Modeling and Seasonal Adjustment of Daily Retail Series

Tucker McElroy (U.S. Census Bureau)
Brian Monsell (U.S. Census Bureau)
Rebecca Hutchinson (U.S. Census Bureau)
Daniel Fernandez (Palantir)

ESMD Seasonal Workshop
November 4, 2016

Disclaimer

This presentation is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Outline

- Daily retail time series
- Goals of analysis
- Structural models with fixed effects
- Computation and algorithms

The largest credit card payment processor in the USA. All of the credit, debit, prepaid, and EBT (Electronic Benefit Transfer) transactions for each merchant that utilizes the FirstData service are recorded, with information on authorizations, settlements, and an exact time stamp. These items, from over 600 merchant categories, are then aggregated into the NAICS codes at a daily frequency, with adjustments for the local time zone of the merchant. The result is a retail database, covering all types of cards, all banks, all networks, and all fifty states, as well as all customer segments and all sizes of merchants.

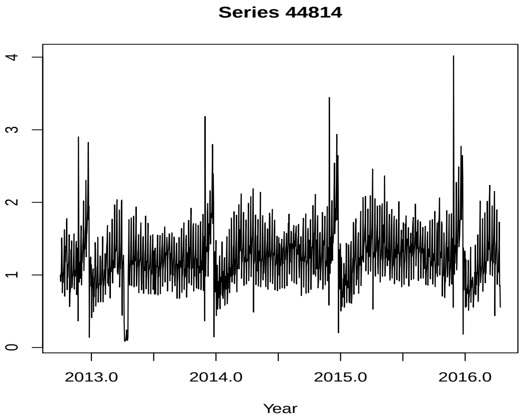

FIGURE: Time series plot of daily retail data 44814 (Family Clothing Stores), covering the period from October 1, 2012 through April 12, 2016.

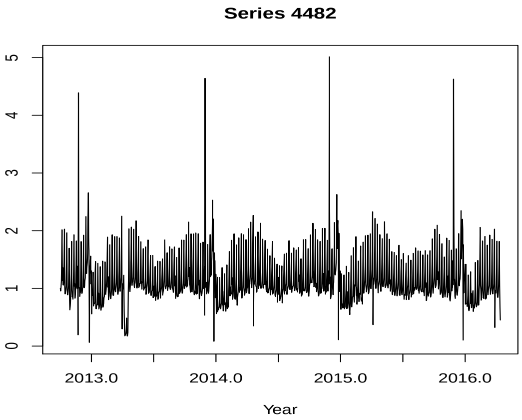

FIGURE: Time series plot of daily retail data 4482 (Shoe Stores), covering the period from October 1, 2012 through April 12, 2016.

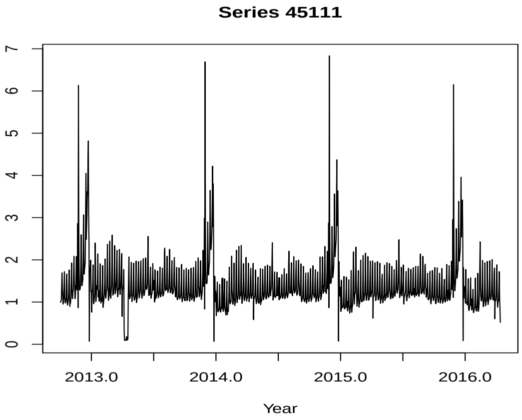

FIGURE: Time series plot of daily retail data 45111 (Sporting Goods Stores), covering the period from October 1, 2012 through April 12, 2016.

Label	Title
44311	Appliance, Television, and Other Electronics
44411	Home Centers
44511	Supermarkets and Other Grocery Stores
44814	Family Clothing Stores
4482	Shoe Stores
45111	Sporting Goods Stores
45211	Department Stores
45291	Warehouse Clubs and Superstores
45411	Electronic Shopping and Mail Order

TABLE: FirstData daily retail series studied. (All series are normalized for disclosure avoidance, such that first value equals one.)

