Electronic Journal of Statistics
ISSN: 1935-7524

Additional Examples for Casting Vector Time Series: Algorithms for Forecasting, Imputation, and Signal Extraction

Tucker McElroy[1]

[1] *Research and Methodology Directorate U.S. Census Bureau, 4600 Silver Hill Road, Washington, D.C. 20233 e-mail:* tucker.s.mcelroy@census.gov

Disclaimer

This paper is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the author and not those of the U.S. Census Bureau. The statistics in this paper were approved for dissemination by the Disclosure Review Board (release number: CBDRB-FY21-CED004-001).

1. Monthly Housing Starts

Here we study "New Residential Construction (1964-2012), Housing Units Started, Single Family Units," or *housing starts* for short[1], corresponding to the four regions of South, West, NorthEast, and MidWest. Our objective with this illustration is to show that the direct matrix approach is identical to the methods of Section 3.4, as we take the filter truncation level m larger and larger. Secondly, we show how growth rates can be generated, along with uncertainty, using either method. The direct matrix formulas for signal extraction (McElroy and Trimbur, 2015) were computed; this was feasible, because there are 49 full years of data (no missing values), covering 1964-2012, so that $T = 588$ is of moderate size.

We proceeded by fitting the structural model discussed in McElroy (2017) to the full data span (after some pre-processing for outliers), disallowing any co-integration constraints. We computed the WK filter for trend, trend-irregular (or seasonally adjusted), and seasonal components, with filter coefficients and frequency response functions computed in the manner described in Section C of the Supplement. With forecasts and aftcasts, we then generated the signal extractions along with uncertainty, using both the direct matrix approach and $m = 250$ casts (see Figure 1). The shading around each extraction indicates a confidence interval of plus or minus two times the square root MSE.

[1]The four series are from the Survey of Construction of the U.S. Census Bureau, available at http://www.census.gov/construction/nrc/how_the_data_are_collected/soc.html.

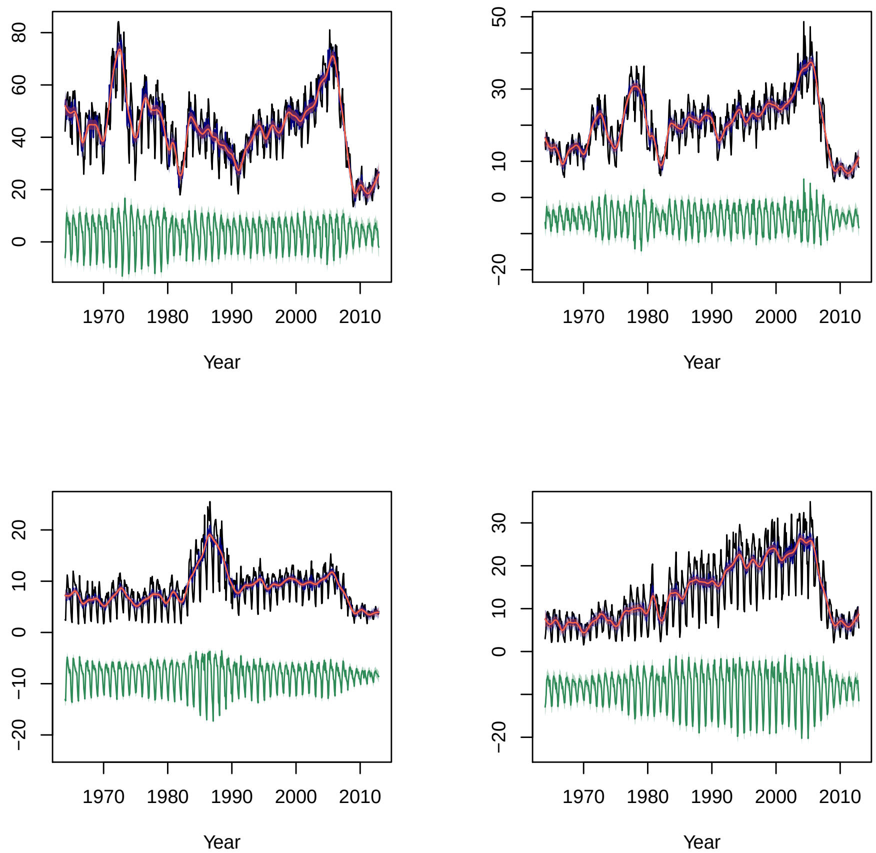

Fig 1. *Extracted components with (shaded) uncertainty bands of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel). The data (black) is overlaid with trend (red), trend-irregular (blue), and seasonal (green) components; the seasonal component has been vertically displaced for easier visualization.*

The time-varying MSE is increased at the sample boundaries, which is reflected by the width of the uncertainty shading being somewhat wider (though this is hard to discern in these plots). To better visualize the signal extraction standard error, we have plotted these for the seasonal adjustment extraction in Figures 2 and 3, corresponding to $m = 10$ and $m = 50$ number of aftcasts and forecasts, respectively. Whereas some discrepancies are apparent in Figure 2 between the direct matrix approach and the truncated filter approach, these discrepancies are largely reduced in Figure 3. There is virtually perfect agreement at $m = 250$ (not shown). Of course, the analyst has complete control over selecting m, the only limitations being on processing time and memory.

