Developing and Improving a Moving Regression Weekly Seasonal Adjustment Program[1]

Revised June 5, 2006

Thomas Evans, Jerry Fields, and Stuart Scott, Bureau of Labor Statistics
Contact: Thomas Evans, Bureau of Labor Statistics, Statistical Methods Staff,
Room 4985, 2 Massachusetts Ave. NE, Washington, DC 20212, USA

Pierce, Grupe, and Cleveland (1984) introduced a fixed regression approach to the problem of seasonal adjustment for weekly time series. Cleveland (1993) expanded this approach by adding locally-weighted regressions to allow for varying seasonal factors, and the Bureau of Labor Statistics adopted his moving regression approach in 2002 for national Unemployment Insurance claims. However, the moving regression program written by Cleveland was incomplete in some areas. Thus, a new program (RegMOVE) is being developed to simplify execution and improve the output, including the addition of diagnostics and high-resolution graphics. The authors hope to provide a useable program that could be available for others.

Key Words: Time Series; Spectral Analysis

Some economic time series that are of interest to the general public are collected weekly instead of monthly or quarterly. Expertise in seasonally adjusting these series is not as widespread and weekly seasonal adjustment programs are not as developed as those for monthly or quarterly data. X-12-ARIMA (Bureau of the Census, 2002) and TRAMO/SEATS (Gomez and Maravall, 1998) are examples of seasonal adjustment programs that assume constant periodicities in the data and are thus not suitable for weekly data.

The Bureau of Labor Statistics (BLS) seasonally adjusts weekly initial claims (IC) and continued claims (CC) data from the Unemployment Insurance (UI) program (Cleveland and Scott 2004). Initially, BLS used a program written by Bill Cleveland of the Board of Governors of the Federal Reserve System. The program is based upon the model given by Pierce, Grupe, and Cleveland (1984) that utilizes a fixed coefficient regression with ARIMA errors. This program is referred to here as "CATS-D" for Calendar Adjustment and Time Series-Deterministic and can be used with either weekly or monthly data. Beginning in 2002, BLS began using a modified version of CATS-D, "CATS-M" (CATS-Moving), also written by Cleveland, which combines a fixed coefficient regression with locally weighted regressions (see Cleveland and Scott, 2004 for details) that allows the seasonal factors to change over time which tends to smooth the seasonally adjusted series.

Another approach for weekly seasonal adjustment is proposed by Harvey, Koopman, and Riani (1997). This method uses a structural time series model with periodic time-varying splines that allows for moving seasonality. As of the writing of this paper, BLS has not tested any other alternatives.

The CATS-M program is written in FORTRAN and needs enhancement in some areas. The output is less polished compared to the earlier fixed regression program and does not provide any diagnostics or graphs. This paper explains what changes were made to the program and how the changes make it easier to execute and analyze the results. It is always critical for a program to produce as many diagnostics and graphics simultaneously that in a piecemeal fashion. Otherwise,

[1] Opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

the process can become complex and time consuming.

The plan of this paper follows. Section 1 briefly explains how the calendar can complicate the seasonal adjustment of weekly data compared to that for monthly or quarterly data. The models for the Pierce, Grupe, and Cleveland (1984) approach and the Cleveland (1993) approach are in Sections 2 and 3 respectively. Section 4 describes the improvements made to the moving program and Section 5 discusses possible future work.

1. The Calendar

To appreciate the problems of seasonal adjustment of weekly data, a brief explanation of the calendar is helpful. Almost everyone in the world now uses the Gregorian calendar which has solved most of the issues affecting previous calendars. The Julian calendar was slightly too long and, over time, the vernal equinox kept moving earlier in the calendar year. To help correct for this, the Gregorian calendar has a 400-year cycle with only 97 leap days instead of 100. Pope Gregory XIII decreed the use of this new calendar in 1582, and it was also around this time that countries actually started to adopt January 1st as the first day of the calendar year.

Each year in the Gregorian calendar has either 365 or 366 days to account for the earth's solar orbit of 365.2424 days which means there are a total of 146,097 days and 20,871 weeks in the cycle. It can be seen from the equation below how often to expect differences in the number of weeks in a year:

20,871 weeks = (400 years x 52 weeks) + 400 ordinary days + 97 leap days
= 20,800 weeks + 497 days
= 20,800 weeks + 497/7 weeks
= 20,800 weeks + 71 weeks

Thus, there are 329 52-week years and 71 53-week years in the 400-year cycle, which means there will be a 53-week year every 5.634 years on average. During the cycle, a 53-week year will occur five years apart 27 times, six years apart 43 times, and seven years apart once.

One can see why seasonal adjustment for weekly data is challenging is that since years can have different numbers of weeks and days, this causes the position of weeks to change from year-to-year. For example, the July 4th holiday is sometimes in week 27 and sometimes in week 28. Such variation has a large impact on seasonality from year-to-year and within months. Monthly and quarterly data do not have the above problems except for certain holidays such as Easter, which can occur in either March or April. Overall, there are 14 possible calendars under the Gregorian system.

