} = \sum_{m\epsilon t} X_{m,1}$, then $\frac{-Cov(P_{t,1}, X_{t,1})}{Var(X_{t,1})} = -\gamma = \beta + \frac{Cov(\sum_{m\epsilon t}\varepsilon_m, X_{t,1})}{Var(X_{t,1})}$. Consequently, if the tracts are formed so that $Cov(\sum_{m\epsilon t}\varepsilon_m, X_{t,1}) = 0$, then $\beta_X^* = \beta$.

(ii) If micro-locales are ex ante identical, where $X_{m,1}$ equals a constant $\hat{x}$, and $P_{m,1}$ equals a constant $\hat{p}$, and if M_t is a random variable that equals the number of micro-locales in the tract, then $\beta_X^* = \frac{Cov(\rho - M_t\hat{p}, M_t\hat{x})}{Var(M_t\hat{x})}$, using the fact that $\rho, \hat{p}$, and $\hat{x}$ are constant, we have $\beta_X^* = -\frac{\hat{x}\hat{p}Cov(M_t, M_t)}{\hat{x}^2 Var(M_t)} = -\frac{\hat{p}.}{\hat{x}}$

(iii) If all micro-locales have the same initial population ($\hat{p}$), and tracts are formed to have both constant population and have a fixed number of micro-locales (M), then $\beta_X^* = \frac{Cov(\rho - M\hat{p}, X_t)}{Var(X_t)} = 0$, since $\rho - M\hat{p}$ is a constant.

Proof of Proposition 3a: The value of $\delta_Z^* = \frac{Cov(\delta X_{t,1} + \vartheta\sum_{m\epsilon t}\varepsilon_m + \sum_{m\epsilon t}\epsilon_m, X_{t,1})}{Var(X_{t,1})} = \delta + \vartheta\frac{Cov(\sum_{m\epsilon t}\varepsilon_m, X_{t,1})}{Var(X_{t,1})}$, as $Cov(\sum_{m\epsilon t}\epsilon_m, X_{t,1}) = 0$, and this equals δ when $\vartheta = 0$ or when $Cov(\sum_{m\epsilon t}\varepsilon_m, X_{t,1})$.

Proof of Proposition 3b: (i) If $X_{m,1}$ equals a constant $\hat{x}$, and $P_{m,1}$ equals a constant $\hat{p}$ and if M_t is a random variable that equals the number of micro-locales in the tract, $\sum_{m\epsilon t}\varepsilon_m = \rho -$

$M_t(\hat{p} + \beta\hat{x})$ and consequently $\frac{Cov(\sum_{m\in t}\varepsilon_m, X_{t,1})}{Var(X_{t,1})} = \frac{Cov(\rho - M_t(\hat{p}+\beta\hat{x}), M_t\hat{x})}{Var(M_t\hat{x})} = -\beta - \frac{\hat{p}.}{\hat{x}}$, and $\delta_Z^* = \delta - \vartheta\beta - \vartheta\frac{\hat{p}.}{\hat{x}}$.

(ii) If all micro-locales have the same initial population ($\hat{p}$), and tracts are formed to have both constant population and a fixed number of micro-locales (M), then $\sum_{m\in t}\varepsilon_m = \rho - M\hat{p} - \beta X_{t,1}$

$\frac{Cov(\sum_{m\in t}\varepsilon_m, X_{t,1})}{Var(X_{t,1})} = \frac{Cov(\rho - M\hat{p} - \beta X_{t,1}, X_{t,1})}{Var(X_{t,1})} = -\beta$ and $\delta_Z^* = \delta - \vartheta\beta$.

Appendix Section B: Tables and Figures

Appendix Table B1: *CBSA-Level Percentage of Population Living Within 1970 Tract Boundaries, 24 Select Major Metropolitan Areas*

CBSA	Statistic	1970	1980	1990	2000	2010	2020	2023
Atlanta	Reverse LTDB Population	1,460,663	1,814,004	2,358,735	3,190,879	3,747,161	4,286,937	4,302,843
	Full CBSA Population	1,842,331	2,326,639	3,069,425	4,247,981	5,268,860	6,069,718	6,141,490
	Population Coverage	0.793	0.780	0.768	0.751	0.711	0.706	0.701
Boston	Reverse LTDB Population	3,493,012	3,398,040	3,484,510	3,664,695	3,709,137	4,038,505	4,007,578
	Full CBSA Population	3,918,092	3,938,585	4,133,895	4,391,344	4,552,402	4,941,632	4,917,661
	Population Coverage	0.892	0.863	0.843	0.835	0.815	0.817	0.815
Charlotte	Reverse LTDB Population	715,218	827,320	998,136	1,301,993	1,719,994	2,076,524	2,114,244
	Full CBSA Population	741,118	855,482	1,024,643	1,330,448	1,758,038	2,111,641	2,149,090
	Population Coverage	0.965	0.967	0.974	0.979	0.978	0.983	0.984
Chicago	Reverse LTDB Population	7,702,578	7,843,073	7,968,184	8,847,140	8,973,878	9,112,693	9,016,764
	Full CBSA Population	7,886,829	8,052,932	8,182,076	9,098,316	9,461,105	9,618,502	9,527,968
	Population Coverage	0.977	0.974	0.974	0.972	0.949	0.947	0.946
Cleveland	Reverse LTDB Population	2,318,837	2,168,714	2,094,990	2,137,133	2,037,532	2,047,927	2,034,395
	Full CBSA Population	2,321,037	2,173,734	2,102,248	2,148,143	2,077,240	2,088,251	2,074,635
	Population Coverage	0.999	0.998	0.997	0.995	0.981	0.981	0.981
Dallas	Reverse LTDB Population	2,369,442	2,950,320	3,910,090	5,065,212	6,212,867	7,459,313	7,618,181
	Full CBSA Population	2,424,486	3,017,178	3,989,294	5,161,544	6,371,773	7,642,617	7,812,900
	Population Coverage	0.977	0.978	0.980	0.981	0.975	0.976	0.975

Appendix Section B: Tables and Figures

Appendix Table B1: *CBSA-Level Percentage of Population Living Within 1970 Tract Boundaries, 24 Select Major Metropolitan Areas*

CBSA	Statistic	1970	1980	1990	2000	2010	2020	2023
Denver	Reverse LTDB Population	1,095,334	1,397,592	1,556,227	1,925,127	2,145,385	2,477,801	2,477,767
	Full CBSA Population	1,116,226	1,450,768	1,650,489	2,157,756	2,487,593	2,889,709	2,901,975
	Population Coverage	0.981	0.963	0.943	0.892	0.862	0.857	0.854
Detroit	Reverse LTDB Population	4,424,052	4,340,311	4,240,299	4,442,378	4,224,691	4,320,970	4,295,961
	Full CBSA Population	4,431,390	4,353,413	4,248,699	4,452,557	4,296,250	4,392,041	4,367,620
	Population Coverage	0.998	0.997	0.998	0.998	0.983	0.984	0.984
Houston	Reverse LTDB Population	2,180,469	3,116,097	3,724,460	4,661,147	5,832,262	7,034,481	7,149,113
	Full CBSA Population	2,201,848	3,148,991	3,767,335	4,715,407	5,946,800	7,149,642	7,274,714
	Population Coverage	0.990	0.990	0.989	0.988	0.981	0.984	0.983
Las Vegas	Reverse LTDB Population	273,288	462,855	740,915	1,374,329	1,936,517	2,248,550	2,276,783
	Full CBSA Population	273,288	463,087	741,459	1,375,765	1,951,269	2,265,461	2,293,764
	Population Coverage	1.000	1.000	0.999	0.999	0.992	0.993	0.993
Los Angeles	Reverse LTDB Population	8,429,609	9,391,462	11,225,466	12,320,578	12,534,968	12,900,701	12,701,813
	Full CBSA Population	8,452,461	9,410,212	11,273,720	12,365,627	12,828,837	13,200,998	13,012,469
	Population Coverage	0.997	0.998	0.996	0.996	0.977	0.977	0.976
Miami	Reverse LTDB Population	2,235,239	3,147,444	4,049,700	4,988,549	5,462,423	6,028,946	6,033,730
	Full CBSA Population	2,236,645	3,147,444	4,056,100	5,007,564	5,564,635	6,138,333	6,138,876
	Population Coverage	0.999	1.000	0.998	0.996	0.982	0.982	0.983

Appendix Section B: Tables and Figures

Appendix Table B1: *CBSA-Level Percentage of Population Living Within 1970 Tract Boundaries, 24 Select Major Metropolitan Areas*

CBSA	Statistic	1970	1980	1990	2000	2010	2020	2023
New York	Reverse LTDB Population	16,721,032	15,823,955	16,175,491	17,533,054	17,300,909	18,442,513	18,075,114
	Full CBSA Population	17,062,309	16,363,540	16,846,046	18,323,002	18,897,109	20,140,470	19,816,413
	Population Coverage	0.980	0.967	0.960	0.957	0.916	0.916	0.912
Orlando	Reverse LTDB Population	522,575	801,504	1,220,834	1,640,384	2,117,378	2,653,934	2,687,930
	Full CBSA Population	522,575	804,925	1,224,852	1,644,561	2,134,411	2,673,376	2,721,022
	Population Coverage	1.000	0.996	0.997	0.997	0.992	0.993	0.988
Philadelphia	Reverse LTDB Population	5,299,574	5,218,421	5,407,998	5,659,291	5,759,145	6,034,059	6,029,411
	Full CBSA Population	5,317,407	5,240,039	5,435,468	5,687,147	5,965,343	6,245,051	6,241,882
	Population Coverage	0.997	0.996	0.995	0.995	0.965	0.966	0.966
Phoenix	Reverse LTDB Population	966,435	1,507,548	2,119,426	3,068,666	3,786,897	4,386,872	4,445,919
	Full CBSA Population	1,035,438	1,599,970	2,238,480	3,251,876	4,192,887	4,845,832	4,941,206
	Population Coverage	0.933	0.942	0.947	0.944	0.903	0.905	0.900
Raleigh	Reverse LTDB Population	228,453	300,720	422,084	625,312	891,874	1,119,082	1,140,743
	Full CBSA Population	317,010	401,981	541,100	797,071	1,130,490	1,413,982	1,449,594
	Population Coverage	0.721	0.748	0.780	0.785	0.789	0.791	0.787
Salt Lake City	Reverse LTDB Population	456,971	614,366	722,092	894,013	1,009,274	1,163,197	1,161,028
	Full CBSA Population	486,031	655,297	768,075	968,858	1,124,197	1,300,293	1,304,046
	Population Coverage	0.940	0.938	0.940	0.923	0.898	0.895	0.890

Appendix Section B: Tables and Figures

Appendix Table B1: *CBSA-Level Percentage of Population Living Within 1970 Tract Boundaries, 24 Select Major Metropolitan Areas*

CBSA	Statistic	1970	1980	1990	2000	2010	2020	2023
San Antonio	Reverse LTDB Population	862,979	1,033,834	1,248,375	1,478,733	1,826,394	2,160,287	2,193,718
	Full CBSA Population	951,876	1,154,648	1,407,745	1,711,703	2,142,508	2,558,143	2,612,802
	Population Coverage	0.907	0.895	0.887	0.864	0.852	0.844	0.840
San Diego	Reverse LTDB Population	1,329,207	1,819,628	2,461,766	2,792,500	3,020,623	3,219,758	3,207,089
	Full CBSA Population	1,357,854	1,861,846	2,498,016	2,813,833	3,095,313	3,298,634	3,282,782
	Population Coverage	0.979	0.977	0.985	0.992	0.976	0.976	0.977
San Francisco	Reverse LTDB Population	3,097,458	3,238,876	3,665,733	4,104,511	4,182,475	4,579,208	4,487,270
	Full CBSA Population	3,109,519	3,250,630	3,686,592	4,123,740	4,335,391	4,749,008	4,653,593
	Population Coverage	0.996	0.996	0.994	0.995	0.965	0.964	0.964
Seattle	Reverse LTDB Population	1,830,320	2,082,941	2,548,573	3,026,878	3,379,152	3,949,299	3,951,014
	Full CBSA Population	1,832,896	2,093,112	2,559,164	3,043,878	3,439,809	4,018,762	4,021,467
	Population Coverage	0.999	0.995	0.996	0.994	0.982	0.983	0.982
Tampa	Reverse LTDB Population	1,012,394	1,372,133	1,680,697	1,915,831	2,102,419	2,372,869	2,403,985
	Full CBSA Population	1,105,553	1,613,603	2,067,959	2,395,997	2,783,243	3,175,275	3,240,469
	Population Coverage	0.916	0.850	0.813	0.800	0.755	0.747	0.742
Washington	Reverse LTDB Population	2,915,051	3,070,760	3,682,565	4,244,955	4,837,226	5,478,157	5,440,004
	Full CBSA Population	2,915,051	3,375,973	4,088,223	4,750,758	5,530,076	6,251,434	6,237,018
	Population Coverage	1.000	0.910	0.901	0.894	0.875	0.876	0.872

Note: We are able to identify the CBSA's true population using the full-coverage county data from 1970-2023. The 'Full CBSA Population' variable is this population value in each year. The Reverse LTDB Population is the sum over all tracts in a year for our set of 1970's tracts.

Appendix Table B2: *Source Year Tracts with Sum of Weights = 1*

Source Year Tracts	**1980**	**1990**	**2000**	**2010**	**2020**
Total Tracts	38406	42582	46987	55923	63633
Total Tracts with $\sum_{b \in B_s} weight_{s,b} = 1$	38381	42534	46902	50687	57904
Percentage Tracts with $\sum_{b \in B_s} weight_{s,b} = 1$	0.999	0.999	0.998	0.906	0.910

Note: Refer to Appendix Section C for a discussion about how the percentage of tracts with weights summing to 1 in 2010 and 2020 is an underestimate because we drop all tract intersections with a weight of less than 5% to account for tiny tract intersections generated by mismatched TIGER/Lines.

Appendix Table B3: Distribution of Correlations Coefficients for Initial Population and Other Initial Variables, All CBSAs

Statistics		*House Price*	*Median Income*	*Non-White Population Share*	*Distance to Center*	*Housing Unit Density*
Percentiles						
	1%	-0.627	-0.575	-0.348	-0.604	-0.172
	5%	-0.426	-0.464	-0.138	-0.546	-0.045
	10%	-0.382	-0.388	-0.065	-0.481	0.071
	25%	-0.279	-0.305	0.049	-0.349	0.231
	Median	-0.139	-0.181	0.165	-0.199	0.427
	75%	-0.001	-0.050	0.291	-0.053	0.561
	90%	0.118	0.088	0.383	0.063	0.655
	99%	0.388	0.451	0.513	0.338	0.864
Moments						
	Mean	-0.132	-0.165	0.162	-0.197	0.394
	Std. Dev.	0.200	0.192	0.178	0.215	0.231
	Min	-0.676	-0.669	-0.405	-0.786	-0.350
	Max	0.495	0.499	0.699	0.455	0.945
Observations	*N*	295	295	295	295	295

Note: This table reports the distribution of correlation coefficients from the LTDB for within-CBSA correlation between initial population and median house prices, median income, non white population shares, distance to center and housing unit density for all 295 CBSAs. The observations are 1970 correlation coefficients at the CBSA level.

Appendix Table B4: Distribution of Correlations Coefficients for Initial Prices and Multiple Variables Correlated with Population, Top 100 CBSAs

Statistics		*Median Income*	*Non-White Population Share*	*Housing Unit Density*
Percentiles				
	1%	0.011	-0.635	-0.582
	5%	0.373	-0.522	-0.504
	10%	0.469	-0.482	-0.469
	25%	0.641	-0.399	-0.395
	Median	0.726	-0.333	-0.229
	75%	0.818	-0.263	-0.038
	90%	0.853	-0.207	0.106
	95%	0.883	-0.127	0.145
	99%	0.915	0.177	0.316
Moments				
	Mean	0.694	-0.323	-0.211
	Std. Dev.	0.172	0.139	0.213
	Min	-0.199	-0.641	-0.605
	Max	0.919	0.184	0.398
Observations	*N*	100	100	100

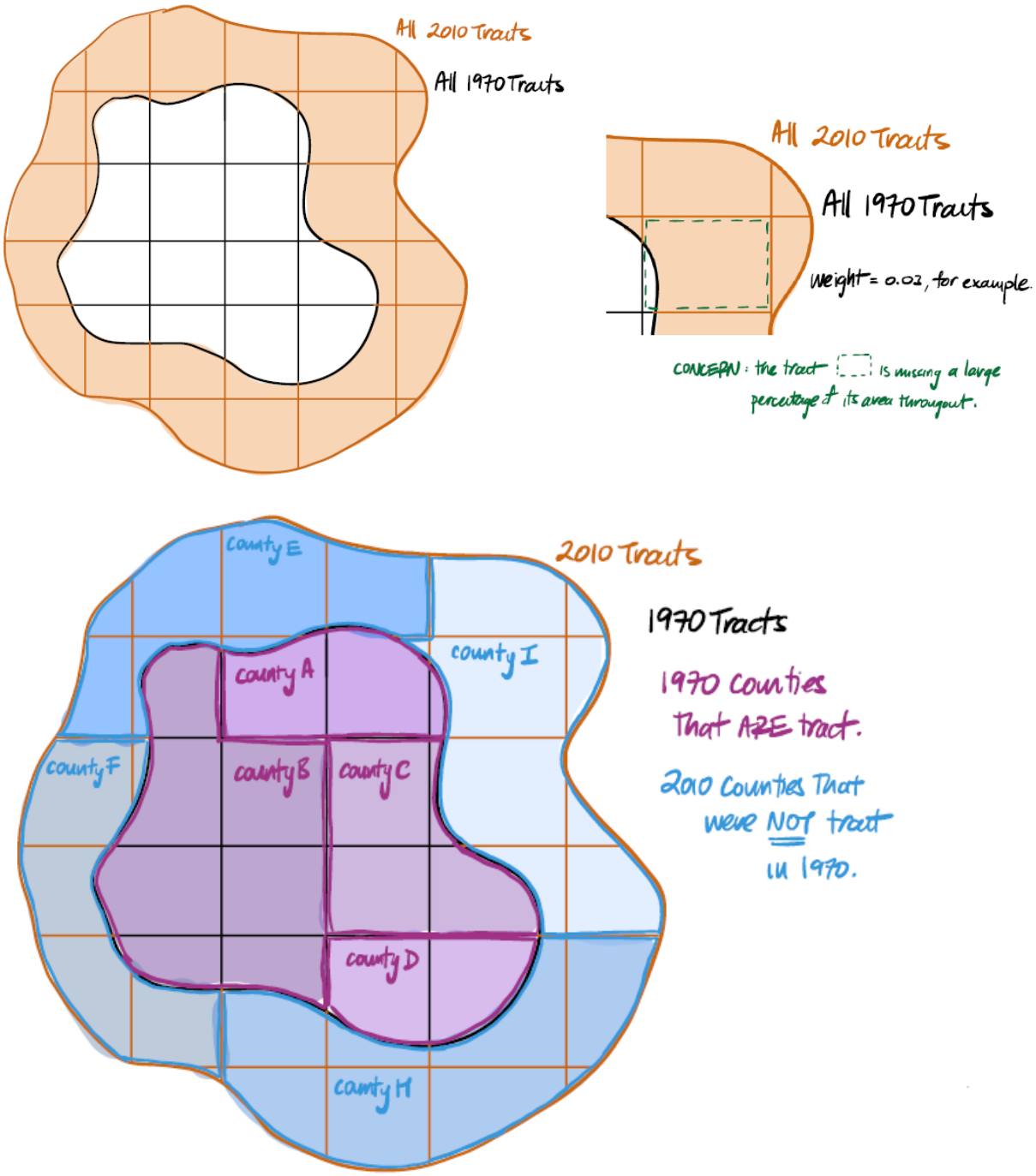
Note: This table reports the distribution of correlation coefficients from the LTDB for within-CBSA correlation between initial prices and median income, non white population shares, and housing unit density for the top 100 most populous CBSAs. The observations are 1970 correlation coefficients at the CBSA level. We compute the correlation coefficient for all CBSAs (with over 20 tracts) in Appendix Table B4.