Finally, suppose we wish to know growth rates for the trend. According to Section 3.4, we select $\varphi(z) = I_N - I_N z$, as this corresponds to a first difference of each component time series. Utilizing *Ecce Signum*, the resulting trend growth rates (with corresponding uncertainty as shasing) are displayed in Figure 4.

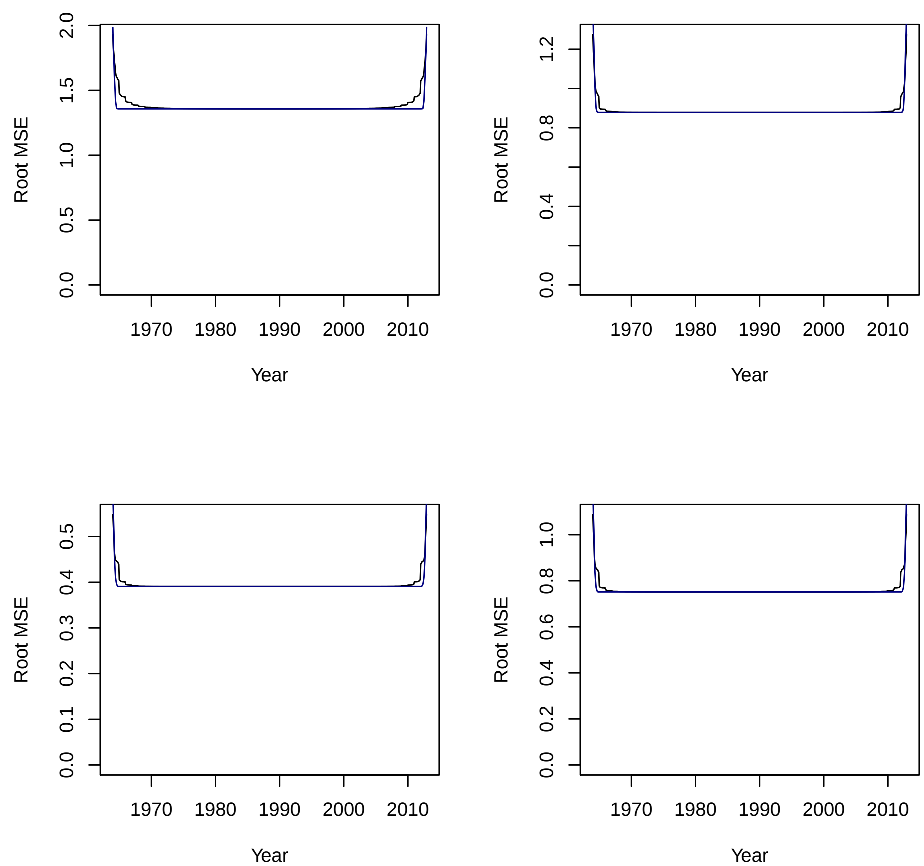

FIG 2. *Square root MSE over time for the trend-irregular component of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel). The exact square root MSE (black) is compared to the approximation (blue) based on truncating the WK filter using $m = 10$ casts.*

