Supplementary Appendix

Daily House Price Indices:
Construction, Modeling, and Longer-Run Predictions

Tim Bollerslev, Andrew J. Patton and Wenjing Wang

March 19, 2015

A.1 Data and data cleaning

The historical transaction records in DataQuick extends from the late 1990s to 2012 (exact dates are given in Talbe A.2 below) with some large metropolitan areas, such as Boston and New York, having transactions recorded as far back as 1987. Properties are uniquely identified by property IDs, which enable us to identify sale pairs. We rely U.S. Standard Use Codes contained in the DataQuick database to identify transactions of single-family residential homes. The specific counties included in each of the ten MSAs are listed in Table A.1.

Our data cleaning rules are based on the same filters used by S&P/Case-Shiller in the construction of their monthly indices. In brief, we remove all transactions that are not "arms length," using a flag for such transactions available in the database. We also remove transactions with "unreasonably" low or high sale prices (below $5000 or above $100 million, and those generating an average annual return of below -50% or above 100%), as well as any sales pair with an interval of less than six months. Sale pairs are also excluded if there are indications that major improvements have been made between the two transactions, although such indications are not always present in the database.

Once these filters are imposed, we use all remaining sale pairs to estimate the repeat-sales model presented in equation (1) using the estimation procedure described in Section 3.1. For the Los Angeles MSA, for example, we have a total of 877,885 "clean" sale pairs, representing an average of 180 *daily* sale pairs over the estimation period. Details for all ten MSAs are provided in Table A.2 below.

Table A.1: Metropolitan Statistical Areas (MSAs)

MSA	Represented counties	Counties in our indices
Los Angeles-Long Beach-Santa Ana, CA Metropolitan Statistical Area (Los Angeles)	Los Angeles CA, Orange CA	Los Angeles CA, Orange CA
Boston-Cambridge-Quincy, MA-NH Metropolitan Statistical Area (Boston)	Essex MA, Middlesex MA, Norfolk MA, Plymouth MA, Suffolk MA, Rockingham NH, Strafford NH	Essex MA, Middlesex MA, Norfolk MA, Plymouth MA, Suffolk MA, Rockingham NH, Strafford NH
Chicago-Naperville-Joliet, IL Metropolitan Division (Chicago)	Cook IL, DeKalb IL, Du Page Il, Kane IL, Kendal IL, McHenry IL, Will IL, Grundy IL	Cook IL, DeKalb IL, Du Page Il, Kane IL, Kendal IL, McHenry IL, Will IL, Grundy IL
Denver-Aurora, CO Metropolitan Statistical Area (Denver)	Adams CO, Arapahoe CO, Broomfield CO, Clear Creek CO, Denver CO, Douglas CO, Elbert CO, Gilpin CO, Jefferson CO, Park CO	Adams CO, Arapahoe CO, Broomfield CO, Clear Creek CO, Denver CO, Douglas CO, Elbert CO, Gilpin CO, Jefferson CO, Park CO
Miami-Fort Lauderdale-Pompanp Beach, FL Metropolitan Statistical Area (Miami)	Broward FL, Miami-Dade FL, Palm Beach FL	Broward FL, Miami-Dade FL, Palm Beach FL
Las Vegas-Paradise, NV Metropolitan Statistical Area (Las Vegas)	Clark NV	Clark NV
San Diego-Carlsbad-San Marcos, CA Metropolitan Statistical Area (San Diego)	San Diego CA	San Diego CA
San Francisco-Oakland-Fremont, CA Metropolitan Statistical Area (San Francisco)	Alameda CA, Contra Costa CA, Marin CA, San Francisco CA, San Mateo CA	Alameda CA, Contra Costa CA, Marin CA, San Francisco CA, San Mateo CA

