High Frequency House Price Indexes with Scarce Data

Hoesli, Martin E.; Bourassa, Steven

Swiss Finance Institute
Research Paper Series N°16-27

High Frequency House Price Indexes with Scarce Data

Steven C. BOURASSA
Florida Atlantic University

Martin HOESLI
University of Geneva, Geneva Finance Research Institute, Swiss Finance Institute,
University of Aberdeen Business School, and Kedge Business School

swiss:finance:institute

High Frequency House Price Indexes with Scarce Data

Steven C. Bourassa

School of Urban and Regional Planning and School of Public Administration, Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, email: sbourassa@fau.edu

Martin Hoesli

Geneva Finance Research Institute and Swiss Finance Institute, University of Geneva, 40 boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland; University of Aberdeen Business School, Scotland; and Kedge Business School, France, email: martin.hoesli@unige.ch

March 7, 2016

Keywords: house prices, high-frequency price indexes, repeat sales method, scarce data

JEL code: R31

High Frequency House Price Indexes with Scarce Data

Abstract

We show how a method that has been applied to commercial real estate markets can be used to produce high frequency house price indexes for a city and for submarkets within a city. Our application of this method involves estimating a set of annual robust repeat sales regressions staggered by start date and then undertaking an annual-to-monthly (ATM) transformation with a generalized inverse estimator. Using transactions data for Louisville, Kentucky, we show that the method substantially reduces the volatility of high frequency indexes at the city and submarket levels. We demonstrate that both volatility and the benefits from using the ATM method are related to sample size.

Introduction

Reliable house price indexes are necessary for understanding the dynamics of urban housing markets. The Federal Housing Finance Agency publishes indexes for metropolitan areas in the United States; however, metropolitan areas are comprised of submarkets and house price dynamics can vary across submarkets (Bourassa, Hoesli, and Peng, 2003). Prices could rise rapidly in one area while they rise only moderately or even decline in another area. Therefore, it is useful to have reliable house price indexes at the submarket level.

The volume of transactions limits the frequency and the degree of geographical disaggregation at which an index can be produced. For example, due to scarce data a monthly index for a city or a less frequent index for a neighborhood might be extremely volatile and, therefore, not particularly useful due to scarce data. Several papers have focused on the issue of index construction with thin data, either by parameterizing the historical time dimension (Schwann, 1998; McMillen and Dombrow, 2001; Francke, 2010) or by making use of the spatial or temporal correlation in real estate markets (Pace, Barry, Clapp, and Rodriguez, 1998; Clapp, 2004).

The focus of this paper is to show how a method that has been applied to commercial real estate price indexes (Bokhari and Geltner, 2012) can be useful in constructing high frequency house price indexes for both cities and submarkets within cities. The method involves two stages. The first stage produces a set of low frequency (typically annual) indexes with staggered start dates using a standard index construction approach. For housing markets, this could be either the hedonic or the repeat sales methods. We apply a robust repeat sales method (Bourassa, Cantoni, and Hoesli, 2013). In the second stage, a generalized inverse procedure is used to convert the staggered series of low frequency indexes into a high frequency index (quarterly or monthly). We apply this frequency conversion method to housing market data for Louisville, Kentucky, for the period from January 1998 through June 2010, producing monthly indexes from a set of annual indexes. We compare our converted monthly indexes with monthly repeat sales indexes produced using monthly time dummy variables. The indexes constructed with the annual-to-monthly (ATM) transformation are much less volatile than the monthly repeat sales indexes; this applies even for time periods and areas with limited numbers of transactions.

Our primary contribution to the literature is to demonstrate the applicability of the frequency conversion method to housing markets, where it has considerable potential in the construction of indexes with high frequency and spatial granularity. We also show how the frequency conversion method can be combined with robust techniques to produce indexes that are relatively free of noise and unbiased by outliers. We further show that both volatility and the benefits from using the ATM method are related to sample size.

The paper is structured as follows. The next section provides a brief review of the literature on index construction with scarce data. We then discuss our empirical strategy. The subsequent section discusses our results, while we provide some concluding remarks in a final section.

Previous Research

In general, previous research has responded to scarce data in housing markets by modeling the relationship between current and previous house prices. In other words, the time dummies used to produce indexes in a repeat sales or hedonic model are replaced with time series models or models that take advantage of temporal and spatial correlation.[1]

Schwann (1998) replaces the time dummies in his estimating equation with a parsimoniously parameterized time-series function. Using data for Vancouver, B.C., his price index is less biased and less noisy than a benchmark hedonic index. McMillen and Dombrow (2001) use a flexible Fourier expansion to account for time in a repeat sales model. Trigonometric terms allow time to be modeled continuously rather than discretely, resulting in relatively smooth indexes. The method enables them to produce indexes for various suburbs of Cook County, Illinois, for which data are scarce. Francke (2010) uses a local linear trend model in which both the level and the drift parameter can vary over time. The model is applied to Dutch housing data for various levels of spatial granularity. Francke's approach considerably reduces the noisiness of the resulting price indexes.

Modeling approaches based on spatial and temporal correlations create three-dimensional price surfaces that can be used to produce indexes for locations with scarce data. For example, Pace, Barry, Clapp, and Rodriguez (1998) combine spatial autoregressive models with a temporal autoregressive process to produce a spatiotemporal autoregressive (STAR) model for Fairfax County, Virginia. The STAR model makes use of the *k* nearest neighbors in space and time. Another example is provided by Clapp (2004), who applies a semiparametric hedonic method to data for Fairfax County. His approach yields a price surface that changes

[1] Another approach in the repeat sales context is to increase the sample size by adding data for properties that sold only once during the sample period. Guntermann, Liu, and Nowak (forthcoming) accomplish this by creating new pairs of transactions based on nearest neighbors.

over time. Clapp's indexes are estimated fairly precisely and perform better out-of-sample than standard hedonic models.

