MAT 8444 Report - Time Series Analysis of Electricity Demand and Related Data

Duane Stanton

May 10, 2019

Upfront Details

Motivation and Data

In order to practice a variety of time series analyses on 'real-world' data, hourly electricity demand from the Energy Information Administration[1] (EIA) was collected from the administration's site. Specific variables included per-hour **actual demand** (the variable of primary interest), per-hour net generation (with any differential coming from imported electricity), and per-hour forecast demand 24 hours in the future. Data was aggregated to the daily level to align with temperature data (discussed further down). Demand was analyzed for the New York ISO and ISO New England (Independent System Operators - organizations tasked with monitoring/coordinating/controlling the electrical grid for a defined region) between 5am, July 1, 2015 and 5am, February 18, 2019 - observations from 2019 were excluded from the modeling process, originally planned due to limits in the economic index data, but ultimately it was held out for evaluating short-term future model performance.

The primary analysis focused on NYISO data, with a very similar modeling process applied to ISONE data for comparison.

The secondary data source of most interest was summary data for daily mean temperature in degrees Fahrenheit across New York counties from the National Oceanic and Atmospheric Administration[2] (NOAA), for the same time period as the EIA data - this data was then population-weighted within years based on data from the U.S. Census Bureau[3] to calculate state-level temperature estimates (since I considered population density an important factor when trying to link location temperatures and electricty consumption, e.g. air conditioner use in the summer).

While I initially planned to use similar data for all counties in the New England, limitations in the NOAA data prevented this. Specifically, county-specific temperature data was provided as measurements from specific weather stations located in a given county, and while each station had a unique identifier, there was no readily accessible key on the NOAA site or elsewhere indicating in which county a given station resided - this would mean either manually looking up the location of each station (as I did for New York data), or iteratively downloading data for a single county at a time, manually appending county labels for each data subset.

A final secondary data source considered was a composite index of monthly state-level economic conditions from the Federal Reserve Bank of Philadelphia[4], with 100 aligning with the annual mean of 2007 (considered the peak of the previous economic cycle). This 'coincident index' factored in "nonfarm payroll employment, average hours worked in manufacturing by production workers, the unemployment rate, and wage and salary

---

[1] "Bulk download facility." *U.S. Energy Information Administration.* U.S. Electrical System Operating Data. https://www.eia.gov/opendata/bulkfiles.php. Accessed 25 February 2019.

[2] "Climate Data Online Search." *NOAA - National Centers for Environmental Information.* https://www.ncdc.noaa.gov/cdo-web/search. Accessed 24 March 2019.

[3] "PEPAGESEX - Annual Estimates of the Resident Population for Selected Age Groups by Sex for the United States, States, Counties, and Puerto Rico Commonwealth and Municipios." *U.S. Census Bureau American FactFinder.* Accessed 16 March 2019 for years 2015, 2016, 2017 estimates.

[4] "State Coincident Indices." *Federal Reserve Bank of Philadelphia.* https://www.phil.frb.org/research-and-data/regional-economy/indexes/coincident/. Accessed 17 February 2019 ; data accessed from the "Revised Data" linked .xls file at upper-right on the page.

disbursements deflated by the consumer price index (U.S. city average)"; these factors and "an underlying (latent) factor" are input to a system of five equations (one for each factor mentioned) and ultimately generate the index value. There are papers linked in the cited source, but it is not immediately obvious how exactly the equations/indices are structured and weighted.

Summary of Data Processing - missing data imputation approaches

Electricity data was aggregated (summed) to the daily level for most of the analyses, as the related temperature data was available at the daily level as well. Some secondary analyses futher aggregated electricity and temperature data to the monthly level, as the economic index data was avalable at the monthly level.

There were some cases of missing (`NA`) electricity data (117 hourly observations for ISONE demand, 127 for ISONE net generation, 195 for NYISO demand and 205 each for net generation and day-ahead demand forecast (plus 48 single-hour observations 'completely' missing from the dataset, 19 for 01-Jul-2018, 5 for 02-Jul-2018, 18 for 11-Nov-2018, and 6 for 12-Nov-2018); for context, ISONE data had 31,873 rows of hourly data, NYISO had 31,825 rows. For these, the following imputation scheme was applied:

- For actual demand missingness, first model-estimated 'day-ahead demand' from 24 hours prior (from the original dataset) was input; if this was also missing, then the mean of 24 hours prior and 24 hours in the future was input; if one of these but not the other was missing, the non-missing value was input. For cases of *still*-missing data, the 'past and future observations' approach was applied to 48 hours-apart data. Where '2 days out' data was still insufficient, I applied the `na.kalman()` function from the `imputeTS` package to apply Kalman smoothing (with `method="auto.arima"`) *except* for 11 and 12 Nov 2018 still-missing demand data, where this approach yielded a exceptionally low daily-aggregated demand valued - in that case I uses `imputeTS::na.ma()` for moving-average imputation, which aligned much better with the general (nearby) data structure.

- For forecasted demand or net generation missingness, the "24 hours-apart" and (if needed) "48 hours-apart" method described above was used.