Table A.1: Continued

MSA	Represented counties	Counties in our indices
New York City Area (New York)	Fairfield CT, New Haven CT, Bergen NJ, Essex NJ, Hudson NJ, Hunterdon NJ, Mercer NJ, Middlesex NJ, Monmouth NJ, Morris NJ, Ocean NJ, Passaic NJ, Somerset NJ, Sussex NJ, Union NJ, Warren NJ, Bronx NY, Dutchess NY, Kings NY, Nassau NY, New York NY, Orange NY, Putnam NY, Queens NY, Richmond NY, Rockland NY, Suffolk NY, Westchester NY, Pike PA	Fairfield CT, New Haven CT, Bergen NJ, Essex NJ, Hudson NJ, Hunterdon NJ, Mercer NJ, Middlesex NJ, Monmouth NJ, Morris NJ, Ocean NJ, Passaic NJ, Somerset NJ, Sussex NJ, Union NJ, Warren NJ, Bronx NY, Dutchess NY, Kings NY, Nassau NY, New York NY, Orange NY, Putnam NY, Queens NY, Richmond NY, Rockland NY, Suffolk NY, Westchester NY
Washington-Arlington-Alexandria, DC-VA-MD-WV Metropolitan Statistical Area (Washington, D.C.)	District of Columbia DC, Calvert MD, Charles MD, Frederick MD, Montgomery MD, Prince Georges MD, Alexandria City VA, Arlington VA, Clarke VA, Fairfax VA, Fairfax City VA, Falls Church City VA, Fauquier VA, Fredericksburg City VA, Loudoun VA, Manassas City VA, Manassas Park City VA, Prince William VA, Spotsylvania VA, Stafford VA, Warren VA, Jefferson WV	District of Columbia DC, Calvert MD, Charles MD, Frederick MD, Montgomery MD, Prince Georges MD, Alexandria City VA, Arlington VA, Fairfax VA, Falls Church City VA, Fauquier VA, Fredericksburg City VA, Loudoun VA, Manassas City VA, Manassas Park City VA, Prince William VA, Spotsylvania VA, Stafford VA

Note: The table reports the counties included in the ten Metropolitan Statistical Areas (MSAs) underlying the S&P/Case-Shiller indices. The name of each MSA is abbreviated by that of its major city or county, as indicated in parenthesis.

Table A.2: Data summary

Los Angeles	Boston	Chicago	Denver	Miami	Las Vegas	San Diego	San Francisco	New York	Washington, D.C.
Panel A: Data availability									
Full sample start date									
04/01/88	02/01/87	01/08/96	02/01/98	02/01/97	04/01/88	04/01/88	04/01/88	02/01/87	01/10/96
Daily index start date									
03/01/95	03/01/95	09/01/99	05/03/99	04/01/98	01/03/95	01/02/96	01/03/95	01/03/95	06/01/01
Full sample end date									
10/23/12	10/11/12	10/12/12	10/17/12	10/15/12	10/17/12	10/23/12	10/18/12	10/23/12	10/23/12
Panel B: Transactions									
Total transactions									
10,285,770	2,121,471	3,948,706	1,672,669	3,689,159	2,236,138	2,845,804	3,778,446	5,943,114	2,168,018
Single family residential housing transactions									
5,970,536	1,141,930	1,886,433	1,000,785	1,366,745	1,479,872	1,584,732	2,331,860	2,951,031	1,055,537
Arms-length transaction									
2,562,884	975,964	1,157,215	672,512	935,985	915,408	755,440	1,031,261	2,307,079	759,752
After excluding transaction value <= 5000 or >= 100,000,000									
2,555,165	917,039	1,156,042	671,605	935,178	913,682	754,106	1,030,384	2,271,467	757,675
After excluding houses sold only once									
1,980,740	638,577	659,732	475,481	668,552	729,365	579,152	757,379	1,234,074	459,842

Table A.2: Continued

Los Angeles	Boston	Chicago	Denver	Miami	Las Vegas	San Diego	San Francisco	New York	Washington, D.C.
Panel B: Transactions (continued)									
After excluding transactions happen within 6 months									
1,627,149	532,761	561,945	374,045	576,810	645,869	510,450	688,869	1,026,836	325,251
After excluding >= 2× standard deviations and >= 6× median transaction values									
1,578,869	514,356	543,038	360,944	561,805	628,790	494,894	665,537	999,284	313,777
Panel C: Sale pairs									
Total pairs									
939,476	294,101	292,737	198,608	321,358	378,093	296,985	397,229	544,326	162,693
After excluding renovation/reconstruction between two sales									
899,573	286,760	292,737	187,977	287,790	226,701	244,059	350,500	540,235	151,203
After excluding abnormal annual returns (less than -50% or more than 100%)									
878,017	272,858	277,160	181,633	281,393	221,877	239,232	341,878	512,251	143,481
After excluding sale pairs with second transaction on weekends									
878,002	272,727	277,095	180,504	277,442	221,876	239,215	341,858	508,860	143,433
After excluding sale pairs with second transaction on federal holidays									
877,885	272,414	277,079	180,003	276,676	221,554	239,041	341,469	508,548	143,431
Average *daily* sale pairs for the daily index estimation period									
180	55	84	53	77	49	51	70	109	49

A.2 Supplementary tables and figures

Tables A.3-A.5 and Figures A.1-A.3 contain additional empirical results for each of the ten MSAs pertaining to: the noise filter estimates; the daily HAR-X-GARCH-CCC correlation estimates; the unconditional correlations of the monthly Case-Shiller index returns; the sample autocorrelations for the raw daily index returns; time series plots of the raw and filtered daily house price indices; and the unconditional return correlations as a function of the return horizon.