Focusing on commercial real estate price indexes, Bokhari and Geltner (2012) introduce the two-stage frequency conversion method that we apply here to housing markets. They apply the method to different commercial real estate sectors in multiple metropolitan areas and regions in the U.S. In the first stage, they compute annual repeat sales regressions with quarterly staggered starting dates. They then convert the annual indexes to quarterly indexes using a generalized inverse estimator. Their frequency conversion indexes are less volatile than standard repeat sales indexes.

Empirical Strategy

In this section, we describe our methods and data, including the robust technique that we apply when estimating repeat sales models and the frequency conversion method adapted from Bokhari and Geltner (2012). We then explain how we measure the effects of the frequency conversion method on house price index volatility. Finally, we give an overview of our Louisville data.

Robust Repeat Sales Approach

Robust estimators can be used to reduce the influence of outliers in a wide range of contexts, including the construction of house price indexes. In contrast to OLS, robust estimators do not minimize the sum of squared errors; instead, they minimize some other function that reduces the influence of large errors. Bourassa, Cantoni, and Hoesli (2013) show that a robust bisquare estimator does a much better job than OLS in tracking a "true" repeat sales index and in reducing the magnitude of revisions when new data are added. The bisquare estimator in effect gives zero weight to observations with large errors (i.e., large standardized residuals), thereby eliminating their impacts on the resulting price indexes (Beaton and Tukey, 1974).

We use the robust bisquare estimator for two purposes.[2] First, we use it directly to produce monthly repeat sales indexes that serve as our baselines for comparison with the indexes created using the frequency conversion method. Second, we use it to produce the staggered annual repeat sales indexes that are the first stage of the frequency conversion process.

For the monthly baseline indexes, we apply the robust method to a repeat sales model with time dummies defined as in Bailey, Muth, and Nourse (1963): the first sale in each pair of transactions is given a value of -1, the second sale a value of 1, and all other dummies a value of 0. In this case, the estimated coefficients on the time dummies measure the price change from the base period to each subsequent time period. An alternative approach to specifying a repeat sales model is to assign a value of 1 to each period starting with the first sale and ending with the second sale; all other periods are given a value of 0. Shiller (1993) shows that this is equivalent to the Bailey, Muth, and Nourse approach; however, in this case the estimated coefficient for a time period measures only the price change within that period. A variation on this is to time-weight the dummies depending on when the transaction took place. As Bokhari and Geltner (2012) point out, this is particularly important for low-frequency indexes because it eliminates temporal aggregation. With respect to the annual indexes calculated for the first stage of the frequency conversion process, we follow Bokhari and Geltner (2012) and calculate time-weighted dummies that reflect the proportion of each year that is between the two transactions. For example, for an index based on calendar years, a property that sold first on 1 July 2006 and then again on 30 June 2008 would have values of 0.5 for 2006 and 2008, a value of 1 for 2007, and values of 0 for all other years.

Frequency Conversion Method

The first stage of the frequency conversion method involves estimating a staggered series of low frequency robust repeat sales indexes. In our case, we estimate 12 annual models for years defined to start at

[2] We use the *robustreg* procedure in SAS.

the beginning of each January, February, and so forth. For example, the January regression has time-weighted dummy variables that each span the period from January 1 through December 31. We save the coefficients ("returns") from each of the staggered annual estimations for use in the second stage.

In the second stage, we stack the returns from the first stage on the left-hand side of the equation and specify monthly time dummies on the right-hand side. The monthly dummies are set equal to 1 for each month during the 12 months to which the return applies and zero otherwise. For example, for the return applicable to the calendar year 2006, the dummies for each month in 2006 would be set equal to 1, while all other dummies would be set equal to 0. For this equation, we have 139 returns (for the 139 annual periods starting in January 1998 through July 2009) and 150 time dummies (for the 150 months from January 1998 through June 2010). Given that we have fewer observations than unknowns, there is no unique solution to the regression equation. However, as Bokhari and Geltner (2012) point out, one solution is better than the others because it minimizes the variance of the estimates. This is obtained using the generalized inverse estimator (GIE), which we apply here. We use the *iml* procedure in SAS to implement the GIE; a sample of the SAS code is provided in the Appendix.

Measuring Effects on Volatility

To assess the effects of the ATM frequency conversion approach on index volatility, we calculate the standard deviations of monthly price changes for the baseline robust repeat sales index and compare those with the corresponding standard deviations for the robust ATM index. We do this comparison for the city as a whole and for submarkets. We further show how the benefit of using the ATM approach varies with sample size. For this purpose, we regress the ratios of the ATM and baseline standard deviations for submarkets on sample size. In a second regression, we also explore the relationship between the volatility of the robust ATM index and sample size.

Data

We use data for single-family houses that sold in Louisville, Kentucky (Jefferson County), between January 1998 and June 2010. The data were obtained from the Multiple Listing Service (MLS) of the Greater Louisville Association of Realtors. Some 23,906 repeat sales were identified. For analysis purposes, we divide Louisville into two sets of submarkets, one set consisting of nine areas that approximate areas defined by the MLS and the other set consisting of 30 zip codes. The sample sizes (numbers of repeat sales) ranged from 1,759 to 4,720 for the nine areas and from 214 to 1,724 for the zip codes.[3]

Results

Exhibit 1 contains the baseline robust repeat sales index and the robust ATM index for Louisville as a whole. The ATM index is much less volatile than the baseline index. Exhibit 2, Panel A, contains the standard deviations of index changes. The volatility of the ATM index (0.3%) is 83% lower than that of the baseline index (1.7%). Turning now to the analysis of the set of nine submarkets, Exhibits 3 and 4 show the indexes for the areas with the highest (Area 7) and the lowest (Area 4) numbers of repeat sales, respectively. Again, the ATM indexes are far less volatile than the baseline indexes. Although we see substantial improvement in the ATM indexes relative to the baseline indexes (Exhibit 2, Panel B), the level of volatility in the ATM index may nevertheless remain too high. In Area 1, for example, the volatility of the ATM index is more than six times as high as that for the city.[4]

[3] We omitted six zip codes with very small sample sizes as it was not possible to estimate the baseline indexes in those cases.