Appendix Table B5: Distribution of Regression Coefficients from on Residuals from Log Population Changes Regressed on Log House Prices, Log Income and Non-White Share of Population

Independent Variable		***Panel A: Log House Prices***		***Panel B: Log Income***		***Panel C: Non-White Population Share***	
		All CBSAs	*Top 100 CBSAs*	*All CBSAs*	*Top 100 CBSAs*	*All CBSAs*	*Top 100 CBSAs*
Percentiles							
	1%	-0.236	0.089	-0.274	0.103	-0.217	0.122
	5%	0.081	0.280	0.075	0.283	0.096	0.277
	10%	0.229	0.412	0.227	0.399	0.254	0.394
	25%	0.538	0.696	0.527	0.666	0.563	0.681
	Median	0.882	0.935	0.882	0.931	0.902	0.940
	75%	0.979	0.977	0.976	0.973	0.989	0.984
	90%	1.027	1.006	1.026	1.006	1.038	1.014
	99%	1.276	1.123	1.334	1.085	1.362	1.111
Moments							
	Mean	0.745	0.818	0.741	0.812	0.760	0.822
	Std. Dev.	0.342	0.248	0.351	0.248	0.349	0.251
	Min	-1.900	-0.241	-2.217	-0.117	-1.733	-0.131
	Max	1.920	1.186	2.153	1.163	2.098	1.184
Observations	*N*	1818	500	1818	500	1819	500

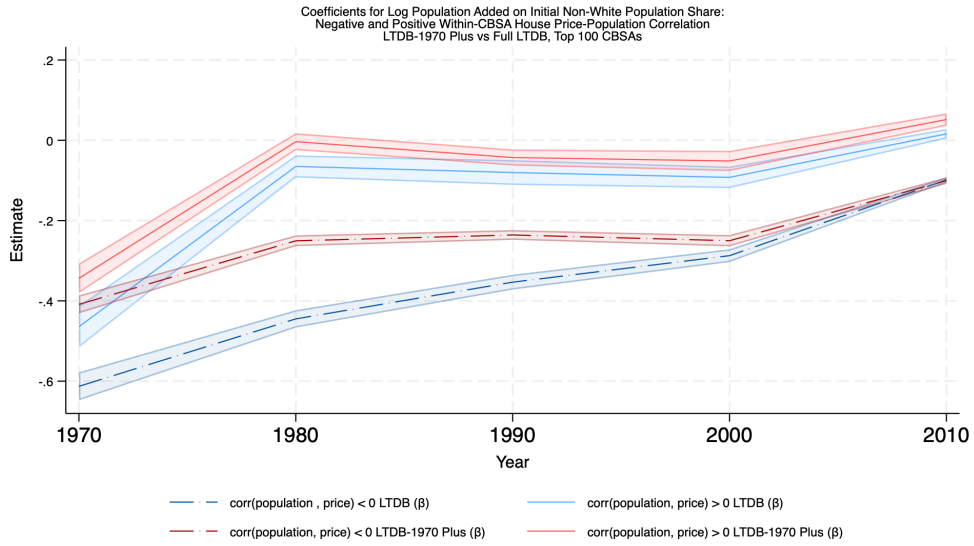
Note: Panel A, Panel B and Panel C of this Table follow the same process outlined in the rest of this note. For illustration, consider Panel A, which uses house prices. This table reports the distribution of regression coefficients from the following sequence of regressions. First, we estimate for each CBSA-decade, a log-log specification of log changes in population on initial log median house prices. We extract the CBSA-decade-specific residuals from this first regression and then run CBSA-decade specific regressions for the log changes in the housing stock on those residuals. The result are regression coefficients for each CBSA-decade. Since there are 5 decades (1970, 1980, 1990, 2000 and 2010), the first sub-column reports the distribution of CBSA-decade specific coefficients for all CBSA-decades with more than 20 observations. The second sub-column reports the distribution of CBSA-decade specific coefficients for the top 100 most populous CBSAs. This follows the specification outlined in equation (1c) and is the decadal CBSA-level distribution of the ϑ parameter.

Appendix Figure B1: *Tract Evolution at the Boundaries of CBSA's*



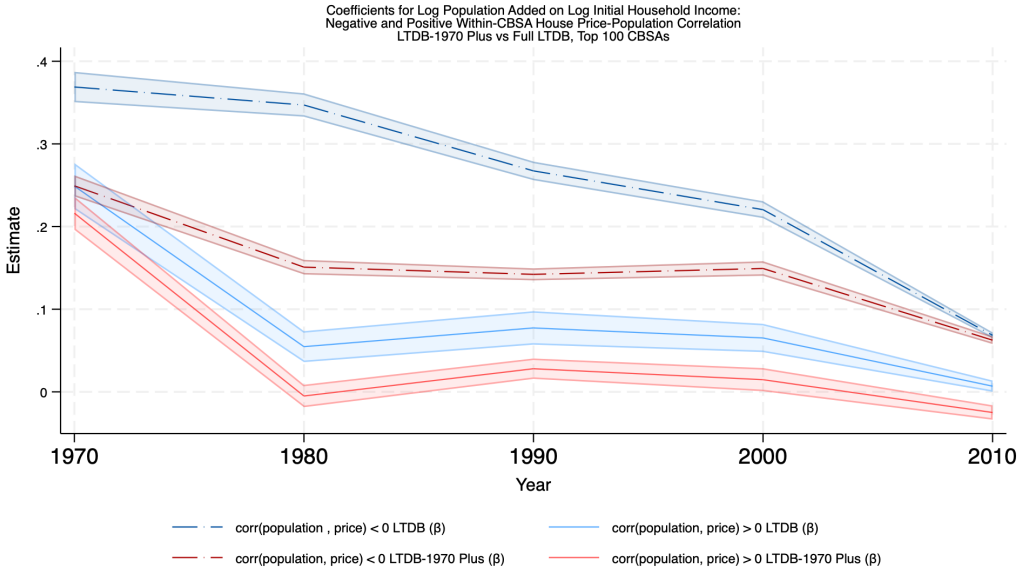
Note: The three panels of this figure are an illustration of how tracts begin in the data. The top left panel shows an example akin to Figure 3 in the main text. There are core 1970 tracts, and additional 2010 tracts within the CBSA. The top right panel shows a potential concern we may have in the data. If the 2010 tract (green dots) is missing 97% of the tract area in 1970, then our weights will be artificially low and undercount at the boundary of the CBSA. However, the bottom panel shows how the tracts actually enter the data. The main 1970 tracts are broken into 4 whole counties, which seldomly change throughout the sample. Those countries are entirely contained in the 1970 definition of the CBSA. However, in 2010, there are 4 additional counties added, and those counties are broken into tracts. This means that a tract on the periphery of a CBSA will never be a small chunk of a larger tract that is cut off by some arbitrary city boundary. Instead, through time it is the counties which define the tracts, and thus no undercounting on the periphery of the CBSAs occurs. A tract only enters in the data once the county enters the data, and at that point all tracts within the county enter the data.

Appendix Figure B3: *Log Population Added on Log Non-White Population Shares for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSA*



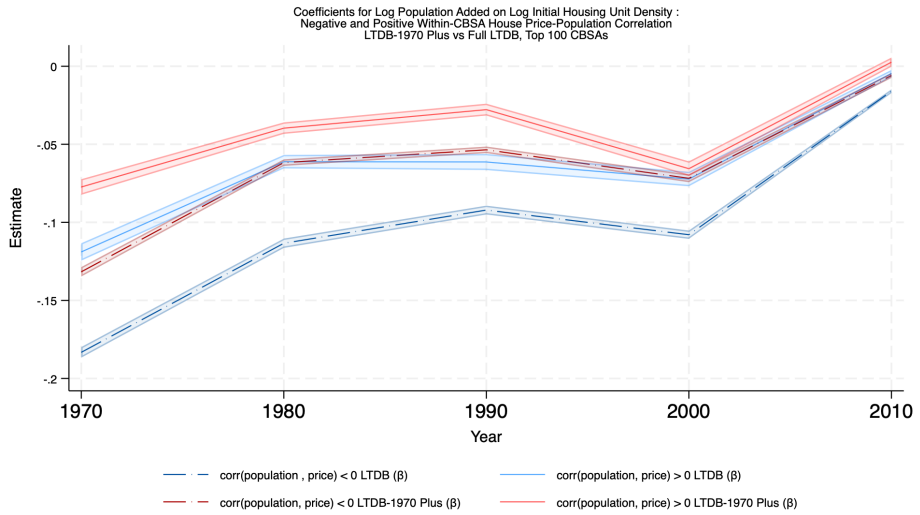
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and distance to center are negative (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure B4: *Log Population Added on Log Housing Unit Density for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSA*



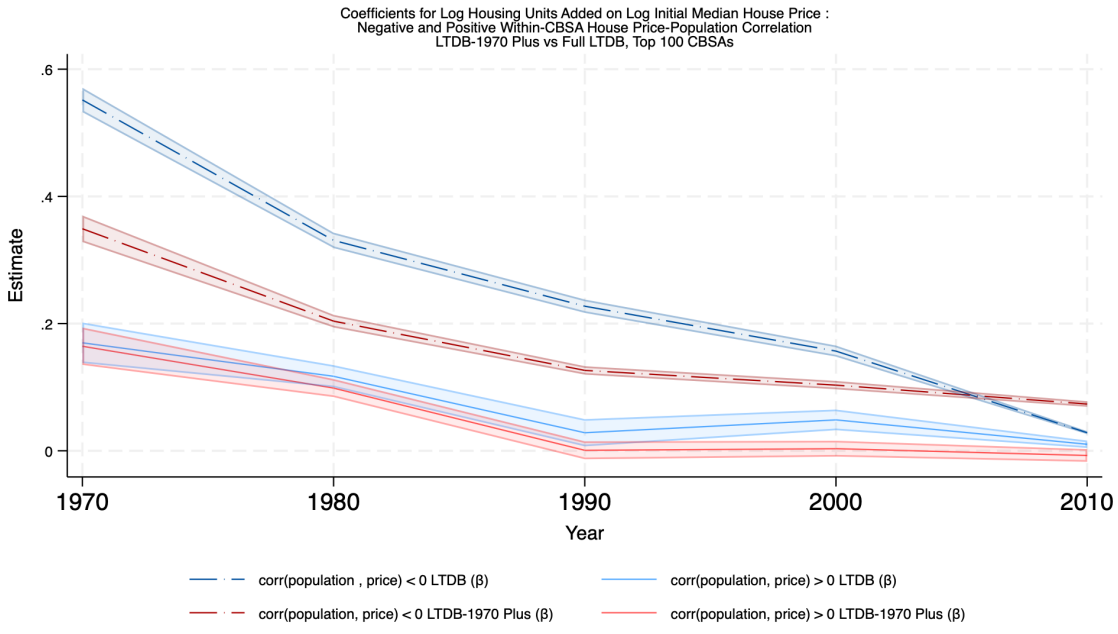
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and housing unit density are negative (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure B5: *Log Population Added on Housing Unit Density for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSA*



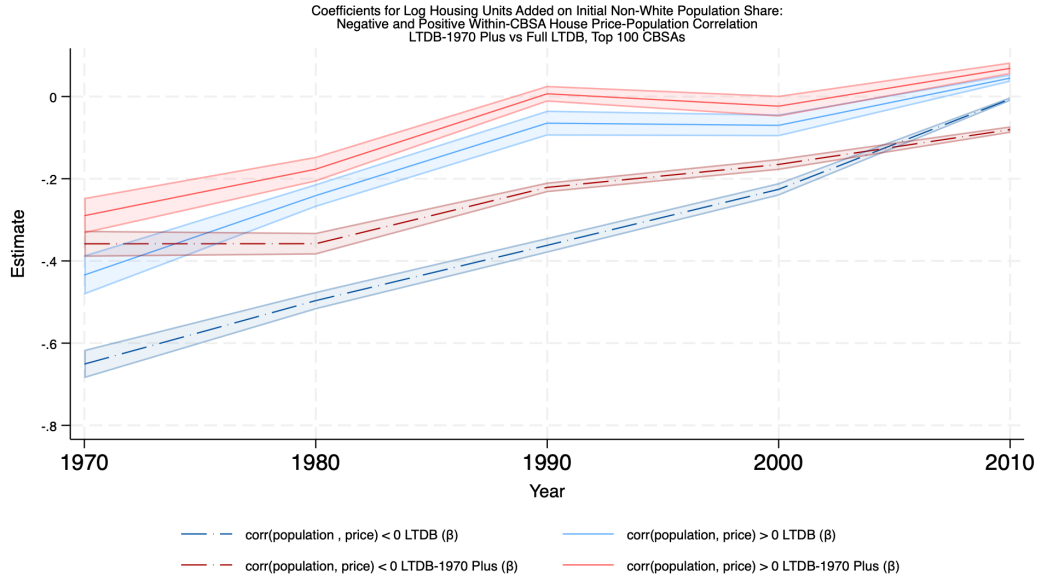
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and income are negative (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure B6: *Log Housing Units Added on Log Prices for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSA*



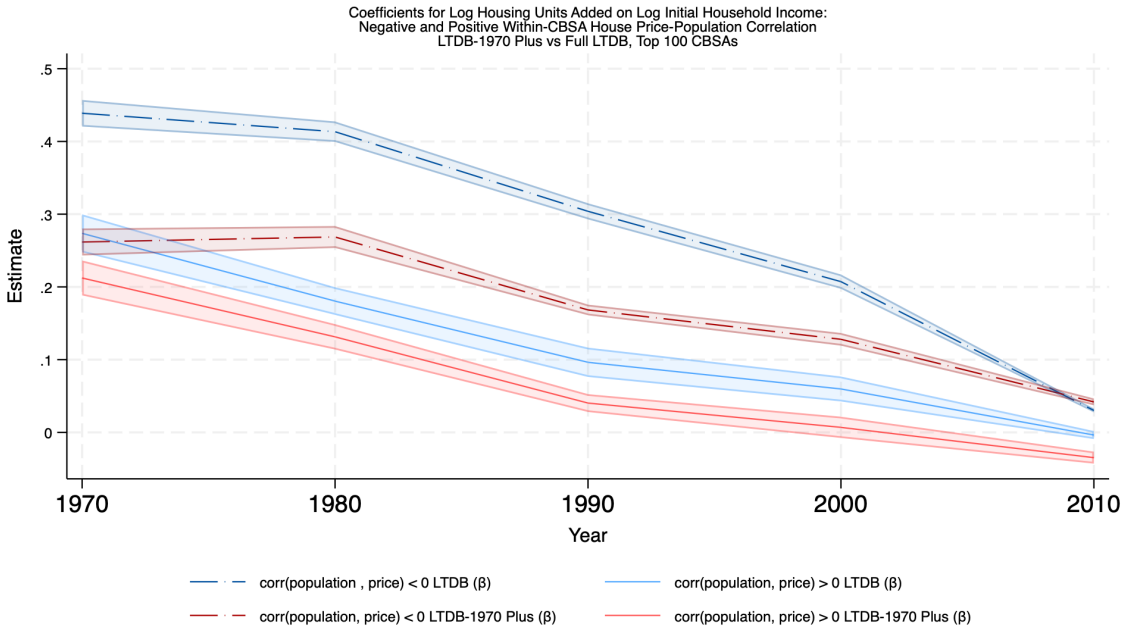
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log housing unit changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and prices are negative (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure B7: *Log Housing Units Added on Log Non-White Population Shares for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSA*



Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log housing units changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and distance to center are negative (dotted lines), and all other CBSAs (solid lines). Housing units are in logs, and population shares are in percentage points – we have no other normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure B8: *Log Housing Units Added on Log Income for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSA*



Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log housing units changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and income are negative (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Section C: Coverage Maps Over Time—National and Select CBSAs

Figure C1: Entire Sample of Tracts in Reverse LTDB, by Decade of Tracts Definitions

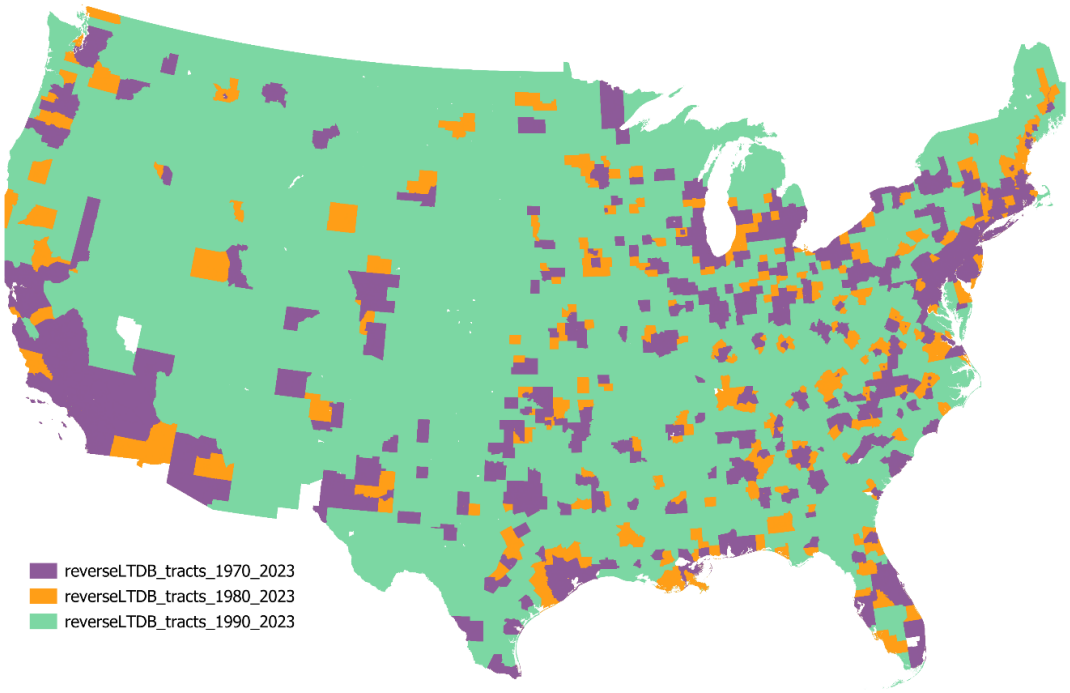


Figure C2: Entire Sample of Tracts in Reverse LTDB, by Decade of Tracts Definitions