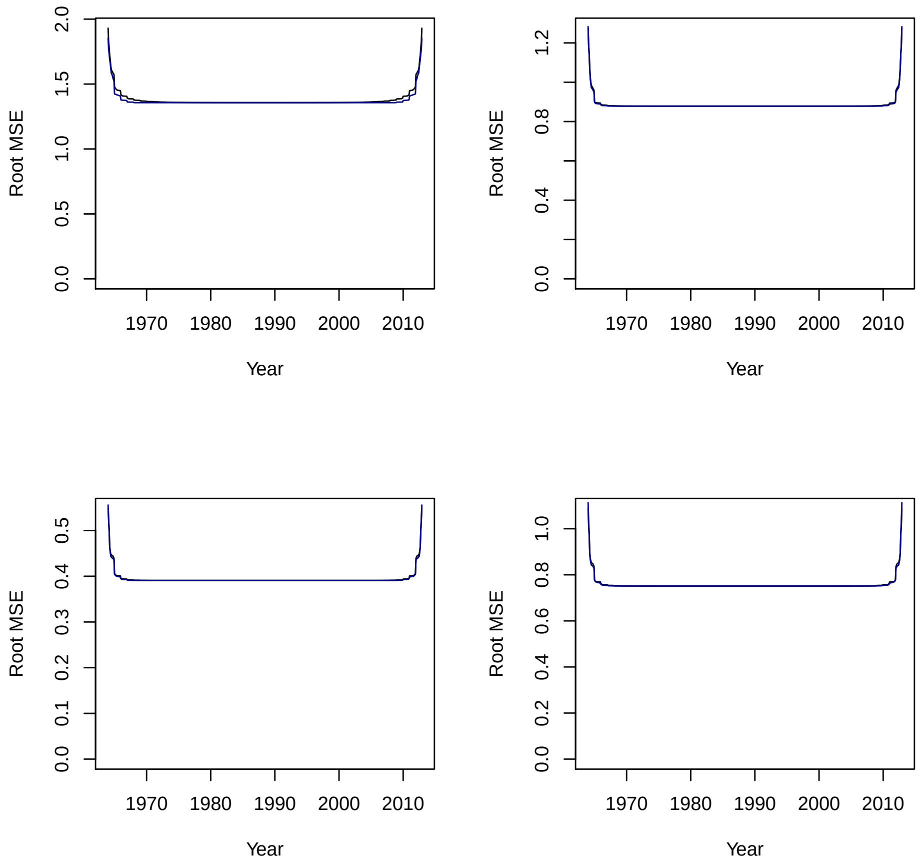

Fig 3. *Square root MSE over time for the trend-irregular component of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel). The exact square root MSE (black) is compared to the approximation (blue) based on truncating the WK filter using $m = 50$ casts.*

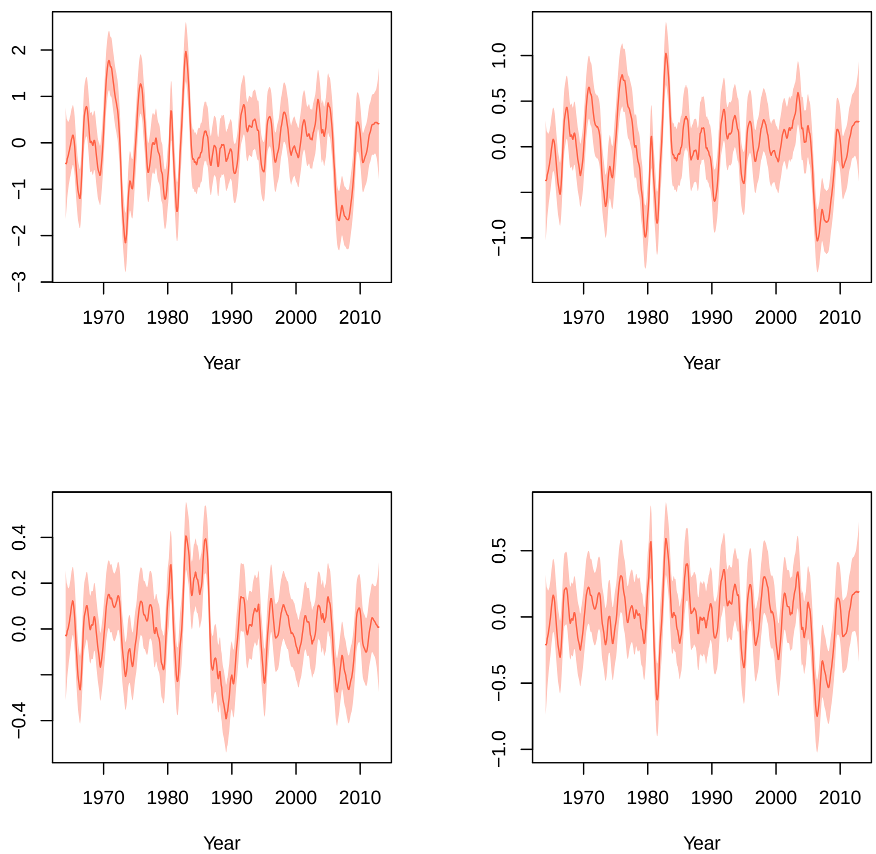

Fig 4. *Trend growth rates (red) with (shaded) uncertainty bands of housing starts, for South (upper left panel), West (upper right panel), NorthEast (lower left panel), and MidWest (lower right panel).*

2. Weekly Retail Products

The next example consists of weekly time series obtained from Dominick's Database, published by the Kilts Center for Marketing (University of Chicago School of Business). This database contains weekly store-level sales from Dominick's Finer Foods from 1989 through 1994 around the Chicago, Illinois area in the U.S.A. For one of the stores we have extracted the sales data for bathroom tissues (tti) and paper towels (ptw), and refer to the bivariate series as *products* for short. The purpose of this example is to illustrate the solution to the ragged edge problem in a low-dimensional case where it is feasible to compare to matrix-based solutions.