[4] Area 1 was a particularly volatile submarket as it experienced considerable investment by landlords during the early part of our study period and large numbers of foreclosures during the latter part of the period.

Exhibits 5 and 6 show the baseline and ATM indexes for the zip codes with the highest and lowest numbers of repeat sales, respectively. The ATM conversion yields an index with little volatility in the zip code with the highest number of repeat sales. In the zip code with the lowest number of sales, however, the ATM index still displays an implausible level of volatility although its level is 92% less than the volatility of the corresponding baseline index. The high level of volatility is presumably attributable to the small number of repeat sales in that area (214) combined with the characteristics of the underlying transactions data. Exhibit 2, Panel C, indicates that the six zip codes with the highest monthly price index volatility all have relatively small numbers of repeat sales.[5] However, the indexes for some zip codes with small numbers of transactions exhibit low volatility. Exhibit 7 shows an example. Although the number of transactions in that case is only 397, the standard deviation of the index changes is 0.5%.

To determine how the scarcity of data affects the benefits from using the ATM conversion, we regress the ratios of the ATM and baseline standard deviations for submarkets (based on zip codes) on sample size. The lower the ratio, the greater the benefit of using the ATM conversion. The regression results indicate that the benefits from the ATM method diminish (the ratio increases) as sample size (n) increases (t-statistics and significance levels are shown in parentheses, with *** = 1%):

$$\underset{}{Ratio} = \underset{(4.421^{***})}{0.032} + \underset{(3.933^{***})}{0.000032}\, n \quad Adj.R^2 = 0.333 \quad N = 30 \tag{1}$$

We also measure the effects of sample size on the volatility of the robust ATM index. Here we regress the standard deviations of the changes in the ATM indexes for zip codes on the natural logarithm of the number of repeat sales. The results show that sample size is negatively related to volatility:

$$\underset{}{Volatility} = \underset{(4.819^{***})}{0.154} - \underset{(-4.396^{***})}{0.021}\ln(n) \quad Adj.R^2 = 0.387 \quad N = 30 \tag{2}$$

[5] In practice, adjacent zip codes could be combined with the aim of reducing excessive volatility.

Concluding Remarks

Although housing markets are more liquid than commercial real estate markets, the construction of price indexes may be hampered by scarce data when one wishes to increase index frequency or spatial granularity. We show how a two-stage method that has been developed for commercial real estate can be useful in a housing context. The first stage consists of constructing a set of staggered annual robust repeat sales indexes, while the second stage applies a conversion method to yield monthly indexes.

Our empirical analyses show that the resulting robust ATM indexes are far less volatile than repeat sales indexes constructed using monthly dummy variables. The benefits of the approach are shown to be greatest when most needed, i.e., when data are scarce. For some submarkets with small numbers of repeat sales, however, index volatility remains relatively high despite the ATM conversion. We also find that the volatility of the ATM index decreases as sample size increases.

References

Bailey, M.J., R.F. Muth, and H.O. Nourse. A Regression Method for Real Estate Index Construction. *Journal of the American Statistical Association*, 1963, 58:304, 933-942.

Beaton, A.E., and J.W. Tukey. The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data. *Technometrics*, 1974, 16:2, 147-185.

Bokhari, S., and D. Geltner. Estimating Real Estate Price Movements for High Frequency Tradable Indexes in a Scarce Data Environment. *Journal of Real Estate Finance and Economics*, 2012, 45:2, 522-543.

Bourassa, S.C., E. Cantoni, and M. Hoesli. Robust Repeat Sales Indexes. *Real Estate Economics*, 2013, 41:3, 517-541.

Bourassa, S.C., M. Hoesli, and V.S. Peng. Do Housing Submarkets Really Matter? *Journal of Housing Economics*, 2003, 12:1, 12-28.

Clapp, J.M. A Semiparametric Method for Estimating Local House Price Indices. *Real Estate Economics*, 2004, 32:1, 127-160.

Francke, M.K. Repeat Sales Index for Thin Markets: A Structural Time Series Approach. *Journal of Real Estate Finance and Economics*, 2010, 41:1, 24-52.

Guntermann, K.L., C. Liu, and A.D. Nowak. Price Indexes for Short Horizons, Thin Markets or Smaller Cities. *Journal of Real Estate Research*, forthcoming.

McMillen, D.P., and J. Dombrow. A Flexible Fourier Approach to Repeat Sales Price Indexes. *Real Estate Economics*, 2001, 29:2, 207-225.