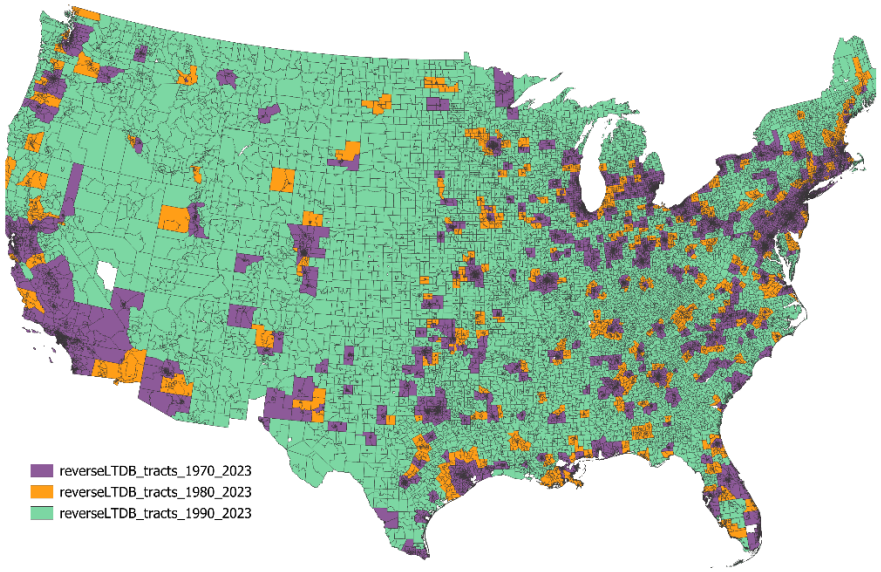


Figure C3: Philadelphia, Baltimore and Washington CBSAs and Surrounding Areas in Reverse LTDB, by Decade of Tracts Definitions

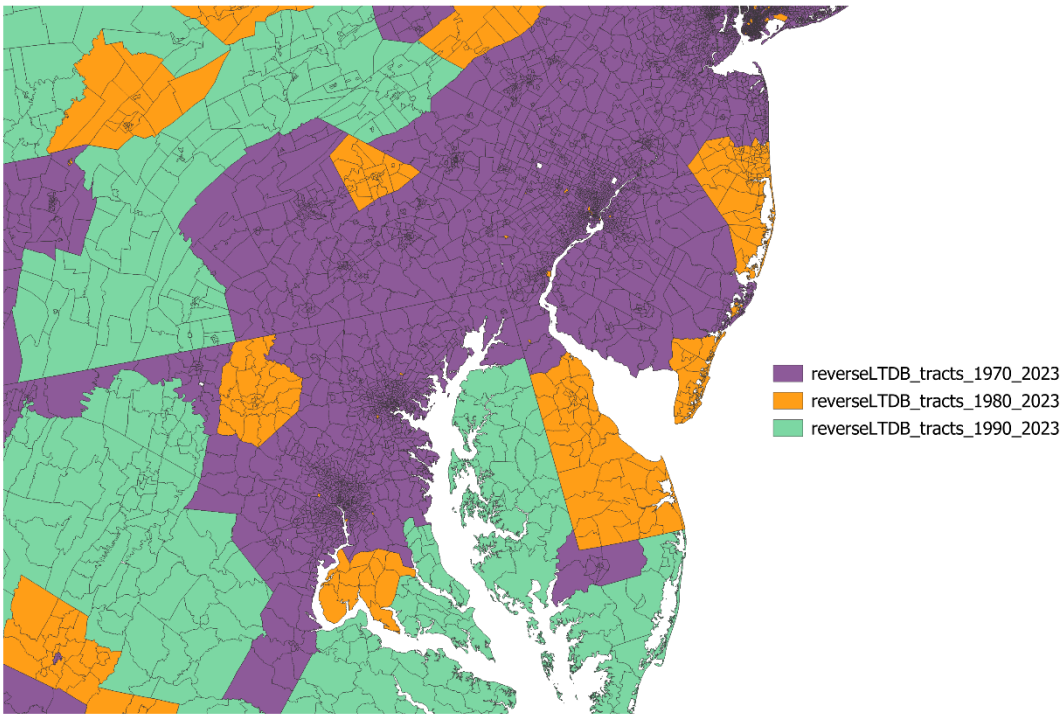


Figure C4: Atlanta in Reverse LTDB, by Decade of Tracts Definitions

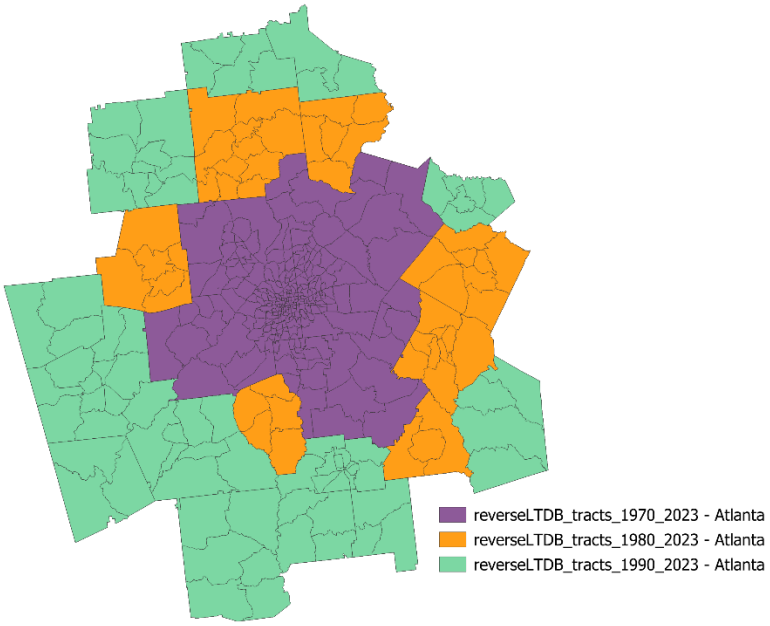


Figure C5: Austin in Reverse LTDB, by Decade of Tracts Definitions

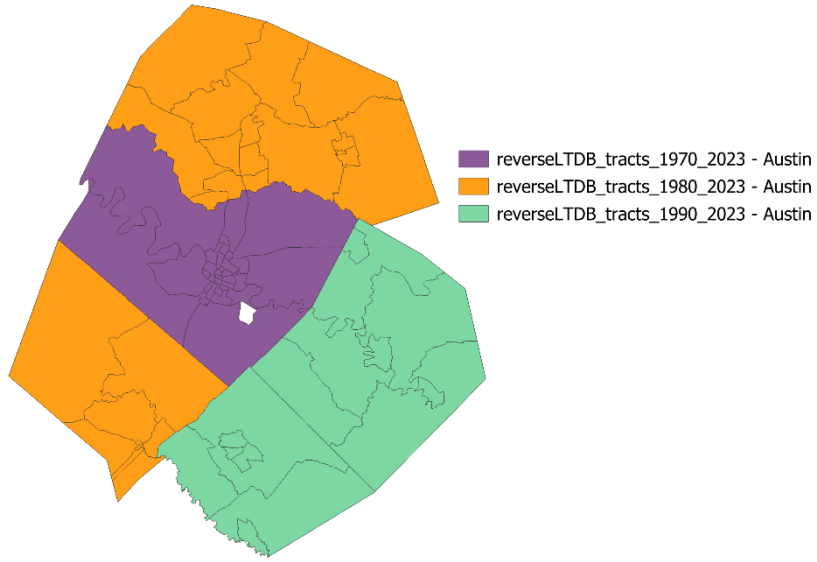


Figure C6: Austin and San Antonio in Reverse LTDB, by Decade of Tracts Definitions

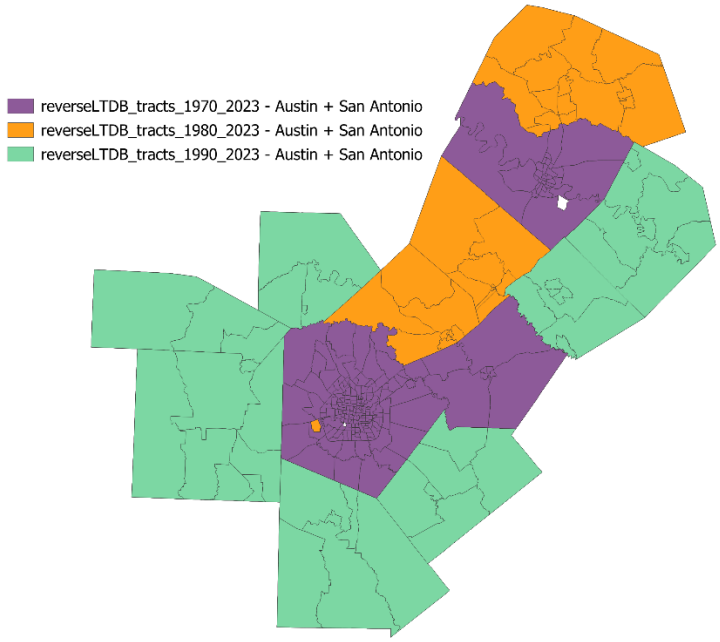
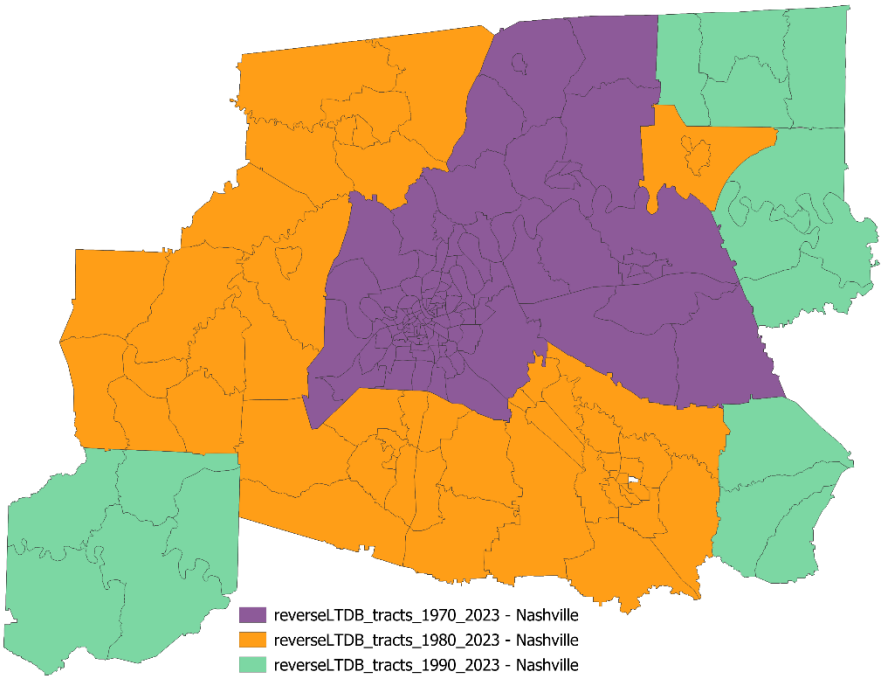


Figure C7: Nashville in Reverse LTDB, by Decade of Tracts Definitions



Appendix Section D: Missing Data Reweighting and TIGER/Line Tract Intersections

In this section, we address two primary concerns about the LTDB-1970 (and tract harmonization methods, in general). This first arises from the potential for data to be missing. Because we use area weighting to construct aggregates of count variables, missing data from source year geometries will cause underestimates for the consistent 1970's tract boundaries. In some year, if a component of a 1970-tract boundary intersection set is missing data for a count variable such as population, and we sum over those components to get the harmonized 1970-boundary value in that year, we will underestimate the 1970-tract population in that year. Distributional variables, on the other hand, turn out to be less sensitive to missing data since most economic variables do not drastically vary across small units of space such as neighboring census tracts. Hence, taking re-weighted averages will not bias the results unless the missing data is correlated with some other variable of interest.[1] A second concern arises from the fact that we have to use different vintages of TIGER/Line shapefiles. This could generate inaccuracies in the spatial intersections discussed above.[2]

We define missing data by the percentage of the 1970 census tract area that is missing source year tract data. In the final dataset, a count variable is considered missing in a later year if over 25% of the 1970's tract area is missing data for that variable. A distribution variable is considered missing if over 75% of the 1970's tract area is missing for that variable. If, for a given 1970 tract boundary and its source-year tract intersections, the area of missing data is less than 75%, the non-missing area becomes the new 1970's tract area for that weighted average. The LTDB, on the other hand, only sets a tracts data to missing if 100% of the tracts data is missing, which leads to systematic undercounting of count variables.[3]

Appendix Figure D1 provides an illustration of these issues. The left panel shows the 1970 tract boundary (in blue) for G4802010032200, which is southeast of Houston's center-city, overlayed on the 2000 tract boundaries for Houston. The right panel shows four intersecting tracts. Appendix Table D1 provides a glimpse into the underlying data. We can see that the northernmost (red) intersecting 2000 tract (G4802010320400) is missing data for

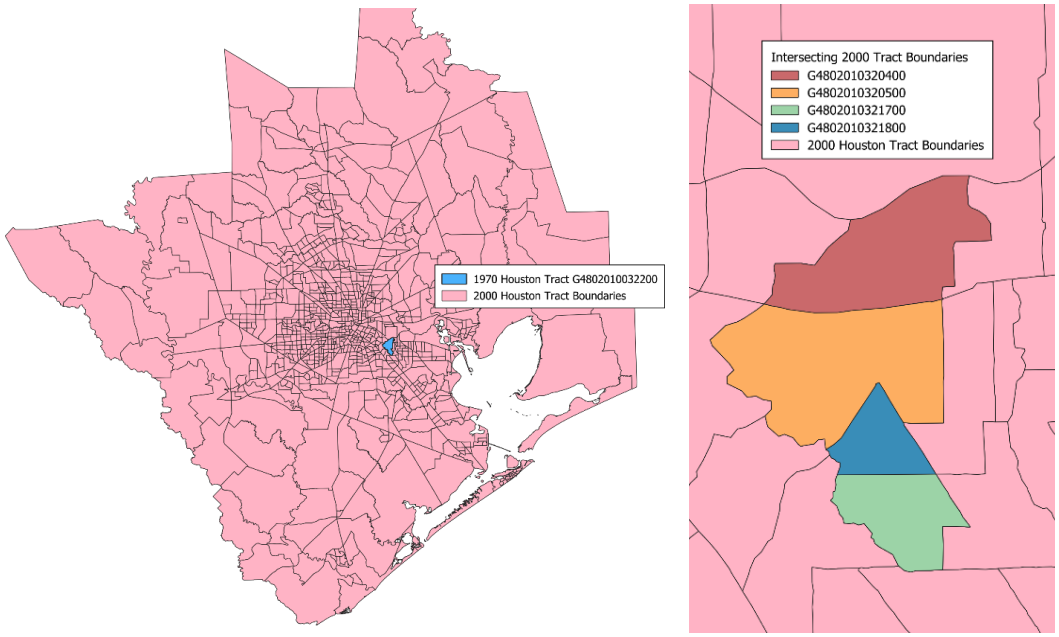
[1] Additional analysis that is available upon request did not find any strong correlations of missing data flags with various other covariates.

[2] More specifically, using the 2000 TIGER/Line shapefiles for our base 1970 census tract layer along with the 2020 vintage TIGER/Line shapefiles for 2010, and 2020 layers generates a significant number of small area weights. This issue is discussed in Appendix Section G, *Sensitivity to Missing Source Year Data*.

[3] See Appendix Section H for a tabulation of the missing source-year tract data in the LTDB (Appendix Tables H1 and H2), as well as a discussion for how those missing source year tracts affect a larger percentage of tracts when harmonizing to the 2010 tract delineation than in the 1970-LTDB

median home value (*2000 Median Home Value ($)*). The *Missing 2000s Area in 1970 Tract (m²)* column shows that 4,061,438.4 square meters of the 1970 tract is missing, which makes up 30.5% (*% of Area Missing Home Value*) of the 1970 tracts total area (*1970 Tract Total Area (m²)* = 13,314,794 square meters).[4] The key variable is '*% of Area Missing Home Value'*. If this value is over 75%, then we would set the 1970 tract to have a missing value and it would not be included in our database. However, because the missing area is less than 75%, we construct the owner-occupied contribution weights with the non-missing areas (*Non Missing Owner-Occupied Area Weight).* These weights sum to one, so taking the weighted average of *2000 Median Home Value ($)*, we get a 1970 tract boundary value of $55,748 for the median home value (*Imputed 1970 Tract 2000-Median Value ($)*). Appendix Tables D2 and C3 then tabulate the counts of missing (and excluded) data, and the percentage of missing (and not excluded) data for key variables in our data. The largest count of missing (and excluded) data is rental prices, reaching 5.4% (1,894/34,440) in 1980.[5]

Appendix Figure D1: *1970 Tract G1302470060300 Intersection with Four 2000's Tracts*



[4] Note that the reason *Missing 2000s Area in 1970 Tract (m²)* does not exactly equal the *2000 Tract Total Area (m²)* column for tract G4802010320400 is because of small differences in the tract definition. The G4802010320400 2000 tract is slightly larger on the north-eastern most component of the 1970's tract. Thus, some of the missing 2000's area from the G4802010320400 tract is not counted towards the total missing 1970's area because it would not be included in the tract intersection.

[5] See Appendix Section H for an in-depth discussion of the missing source-year tract data in the LTDB versus the 1970-LTDB. There, using recreated Appendix Tables D2 and D3 and equivalent tables for the LTDB, we conclude that problems arising from missing source year data for tracts that have intersections with multiple other tracts in years after 1970 are relatively rare in the 1970-LTDB, and are less significant than those in the LTDB.

Appendix Table D1: *1970 Tract G4802010032200 Reweighting for Missing Area*

1970 Tract ID (GISJOIN1970)	2000 Tract ID (GISJOIN2000)	2000 Median Home Value ($)	2000 Tract Total Area (m²)	1970 Tract Total Area (m²)	Missing 2000s Area in 1970 Tract (m²)	% of Area Missing Home Value	Non-missing Owner-Occupied Area Weight	Weighted Median Value Contribution ($)	Imputed 1970 Tract 2000-Median Value ($)
G4802010032200	G4802010321700	67,400	2,101,515.50	13,314,794	4,061,438.40	30.50%	0.3212	21,649.85	55,748
G4802010032200	G4802010320500	51,800	5,848,905.10	13,314,794	4,061,438.40	30.50%	0.3661	18,968.74	55,748
G4802010032200	G4802010321800	48,400	1,302,953.10	13,314,794	4,061,438.40	30.50%	0.3126	15,129.54	55,748
G4802010032200	G4802010320400	–	4,191,239.90	13,314,794	4,061,438.40	30.50%	0	–	55,748

Note: *GISJOIN1970* is the unique identifier for the 1970 tract boundaries. *GISJOIN2000* is the unique identifier for the 2000 tract boundaries. In this screenshot, one 1970's tract intersects with six 2000's tracts. *2000 Tract Total Area* measures the area (m^2) for the 2000 tracts, while *1970 Tract Total Area* measures the area (m^2) for the 1970 tract. *Missing 2000s Area in 1970 Tract (m^2)* provides the total area within the 1970's tract that is missing the median home value variable. The subsequent column *% of Area Missing Home Value* divides the total 1970 tract area by the total missing area for the median home value variable. This is the most important column, because if it is over 75% for the distribution variables, or over 25% for count variables, we set the 1970's tract to missing for that year and variable. Since 4,061,438.40 m^2 is missing, the new total 1970 tract area used in the estimate for the home value is 13,314,794 - 4,061,438.40 = 9,253,355.6. The *Non-Missing Owner-Occupied Area Weight* is the owner-occupied contribution area weight generated for the crosswalk. Notice that for the tract with missing data, the weight is zero. *2000 Median Home Value ($)* is the 2000's tract median home value, which when multiplied by the *Non-Missing Area Weight* results in the *Weighted Median Value Contribution ($)*. This ensures that places with more owner occupied unites are weighted heavier than those with less owner-occupied units. Notice that for the tract with missing data, the weighted median value contribution is null. Summing over all 2000 tracts, we get the estimate for the median home value within the 1970 tract boundary G4802010032200 for the year 2000 (*Imputed 1970 Tract 2000-Median Value ($)*)

Appendix Table D2: *Counts of Tracts by Percentage of Area Missing in 1970-LTDB (count variables)*

Missing Category	Year	Tract Count	Total Units	Owner Occupied	Renter Occupied
Panel A: 0-25% 1970-Tract Area Missing	1980	34,440	0	186	178
	1990	34,424	0	373	388
	2000	34,390	0	77	63
	2010	34,359	0	203	166
	2020	34,350	0	0	0
Panel B: >25% 1970-Tract Area Missing	1980	34,440	0	1,530	1,388
	1990	34,424	0	344	175
	2000	34,390	0	240	156
	2010	34,359	0	335	227
	2020	34,350	0	0	0

Note: This Table reports the total count of tracts in the LTDB-1970 by year, and the tract-year count of tracts with a percentage of area missing between 0 and 25% (Panel A), and over 25% (Panel B) for important count variables. The columns for total units, owner occupied and renter occupied tabulate the tract counts. The cutoff for dropping harmonized tracts for count variables is 25%, which means that the number of tracts dropped in each year due to missing raw source year data is equal to Panel B. For example, in 1980, we have 178+1,388=1,566 total 1970's tracts that have some amount of missing source-year data in their tract intersections for renter occupied units. 178 of those tracts had less than 25% of the 1970's tract area missing the renter occupied units variable and are not dropped. 1,388 of the tracts had over 25% of the data missing, and are dropped from the sample. In comparison, observations within the LTDB are only dropped if exactly 100% of the area is harmonized tract area is missing.