Table A.3: Noise filter estimates

	Los Angeles	Boston	Chicago	Denver	Miami	Las Vegas	San Diego	San Francisco	New York	Washington, D.C.
μ	0.018	0.019	0.002	0.009	0.013	0.001	0.020	0.018	0.016	0.017
	(0.006)	(0.007)	(0.001)	(0.008)	(0.010)	(0.000)	(0.006)	(0.007)	(0.005)	(0.009)
σ_η	2.457	5.888	4.668	3.779	4.113	5.362	3.746	4.925	4.349	4.612
	(0.039)	(0.140)	(0.155)	(0.139)	(0.076)	(0.212)	(0.058)	(0.105)	(0.132)	(0.108)
σ_u	0.379	0.388	0.593	0.327	0.497	0.568	0.407	0.525	0.376	0.501
	(0.022)	(0.034)	(0.057)	(0.040)	(0.035)	(0.056)	(0.022)	(0.029)	(0.034)	(0.037)
σ_η/σ_u	6.478	15.180	7.866	11.544	8.273	9.448	9.204	9.376	11.576	9.200

Note: Quasi Maximum Likelihood Estimates (QMLE) with robust standard errors in parentheses.

Table A.4: Daily HAR-X-GARCH-CCC correlations

	Los Angeles	Boston	Chicago	Denver	Miami	Las Vegas	San Diego	San Francisco	New York	Washington, D.C.
Los Angeles	1.000	-0.001	-0.004	-0.009	0.049	0.033	0.056	0.177	0.011	0.031
Boston		1.000	-0.013	0.014	0.023	0.025	0.041	0.023	-0.010	0.022
Chicago			1.000	-0.004	0.018	0.006	-0.014	0.036	0.048	0.015
Denver				1.000	0.030	0.024	0.031	-0.004	-0.022	0.023
Miami					1.000	0.017	0.025	0.038	0.038	0.018
Las Vegas						1.000	0.028	0.030	-0.023	0.026
San Diego							1.000	0.054	0.002	0.011
San Francisco								1.000	0.005	0.023
New York									1.000	0.025
Washington, D.C.										1.000

Note: Conditional daily correlations estimated from the HAR-X-GARCH-CCC model.

Table A.5: Unconditional correlations of monthly S&P/Case-Shiller index returns

	Los Angeles	Boston	Chicago	Denver	Miami	Las Vegas	San Diego	San Francisco	New York	Washington, D.C.
Los Angeles	1.000	0.651	0.658	0.543	0.870	0.875	0.926	0.835	0.778	0.881
Boston		1.000	0.767	0.749	0.527	0.495	0.672	0.693	0.725	0.773
Chicago			1.000	0.679	0.637	0.544	0.567	0.688	0.818	0.762
Denver				1.000	0.398	0.382	0.545	0.693	0.496	0.666
Miami					1.000	0.799	0.782	0.743	0.795	0.802
Las Vegas						1.000	0.819	0.663	0.684	0.748
San Diego							1.000	0.833	0.712	0.839
San Francisco								1.000	0.659	0.855
New York									1.000	0.816
Washington, D.C.										1.000

Note: The correlations are based on the same June 2001 to September 2012 sample period used in the estimation of the daily HAR-X-GARCH-CCC model.