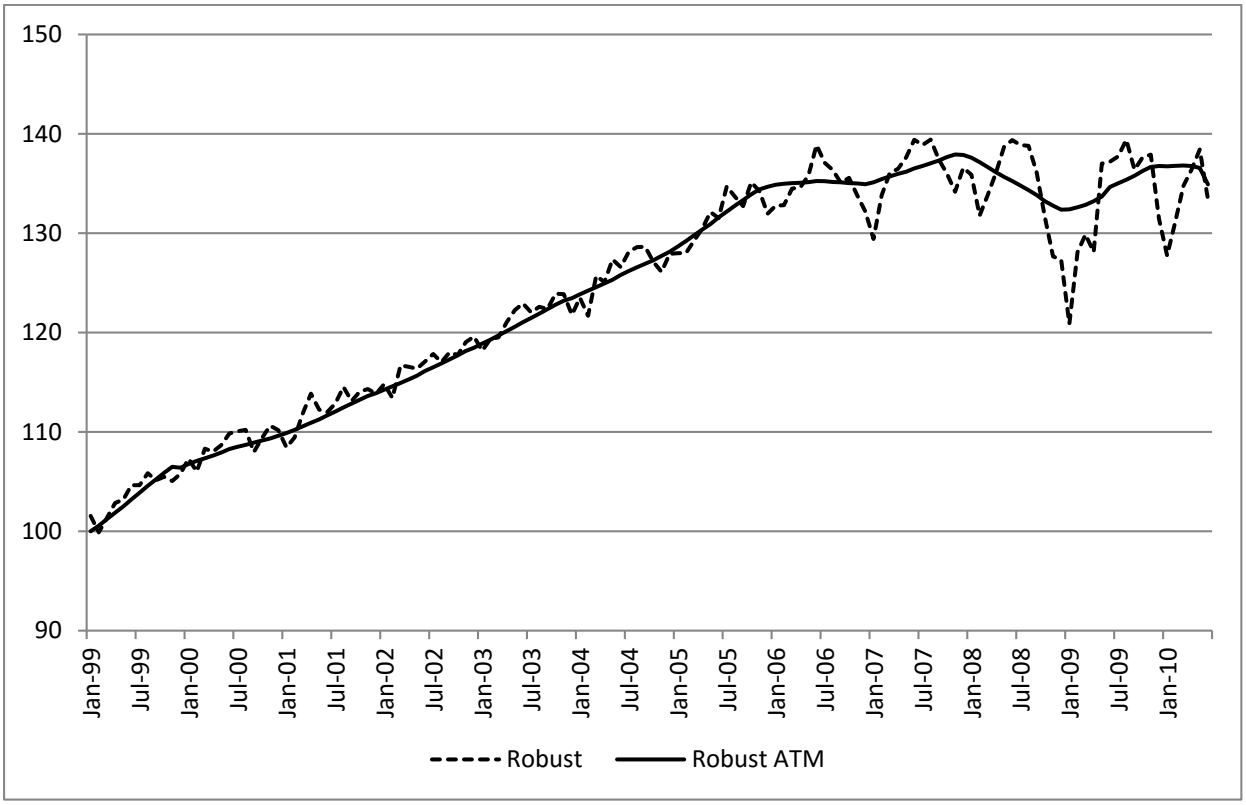
Pace, R.K., R. Barry, J.M. Clapp, and M. Rodriguez. Spatiotemporal Autoregressive Models of Neighborhood Effects. *Journal of Real Estate Finance and Economics*, 1998, 17:1, 15-33.

Schwann, G.M. A Real Estate Price Index for Thin Markets. *Journal of Real Estate Finance and Economics*, 1998, 16:3, 269-287.

Shiller, R.J. *Macro Markets: Creating Institutions for Managing Society's Largest Economic Risks*. Oxford: Oxford University Press, 1993.

Exhibit 1

Robust and Robust ATM Indexes for Louisville, January 1999-June 2010



Note: January 1999 = 100 for the robust ATM index. Due to the volatility in the baseline robust index, the starting point for that index is modified so that the mean of the baseline robust index numbers is equal to the mean of the robust ATM index numbers. The indexes shown here start in 1999 rather than 1998 because the price dynamics estimated for 1998 are sometimes implausible; this is presumably due to the low numbers of transactions in the first year of data (limited mainly to first sales).

Exhibit 2

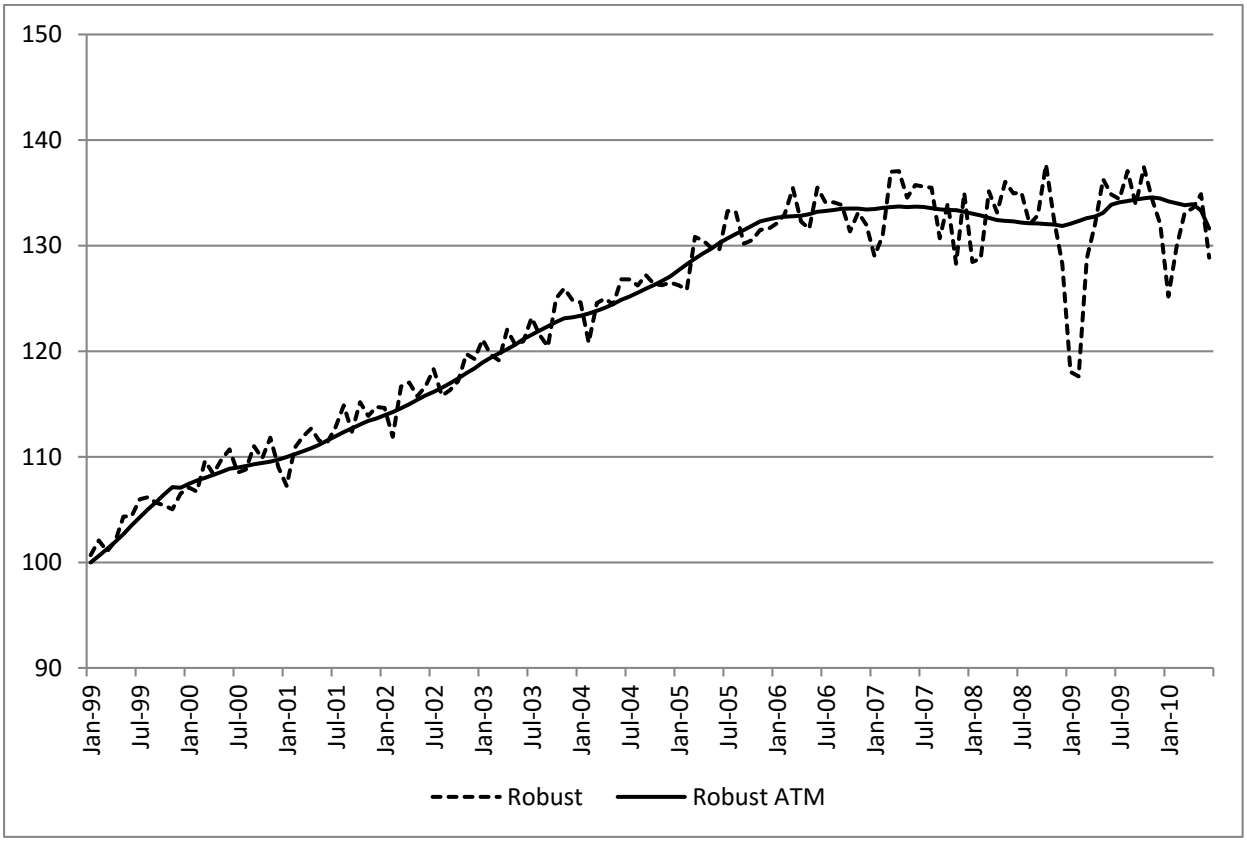
Volatility of Robust and Robust ATM Index Changes