Appendix Table D3: *Counts of Tracts by Percentage of Area Missing in 1970-LTDB (distribution variables)*

Missing Category	Year	Tract Count	House Prices	Median Household Income	Rental Prices
Panel A: 0-75% 1970-Tract Area Missing	1980	34,440	580	420	647
	1990	34,424	453	435	565
	2000	34,390	421	119	355
	2010	34,359	812	450	1,032
	2020	34,350	1,666	850	2,136
Panel B: >75% 1970-Tract Area Missing	1980	34,440	1,815	1,117	1,894
	1990	34,424	549	243	349
	2000	34,390	1,054	199	508
	2010	34,359	582	228	411
	2020	34,350	1,002	357	781

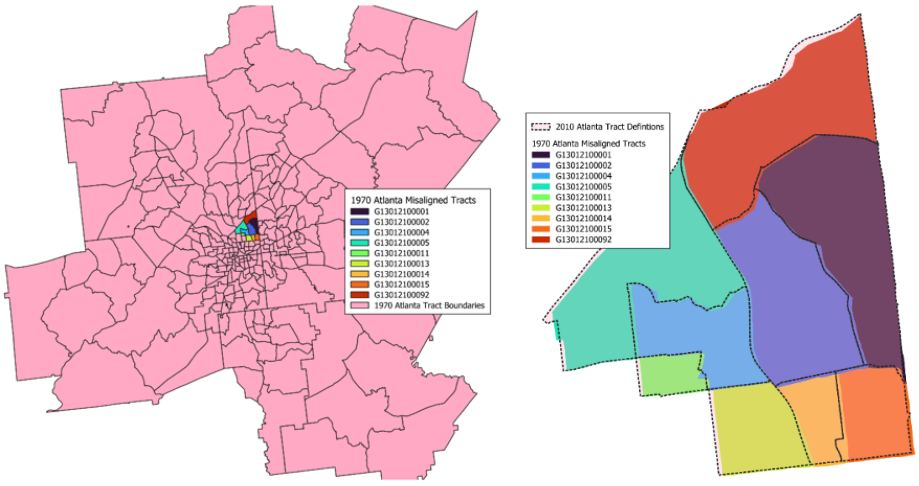
Note: This Table reports the total count of tracts in the LTDB-1970 by year, and the tract-year count of tracts with a percentage of area missing between 0 and 75% (Panel A), and over 75% (Panel B) for important distribution variables. The columns for house price, median household income and rental prices tabulate the tract counts. The cutoff for dropping harmonized tracts for distribution variables is 75%, which means that the number of tracts dropped in each year due to missing raw source year data is equal to Panel B. For example, in 2020, we have 1,666+1,002=2,668 total 1970's tracts that have some amount of missing data in their tract intersections for house prices. 1,668 of those tracts had less than 75% of the 1970's tract area missing the house price variable and are not dropped. 1,002 of the tracts had over 75% of the data missing, and are dropped from the sample. In comparison, observations within the LTDB are only dropped if exactly 100% of the area is harmonized tract area is missing.

The second potential problem with the Reverse LTDB arises from the use of different TIGER/Line vintages.[6] As mentioned in the text, using a common TIGER/Line vintage allows for precise geographical matching of tracts that do not change. However, when we use different vintages, even for unchanging tract boundaries, the overlapping areas can change. While the occurrences are numerous, the weights generated are miniscule, and thus the impact is quite small in the vast majority of cases.

Appendix Figure D2 overlays a set of tracts in Atlanta that do not change from 1970-2010. However, upon closer inspection, the right panel shows that there are deviations in the 2010 census tract boundary lines – which uses the 2020 TIGER/Lines. This is not due to mismatched tracts; rather, it is the shapefiles themselves that are drawn slightly differently. Mechanically, this does two things. First, it generates *many* more tract-intersections. Second, those intersecting areas are very small, resulting in equally small weights.

Appendix Table D4 tabulates the first fact. We have a stark increase in the number of tract intersections once we switch to the 2020 TIGER/Lines for 2010 – from 66,899 in 2000 to 227,651 in 2010. This is not driven by a markedly large increase in the number of source-year tracts available. The number of intersections per source year tract goes from 1.14 (44,101/38,406) in 1980 to 3.77 (240,152/63,633) in 2020. Appendix Table D4 also documents the second fact – that the distribution of weights within those decades are heavily skewed towards zero. The mean of an intersection's calculated weight in 2000 goes from 0.7014 to 0.2295 in 2010. From 1980-2000, the median weight was 1, while from 2010 onwards, the median weight is 0.005 and 0.006.

Appendix Figure D2: *Unchanging 1970 Tracts with Misaligned Tract Boundaries in 2010*



[6] Logan, et. al. (2014) provide no information on any possible differences in TIGER/Line vintages used in the construction of their crosswalks across years.

Appendix Table D4: *Source Year Tract Counts, Tract Intersections and Distribution of Weights in the 1970-LTDB*

	1980	1990	2000	2010	2020
Unique Source Year Tracts	38,406	42,582	46,987	55,923	63,633
Total Tract Intersections	44,101	53,859	66,899	227,651	240,152
Intersections per Source Year Tract	1.148	1.265	1.424	4.071	3.774
Distribution of Weights	**1980**	**1990**	**2000**	**2010**	**2020**
Mean	0.8705	0.7900	0.7014	0.2295	0.2483
1st p-tile	0.0011	0.0003	0.0000	0.0000	0.0000
10th p-tile	0.1747	0.0269	0.0045	0.0000	0.0001
25th p-tile	1	0.8031	0.1828	0.0006	0.0007
50th p-tile	1	1	1	0.0050	0.0060
75th p-tile	1	1	1	0.1923	0.4375
90th p-tile	1	1	1	0.9869	0.9902
90th p-tile	1	1	1	1	1

Note: The top panel of this table provides a tabulation of the unique source year tract count by year (row 1), and the total tract intersections by year (row 2). Row 3 divides the total tract intersections by the number of unique source year tracts to show how often a source-year tract intersects with an underlying harmonized 1970 tract. As mentioned in the text, notice the jump in tract intersections and intersectionsper source year tract in 2010 and 2020, when we use the 2020 TIGER/Line vintage. The table's lower panel provides a distribution of the weights for each source-year tract. In 1980, over 75% of the tracts have a weight of 1, which means that they're falling entirely within 1970 tract boundaries. The median weight remains 1 until 2010, when it falls to 0.005 – driven entirely by the significantly increased number of tract intersections due to mismatched TIGER/Line vintages. Since all of these extra tract intersections generate really small weights, the median is pulled down, and so is the mean.

This problem is two-sided. When we have an incorrect overlapping area, a tract that otherwise should have had a weight of 1 will be less than 1. For example, look at the darkest purple tract (G13012100001) in Appendix Figure D2. This tract should have a weight of 1, but instead it is generating a weight less than 1, and then a very small weight with the remaining area that is incorrectly overlapping with the neighboring orange tract. Appendix Table D5 then shows that the instances of source year tracts that lay *entirely* within a 1970 tract boundary falls by over half, from 81% in 2000 to 35% in 2010. Because it is possible (as in Figure 5 in the main text) that source year tracts span multiple 1970's tracts because of a valid redefinition, the drop in 'perfect match' tracts is not solely driven by the TIGER/Line misalignment. However, if we consider anything with a weight greater than 0.95 a 'perfect match', we can see that the rate of decline in perfect matches is not near as stark. Instead of falling by half, in 2000 the percentage of unique source year tracts with a weight over 0.95 is 87.5%, while in 2010, that percentage falls to only 70.3%. This is important because we

believe that a key strength of the 1970-LTDB is that it mechanically has less noise than the LTDB simply because there are less instances of reweighting.[7]

Appendix Table D5: *Source Year Tract Counts, and Tabulating Small and Perfect Tract Intersections in the 1970-LTDB*

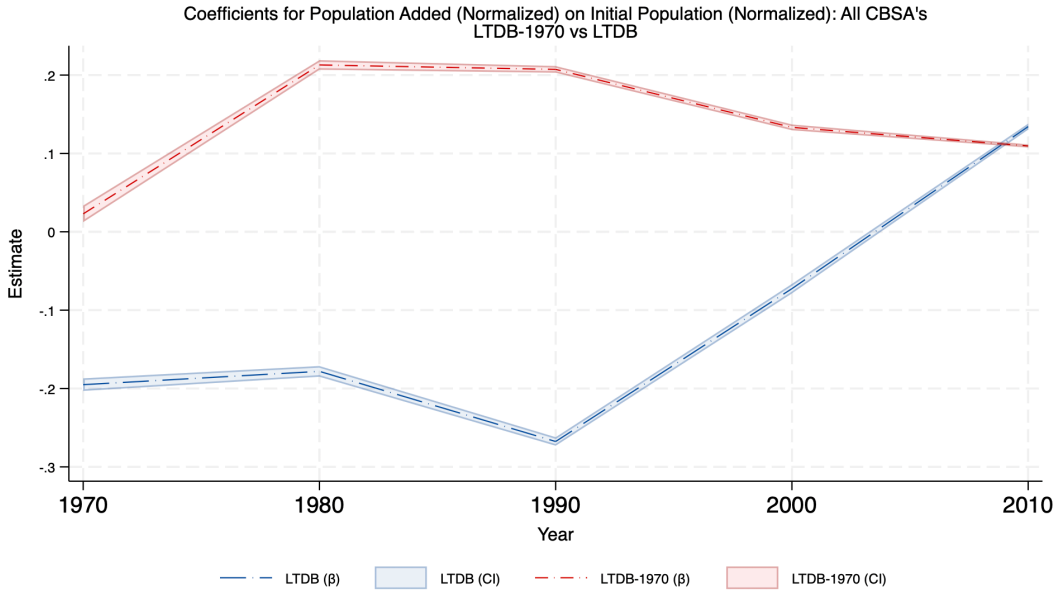
Year	1980	1990	2000	2010	2020
Unique Source-Year Tract Count	38,406	42,582	46,987	55,923	63,633
Source-Year Tracts with weight = 1	35,045	36,579	38,009	19,709	24,219
Percent of Source-Year Tracts with weight = 1	0.912	0.859	0.809	0.352	0.381
Source-Year Tracts with weight > 0.95	36,080	38,533	41,137	39,330	45,336
Percent of Source-Year Tracts with weight > 0.95	0.939	0.905	0.875	0.703	0.712
Total Tract Intersections	44,101	53,859	66,899	227,651	240,152
Intersections with weight < 0.05	3,051	6,576	13,226	162,848	166,530
Intersections with 0.05 < weight < 0.95	4,970	8,750	12,536	25,473	28,286
Intersections with weight > 0.95	36,080	38,533	41,137	39,330	45,336
Percent of Intersections Dropped	0.069	0.122	0.198	0.715	0.693
Final Count of Unique Source Year Tracts after Dropping Intersections with weight < 0.05	38,393	42,556	46,935	52,368	59,780
Unique source year tracts dropped	13	26	52	3,555	3,853
Percentage of unique tracts dropped because they had a weight < 0.05	0.03%	0.06%	0.11%	6.36%	6.06%

Note: There are 5 sections to this table. The first section is equivalent to row 1 of Appendix Table D4. The second section counts the number of source-year tracts which fall *entirely* in a 1970 tract boundary, and computes the percentage of all source year tracts as well. Notice the fall from 0.809 to 0.352 – this is coming from the fact that the mismatched TIGER/Line problem is two-sided. Since many small tract intersections are generated, it is also the case that perfectly aligned tracts will receive weights less than 1. The next section provides a correction in line with our decision rule to drop any intersection with a weight less than 5%. Notably, the source year tract count and percentage with a weight over 95% drastically increases for 2010 and 2020, without increasing 1980, 1990, or 2000 by more than 7%. The fourth section breaks down the tract intersections into three different weight categories. In this case, the number of intersections with a weight less than 0.05 jumps from 13,226 in 2000 to 162,848 in 2010 and 166,530 in 2020. The first row of the fifth section takes the number of tract intersections with a weight less than 0.05 (row 2 of section 4) and divides it by the total tract intersections (row 1 of section 4) to get the percentage of tract intersections dropped before crosswalking in the 1970-LTDB. The second, third, and fourth row of the fifth section count the number of source-year tracts that are dropped as a result of dropping the small tract intersections. A notable flaw of our 5% rule is that when a source year tract is extremely small relative to the original tract (ie, it experiences a significant growth in population), we drop the tract intersection, and thus the very small tract. However, the extent of this is effectively null in 2000 (0.11%), and only increases marginally in 2010 and 2020 to just above 6%.

[7] It is noteworthy that the increase in the percentage of intersections dropped from 6.9% in 1980 to 19.8% in 2000 in Appendix Table D5 does not imply that we are dropping nearly one-fifth of our already lower number of tracts (compared to the LTDB) in 2000. Rather, we are dropping nearly 20% of the instances where a 2000 tract intersects with a 1970 tract and the weight calculated is less than 0.05 (5%). The number of unique source year tracts being dropped is only 52 (column 3 of the next to last row in Appendix Table D5). This is only 0.11% of the 38,009 total source year tracts that year.

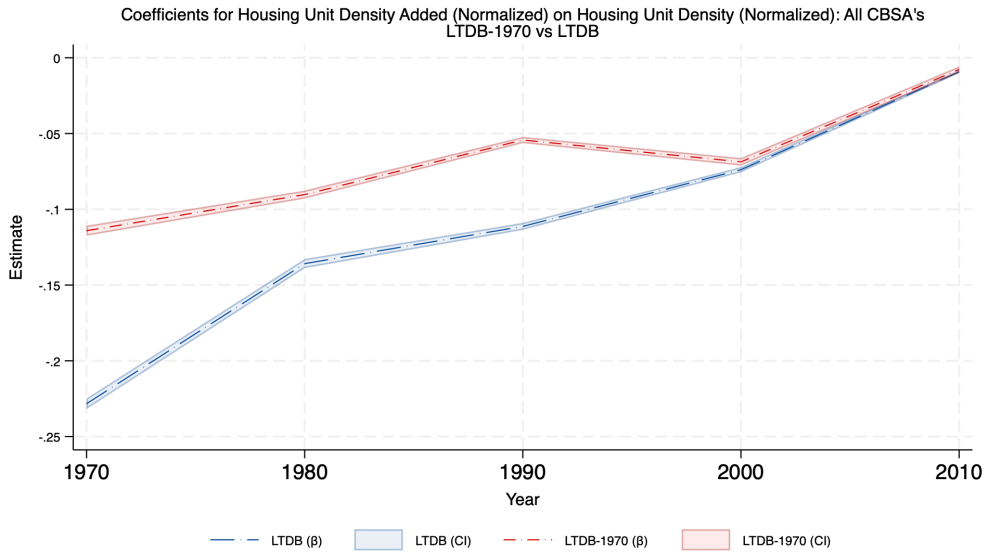
Appendix Section E: Regression Results for 1970-LTDB

Appendix Figure E1: *LTDB-1970 Version of Figure 7 – Coefficients from Regression of Population Changes on Initial Population*



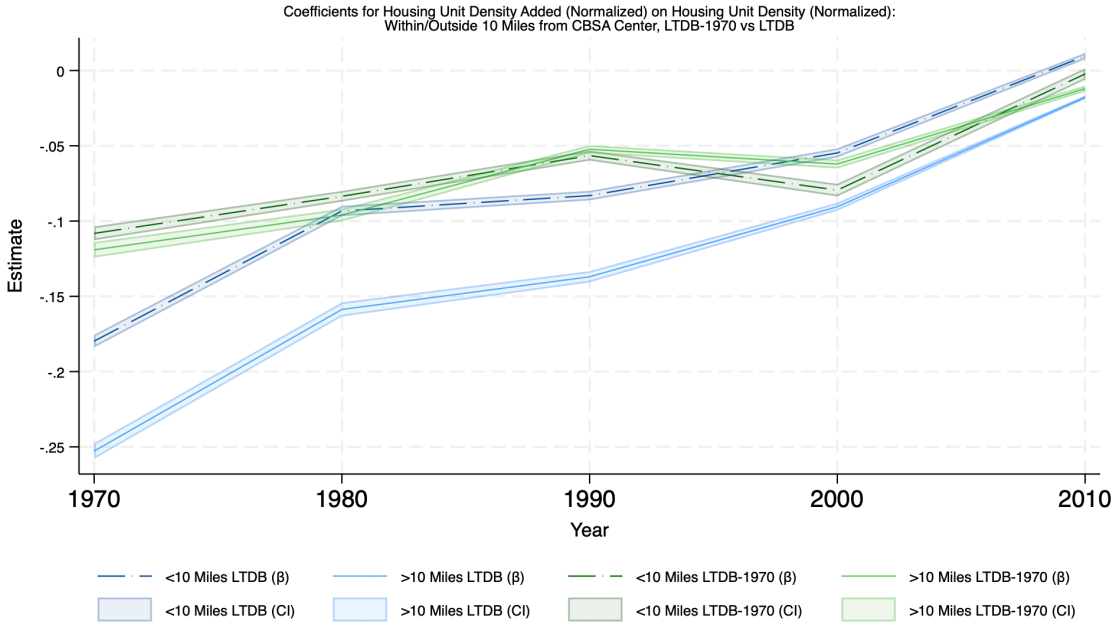
Note: The lines in this figure report the coefficients from normalized decadal tract-level regressions from 1970-2010 for decadal population changes on initial population in the LTDB and LTDB-1970, separately. Changes in population and initial population are normalized by CBSA-specific base-year suburban/city-center mean population. The shaded areas are 95% confidence intervals. The sample includes all 309 CBSAs, and all regressions include CBSA fixed effects.