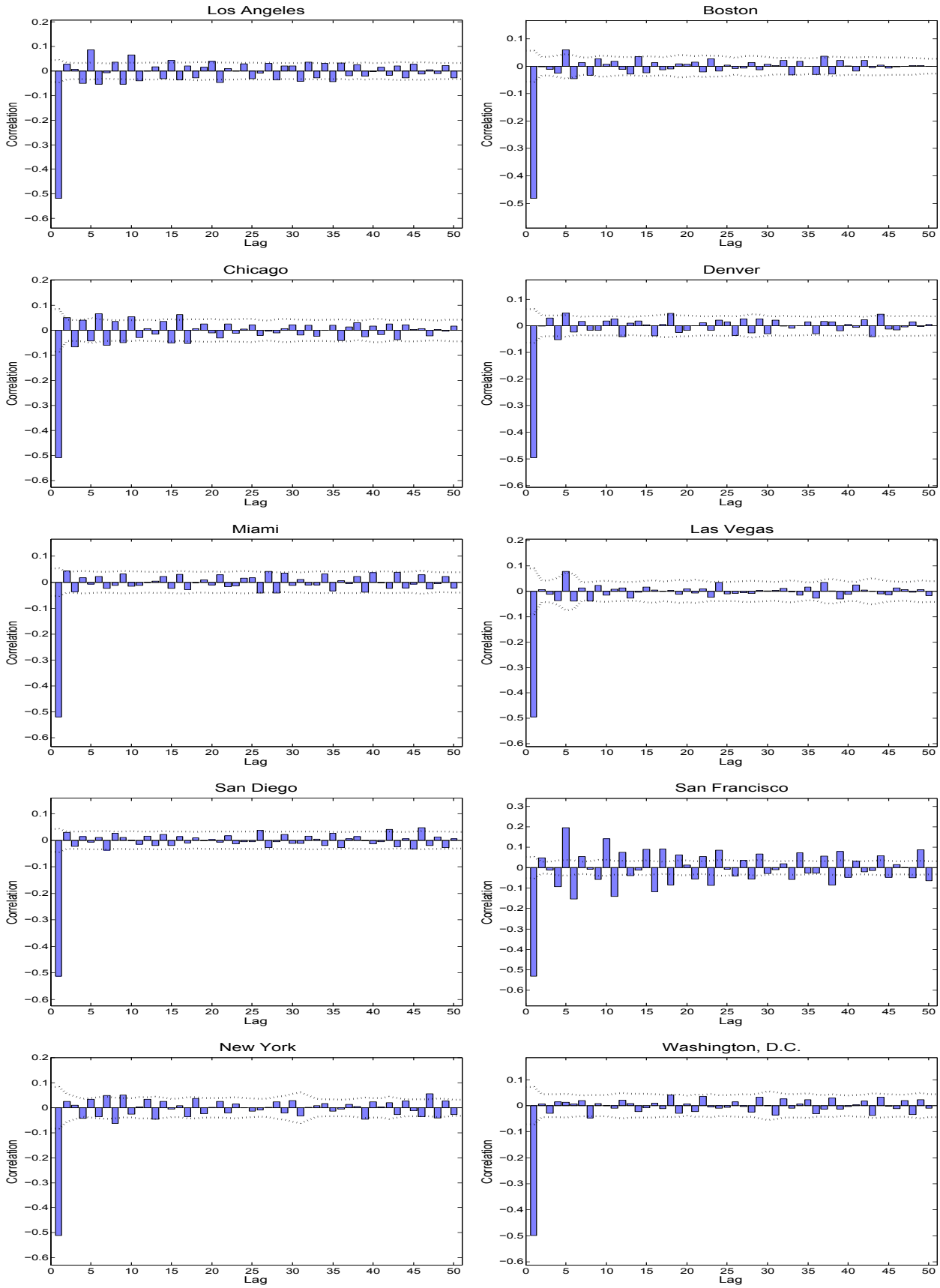


Figure A.1: Sample autocorrelations for raw daily index returns, with 95% confidence intervals