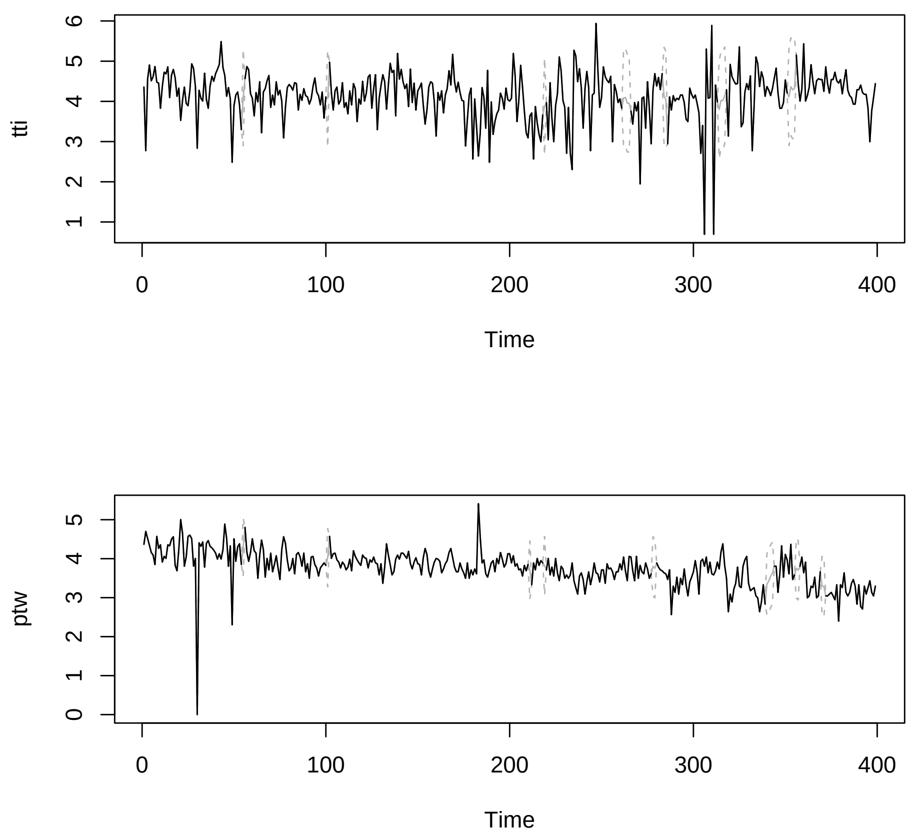

Fig 5. *Plots of bathroom tissues (tti) and paper towels (ptw) by week (black), with imputed values (solid grey) and confidence intervals (dashed grey) for NAs.*

In addition to some missing values (which are missing for both tti and ptw), there are meager values corresponding to virtually no sales activity. We replace any non-positive values with an NA, treating them as missing so that they can be imputed – this is akin to treating such meager values as (small) extremes, and facilitates taking log transforms of the data. However, these meager values occur in a ragged pattern, so that there are times for which only one variable is an NA. In particular these times are

$$t = 55, 101, 219, 262, 263, 264, 265, 284, 285, 314, 315, 316, 317, 352, 353, 354, 355$$

for tti and

$$t = 55, 101, 211, 219, 278, 279, 340, 341, 342, 343, 356, 357, 370, 371$$

for ptw. Based on exploratory analysis of this data, a VAR model was identified and various orders of fit were attempted. The likelihood evaluation uses the exact methods of this article, implicitly casting and evaluating throughout the optimization routine. A stable VAR parameterization was utilized, because the data does not display non-stationarity. The final model was of order 3 with a mean and no other regressors. The final value of the likelihood was -317.8978, which was verified by a direct matrix approach based on Section 2 – this is fairly easy to implement because there are no differencing operators. The diagnostics for the residuals indicate that the model is adequate for illustrative purposes. Finally, casts (with casting MSEs) were generated for all the NAs, and plotted in Figure 5. Available data is in black, whereas imputed NAs are solid grey, with dashed grey lines denoting a confidence interval of plus or minus two times the square root MSE.