2. The CATS-D Program

To tackle the problems with weekly data, Pierce, Grupe, and Cleveland (1984) utilize a fixed regression method with ARIMA errors. The first step is to apply a fixed regression to estimate outlier and holiday effects and a deterministic seasonal component. Their basic model is

$$y_t = s_{1t} + p_{1t} + s_{2t} + p_{2t} + e_t \quad (2.1)$$

where y_t is the observed series at time *t*, *s* is the seasonal component, *p* is the trend component and *e* is the irregular component. The subscript *1* refers to the deterministic component and *2* is for the stochastic component.

As dummy variables cannot capture deterministic seasonality for weekly data, CATS-D uses

trigonometric variables instead:

$$w_t = \sum_{i=1}^{k}\left(a_i \sin\frac{2\pi i Y(t)}{N_y} + b_i \cos\frac{2\pi i Y(t)}{N_y} \right) + \sum_{i=1}^{l}\left(c_i \sin\frac{2\pi i M(t)}{N_m} + d_i \cos\frac{2\pi i M(t)}{N_m} \right) \quad (2.2)$$

where w_t is the deterministic seasonal component at week t, N_y is the number of days in the year (365 or 366), N_m is the number of days in the month (28, 29, 30, or 31), *Y(t)* is the index for the day of the year in which week *t* ends, *M(t)* is the index for the day of the month in which week *t* ends, a, b, c, d are the respective sine or cosine coefficients at frequency *i*, and *k* and *l* are the number of seasonal frequencies. (The within-month or trading day effects in (2.2) are not used with weekly UI claims data.) Holiday and other seasonal effects are captured by indicator variables. Several U.S. holidays are included in the program and others can be user defined.

The overall model can be shown by

$$\Delta(y_t - x_t'\beta) = \theta(L)e_t \quad (2.3)$$

where Δ is the first-differencing operator, x_t' is a row of the design matrix for the deterministic parts, and $\theta(L)e_t$ is an IMA representation of the preliminary residuals. UI claims data are logged and differenced. At this time, the program limits the nonseasonal IMA order to (0 1 2). Since modeling any stochastic seasonal effects would be through a seasonal model of the preliminary residuals, the program thus does not attempt to estimate any stochastic seasonal components. Regardless, it is expected that modeling the stochastic seasonal part in this manner would not work too well anyway.

3. The CATS-M Program

Details for the CATS-M methodology and how weights are selected are explained in Cleveland and Scott (2004) in detail. Briefly, the second program starts out similar to the first since a global regression is performed to estimate the calendar effects. The model also uses the seasonal variables in (2.2) and is similar to (2.3)

$$\Delta(y_t - x_t'\beta) = e_t \quad (3.1)$$

except there is no modeling of the errors, and that linear trend and slope terms were added to the design matrix. The weighted regression for a particular year in matrix form is

$$y = X\beta + e \quad (3.2)$$

with weighted least squares estimator

$$\hat{\beta} = (X'WX)^{-1}X'Wy \quad (3.3)$$

where *W* is a diagonal weight matrix. The *W* matrix is found using the model

$$y_t = x_t + e_t \quad (3.4)$$

$$(1-L)(1-\phi L)x_t = a_t \quad (3.5)$$

where ϕ is an autoregressive coefficient. The weights come from

$$E[x \mid y] = \left[I + \upsilon\Sigma_x^{-1} \right]^{-1} y = Wy \quad (3.6)$$

where I is an identity matrix, Σ_x is an autocovariance matrix for the AR process, and $\upsilon = \sigma_e^2 / \sigma_a^2$.

Determining the various parameters required by the program can take some extra work. Scott and Cleveland (2004) explain how this was done for the UI claims data and for steel production. A fixed weight regression will be similar to a high value for υ, while $\upsilon = 10$ and $\upsilon = 24$ are similar to

3x5 and 3x9 seasonal filters respectively, while changing the value of ϕ will affect how quickly the seasonal factors move. One could think of υ as a noise-to-signal ratio: as υ gets larger, more observations are used in the weighting scheme and the seasonally adjusted series will be smoother. For UI claims data, setting $\phi = 0.4$ and $\upsilon = 16$ works well in terms of smoothness, residual seasonality, and revisions, while $\phi = 0.5$ and $\upsilon = 20$ is used for the steel data. About 80 percent of the weight for the first year come from the first three years when $\upsilon = 10$, compared to about two-thirds of the weight when $\upsilon = 24$.

Further, for UI, only the first 30 sine and cosine pairs are necessary for the seasonal term, which is enough to capture periodic effects as short as two weeks. However, the first 48 sine and cosine pairs appear optimal for steel.

Known outliers can be tested by the usual intervention analysis. For example, several consecutive weeks of additive outliers (AOs) are used to handle 9-11 and Hurricane Katrina effects in the UI data. A modified version of X-12-ARIMA that handles very long series actually works surprisingly well for general outlier detection using the residuals from the fixed global regression.