Appendix Figure E2: *LTDB-1970 Version of Figure 8 – Housing Unit Density Change and Initial Housing Unit Density*



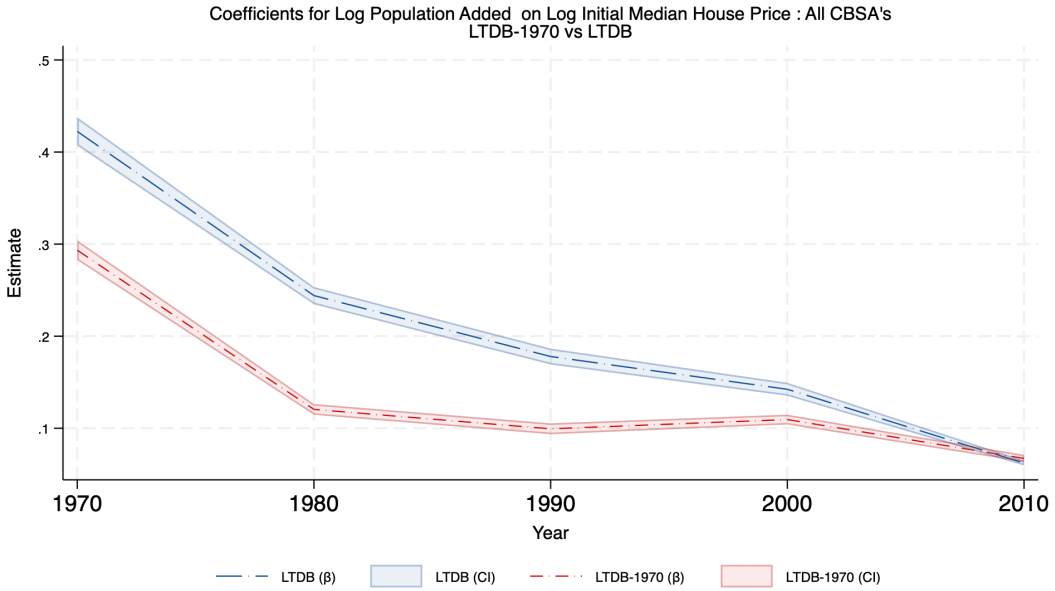
Note: The lines in this figure report the coefficients from normalized decadal tract-level regressions from 1970-2010 for decadal housing density changes on initial housing unit density for the LTDB and LTDB-1970, separately. Changes in housing unit density and initial housing unit density are normalized by CBSA-specific base-year suburban/city-center mean population. The shaded areas are 95% confidence intervals. The sample includes all 295 CBSAs, and all regressions include CBSA fixed effects.

Appendix Figure E3: *LTDB-1970 Version of Figure 9 – Regression Coefficients for Housing Unit Density Change on Initial Housing Unit Density in Suburbs*



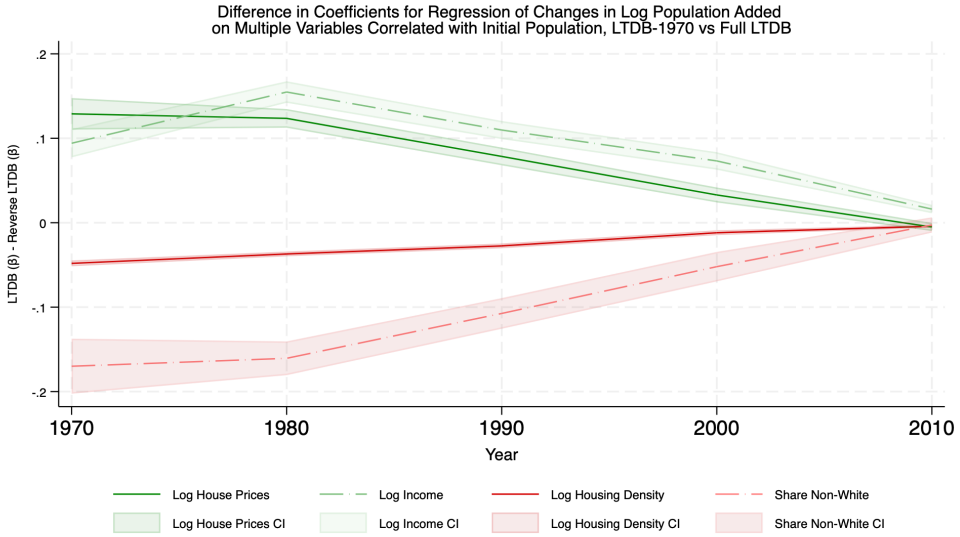
Note: The lines in this figure report the coefficients from normalized decadal tract-level regressions from 1970-2010 for decadal housing density changes on initial housing unit density for the LTDB and LTDB-1970, separately. Within the datasets, the sample is split into tracts greater or less than 10 miles from the CBSA center, and regressions are run separately. Changes in housing unit density and initial housing unit density are normalized by CBSA-specific base-year suburban/city-center mean population. The shaded areas are 95% confidence intervals. The sample includes all 295 CBSAs, and all regressions include CBSA fixed effects.

Appendix Figure E4: *LTDB-1970 Version of Figure 10 – Regression Coefficients for Log Population Added on Log Initial Prices for All CBSAs*



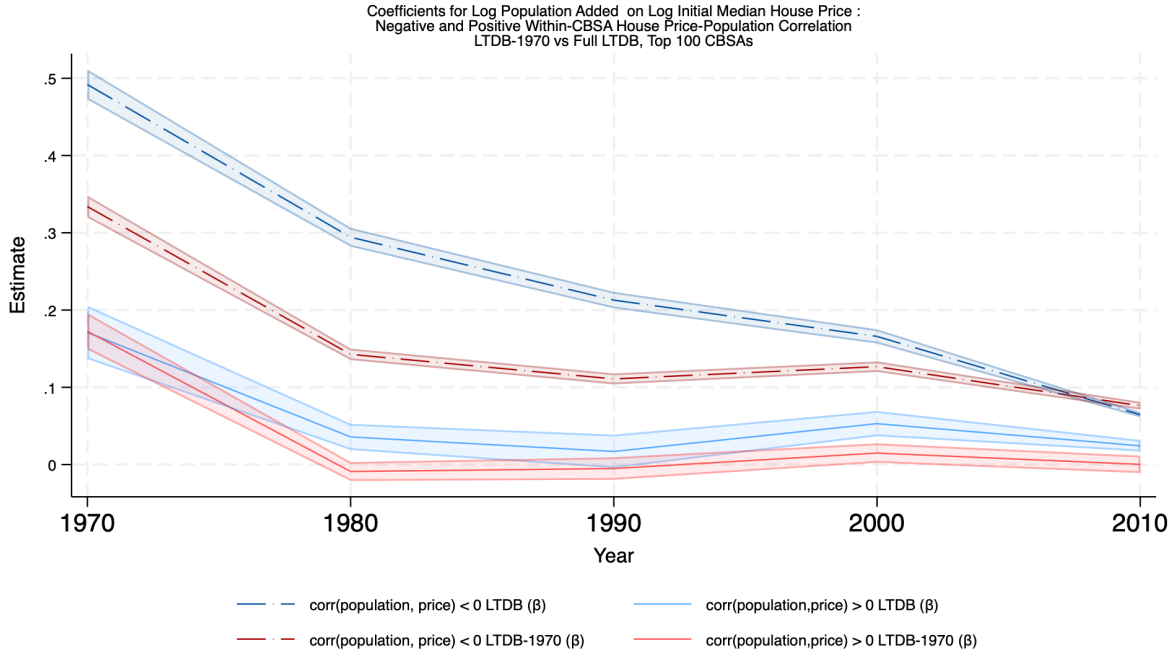
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median house values for the LTDB and LTDB-1970, separately. The samples for these regressions are all CBSAs. The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure E5: *LTDB-1970 Version of Figure 11 – Regression Coefficients for Log Population Added on Multiple Variables Correlated with Population*



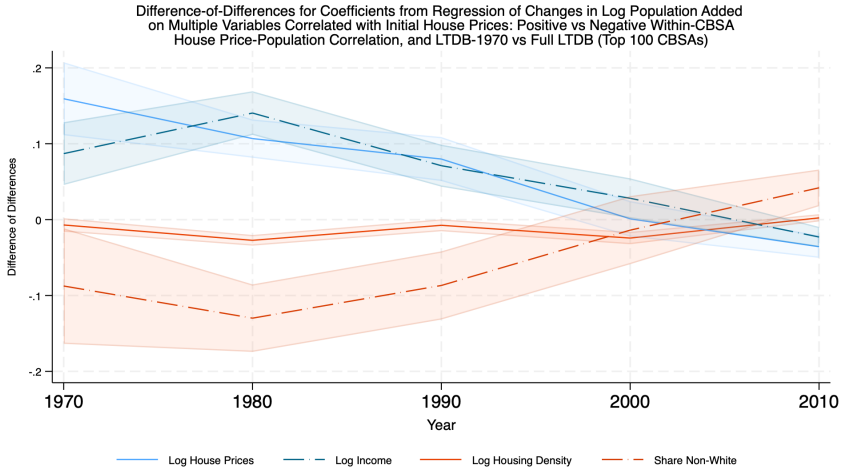
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log prices, log initial income, log housing units density and the non-white population share for the LTDB and LTDB-1970, separately. The samples for these regressions are all CBSAs. The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure E6: *LTDB-1970 Version of Figure 12 – Log Population Added and Log Prices for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSAs*



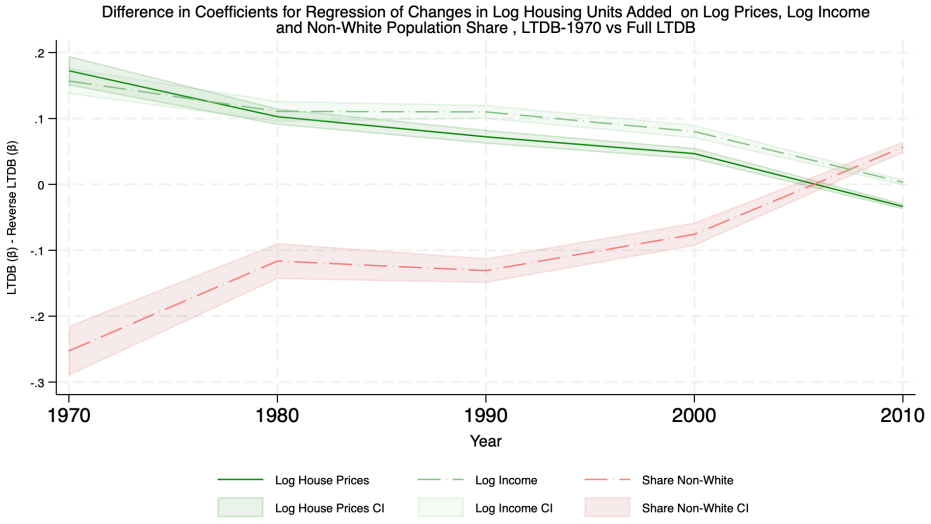
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial prices for the LTDB and LTDB-1970, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and income are negative (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure E7: *LTDB-1970 Version of Figure 13 – Difference of Differences Between Regression Coefficients from Log Population Added on Multiple Variables Correlated with Initial Population in LTDB vs LTDB-1970 for CBSAs with Positive vs Negative Within-CBSA Population-Price Correlation*



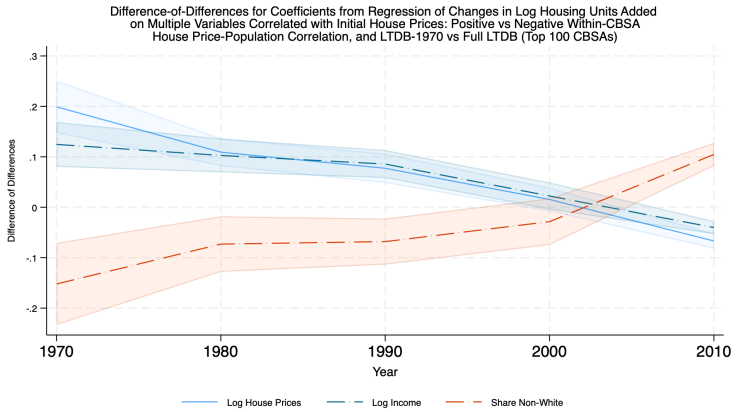
Note: The lines in this figure report the difference of differences between coefficients from decadal tract-level regressions from 1970-2010 for log population changes on multiple independent regressors for the LTDB and LTDB-1970, separately. We first take the regression coefficients from log population changes on each independent regressor for the CBSAs with negative price-population correlation in the LTDB and subtract from them, the LTDB-1970 estimates. Then we do the same for the CBSAs with positive price-population correlation. Taking the difference of these two differences gives a decadal 'difference of differences' coefficient for log population changes on log income, log house prices, log housing unit density and the non-white population share. Population changes, prices, income, and housing unit density are in logs, and the share of population that is non-white is in percentage points. We have no other normalizations The sample for these regressions includes only the top 100 most populous CBSAs. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals, constructed from the standard errors of each independent regression coefficient.

Appendix Figure E8: *LTDB-1970 Version of Figure 14 – Regression Coefficients for Log Population Added on Multiple Variables Correlated with Population*



Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log housing unit changes on log prices, log income and the non-white population share for the LTDB and LTDB-1970, separately. The samples for these regressions are all CBSAs. Housing unit changes and prices are in logs, and the share of population that is non-white is in percentage points. We have no other normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure E9: *LTDB-1970 Version of Figure 15 – Difference of Differences Between Regression Coefficients from Log Housing Units Added on Multiple Variables Correlated with Initial Population in LTDB vs LTDB-1970 for CBSAs with Positive vs Negative Within-CBSA Population-Price Correlation*



Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for changes in the log housing stock on log initial prices for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and prices are negative (dotted lines), and all other CBSAs (solid lines). Housing unit changes and prices are in logs, and thus we have no other normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

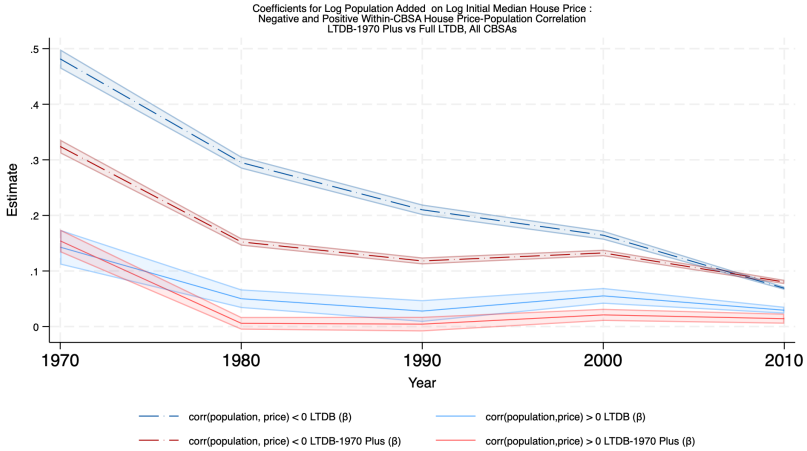
Appendix Table E1: Distribution of Regression Coefficients from on Residuals from Log Population Changes Regressed on Log House Prices, Log Income and Non-White Share of Population, All CBSAs + Top 100 CBSAs

Independent Variable		***Panel A: Log House Prices***		***Panel B: Log Income***		***Panel C: Non-White Population Share***	
		All CBSAs	*Top 100 CBSAs*	*All CBSAs*	*Top 100 CBSAs*	*All CBSAs*	*Top 100 CBSAs*
Percentiles							
	1%	-0.048	0.088	-0.115	0.103	-0.131	0.123
	5%	0.129	0.280	0.127	0.283	0.134	0.277
	10%	0.277	0.412	0.281	0.399	0.282	0.391
	25%	0.594	0.697	0.585	0.663	0.609	0.680
	Median	0.905	0.935	0.900	0.929	0.914	0.940
	75%	0.980	0.975	0.977	0.973	0.991	0.983
	90%	1.023	1.006	1.023	1.006	1.037	1.014
	99%	1.248	1.123	1.289	1.085	1.276	1.111
Moments							
	Mean	0.769	0.818	0.766	0.812	0.782	0.822
	Std. Dev.	0.313	0.247	0.320	0.248	0.327	0.251
	Min	-0.750	-0.241	-0.971	-0.117	-1.644	-0.131
	Max	1.920	1.186	2.153	1.163	2.098	1.184
Observations	*N*	1530	500	1530	500	1531	500

Note: Panel A and Panel B of this Table follow the same process outlined in the rest of this note. For illustration, consider Panel A, which uses house prices. This table reports the distribution of regression coefficients from the following sequence of regressions. First, we estimate for each CBSA-decade, a log-log specification of log changes in population on initial log median house prices. We extract the CBSA-decade-specific residuals from this first regression and then run CBSA-decade specific regressions for the log changes in the housing stock on those residuals. The result are regression coefficients for each CBSA-decade. Since there are 5 decades (1970, 1980, 1990, 2000 and 2010), the first column reports the distribution of CBSA-decade specific coefficients for all CBSA-decades with more than 20 observations. The second column reports the distribution of CBSA-decade specific coefficients for the top 100 most populous CBSAs. This follows the specification outlined in equation (1c) and is the decadal CBSA-level distribution of the ϑ parameter.

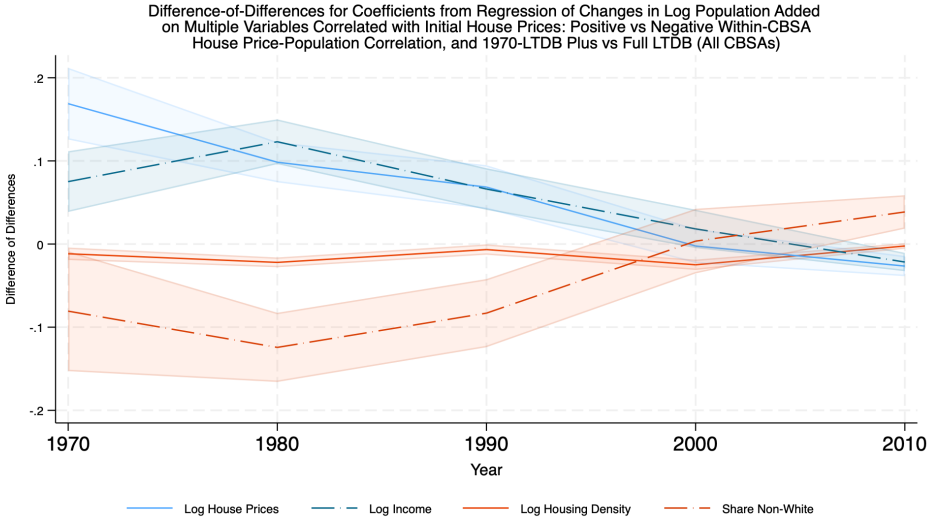
Appendix Section F: Regression Results for Figures 12, 13, and 15 with All CBSAs

Appendix Figure F1: Replicated Figure 12 – Log Population Added and Log Prices for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSAs, All CBSAs



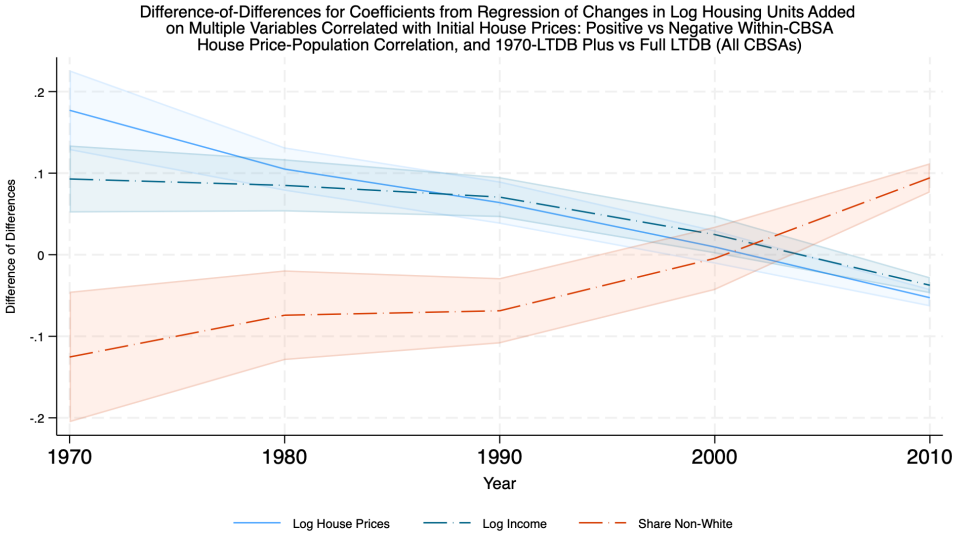
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions include all CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and prices are negative (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Figure F2: Replicated Figure 13 – Difference of Differences Between Regression Coefficients from Log Population Added on Multiple Variables Correlated with Initial Population in LTDB vs LTDB-1970 Plus for CBSAs with Positive vs Negative Within-CBSA Population-Price Correlation, All CBSAs



Note: The lines in this figure report the difference of differences between coefficients from decadal tract-level regressions from 1970-2010 for log population changes on multiple independent regressors for the LTDB and LTDB-1970 Plus, separately. We first take the regression coefficients from log population changes on each independent regressor for the CBSAs with negative price-population correlation in the LTDB and subtract from them, the LTDB-1970 Plus estimates. Then we do the same for the CBSAs with positive price-population correlation. Taking the difference of these two differences gives a decadal 'difference of differences' coefficient for log population changes on log income, log house prices, log housing unit density and the non-white population share. Population changes, prices, income, and housing unit density are in logs, and the share of population that is non-white is in percentage points. We have no other normalizations The sample for these regressions includes all CBSAs. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals, constructed from the standard errors of each independent regression coefficient.