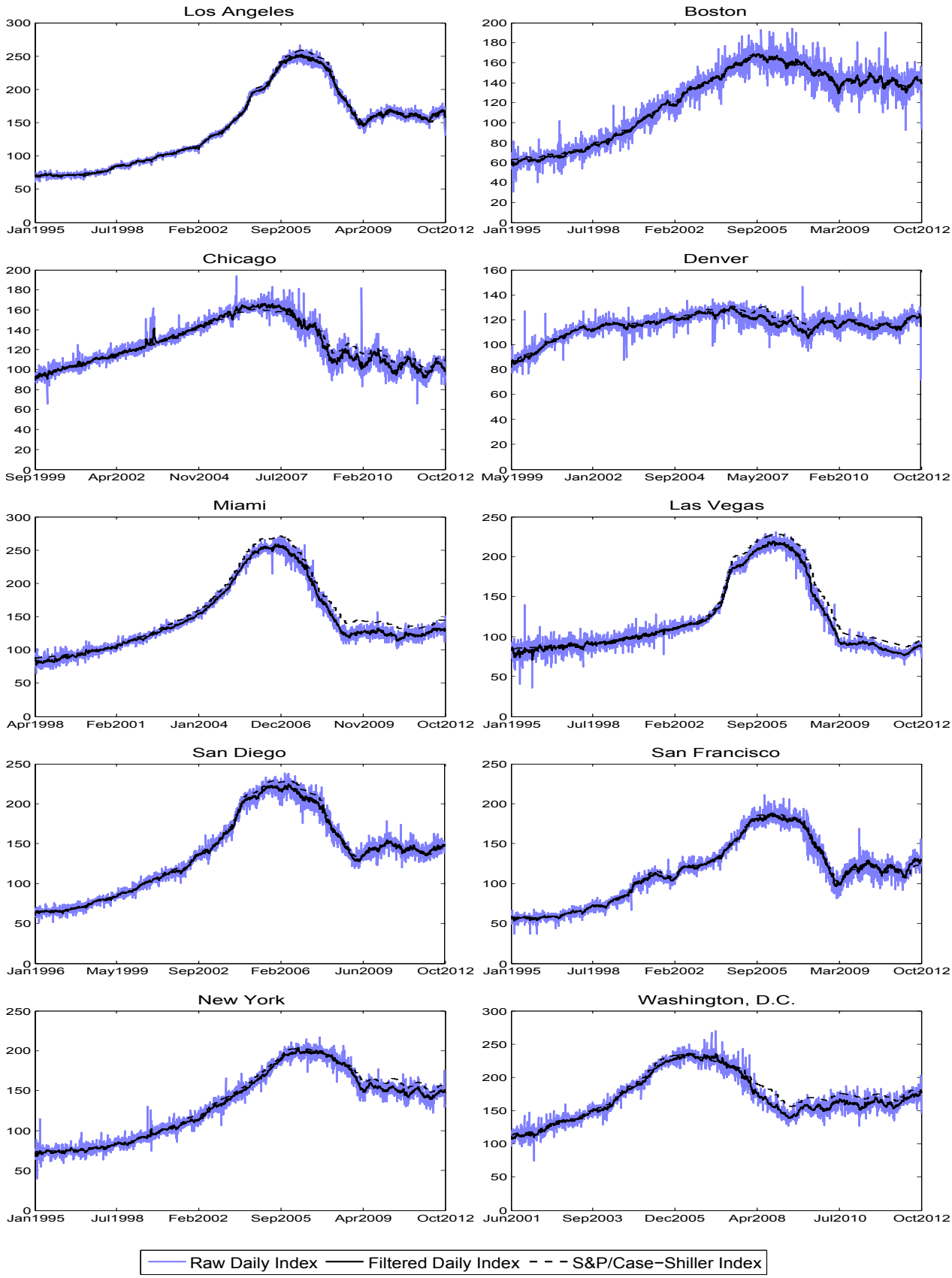


Figure A.2: Raw and filtered daily house price indices for ten MSAs

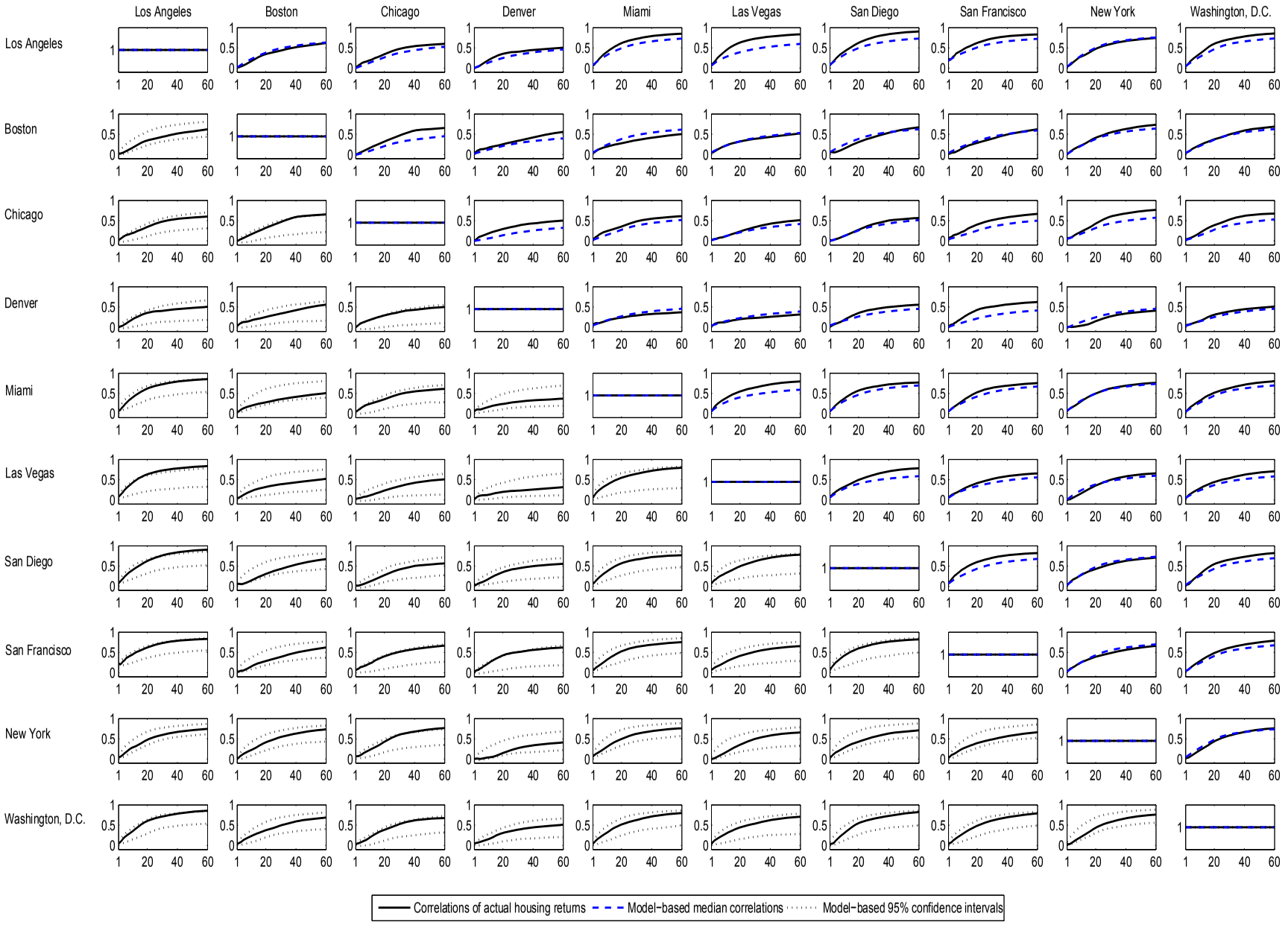


Figure A.3: Unconditional return correlations as a function of return horizon

A.3 Frequency-based comparisons with monthly S&P/Case-Shiller index

In parallel to the monthly S&P/Case-Shiller indices, our daily house price indices are based on all publicly available property transactions. However, the complicated non-linear transformations of the data used in the construction of the indices prevent us from expressing the monthly indices as explicit functions of the corresponding daily indices. Instead, as a simple way to help gauge the relationship between the indices, and the potential loss of information in going from the daily to the monthly frequency, we consider the linear projection of the monthly S&P/Case-Shiller returns for MSA i, denoted $r_{i,t}^{S\&P}$, on 60 lagged values of the corresponding daily index returns,

$$r_{i,t}^{S\&P} = \delta(L) r_{i,t} + \varepsilon_{i,t} \equiv \sum_{j=0}^{59} \delta_j L^j r_{i,t} + \varepsilon_{i,t}, \tag{6.1}$$

where $L^j r_{i,t}$ refers to the daily return on the j^{th} day before the last day of month t. Since all of the price series appear to be non-stationary, we formulate the projection in terms of returns as opposed to the price levels. The inclusion of 60 daily lags match the three-month smoothing window used in the construction of the monthly S&P/Case-Shiller indices, discussed in Section 2. The true population coefficients in the linear $\delta(L)$ filter are, of course, unknown, however they are readily estimated by ordinary least squares (OLS).