Goals

1. Model trend, annual seasonality, and weekly seasonality
2. Extract these components (multivariately)
3. Estimate moving holidays: Easter, Labor Day (LD), Black Friday (BF), Cyber Monday (CM), Super Bowl Sunday (SBS)

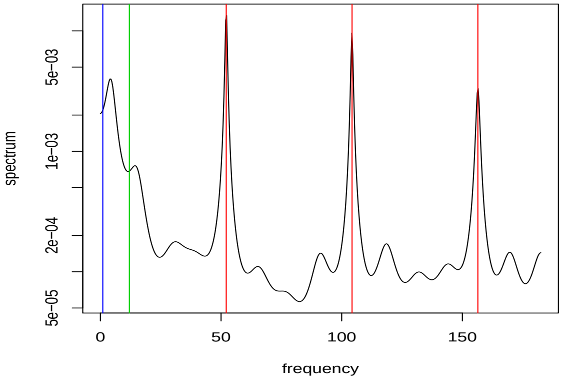

FIGURE: AR spectral density estimate of daily retail data (Series 44814). Vertical red lines correspond to once a week, twice a week, and thrice a week phenomena; the blue line corresponds to annual phenomena, and the green line corresponds to monthly phenomena.

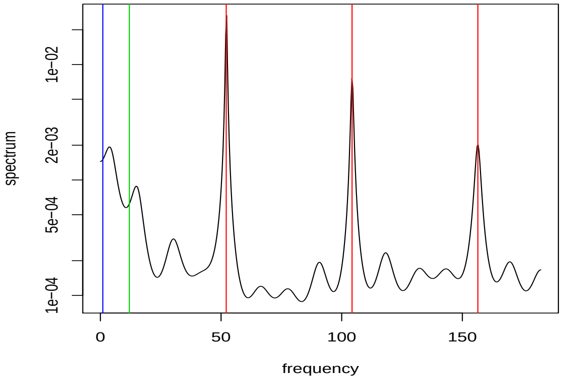

FIGURE: AR spectral density estimate of daily retail data (Series 4482). Vertical red lines correspond to once a week, twice a week, and thrice a week phenomena; the blue line corresponds to annual phenomena, and the green line corresponds to monthly phenomena.

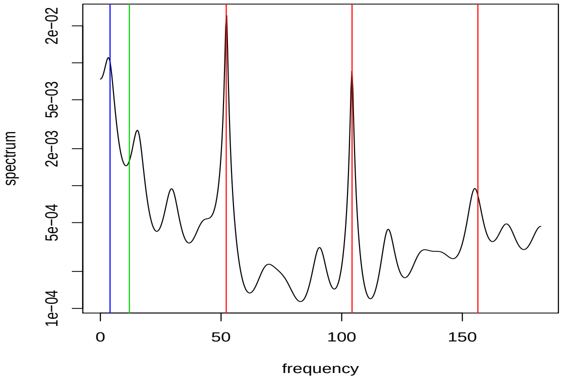

FIGURE: AR spectral density estimate of daily retail data (Series 45111). Vertical red lines correspond to once a week, twice a week, and thrice a week phenomena; the blue line corresponds to annual phenomena, and the green line corresponds to monthly phenomena.

Stochastic unobserved components structure for multivariate $\{y_t\}$, with trend $\{\mu_t\}$, annual seasonal $\{\xi_t^a\}$, weekly seasonal $\{\xi_t^w\}$ (split into three atomic seasonals), irregular $\{\iota_t\}$, and holiday effects $\{z_t\}$:

$$y_t = \mu_t + \xi_t^a + \xi_t^w + \iota_t + z_t$$
$$\xi_t^w = \xi_t^{(1)} + \xi_t^{(2)} + \xi_t^{(3)}.$$

Set $\delta^{\omega}(B) = 1 - 2\cos(\omega)B + B^2$. Each component is difference stationary, with prefix ∂ denoting a differenced component:

$$
\begin{aligned}
\partial \mu_t = (1 - B)\mu_t &\sim \mathrm{WN}(0, \Sigma_\mu) \\
\partial \xi_t^a = \delta^{2\pi/365}(B)\xi_t^a &\sim \mathrm{WN}(0, \Sigma_a) \\
\partial \xi_t^{(1)} = \delta^{2\pi/7}(B)\xi_t^{(1)} &\sim \mathrm{WN}(0, \Sigma_1) \\
\partial \xi_t^{(2)} = \delta^{4\pi/7}(B)\xi_t^{(2)} &\sim \mathrm{WN}(0, \Sigma_2) \\
\partial \xi_t^{(3)} = \delta^{6\pi/7}(B)\xi_t^{(3)} &\sim \mathrm{WN}(0, \Sigma_3) \\
\iota_t &\sim \mathrm{WN}(0, \Sigma_\iota).
\end{aligned}
$$

Regressors are constructed from calendar dates for day index t_* via a window

$$1_{[t_* - b, t_* + f]}$$

for b days back and f days forward. (Minus long-term average.)