Appendix Figure F3: Replicated Figure 15 – Difference of Differences Between Regression Coefficients from Log Housing Units Added on Multiple Variables Correlated with Initial Population in LTDB vs LTDB-1970 Plus for CBSAs with Positive vs Negative Within-CBSA Population-Price Correlation



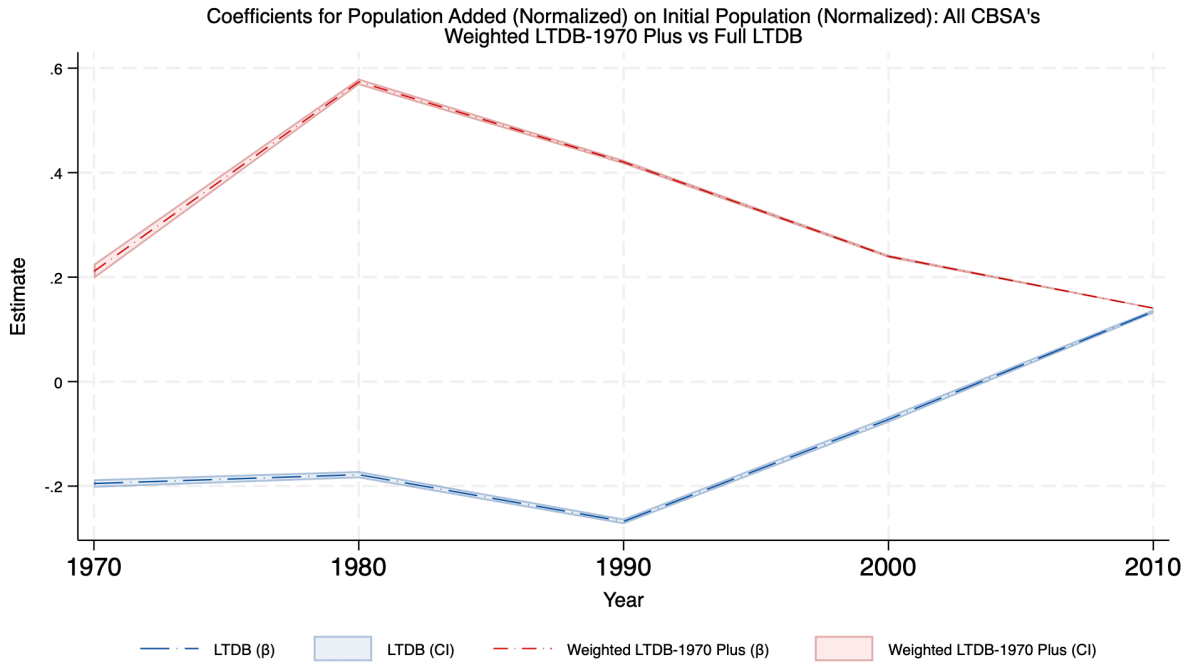
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for changes in the log housing stock on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes all CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and prices are negative (dotted lines), and all other CBSAs (solid lines). Housing unit changes, prices, income, and housing unit density are in logs, and the share of population that is non-white is in percentage points. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals.

Appendix Section G: Regression Results Weighting by the Number of Tract Intersections

Weighting Scheme 1: Frequency Weights on the Reverse LTDB

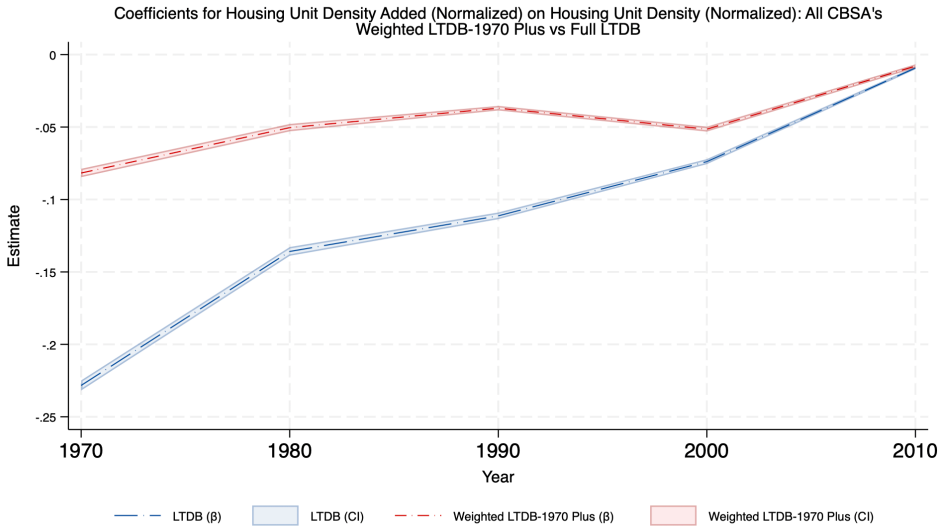
- We assign each 1970-based tract a weight equal to the number of later-year tract intersections it produces.
- In the 1970-LTDB regression, this means sparsely populated tracts that have fragmented into many pieces are effectively "counted" multiple times.
- This mimics the LTDB's tendency to over-count small, sparse suburban tracts.
- We do **not** use analytic weights here because:
    1. The splitting process is not a collection of independent draws, and
    2. Those fragmented pieces do not exhibit lower sampling variance simply by virtue of being smaller.

Appendix Figure G1: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 7 – Coefficients from Regression of Population Changes on Initial Population*



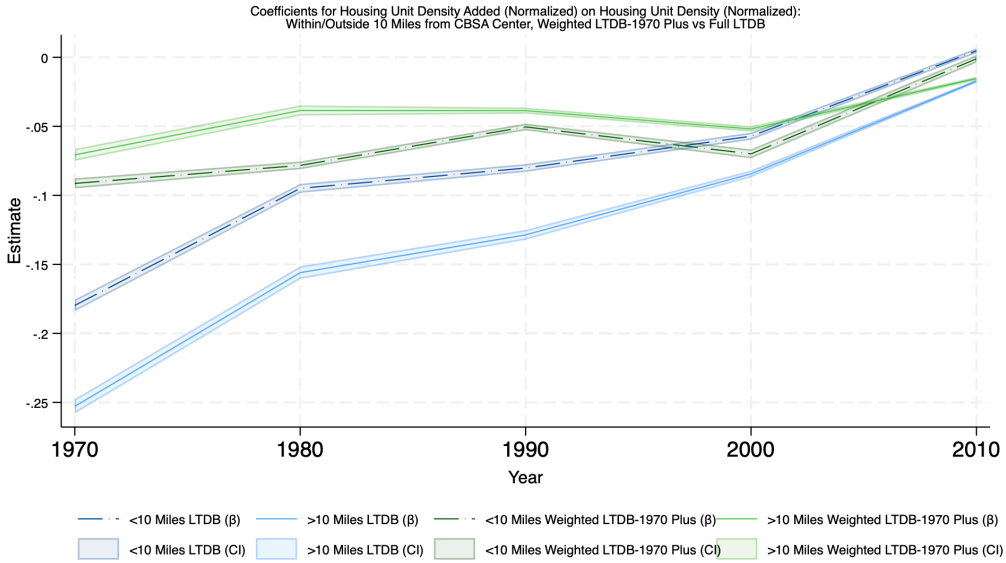
Note: The lines in this figure report the coefficients from normalized decadal tract-level regressions from 1970-2010 for decadal population changes on initial population in the LTDB and LTDB-1970 Plus, separately. Changes in population and initial population are normalized by CBSA-specific base-year suburban/city-center mean population. The shaded areas are 95% confidence intervals. The sample includes all 309 CBSAs, and all regressions include CBSA fixed effects. The Weighted LTDB-1970 Plus has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G2: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 8 – Housing Unit Density Change and Initial Housing Unit Density*



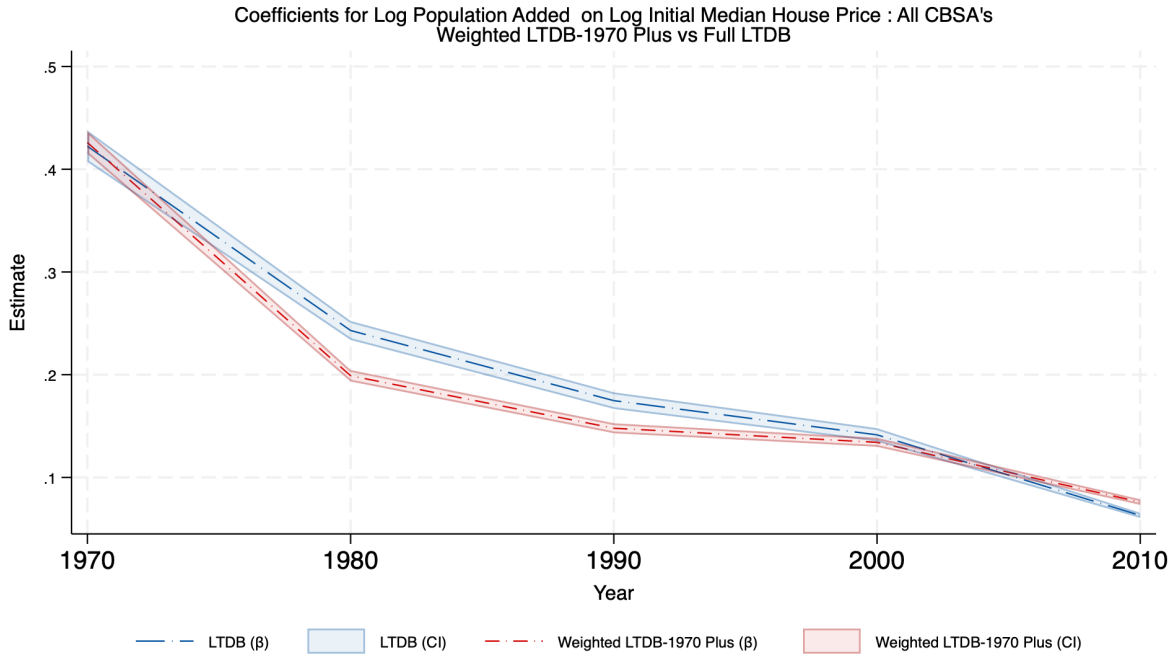
Note: The lines in this figure report the coefficients from normalized decadal tract-level regressions from 1970-2010 for decadal housing density changes on initial housing unit density for the LTDB and LTDB-1970 Plus, separately. Changes in housing unit density and initial housing unit density are normalized by CBSA-specific base-year suburban/city-center mean population. The shaded areas are 95% confidence intervals. The sample includes only the top 100 most populous CBSAs, and all regressions include CBSA fixed effects. The LTDB-1970 Plus has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G3: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 9 – Regression Coefficients for Housing Unit Density Change on Initial Housing Unit Density in Suburbs*



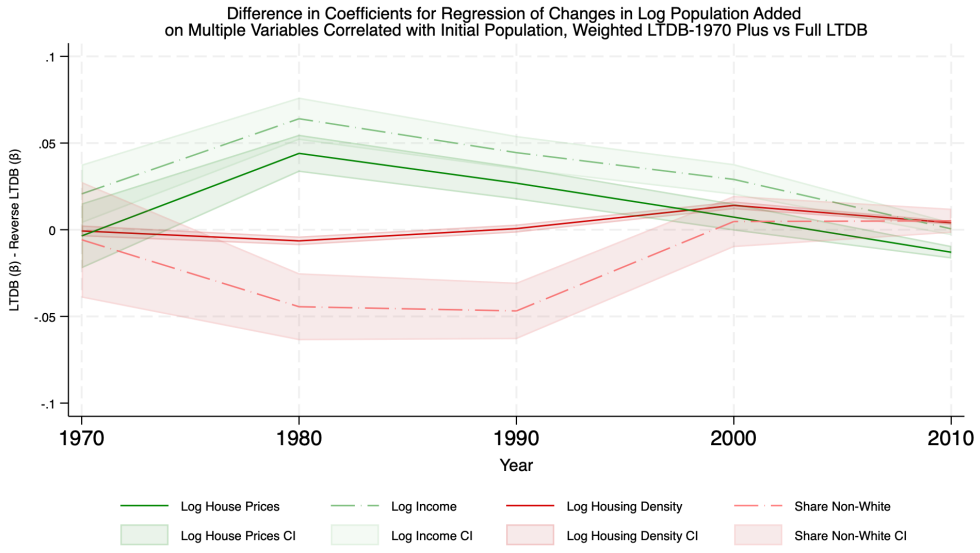
Note: The lines in this figure report the coefficients from normalized decadal tract-level regressions from 1970-2010 for decadal housing density changes on initial housing unit density for the LTDB and LTDB-1970 Plus, separately. Within the datasets, the sample is split into tracts greater or less than 10 miles from the CBSA center, and regressions are run separately. Changes in housing unit density and initial housing unit density are normalized by CBSA-specific base-year suburban/city-center mean population. The shaded areas are 95% confidence intervals. The sample includes only the top 100 most populous CBSAs, and all regressions include CBSA fixed effects. The Weighted LTDB-1970 Plus has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G4: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 10 – Regression Coefficients for Log Population Added Log on Initial Prices for All CBSAs*



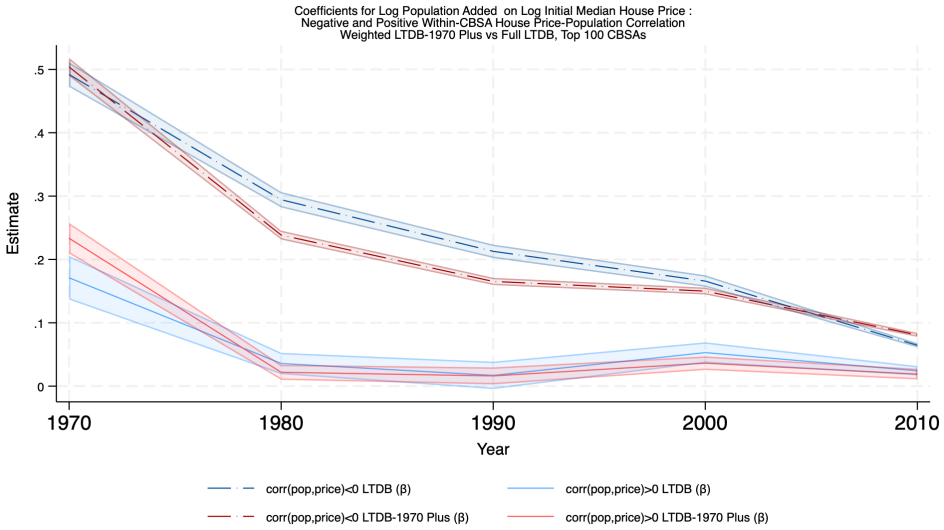
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median house values for the LTDB and LTDB-1970 Plus, separately. The samples for these regressions are all CBSAs. The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals. The LTDB-1970 Plus has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G5: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 11 – Difference in Regression Coefficients for Log Population Added on Multiple Other Variables*



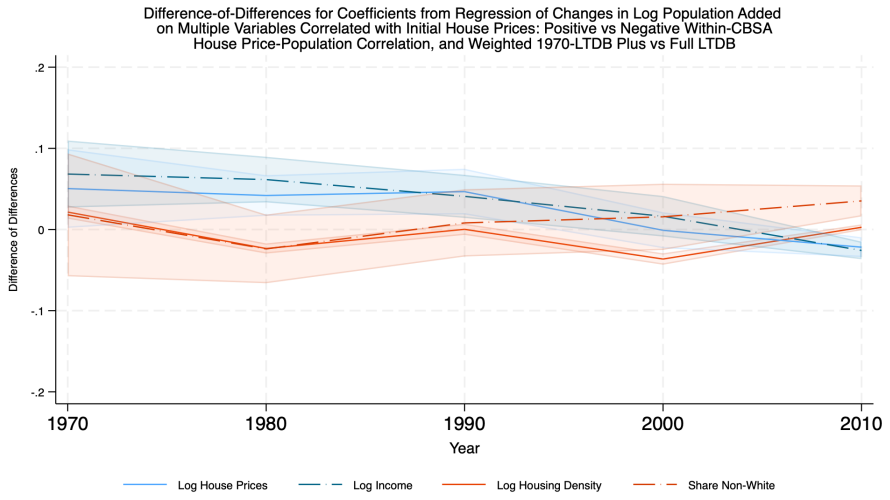
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median income, log prices, log housing unit density, and share of population that is non-white for the LTDB and LTDB-1970Plus, separately. The samples for these regressions are all CBSAs. Population changes, prices, income, and housing unit density are in logs, and the share of population that is non-white is in percentage points. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals. The Weighted Reverse LTDB has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G6: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 12 – Log Population Added and Log Prices for Negative and Non-Negative Within-CBSA Price-Population Correlated CBSAs*



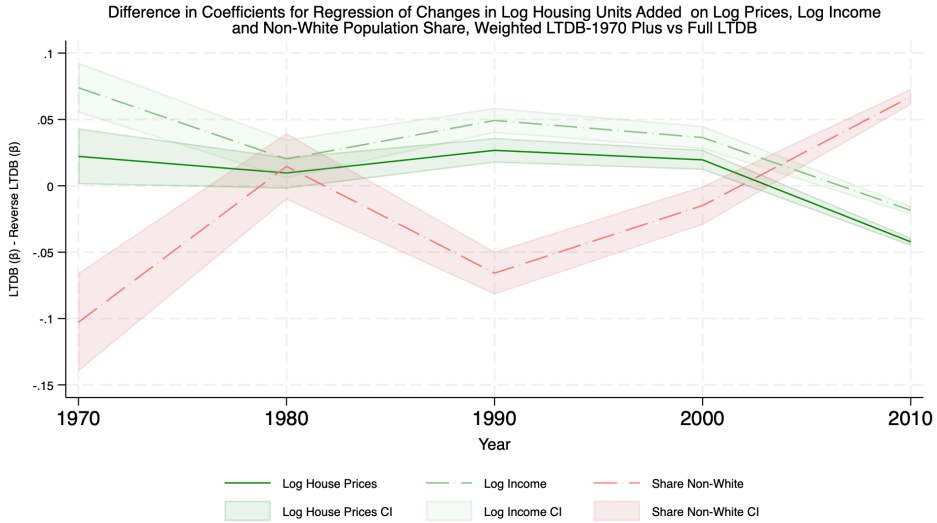
Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log population changes on log initial median prices for the LTDB and LTDB-1970 Plus, separately. The sample for these regressions includes only the top 100 most populous CBSAs. The sample is also split by whether the 1970's correlation coefficient between initial population and income are negative and statistically significant (dotted lines), and all other CBSAs (solid lines). The regressions are run in logs and thus have no normalizations. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals. The Weighted Reverse LTDB has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G7: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 13 – Difference of Differences Between Regression Coefficients from Log Population Added on Multiple Variables Correlated with Initial Population in LTDB vs LTDB-1970 Plus for CBSAs with Positive vs Negative Within-CBSA Population-Price Correlation*