The OLS estimates for $\delta_{j=0,\ldots,59}$ obtained from the single regression that pools the returns for all ten MSAs are reported in the top panel of Figure A.4. Each of the individual coefficients are obviously subject to a fair amount of estimation error. At the same time, there is a clear pattern in the estimates for δ_j across lags, naturally suggesting the use of a polynomial approximation in j to help smooth out the estimation error. The solid line in the figure shows the resulting nonlinear least squares (NLS) estimates obtained from a simple quadratic approximation. The corresponding R^2s for the unrestricted OLS and the NLS fit ($\hat{\delta}_j = 0.1807 + 0.0101j - 0.0002j^2$) are 0.860 and 0.851, indicating only a slight deterioration in the accuracy of the fit by imposing a quadratic

approximation to the lag coefficients. Moreover, even though the monthly S&P/Case-Shiller returns are not an exact linear function of the daily returns, the simple relationship dictated by $\delta(L)$ accounts for the majority of the monthly variation.

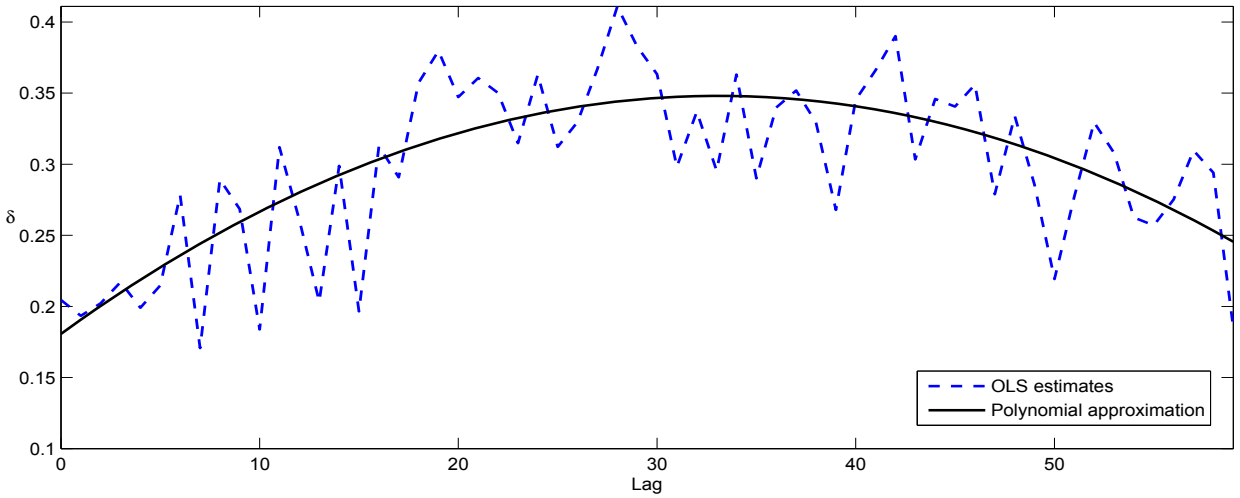
To further illuminate the features of the approximate linear filter linking the monthly returns to the daily returns, consider the gain,

$$G(\omega) = \left[\sum_{j=0}^{59}\sum_{k=0}^{59}\delta_j\delta_k cos(|j-k|\omega)\right]^{1/2}, \quad \omega \in (0,\pi), \tag{6.2}$$

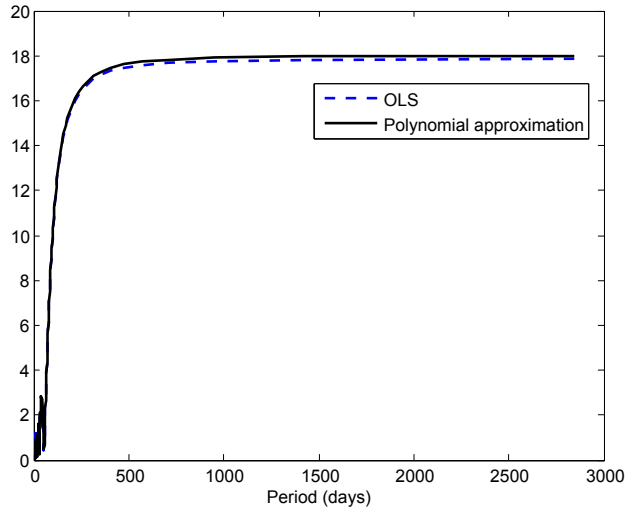
and the phase

$$\theta(\omega) = tan^{-1}\left(\frac{\sum_{j=0}^{59}\delta_j sin(j\omega)}{\sum_{j=0}^{59}\delta_j cos(j\omega)}\right), \quad \omega \in (0,\pi), \tag{6.3}$$