Once the trend, holiday, and outlier effects are removed from the original series, separate regressions are run for each year with the same seasonal model but with different weights. Projected seasonal factors are extrapolated out for a year to conduct seasonal adjustment for the UI claims data. There is some evidence that concurrent seasonality would reduce revisions, but running the program weekly can be a burden. Creating projected factors twice a year may seem more reasonable at this point.

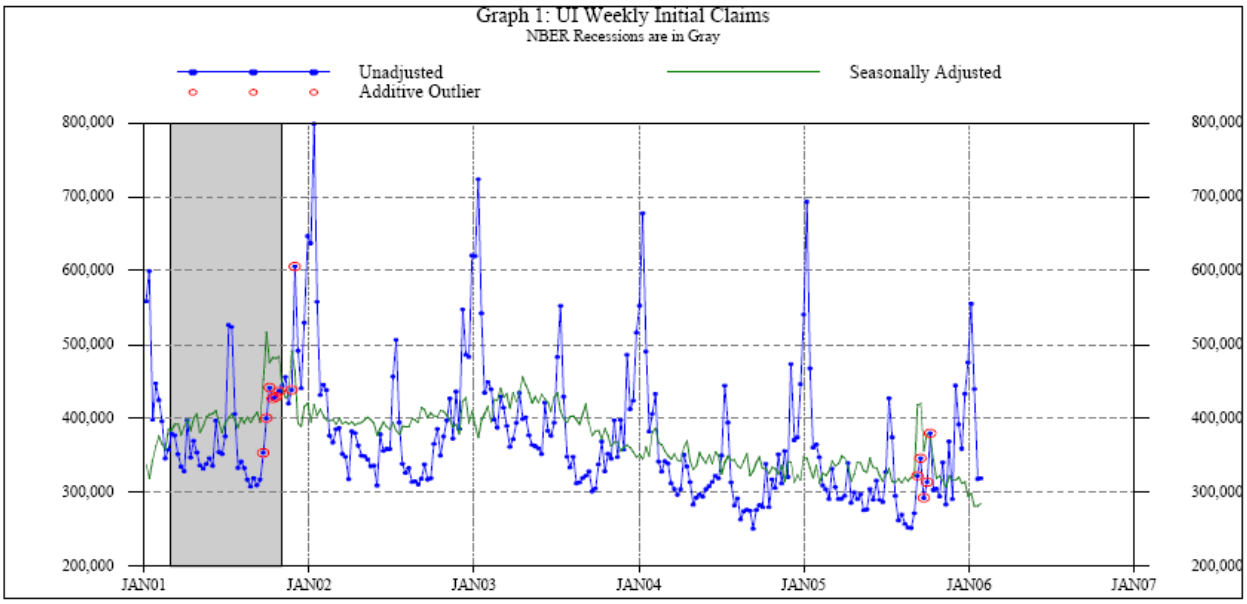
4. The RegMove Program

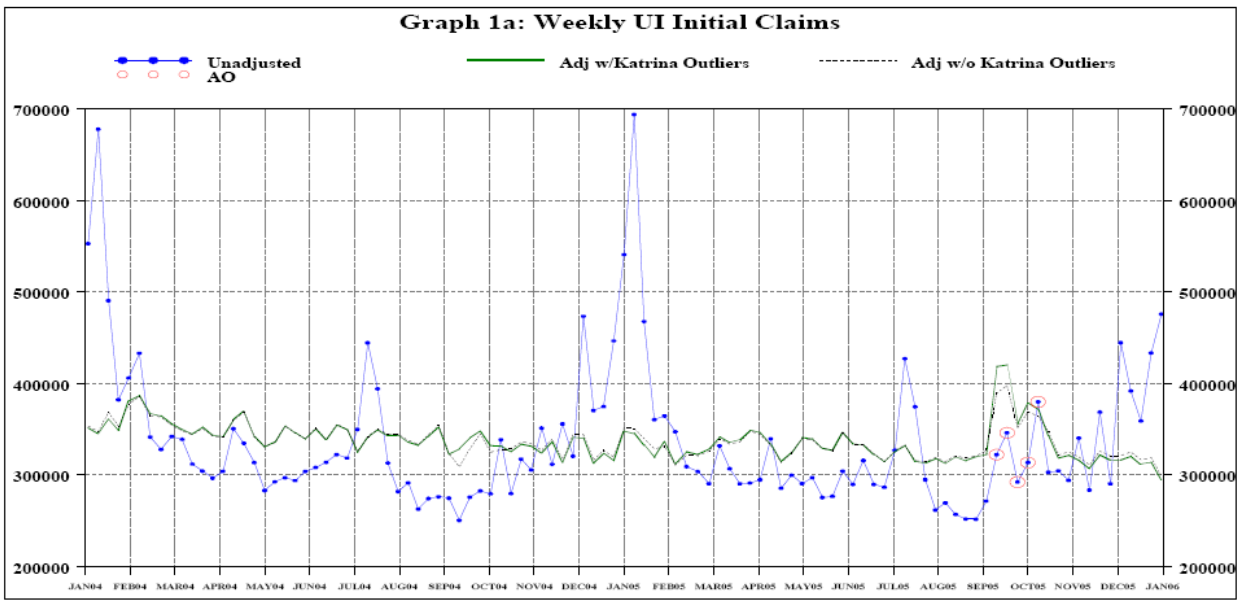
Many changes were made by the authors to the CATS-M program which is now called RegMove which is now a SAS program that calls the FORTRAN code. Regression diagnostics were not available in CATS-M, so tables and machine-readable output files are added in RegMove. Table 1 shows new output for the global regression in RegMOVE. Information for each of the coefficients is now printed out, along with Ljung-Box chi-squared statistics at seasonal lags, an R^2 statistic for the differenced series, outlier, holiday, and individual sine/cosine statistics. Note that the outlier and holiday coefficients are still in multiplicative form. Most of the significant autocorrelations of the differenced unadjusted series come at lags that are multiples of 26. The largest autocorrelations are at the expected lags of 52 and 104. Table 2 has part of another output file that has the original data along with the global fixed regression residuals, t-stats, and p-values. Whenever an outlier or holiday occurs, there is an extra line showing the event's information. Trig seasonal coefficients from the yearly regressions are in another output file displayed in Table 3. Table 4 shows part of the results file with the seasonally adjusted series, seasonal factors, etc.