Geographical Area	Standard Deviations for Robust Index Changes (%)	Standard Deviation for Robust ATM Index Changes (%)	Ratio of Standard Deviations for Robust ATM and Robust Index Changes (%)	Sample Size (Number of Repeat Sales)
Panel A: Citywide Results				
Louisville	1.7	0.3	16.6	23,906
Panel B: Results for Nine Submarkets				
Area 1	22.7	1.9	8.2	2,181
Area 2	6.0	0.4	6.2	2,996
Area 3	3.8	0.3	8.3	1,916
Area 4	12.3	1.2	9.8	1,759
Area 5	10.6	0.9	8.5	2,037
Area 6	5.8	0.5	8.3	2,278
Area 7	2.2	0.3	11.6	4,720
Area 8	2.2	0.3	12.1	3,680
Area 9	2.4	0.2	10.2	2,339
Panel C: Results for Zip Codes				
40059	11.3	0.7	6.1	435
40203	97.5	7.8	8.0	214
40204	22.3	0.6	2.9	681
40205	7.2	0.5	6.9	1,110
40206	17.4	0.6	3.7	609
40207	4.5	0.3	6.8	1,307
40208	84.9	2.7	3.1	238
40210	110.7	4.2	3.8	251
40211	70.1	6.8	9.7	463
40212	92.5	3.7	4.0	336
40213	28.9	1.1	3.9	488
40214	17.3	1.1	6.1	925
40215	42.1	2.0	4.8	661
40216	17.5	1.6	9.3	1,013
40217	21.4	0.6	3.0	717
40218	17.1	0.6	3.8	703
40219	13.1	1.0	7.4	876
40220	4.2	0.3	7.4	1,125
40222	7.9	0.5	6.9	715
40223	6.0	0.3	4.6	840
40228	11.5	0.5	4.4	526
40229	11.1	0.5	4.2	876
40241	3.0	0.2	8.2	1,500
40242	14.5	0.5	3.1	397
40243	10.6	0.4	3.4	401
40245	3.0	0.3	9.4	1,724

Geographical Area	Standard Deviations for Robust Index Changes (%)	Standard Deviation for Robust ATM Index Changes (%)	Ratio of Standard Deviations for Robust ATM and Robust Index Changes (%)	Sample Size (Number of Repeat Sales)
40258	17.4	0.7	4.1	746
40272	16.0	1.1	7.0	973
40291	3.9	0.3	8.6	1,219
40299	3.7	0.3	8.1	1,643