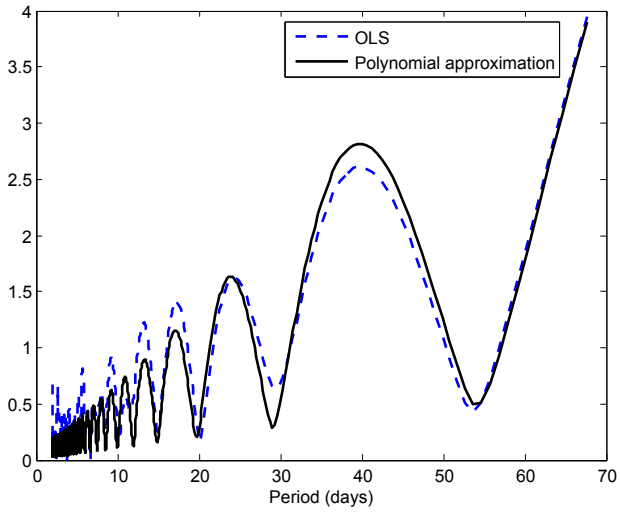
of $\delta(L)$. Looking first at the gains in Figures A.4b and A.4c, the unrestricted OLS estimates and the polynomial NLS estimates give rise to similar conclusions. The filter effectively down-weights all of the high-frequency variation (corresponding to periods less than around 70 days), while keeping all of the low-frequency information (corresponding to periods in excess of 100 days). As such, potentially valuable information for forecasting changes in house prices is obviously lost in the monthly aggregate. Further along these lines, Figures A.4d and A.4e show the estimates of $\frac{\theta(\omega)}{\omega}$, or the number of days that the filter shifts the daily returns back in time across frequencies. Although the OLS and NLS estimates differ somewhat for the very highest frequencies, for the lower frequencies (periods in excess of 60 days) the filter systematically shifts the daily returns back in time by about 30 days. This corresponds roughly to one-half of the three month (60 business days) smoothing window used in the construction of the monthly S&P/Case-Shiller index.



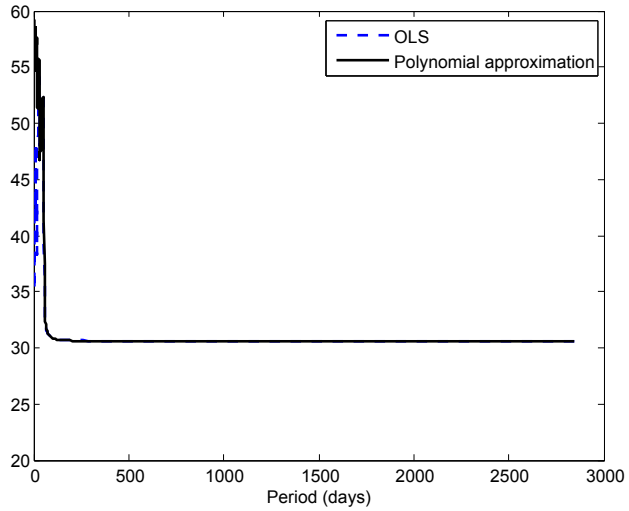
(a) Estimated $\delta(L)$ filter coefficients



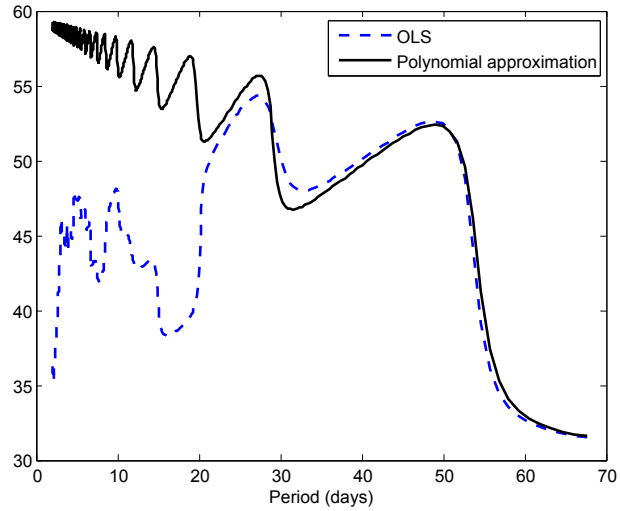
(b) Gain (all periods)



(c) Gain (shorter-run periodicities)



(d) Shift (all periods)



(e) Shift (shorter-run periodicities)

Figure A.4: Characteristics of the $\delta(L)$ filter