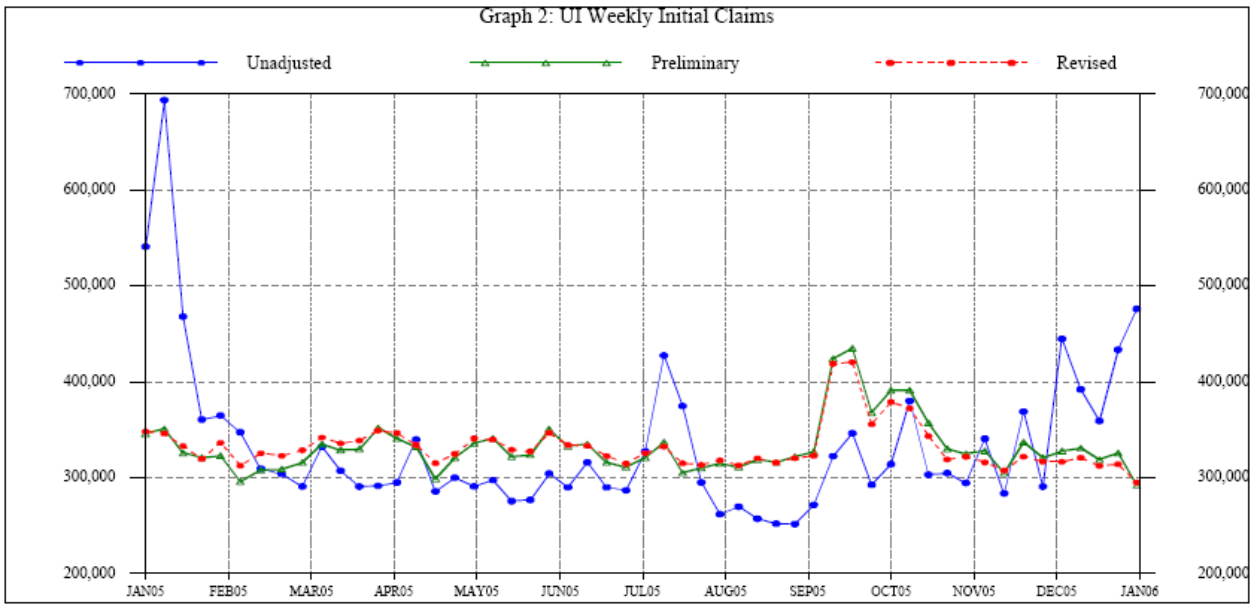
Calling the FORTRAN code under SAS allows high-resolution graphics to be produced only in one run. The graphics can be just as important as the statistics in determining the adequacy of the seasonal adjustment. Below are some of the graphics produced in RegMOVE for the UI IC series.