Exhibit 3

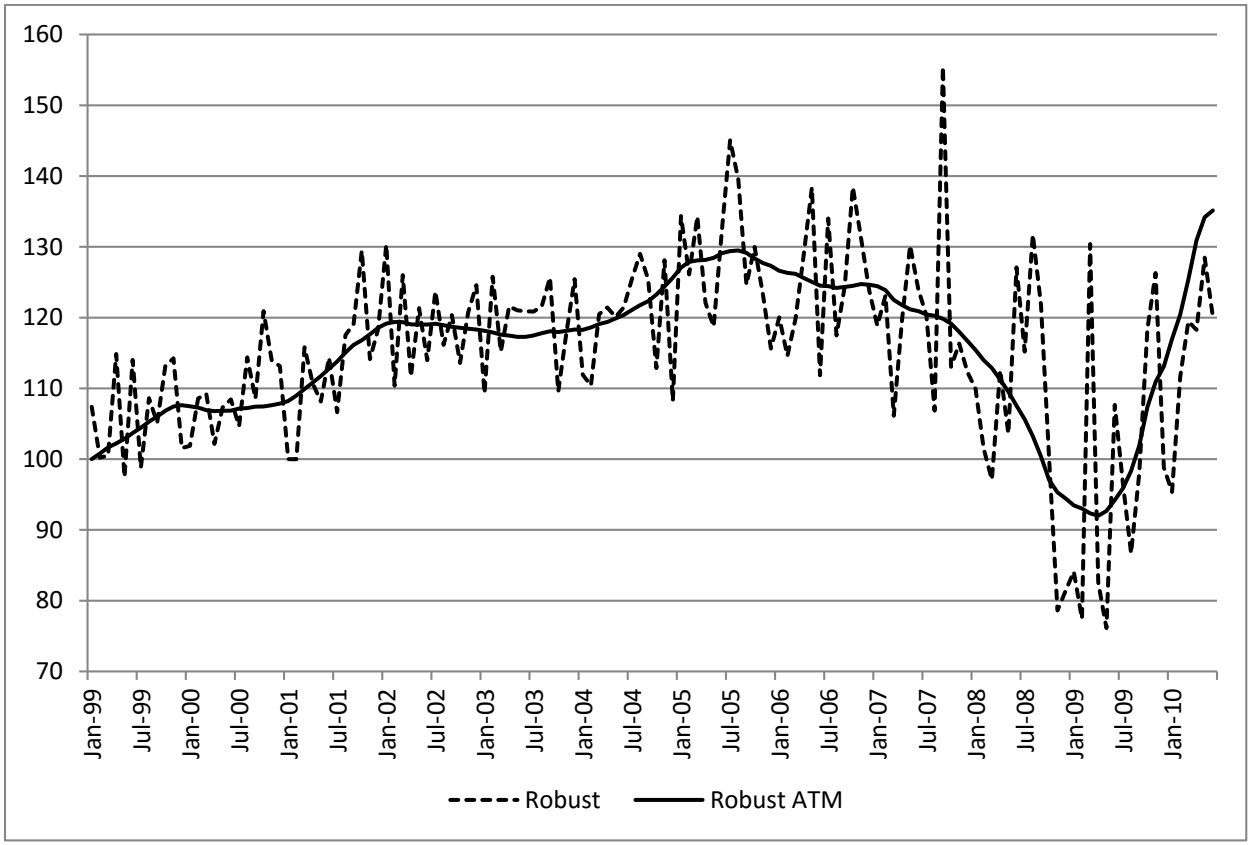
Robust and Robust ATM Indexes for Area 7 (*n* = 4,720)



Note: See Exhibit 1.

Exhibit 4

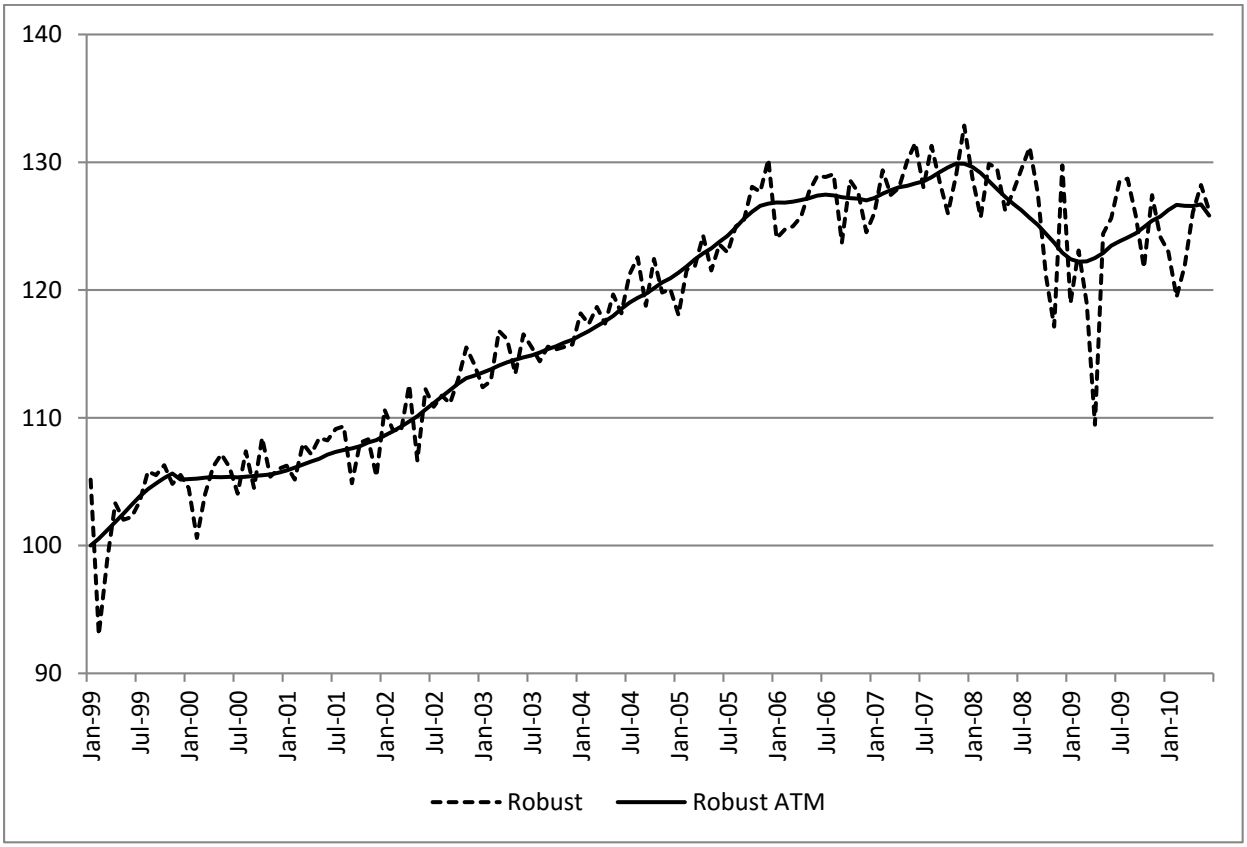
Robust and Robust ATM Indexes for Area 4 (*n* = 1,759)



Note: See Exhibit 1.