Example 1: Easter-day $b = f = 0$.

Example 2: pre-Easter $b = 8$, $f = -1$.

Signals

1. Trend and annual seasonality are hard to distinguish (unit roots are approximately equal); this is not merely an issue of sample size. Can combine into single component, as a trend-seasonal
2. Weekly seasonality is dynamic, and corresponds to “trading day” in monthly series
3. Signals are co-integrated when Σ is reduced rank

Computation

1. Model fitting via Durbin-Levinson algorithm and GLS for Gaussian likelihood (univariate analysis only)
2. Model fitting via Method-of-Moments (MOM) based on spectral density equations (multivariate analysis)
3. Signal extraction via matrix (univariate case) or Wiener-Kolmogorov (WK) filtering of fore-cast extended series (multivariate case)
4. Algorithms encoded in *sigex*, custom R code for multivariate time series forecasting and signal extraction

Big Data Challenges

1. 10,000 data points; matrix methods infeasible. Windowing signal extraction is slow (weeks), but WK filtering is fast (minutes)
2. Hundreds of parameters; MLE is infeasible. A single evaluation of likelihood is slow (minutes), but MOM is fast (1-2 seconds) and stable (no matrix inversions)
3. OLS estimation of fixed effects is inaccurate, and GLS takes too long for multivariate series. We fit univariate models via MLE to get fixed effects, linearize (subtract the predicted values), and multivariately model the de-holidayed series

Label	Holiday Effects
44311	pre-Easter, Easter-day, LD, BF, CM
44411	Easter-day, LD, CM
44511	pre-Easter, LD, BF
44814	Easter-day, pre-Easter, LD, BF
4482	Easter-day, pre-Easter, LD, BF
45111	SBS, Easter-day, BF, CM
45211	Easter-day, BF
45291	Easter-day, LD, CM
45411	pre-Easter, LD, BF, CM

Table: FirstData daily retail series studied, with identified holiday effects (blue for positive, red for negative).

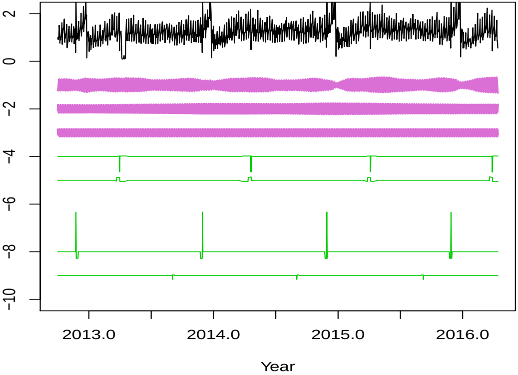

FIGURE: Weekly atomic seasonals (purple) with holiday effects in green, for Series 44814 (Family Clothing Stores). Holidays: Easter-day, pre-Easter, BF, LD.

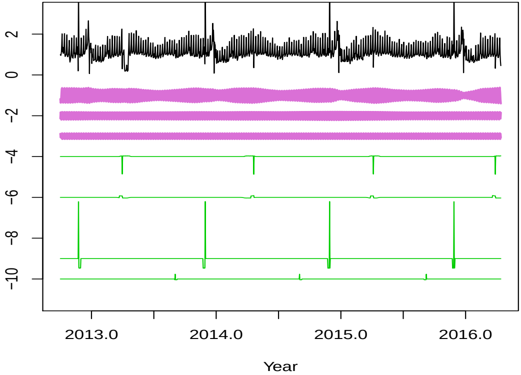

FIGURE: Weekly atomic seasonals (purple) with holiday effects in green, for Series 4482 (Shoe Stores). Holidays: Easter-day, pre-Easter, BF, LD.