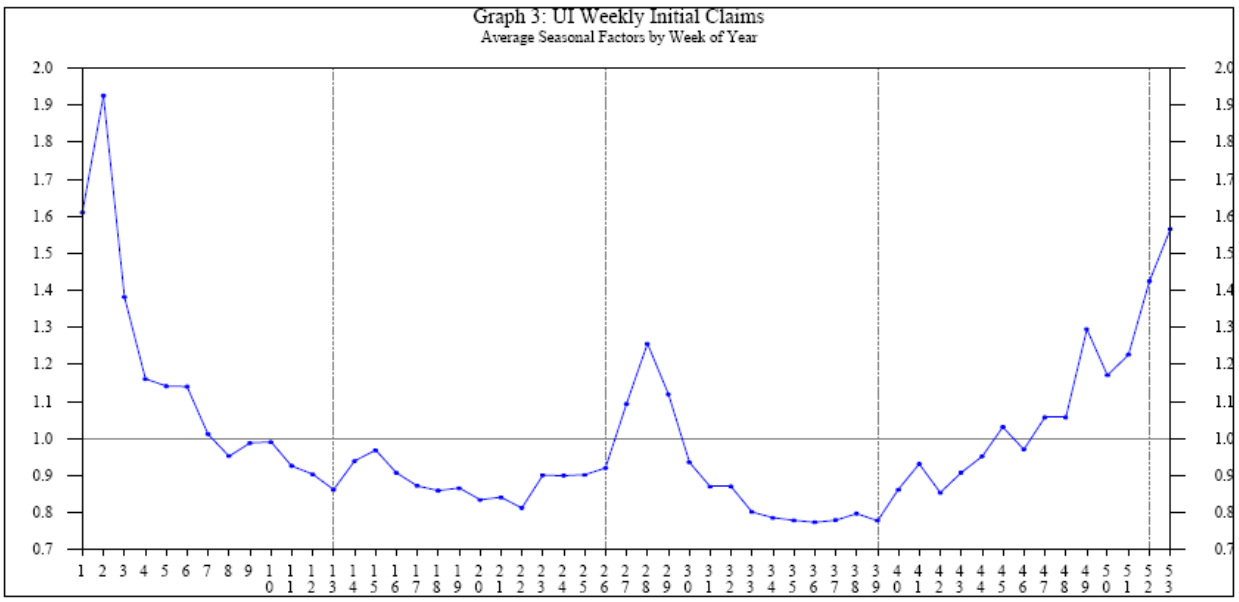
Graph 1 displays the unadjusted and seasonally adjusted series with additive outliers marked. To help evaluate the effects of outliers for Hurricane Katrina in September and October of 2005, Graph 1a plots the seasonally adjusted IC series with outliers for Katrina against ignoring the effects altogether. The AOs not only help smooth the seasonally adjusted series during the affected period, but also in September 2004 as well.

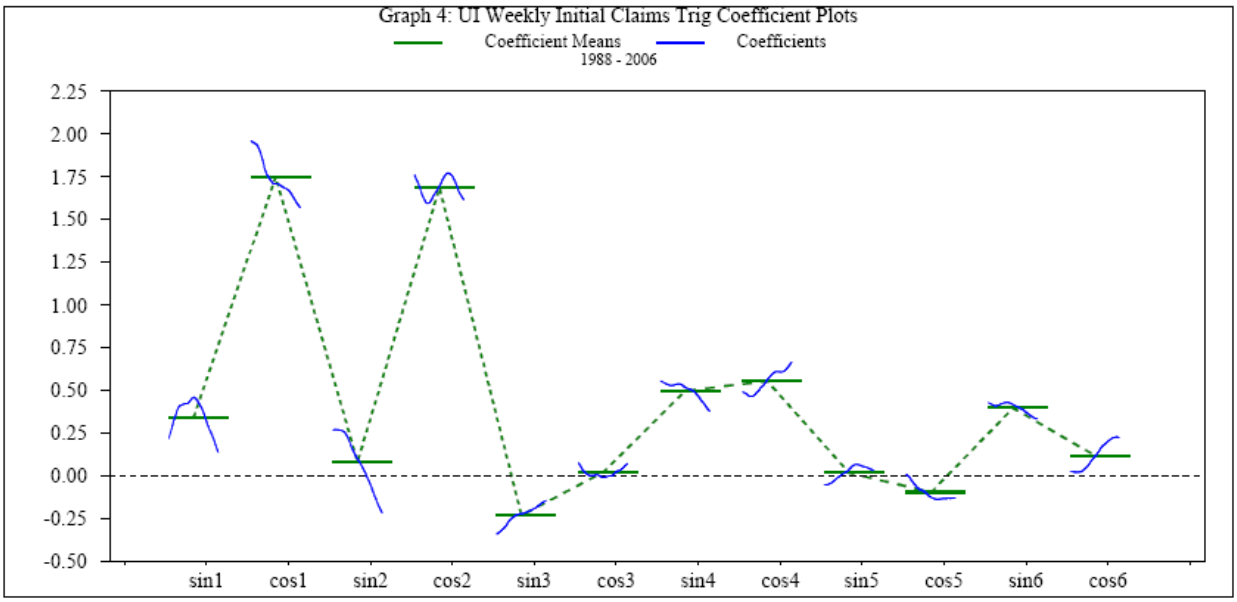
As projected factors are used to create the seasonally adjusted series in the current year, Graph 2 shows the effect of revisions. The moving approach also helped here by typically providing lower revisions. Graph 3 shows the seasonal pattern throughout a typical year; and Graph 4 shows the subplots by sine/cosine pairs that are helpful to determine the number of frequencies needed for the seasonal component.