Exhibit 5

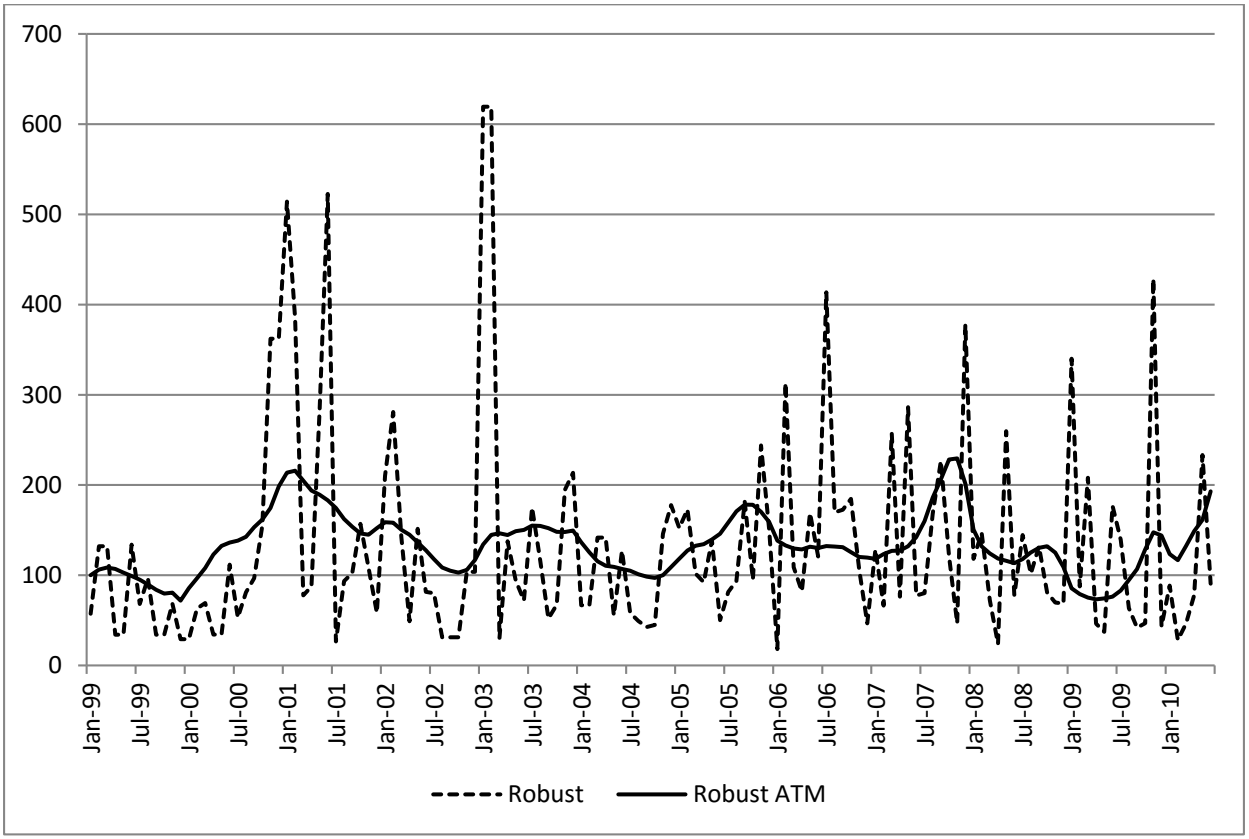
Robust and Robust ATM Indexes for Zip Code 40245 (*n* = 1,724)



Note: See Exhibit 1.

Exhibit 6

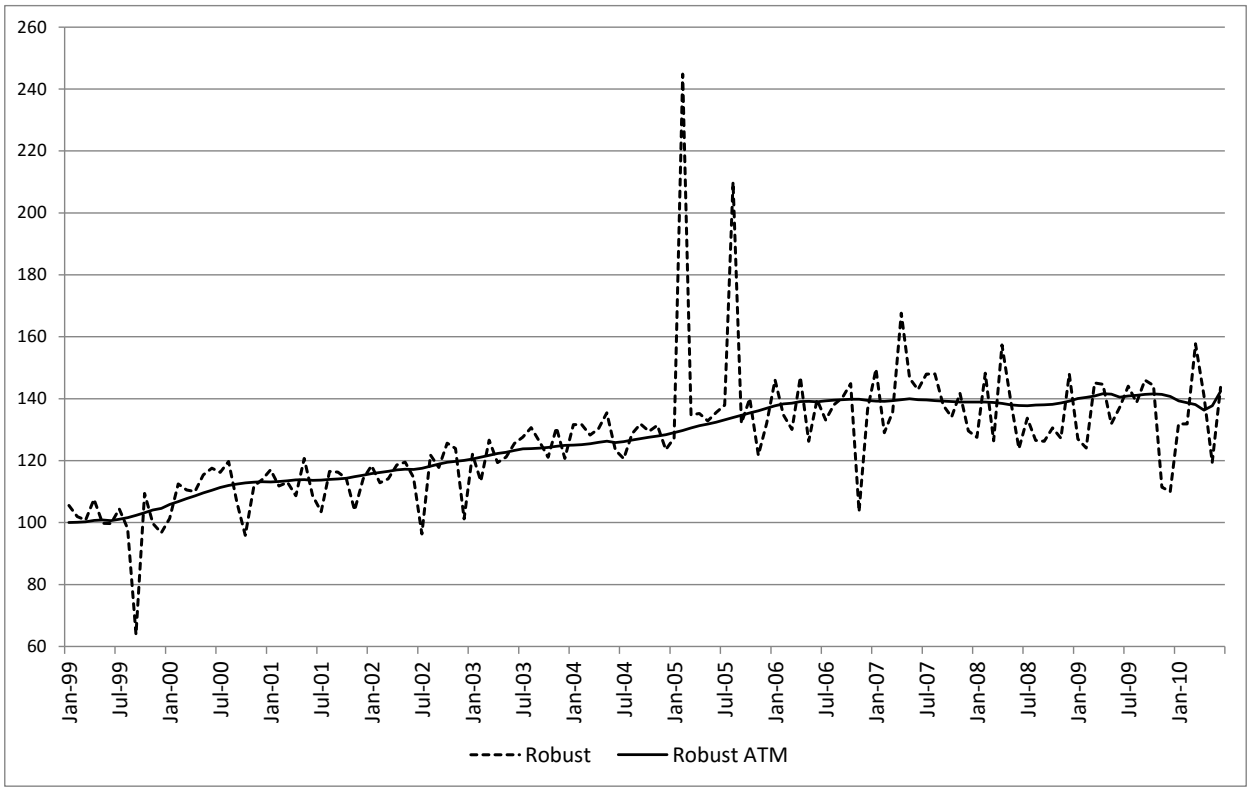
Robust and Robust ATM Indexes for Zip Code 40203 (*n* = 214)