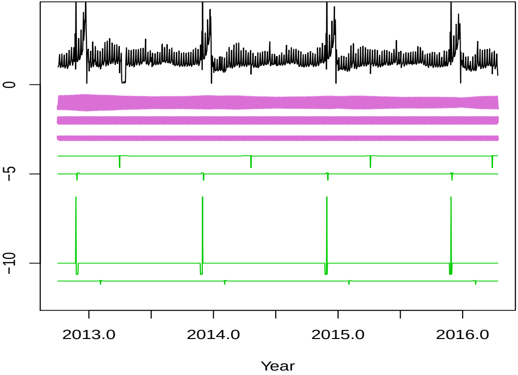

FIGURE: Weekly atomic seasonals (purple) with holiday effects in green, for Series 45111 (Sporting Goods Stores). Holidays: Easter-day, CM, BF, SBS.

Multivariate Results

1. Co-integration Ranks: trend-seasonal (9), first weekly seasonal (7), second weekly seasonal (6), third weekly seasonal (4), irregular (9)
2. No signal extraction uncertainty yet. (Matrix method provides the MSE, but WK filtering requires recursive calculation of multi-step ahead forecast error covariances – not yet encoded.)
3. Trend-seasonal could be further separated by *ad hoc* filter
4. Results presented with aggregate weekly seasonal (the sum of the three atomic weekly components)

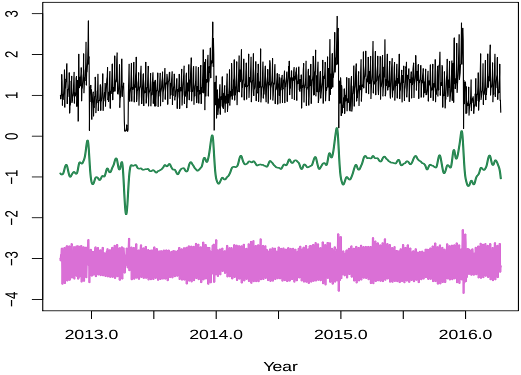

FIGURE: Trend-annual signal (olive) with aggregate weekly seasonal (purple), for Series 44814 (Family Clothing Stores).

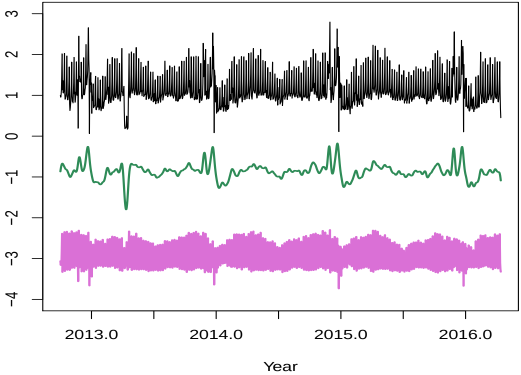

FIGURE: Trend-annual signal (olive) with aggregate weekly seasonal (purple), for Series 4482 (Shoe Stores).

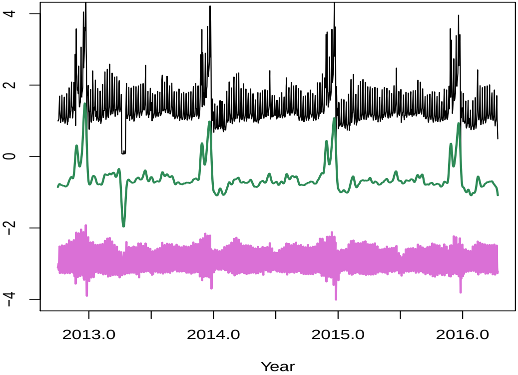

FIGURE: Trend-annual signal (olive) with aggregate weekly seasonal (purple), , for Series 45111 (Sproting Goods Stores).

Discussion

- Why multivariate seasonal adjustment? Hope to reduce uncertainty of estimation by utilizing closely related series, but computational burden increases greatly
- Nuanced holiday effects teased out via daily data (as opposed to monthly data)
- Fixed (or time-varying) trading day is replaced by stochastic weekly effect, which is more adaptive to data features
- With longer series and improved methodology, trend and annual seasonality may be separable

Current and Future Work

- Application to constant-period reporting problem: weekly data that needs to be converted to monthly frequency for publishing. We're currently investigating the use of daily time series to assist
- Extreme-value adjustment and missing values: outliers can be handled as a fixed effect, or as a stochastic anomaly. Missing values are a fore-casting problem
- MSE for WK Method: the MSE for a bi-infinite sample is known (and encoded), but for finite samples we need multi-step ahead forecast error covariances

Contact: tucker.s.mcelroy@census.gov