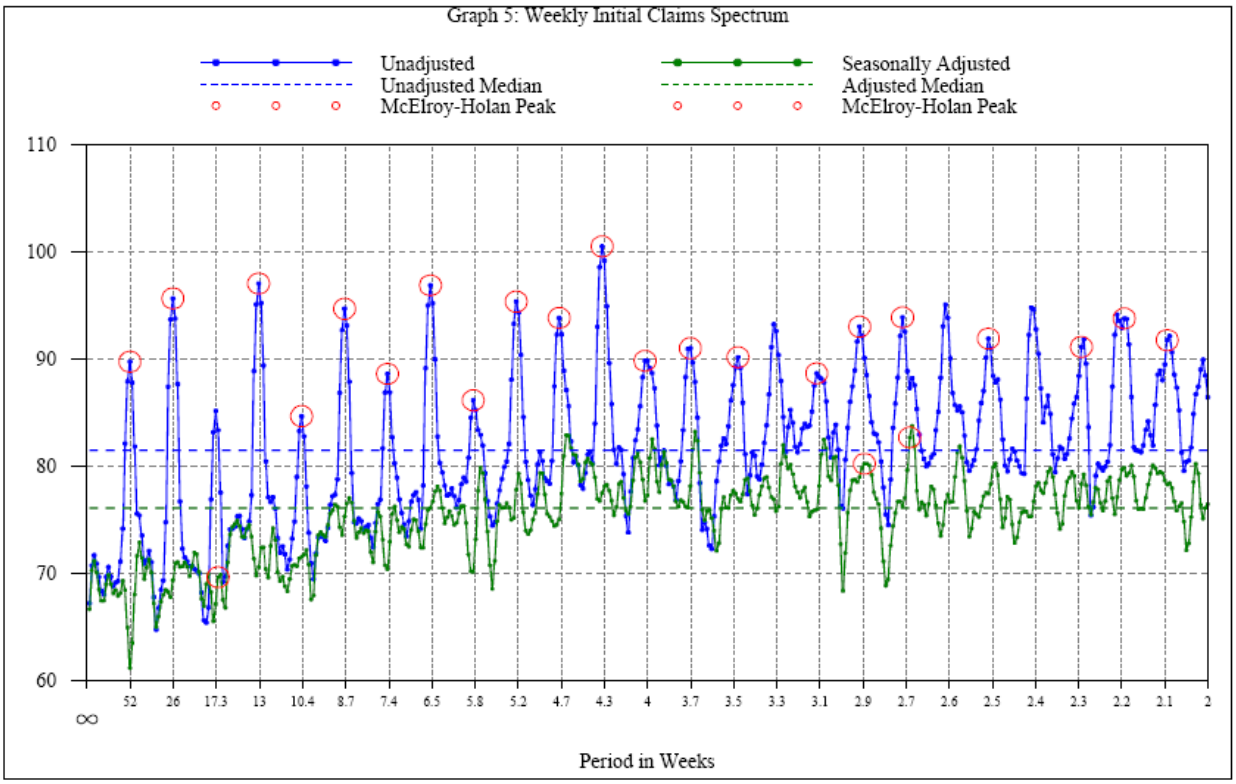



Graph 2: UI Weekly Initial Claims
Unadjusted
Preliminary
Revised
700,000
600,000
500,000
400,000
300,000
200,000
JAN05
FEB05
MAR05
APR05
MAY05
JUN05
JUL05
AUG05
SEP05
OCT05
NOV05
DEC05
JAN06


Graph 3: UI Weekly Initial Claims
Average Seasonal Factors by Week of Year
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7