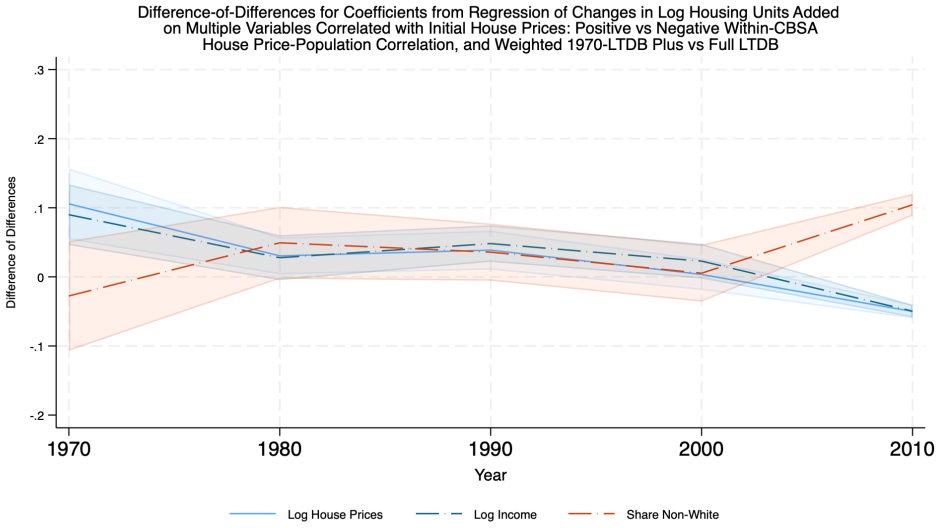
Note: The lines in this figure report the difference of differences between coefficients from decadal tract-level regressions from 1970-2010 for log population changes on multiple independent regressors for the LTDB and LTDB-1970 Plus, separately. We first take the regression coefficients from log population changes on each independent regressor for the CBSAs with negative price-population correlation in the LTDB and subtract from them, the LTDB-1970 Plus estimates. Then we do the same for the CBSAs with positive price-population correlation. Taking the difference of these two differences gives a decadal 'difference of differences' coefficient for log population changes on log income, log house prices, log housing unit density and the non-white population share. Population changes, prices, income, and housing unit density are in logs, and the share of population that is non-white is in percentage points. We have no other normalizations The sample for these regressions includes only the top 100 most populous CBSAs. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals, constructed from the standard errors of each independent regression coefficient. The Weighted Reverse LTDB has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G8: *LTDB-1970 Weighted by Tract Intersections Version of Figure 14 – Difference in Regression Coefficients for Log Housing Units Added on Multiple Other Variables*



Note: The lines in this figure report the coefficients from decadal tract-level regressions from 1970-2010 for log housing unit changes on prices, and share of population that is non-white for the LTDB and LTDB-1970 Plus, separately. The samples for these regressions are all CBSAs. Housing unit changes, prices, and income are in logs, and the share of population that is non-white is in percentage points. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals. The Weighted Reverse LTDB has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Figure G9: *LTDB-1970 Plus Weighted by Tract Intersections Version of Figure 15 – Difference of Differences Between Regression Coefficients from Log Housing Units Added on Multiple Variables Correlated with Initial Population in LTDB vs LTDB-1970 Plus for CBSAs with Positive vs Negative Within-CBSA Population-Price Correlation*



Note: The lines in this figure report the difference of differences between coefficients from decadal tract-level regressions from 1970-2010 for log housing unit changes on multiple independent regressors for the LTDB and LTDB-1970 Plus, separately. We first take the regression coefficients from log population changes on each independent regressor for the CBSAs with negative price-population correlation in the LTDB and subtract from them, the LTDB-1970 Plus estimates. Then we do the same for the CBSAs with positive price-population correlation. Taking the difference of these two differences gives a decadal 'difference of differences' coefficient for log housing unit changes on log income, log house prices, and the non-white population share. Population changes, prices, and income are in logs, and the share of population that is non-white is in percentage points. We have no other normalizations The sample for these regressions includes only the top 100 most populous CBSAs. All regressions include CBSA fixed effects. The shaded areas are 95% confidence intervals, constructed from the standard errors of each independent regression coefficient. The Weighted Reverse LTDB has frequency weights equal to the number of tract intersections each 1970 tract has in 2010.

Appendix Section H: Comparing the 1970-LTDB and LTDB for Missing Data Pass-Through and Outliers

Missing Data: LTDB vs 1970-LTDB

As noted in Section III of the main text and commented on in Appendix Section D, we avoid systematically undercounting count variables in the LTDB-1970 by setting a harmonized tract data point to missing if over 25% of that tract's area has missing source year data. In contrast, the LTDB considers a harmonized tract data point to be missing only if *all* of the tract intersections within the 2010 area are missing. Thus, if a tract has only 1% of its total area, the associated count variable for that tract is not flagged or set to missing in the LTDB.

Since older data are more likely to be missing, the LTDB 'passes through' missing data at a higher rate than the Reverse LTDBs. For example, suppose the 1970 definition tract G04001300303 in Phoenix (see the purple tract in Figure 17 of the main text) was missing data on the housing units variable in 1970. Using 1970-harmonized boundaries, we lose only 1 observation, but using 2010 tract boundaries we lose or undercount 125 observations. In general, if the tracts with missing source year data grow more than the tracts with non-missing source year data (ie, have more tract intersections), the LTDB will have a higher percentage of harmonized tracts with missing data.

Because we lack the tract areas used to construct the weights in the LTDB, we could not parse out whether the missing data influence a small area of many tracts, or disproportionately large areas of many tracts. For the LTDB-1970, we provide a straightforward comparison for the extent of undercounting and reweighting in Appendix Tables D2 and D3. For example, summing rows 1 and 6 for the renter occupied units for 1980 in Appendix Table D2 implies that 4.5% ((1,388+178)/34,440) of the LTDB-1970 tracts will undercount renter occupied units. This is the largest amount of missing data in our sample. By setting 1,388 (4.0%) to missing because they have over 25% of their tract area missing, we only undercount in 178, or 0.5% of the non-missing tracts.

Appendix Table H1 tabulates the missing source year, and harmonized tracts in the LTDB. In 1980, there are 2,320 tracts missing data on renter occupied units (row 12, column 5), which intersect with 6,967 2010 harmonized boundary tracts (row 12, column 5). This means that the LTDB undercounts in 6,967/59,188=11.8% of the tracts if no adjustment is made for missing data. However, in the final dataset there are 1,832 missing tracts (presumably because the entire tract is missing – as this is the only criterion for missing data

in the LTDB). This means that the LTDB undercounts in 9% ((6,967-1,832)/(59,188-1,832)) of the tracts *not* set to missing, relative to 0.5% for the 1970-LTDB.

Appendix Table H2 provides insight into the potential impact of missing data for rents, house prices and income in 1970 and 1980. For 1980, column 6 shows that there are 8.3% of source year tracts missing data on rents (row 4), 6.9% of source year tracts missing data on home value (row 9), and 4.9% of source year tracts missing data on median income (row 14). Column 8 then shows that those missing tracts split at a higher rate than the rest of the sample. The source year missing values affect 19.4% of 2010 tracts for rents (row 5), 15.5% of tracts for home values (row 9), and 11% of tracts for median income (row 14).

Comparing this to the LTDB-1970, Appendix Table D2 shows that 7.3% of tracts are affected by missing rent data (substantially less than the LTDB), but only 5.6% of non-missing tracts are reweighted due to missing data for distribution data.[8] Only 4,794 of the 11,462 tracts with missing rent data are set to missing in the LTDB, implying that 12.3% ((11,462-4,794)/(59,188-4,794)) of all non-missing tracts are reweighted due to missing data for distribution variables, which is substantially higher than the LTDB-1970.

[8] Summing rows 1 and 6 of Appendix Table C2 for 1980 and then dividing by the total number of tracts yields (1,894+647)/34,440, or 7.3%, influenced by missing data. Only 1,894/(34,440-647), or 5.6%, of nonmissing tracts were reweighted.

Appendix Table H1: *LTDB Counts of Source Year Tracts and Harmonized Tracts Missing (count variables)*

Variable	Year	Source Year Tract Count	2010 Harmonized Tract Count	Source Year Tracts		2010 Harmonized Tracts		LTDB Data	
				Missing Tracts	Percent Missing	Missing Tracts	Percent Missing	Missing Tracts	Percent Missing
All Housing Units	1970	34,647	52,772	0	0	0	0	0	0
	1980	46,728	59,188	0	0	0	0	0	0
	1990	61,258	72,694	0	0	0	0	0	0
	2000	65,443	72,694	0	0	0	0	0	0
Owner Occupied	1970	34,647	52,772	350	0.01	501	0.009	28	0.0005
	1980	46,728	59,188	2,320	0.05	6,967	0.118	1,832	0.031
	1990	61,258	72,694	0	0	0	0	0	0
	2000	65,443	72,694	0	0	0	0	0	0
Renter Occupied	1970	34,647	52,772	350	0.01	501	0.009	28	0.0005
	1980	46,728	59,188	2,320	0.05	6,967	0.118	1,832	0.031
	1990	61,258	72,694	0	0	0	0	0	0
	2000	65,443	72,694	0	0	0	0	0	0

Note: This table tabulates the missing data in three separate ways for housing units, owner occupied units and renter occupied units in the LTDB. It is directly comparable to Appendix Table C2, which does the same for the 1970-LTDB. Column 2 provides a count of the source-year tracts in each year, and column 3, the 2010 harmonized tract count. Column 4 counts, for each variable, the number of source-year tracts with missing data, and column 5 divides the total source-year tract count by that number to get the percentage of source year tracts with missing data. Then after crosswalking into the LTDB 2010 harmonized tracts, column 6 tabulates how many unique 2010 tracts have at least one tract intersection with a source-year tract that has a missing value for that variable. Each one of these tracts will be undercounted because a component of the 2010 tract is missing, and count variables are summed. Column 7 divides the total number of 2010 harmonized tracts by that value to get the percentage of 2010 harmonized tracts potentially affected by the missing source-year data. Recall that in the LTDB, the only way for a datapoint to be set to missing is if *all* of the tract intersections (ie, the entire area of the 2010 tract) is missing that value. In contrast, the 1970-LTDB will set a harmonized tract datapoint to missing for count variables if over 25% of the tract area is missing. Finally, column 8 provides a tabulation in the final LTDB dataset of the count of missing tracts for a given variable, and column 9 calculates the percentage of all harmonized tracts that are missing for that variable. These tracts are the one which have *all* of the tract intersections (ie, the entire area of the 2010 tract) missing that variable. This means that since column 6 is larger than column 8, there are tracts which have missing components and thus undercounted, but that are not set to missing, nor are flagged as potential undercounts.

Appendix Table H2: *LTDB Counts of Source Year Tracts and Harmonized Tracts Missing (Distribution variables)*

Variable	Year	Source Year Tract Count	2010 Harmonized Tract Count	Source Year Tracts		2010 Harmonized Tracts		LTDB Data	
				Missing Tracts	Percent Missing	Missing Tracts	Percent Missing	Missing Tracts	Percent Missing
Median Rent	1970	34,647	52,772	490	0.014	959	0.018	352	0.007
	1980	46,728	59,188	3,900	0.083	11,462	0.194	4,794	0.081
	1990	61,258	72,694	150	0.002	332	0.005	165	0.002
	2000	65,443	72,694	100	0.001	268	0.004	187	0.003
Median Home Value	1970	34,647	52,772	1,479	0.043	2,774	0.053	861	0.016
	1980	46,728	59,188	3,218	0.069	9,194	0.155	3,181	0.054
	1990	61,258	72,694	150	0.002	332	0.005	232	0.003
	2000	65,443	72,694	100	0.001	268	0.004	228	0.003
Median Household Income	1970	34,647	52,772	104	0.003	231	0.004	231	0.004
	1980	46,728	59,188	2,304	0.049	6,499	0.110	1,984	0.034
	1990	61,258	72,694	132	0.002	250	0.003	197	0.003
	2000	65,443	72,694	101	0.001	170	0.002	225	0.003

Note: This table tabulates the missing data in three separate ways for medina rent, median home value, and median household income in the LTDB. It is directly comparable to Appendix Table C3, which does the same for the 1970-LTDB. Column 2 provides a count of the source-year tracts in each year, and column 3, the 2010 harmonized tract count. Column 4 counts, for each variable, the number of source-year tracts with missing data, and column 5 divides the total source-year tract count by that number to get the percentage of source year tracts with missing data. Then after crosswalking into the LTDB 2010 harmonized tracts, column 6 tabulates how many unique 2010 tracts have at least one tract intersection with a source-year tract that has a missing value for that variable. Each one of these tracts will be measured with error because a component of the 2010 tract is missing, and distribution variables are constructed with a weighted average. Column 7 divides the total number of 2010 harmonized tracts by that value to get the percentage of 2010 harmonized tracts potentially affected by the missing source-year data. Recall that in the LTDB, the only way for a datapoint to be set to missing is if *all* of the tract intersections (ie, the entire area of the 2010 tract) is missing that value. In contrast, the 1970-LTDB will set a harmonized tract datapoint to missing for distribution variables if over 75% of the tract area is missing. Finally, column 8 provides a tabulation in the final LTDB dataset of the count of missing tracts for a given variable, and column 9 calculates the percentage of all harmonized tracts that are missing for that variable. These tracts are the one which have *all* of the tract intersections (ie, the entire area of the 2010 tract) missing that variable. This means that since column 6 is larger than column 8, there are tracts which have missing components and thus mismeasured, but that are not set to missing, nor are flagged as potential error.

Outliers in the Two Datasets

All large databases, especially those with complex variable construction processes like ours and the LTDB's, have outliers. The data reported above on weighting suggests that the LTDB-1970 will be less sensitive to missing data (i.e., when a 1970 tract is subdivided into many 2010 tracts) compared to the LTDB in which one missing 1970 tract can have a major impact on multiple 2010 boundaries. The LTDB-1970 has many fewer interactions with source year tracts in general and that its weighting outcomes are less likely to generate extreme values. These two facts suggest that the LTDB-1970 should have fewer unexplained outliers than the LTDB. That is, indeed, the case, as we document in in this section. This presentation is lengthy because the analysis includes maps, satellite photos, and the like.

We first consider the issue of large positive outliers, which are of potential concern in the LTDB-1970. These figures generally make sense because of the large tract sizes in 1970 and substantial growth in some previously low-density areas. We find that outliers in the LTDB-1970 typically result from some type of redevelopment after a natural disaster, the consequence of an identifiable re-development plan by a local or state government, or associated with a major infrastructure investment. We have more difficulty explaining outliers in the LTDB. We also cannot explain instances in which large shifts in one direction during one decade are largely reversed in the next decade. We conclude with an analysis and discussion of the most extreme outliers in both data bases.

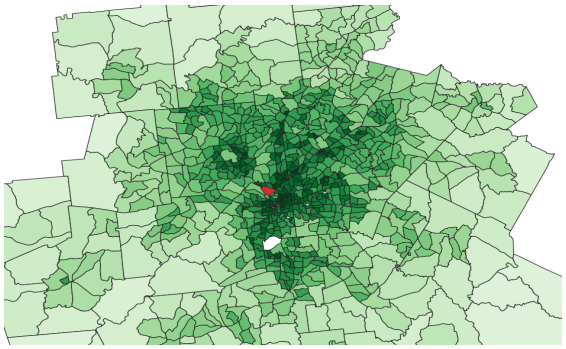
Plausible Outliers and the Reverse LTDB

Appendix Figures H1-H4 depict two of the ten largest negative outliers across all years in the Reverse LTDB. The first is the red tract area (G1301210008902) in Appendix Figure H1 that is just outside of Atlanta's urban core, with Appendix Figure H2 showing the same tract area via satellite imagery. Appendix Figure H3 then shows another nearby tract (G1301210008700) in Atlanta that also lost a large number of units, with Appendix Figure H4 containing the analogous satellite-based photo. Both tracts are located around the NS-Inman yard, which experienced a major redevelopment following a destructive tornado.

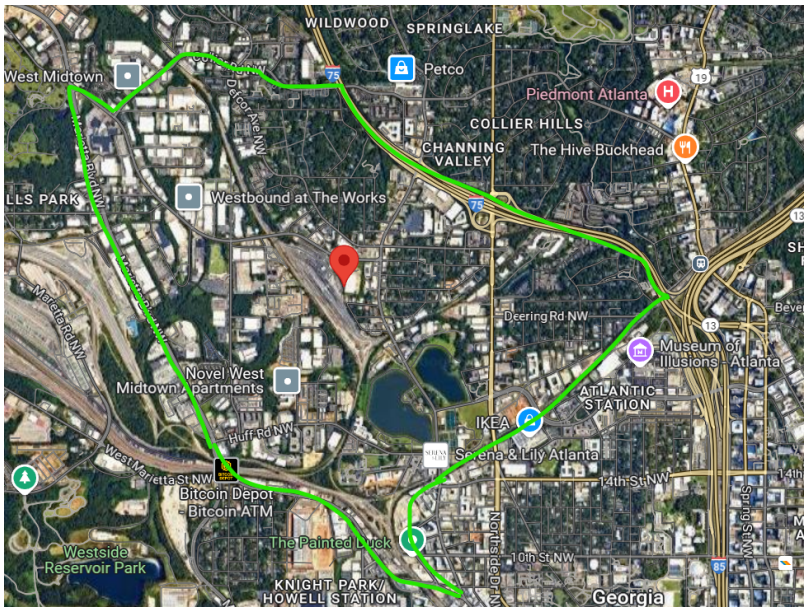
More specifically, this West Highlands neighborhood in Atlanta is the site of the Herman E. Perry Homes public housing project. This project was completed in 1954 and contained 1,100 housing units for African-American families. It suffered significant damage from a tornado in 1975 that destroyed at least 100 apartments. In 1996, the Atlanta Housing Authority received a HOPE VI grant to revitalize the area, leading to the creation of the West Highlands community. Development of West Highlands began in the early 2000s,

transforming the site into a mixed-income community with plans for approximately 784 homes. Thus, a natural disaster started the reduction in units that built over time, and then was followed in the first decade of the 2000s with many hundreds of new units, along with new amenities including parks and the like. At the least, this time pattern cannot be considered measurement error on its face.[9]

Appendix Figure H1: *Tract G1301210008902 in Atlanta, 1,168 units lost from 1990-2000, Reverse LTDB*

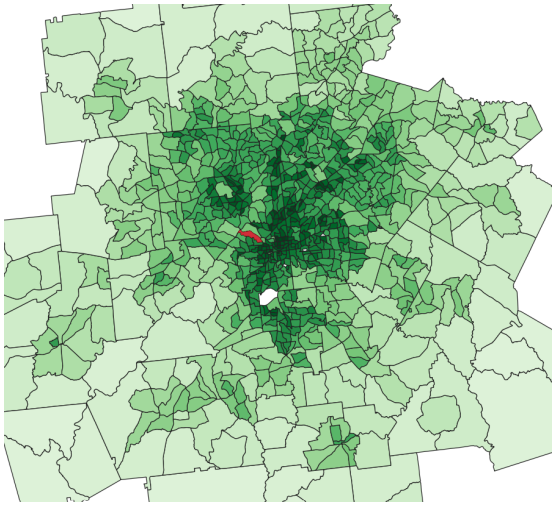


Appendix Figure H2: *Tract G1301210008902, Next to NS-Inman Yard*

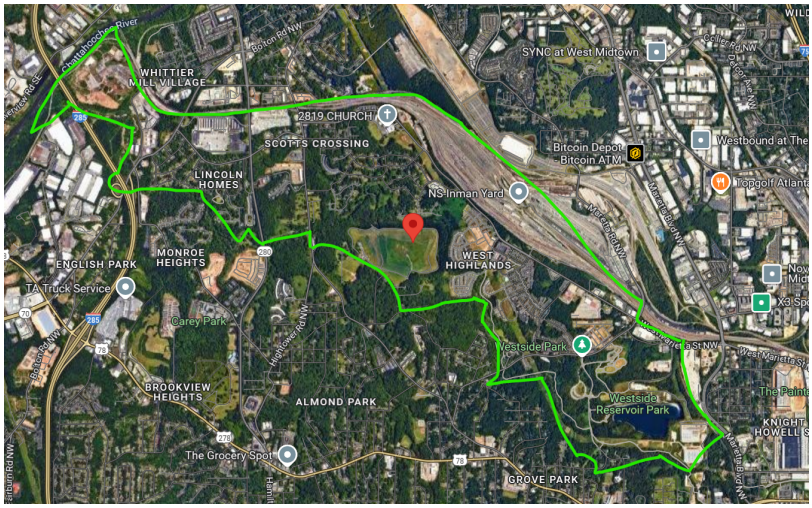


Appendix Figure H3: *Tract G1301210008700 in Atlanta, 1,045 Units Lost from 1990-2000, Reverse LTDB*

[9] See the Atlanta Housing report on this event and neighborhood at the following link: https://www.atlantahousing.org/wp-content/uploads/2019/05/2015.0024-Heman-E.-Perry-Homes-and-West-Highlands-at-Perry-Boulevard-records.pdf. The project also included significant public infrastructure improvements, with costs estimated at $22.3 million, funded by tax allocation district (TAD) bonds.