Note: See Exhibit 1.

Exhibit 7

Robust and Robust ATM Indexes for Zip Code 40242 (*n* = 397)



Note: See Exhibit 1.

Appendix

The following SAS code shows how to implement the second stage of the frequency conversion method. The first data step "stacks" the returns (previously saved to the files *janout* through *decout*) from the first stage staggered annual regressions. The second data step sorts the data by year (it is already sorted by month). The third data step creates the dummy variables, sstd1-sstd150, for the right-hand side of the second stage equation. The *proc iml* step applies the generalized inverse estimator (*ginv*) to estimate the second stage equation. The final step outputs the results to an external file.

```
data temp1;
   merge janout febout marout aprout mayout junout julout augout sepout octout novout decout;
   proc transpose data=temp1 out=temp2;
run;

data temp3;
   set temp2;
   yearseqno=substr(_NAME_,7);
   return=exp(col1)-1;
   proc sort; by yearseqno;
run;

data temp4;
   set temp3;
   array sstd{150} sstd1-sstd150;
   do i=1 to 150;
      if i<_N_ then sstd{i}=0;
      else if i<=(_N_+11) then sstd{i}=1;
      else sstd{i}=0;
   end;
run;

proc iml;
   use temp4;
   read all var{sstd1 sstd2 sstd3 sstd4 sstd5 sstd6 sstd7 sstd8 sstd9 sstd10
      sstd11 sstd12 sstd13 sstd14 sstd15 sstd16 sstd17 sstd18 sstd19 sstd20
      sstd21 sstd22 sstd23 sstd24 sstd25 sstd26 sstd27 sstd28 sstd29 sstd30
      sstd31 sstd32 sstd33 sstd34 sstd35 sstd36 sstd37 sstd38 sstd39 sstd40
      sstd41 sstd42 sstd43 sstd44 sstd45 sstd46 sstd47 sstd48 sstd49 sstd50
      sstd51 sstd52 sstd53 sstd54 sstd55 sstd56 sstd57 sstd58 sstd59 sstd60
      sstd61 sstd62 sstd63 sstd64 sstd65 sstd66 sstd67 sstd68 sstd69 sstd70
      sstd71 sstd72 sstd73 sstd74 sstd75 sstd76 sstd77 sstd78 sstd79 sstd80
      sstd81 sstd82 sstd83 sstd84 sstd85 sstd86 sstd87 sstd88 sstd89 sstd90
      sstd91 sstd92 sstd93 sstd94 sstd95 sstd96 sstd97 sstd98 sstd99 sstd100
      sstd101 sstd102 sstd103 sstd104 sstd105 sstd106 sstd107 sstd108 sstd109 sstd110
      sstd111 sstd112 sstd113 sstd114 sstd115 sstd116 sstd117 sstd118 sstd119 sstd120
      sstd121 sstd122 sstd123 sstd124 sstd125 sstd126 sstd127 sstd128 sstd129 sstd130
      sstd131 sstd132 sstd133 sstd134 sstd135 sstd136 sstd137 sstd138 sstd139 sstd140
      sstd141 sstd142 sstd143 sstd144 sstd145 sstd146 sstd147 sstd148 sstd149 sstd150} into x;
   read all var{return} into y;
   xinv=ginv(x);
   b=xinv*y;
   create atm_out from b;
   append from b;
   show contents;
   close atm_out;
quit;

data _null_;
   %let _EFIERR_ = 0; /* set the ERROR detection macro variable */
   %let _EFIREC_ = 0; /* clear export record count macro variable */
   file '<file reference for output file>' delimiter=',' dsd dropover lrecl=32767;
   if _n_ = 1 then do; /* write column names or labels */
      put "COL1";
   end;
```

```
   set work.atm_out end=EFIEOD;
   format COL1 best12.;
   do;
      EFIOUT + 1;
      put COL1;
      ;
   end;
   if _ERROR_ then call symputx('_EFIERR_',1); /* set ERROR detection macro variable */
   if EFIEOD then call symputx('_EFIREC_',EFIOUT);
run;
```

swiss:finance:institute

c/o University of Geneva
40 bd du Pont d'Arve
1211 Geneva 4
Switzerland

T +41 22 379 84 71
F +41 22 379 82 77
RPS@sfi.ch
www.SwissFinanceInstitute.ch