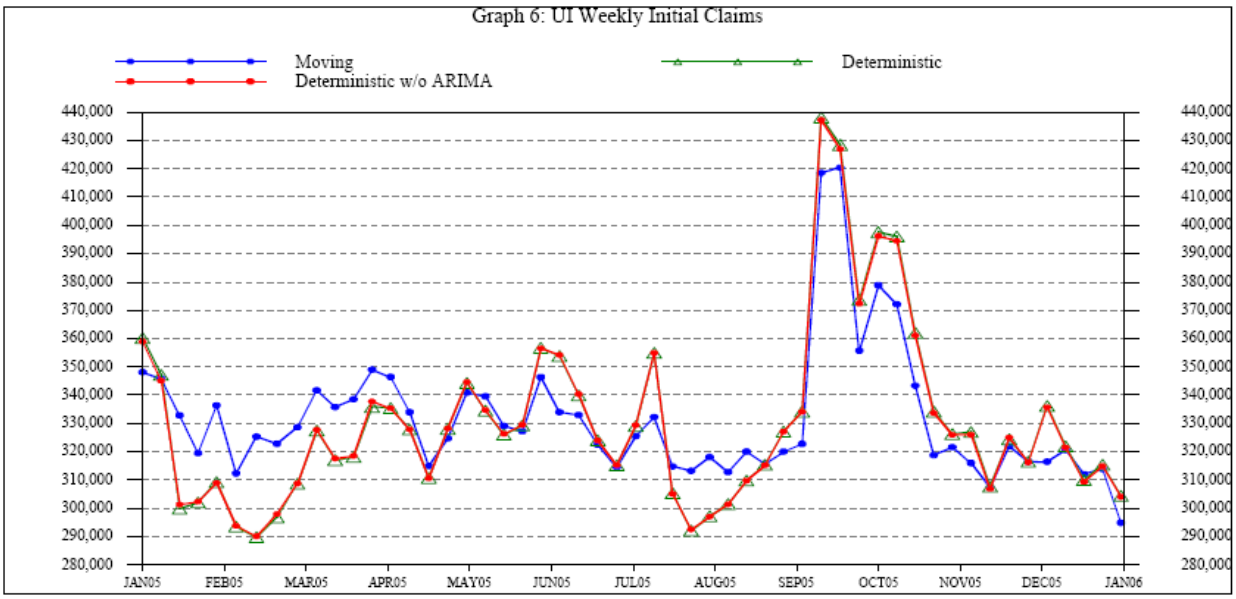


An attempt to show the spectral estimates for the unadjusted and adjusted series is in Graph 5, as an absence of residual seasonality in the adjusted series is important. The circles on the plot indicate peaks as determined by the nonparametric test using the quartic kernel in McElroy and Holan (2005). Wherever there is a red circle in the plot, this indicates that the McElroy-Holan test statistic is significant at the 5% level. Twenty-two of the 25 testable seasonal frequencies (only one side of the Nyquist frequency is observable) for the unadjusted series are determined to have peaks, and the other three are significant at the 10% level. Thus, the series exhibits strong seasonality. Four of the seasonal frequencies for the seasonally adjusted series also have peaks at the 5% level. There is some evidence of residual seasonality in the autocorrelations of the global residuals as well, although correlations at seasonal lags are no more than barely significant. While the presence of residual seasonality is a concern, the program actually seems to remove almost all of the identifiable seasonality.



5. Future Work

There are still several areas where improvements can be made to RegMove. An obvious one would be to add an ARIMA model for the residuals. However, as seen in Graph 6, the IMA error term actually does little to change the final seasonally adjusted estimates produced from the CATS-D program. The Ljung-Box statistics are reduced with the IMA process, but it seems to have more of an effect on the coefficients than the seasonally adjusted estimates. Improvements to the documentation are needed. More diagnostics will be added and other transformation choices are being considered. As there was little documentation for CATS-M originally, much work has been done to expand this, and the effort is continuing.



Finally, there is also the possibility of recoding parts of the program. For example, some options in the control card are not working which can obviously cause some confusion. A final and bold thought could be to rewrite the code completely in FORTRAN and simplify it overall or to rewrite the code in another programming language such as SAS-IML. All of these options are still under evaluation.

References

Bureau of the Census (2002*), X-12-ARIMA Reference Manual* (Version 0.2.10), Washington, DC: Author.

Cleveland, W.P. (1993), "A Time-Varying Least Squares Approach to Seasonal Adjustment of Weekly Time Series," unpublished paper.

Cleveland, W.P., and Scott, S. (2004), "Seasonal Adjustment of Weekly Time Series with Application to Unemployment Insurance Claims and Steel Production," in *American Statistical Association Proceedings of the Business and Economics Section*, pp. 1108-1115

Gomez, V., and Maravall, A. (1998), "Automatic Modeling Methods for Univariate Series," Working Paper 9808, Bank of Spain.

Harvey, A., Koopman, S.J., and Riani, M. (1997), "The Modeling and Seasonal Adjustment of Weekly Observations," *Journal of Business & Economics Statistics*, 15, 354-368.

McElroy, T., and Holan, S. (2005), "A Nonparametric Test for Assessing Spectral Peaks," SRD Research Report: Statistics #2005-10, Bureau of the Census.

Pierce, D.A., Grupe, M.R., and Cleveland, W.P. (1984), "Seasonal Adjustment of Weekly Monetary Aggregates: A Model-Based Approach," *Journal of Business & Economics Statistics*, 2, 260-270.