Appendix Figure H4: *Tract G1301210008902, West Highlands – Other Side of NS-Inman Yard*

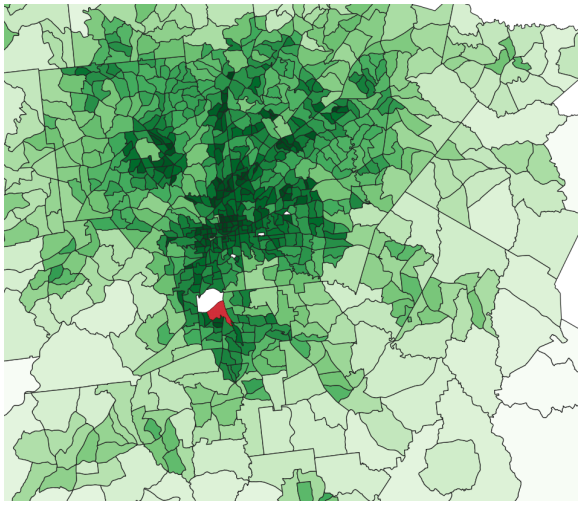


Another example of a significant reduction in units observed in the 1970-LTDB is the 1,163 decline from 1990-2000 in tract G1300630040407. Appendix Figure H5 shows that the tract is located directly south of Hartsfield-Jackson Atlanta International Airport, with Appendix Figure H6 providing the satellite view of the area. The housing unit loss coincided with the opening of the world's largest air passenger terminal covering 2.5 million square feet.[10] We cannot confirm whether the housing losses were caused by the use of eminent domain or whether those homes disappeared after the amenity structure of the neighborhood changed. What is clear is that a major public infrastructure investment occurred that could

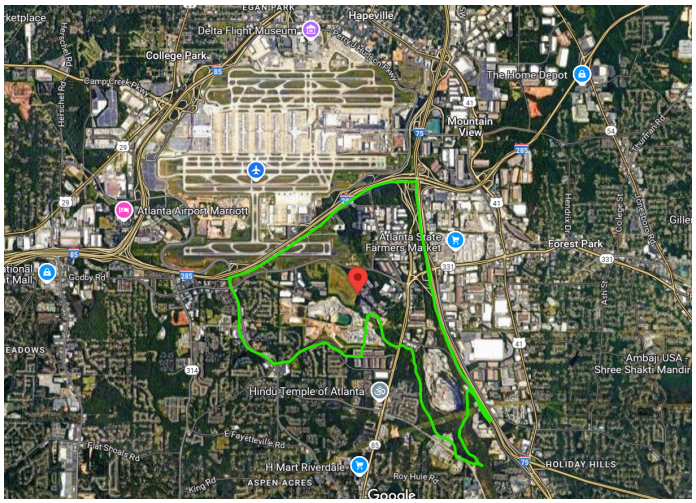
[10] See this URL (https://www.atl.com/about-atl/history-of-atl/) for a detailed history of this episode, as well as of the Atlanta airport in general.

have generated the time pattern observed in the data, either directly by demolishing houses or via an endogenous amenity change.

Appendix Figure H5: *Tract G1300630040407 in Atlanta, 1,163 Units Lost from 1990-2000*



Appendix Figure H6: *Tract G1300630040407 – South of Hartfield-Jackson Atlanta International Airport*

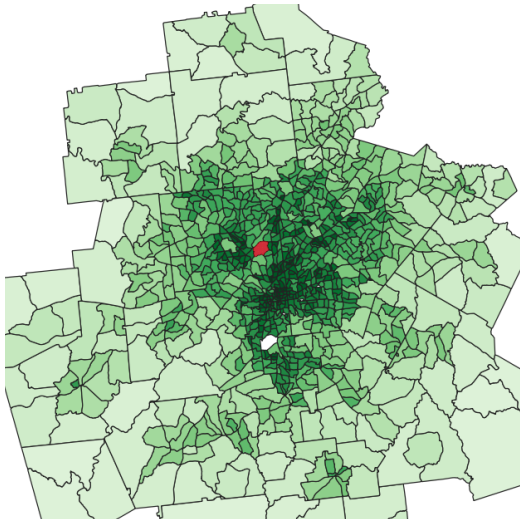


Unexplained Outliers in the LTDB

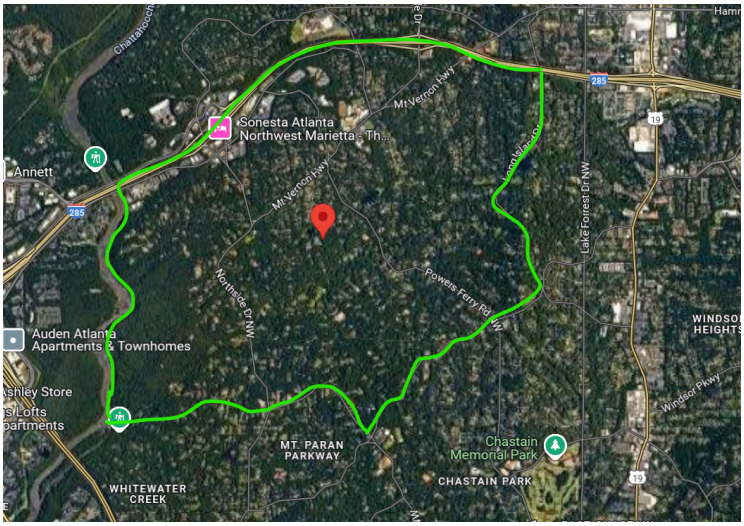
While the online appendix provides other examples of seemingly valid reductions in the Atlanta and Phoenix housing stocks in the 1970-LTDB, large outlier are more numerous in the LTDB and we find them more difficult to explain. Here, we analyze 3 of the 5 largest reductions of units in Atlanta across any decade, but which happen to have occurred in the 1990s in the LTDB data.

The first is tract G1301210010206 (Appendix Figure H7), which lost 1,664 units from 1990-2000. The tract is located in a well-to-do area called Sandy Springs.[11] There is no noticeable amenity change in close proximity to the tract, and there is nothing online showing us that there was mass redevelopment. Appendix Figure H8 shows the tract via satellite. We can find no changes in demographics or other economic variables that plausibly could account for this example of mean reversion.

Appendix Figure H7: *Tract G1301210010206 – Sandy Springs, Atlanta with 1,664 units lost 1990-2000*



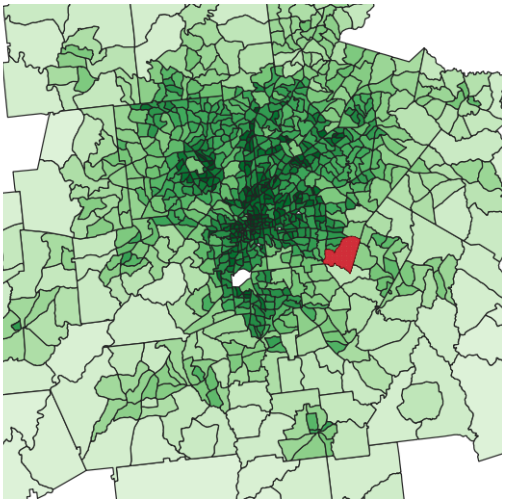
Appendix Figure H8: *Tract G1301210010206 – Sandy Springs, Atlanta*



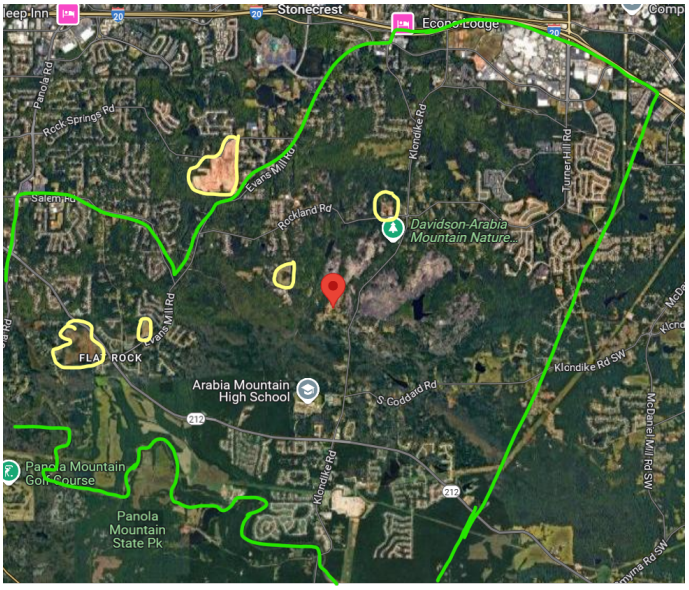
[11] Median household income was $101,593 in 2023 according to U.S. Census data available at this URL: (https://datacommons.org/place/geoId/1368516?utm_medium=explore&mprop=income&popt=Person&cpv=age,Years15Onwards&hl=en).

Appendix Figure H9 and H10 depict a suburb to the east of Atlanta for which there is no evidence of a natural disaster or large infrastructure investment that could plausibly have changed the amenity environment to cause this severely negative change amidst an upward trend in unit growth. The highlighted yellow areas are clear instances of teardowns, however there are only a few small instances of these within the tract boundaries. In fact, the largest teardown occurred just outside of the tract, to the north-west. This is another example of unexplained change involving unit loss between decades of growth.

Appendix Figure H9: *Tract G1300890023418 – South of Stonecrest, 1,234 Units Lost 1990-2000*

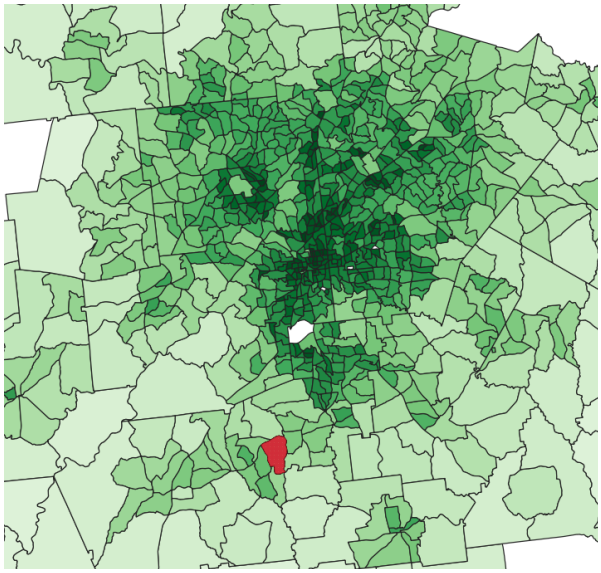


Appendix Figure H10: *Tract G1300890023418 – South of Stonecrest*

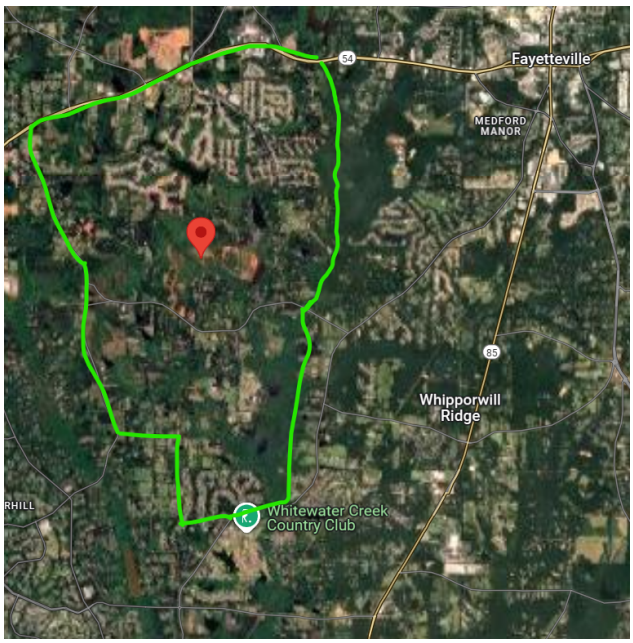


Our next example is for a census tract to the west of Fayetteville in the far southern part of the Atlanta region (Appendix Figures H11 and H12). This tract saw a loss of 1,184 units between 1990-2000 and there is no detectable reason why. We certainly cannot find any infrastructure change or major amenity change.

Appendix Figure H11: *Tract G1301130140305 – East of Fayetteville, 1184 Units Lost 1990-2000*



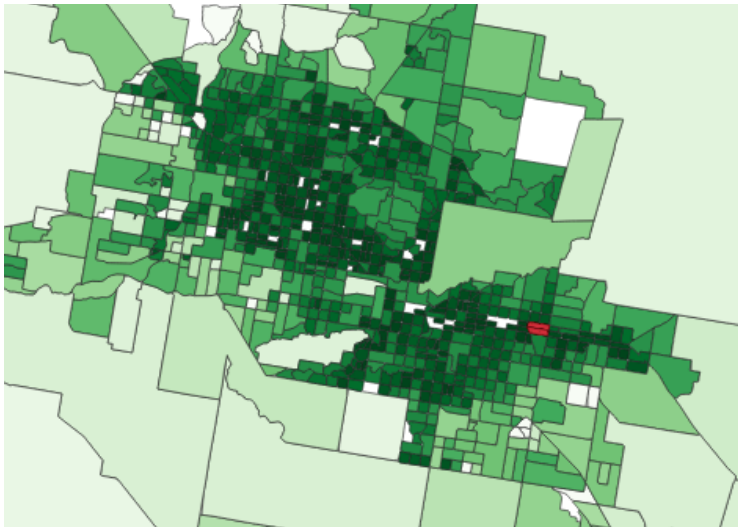
Appendix Figure H12: *Tract G1301130140305 – East of Fayetteville*



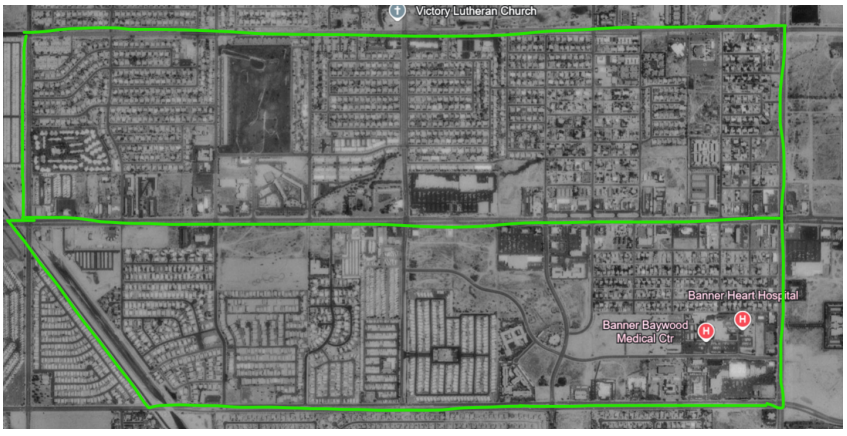
Our final example of unexplained outliers arises from two tracts in the Phoenix metropolitan area. The 7th and 10th largest reduction in units in Phoenix over the entire

sample period of the LTDB occur in tract G0400130420214 (top) and tract G0400130422625 (bottom), which are located in the East Mesa part of the metro area as identified by the red rectangles in Appendix Figure H13. This is a place that experienced rapid growth from 1990 to 2020. However, the LTDB shows losses of 1,059 and 866 units, respectively, from 2000-2010. Appendix Figure H14 is taken from Historical Google Earth on July 27th, 1992; Appendix Figure H15 was taken on October 19th, 2003; and Appendix Figure H16 is from March 17th, 2020. Inspecting the land area and physical structures in these two tracts yields no evidence of a large reduction in their housing stocks between 2000 and 2010.

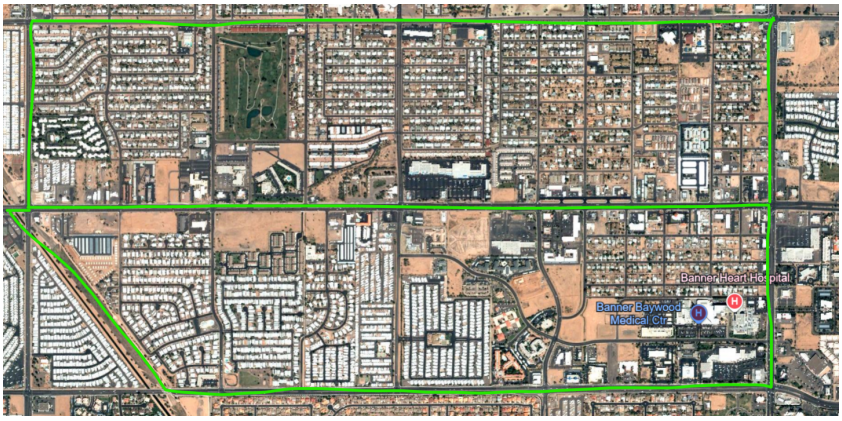
Appendix Figure H13: Tract G0400130420214 and G0400130422625 – Phoenix



Appendix Figure H14: *Tract G0400130420214 and G0400130422625 – Phoenix on July 27th, 1992*



Appendix Figure H15: *Tract G0400130420214 and G0400130422625 – Phoenix on October 19th, 2003*



Appendix Figure H16: *Tract G0400130420214 and G0400130422625 – Phoenix on March 17th, 2010*