Table 1: Global Fixed Regression Diagnostics

```
RegMOVE  ver 2.01

Execution at 13:05, 18-APR-2006
series: iclaims
series begins: 1988 week: 05
series ends:   2005 week: 05
output file: anova.dat


COMPONENT        DoF        SS          MSS          F          p-value
  Holiday         13       5.7408       0.4416     208.6801       0.0000
  Outliers        14       0.6662       0.0476      22.4855       0.0000
  Seasonal        60       6.5367       0.1089      51.4827       0.0000
  Linear trend     2       0.0001       0.0001       0.0278       0.9726
Model             89      13.3453       0.1499      70.8585       0.0000
Error            798       1.6887       0.0021
Total            887      15.0339       0.0169

R-Square= 88.77%
Box-Ljung statistic (approx. chi-square)
   Lag     Q               p-value
    52     261.8703         0.0000
   104     398.6564         0.0000

Outlier estimates
WK  YEAR         FACTOR     STD. ERR.        T            p-value
40  1989         1.1819        0.0341        4.9071        0.0000
30  1992         1.3985        0.0340        9.8651        0.0000
30  1993         1.3369        0.0341        8.5235        0.0000
52  1993         0.8780        0.0345       -3.7721        0.0001
 5  1994         1.1249        0.0340        3.4606        0.0003
 3  1996         1.1737        0.0351        4.5618        0.0000
38  2001         1.1314        0.0442        2.7954        0.0027
39  2001         1.2863        0.0568        4.4354        0.0000
40  2001         1.1745        0.0623        2.5833        0.0050
41  2001         1.1738        0.0626        2.5594        0.0053
42  2001         1.1302        0.0576        2.1269        0.0169
43  2001         1.1136        0.0442        2.4328        0.0076
47  2001         1.1716        0.0393        4.0335        0.0000
48  2001         1.1736        0.0406        3.9422        0.0000

Holiday estimates
HOLIDAY              FACTOR     STD. ERR.        T            p-value
User                 1.1065        0.0251        4.0368        0.0000
User                 1.0479        0.0220        2.1301        0.0167
User                 0.9185        0.0257       -3.3056        0.0005
New Years            1.0850        0.0103        7.8931        0.0000
MLK Day              0.8286        0.0169      -11.0996        0.0000
Presidential         0.9330        0.0166       -4.1875        0.0000
Easter               0.9539        0.0082       -5.7615        0.0000
Memorial Day         0.8947        0.0163       -6.8484        0.0000
4th of July          0.9544        0.0164       -2.8500        0.0022
Labor Day            0.8912        0.0166       -6.9222        0.0000
Columbus Day         0.9532        0.0178       -2.6927        0.0036
Veterans Day         0.8762        0.0166       -7.9629        0.0000
Thanksgiving         0.7996        0.0178      -12.5724        0.0000

Seasonal  estimates
TERM        FACTOR     STD. ERR.        T            p-value
  1         0.3486        0.1822        1.9136        0.0280
  2         1.7341        0.1823        9.5126        0.0000
etc.
```

Table 2: Global Fixed Regression Output

```
RegMOVE  ver 2.01
Execution at 11:47, 18-APR-2006
series: iclaims
series begins: 1988 week: 05
series ends:   2005 week: 05
output file: global.dat


        OBSERVATION    ESTIMATE        RESIDUAL      STD. ERR.          T         p-value
    [EFFECT         FACTOR     STD. ERR.        T             p-value]
   1       395000.       395001.       -0.0645        0.0128       -5.0332        0.0000
   2       381000.       406394.       -0.0025        0.0129       -0.1973        0.4218
   3       335000.       358236.        0.0333        0.0129        2.5723        0.0051
   4       316000.       326866.       -0.0246        0.0180       -1.3646        0.0864
    [Holiday         0.9330        0.0166       -4.1875        0.0000]

   5       324000.       343474.       -0.0411        0.0153       -2.6853        0.0037
   6       312000.       344613.       -0.0063        0.0124       -0.5080        0.3058
   7       294000.       326789.       -0.0145        0.0124       -1.1710        0.1210
   8       276000.       311262.       -0.0028        0.0125       -0.2200        0.4129
   9       269000.       304204.        0.0002        0.0145        0.0114        0.4952
  10       257000.       290585.        0.0255        0.0147        1.7313        0.0419
    [Holiday         0.9539        0.0082       -5.7615        0.0000]
.
.
.
 233       507900.       472963.        0.0033        0.0148        0.2216        0.4123
 234       453900.       421288.        0.0535        0.0335        1.5954        0.0555
 235       556300.       489450.        0.0535        0.0335        1.5954        0.0555
    [Outlier         1.3985        0.0340        9.8651        0.0000]
etc.
```

Table 3: Sine and Cosine Coefficients from Locally Weighted Regressions

```
RegMOVE  ver 2.01
Execution at 11:47, 18-APR-2006
series: iclaims
series begins: 1988 week: 05
series ends:   2005 week: 05
output file: coefs.dat

 1  1988    0.3010
 1  1989    0.3283
 1  1990    0.3502
 1  1991    0.3732
 1  1992    0.3843
 1  1993    0.4029
 1  1994    0.4119
 1  1995    0.4196
 1  1996    0.4175
 1  1997    0.4036
 1  1998    0.3851
 1  1999    0.3604
 1  2000    0.3257
 1  2001    0.2905
 1  2002    0.2703
 1  2003    0.2615
 1  2004    0.2465
 2  1988    1.9176
 2  1989    1.9252
 2  1990    1.9308
etc.
```

Table 4: General Program Output

```
RegMOVE  ver 2.01
Execution at 11:47, 18-APR-2006
series: iclaims
series begins: 1988 week: 05
series ends:   2005 week: 05
output file: results.dat

wk  year          sa   safactor        obs    outlier    holiday
 5  1988     331656.     1.1910    395000.     1.0000     1.0000
 6  1988     315780.     1.2065    381000.     1.0000     1.0000
 7  1988     304999.     1.0984    335000.     1.0000     1.0000
 8  1988     323660.     0.9763    316000.     1.0000     0.9330
 9  1988     313169.     1.0346    324000.     1.0000     1.0000
10  1988     303308.     1.0287    312000.     1.0000     1.0000
11  1988     299175.     0.9827    294000.     1.0000     1.0000
12  1988     304307.     0.9070    276000.     1.0000     1.0000
13  1988     299767.     0.8974    269000.     1.0000     1.0000
14  1988     306346.     0.8389    257000.     1.0000     0.9539
etc.
```