- Temperature data was calculated at the county level by taking the mean of all observations for a given date in that county for the daily mean temperature. If this daily mean was not available, then the mean of the daily high and the daily low was used instead. This second method seemed to fall within 1 degree Fahrenheit of the recorded daily mean for cases where the daily mean was available, and seemed acceptable for imputation. For the minority of cases still missing a county mean temperature for a given date (43 of 1,329 days), the dataset-wide mean temperature for that date was used. For missing daily maximum or minimum data, the corresponding statewide mean value was imputed - this followed the mean temperature imputation process.

- Not all counties had weather stations and thus some counties lacked temperature observations. In order to apply population weighting of temperature data, counties with unavailable data were assigned to have the same temperature data as a neighboring county as the same general longitude (e.g. Kings County, NY data was also assigned to Queens County, NY as Kings County (Brooklyn) is adjacent to Queens County (Queens)).

- Census data was not available for 2018, so for county population weighting of temperature data, population estimates from 2017 were re-used. Based on comparison of 2015/16/17 county shifts, this should not introduce much bias in population weights, as New York counties had very stable population proportions (proportion of the total NY population living in a given county) in each of the three years where data was available.

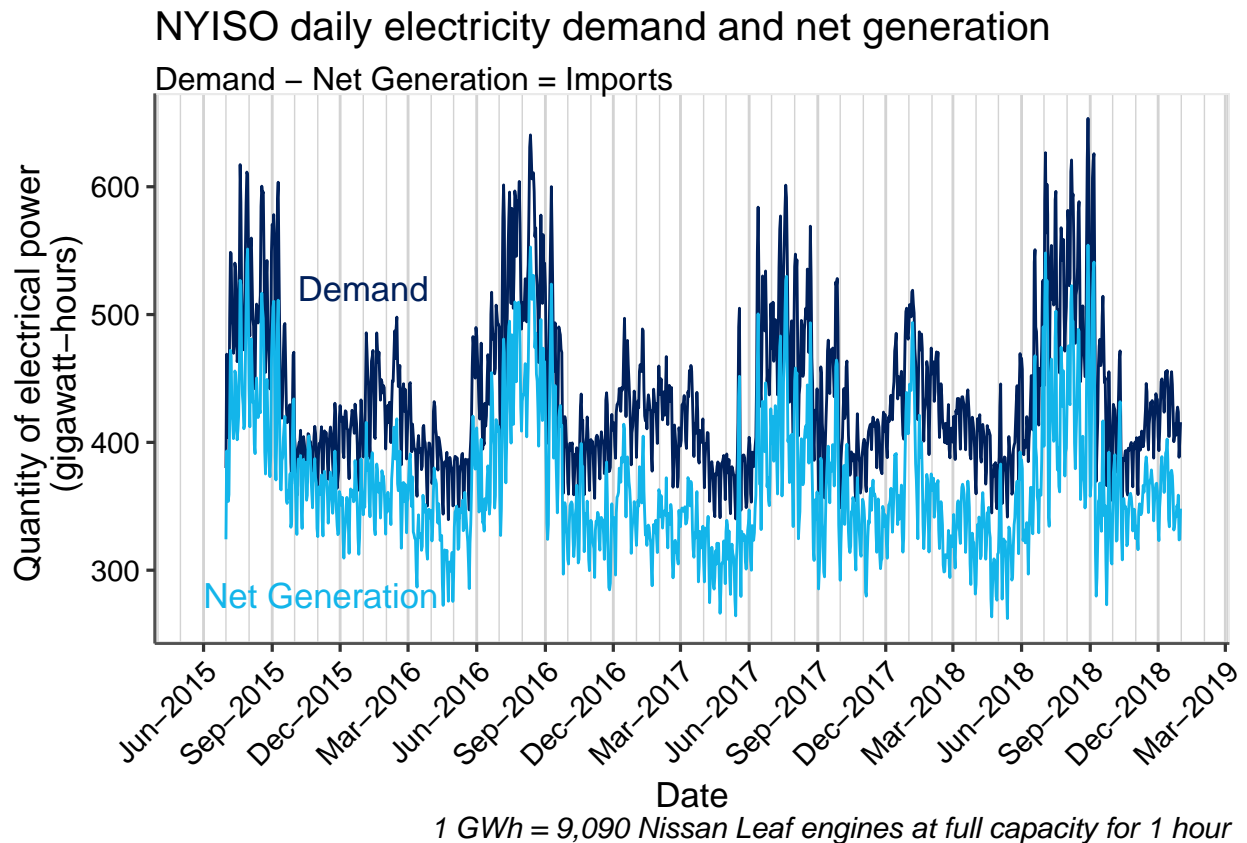## Analyses Conducted and Results

This analysis largely focused on regression with ARMA errors, with spectral analysis indicating good regression coefficients for sine/cosine pairs. While NYISO daily electricity demand was the primary interest and

yielded the better-performing model, a closely related ISONE model yielded fairly good results in its own right.

Detailed Analysis

Exploratory Data Analysis

NYISO data

NYISO daily electricity demand and net generation

Demand – Net Generation = Imports

1 GWh = 9,090 Nissan Leaf engines at full capacity for 1 hour

There seems to be a pretty clear seasonal component, with demand/net generation spiking primarily in the summer (July, August, September) and secondarily in winter (late December to late March). The greatest gap between demand and net generation occurs during this secondary increase. At least, this is the case when considering net generation as a percentage of demand; in absolute terms the discrepancies are fairly stable across months.

**NYISO daily electricity demand and temperature summaries**
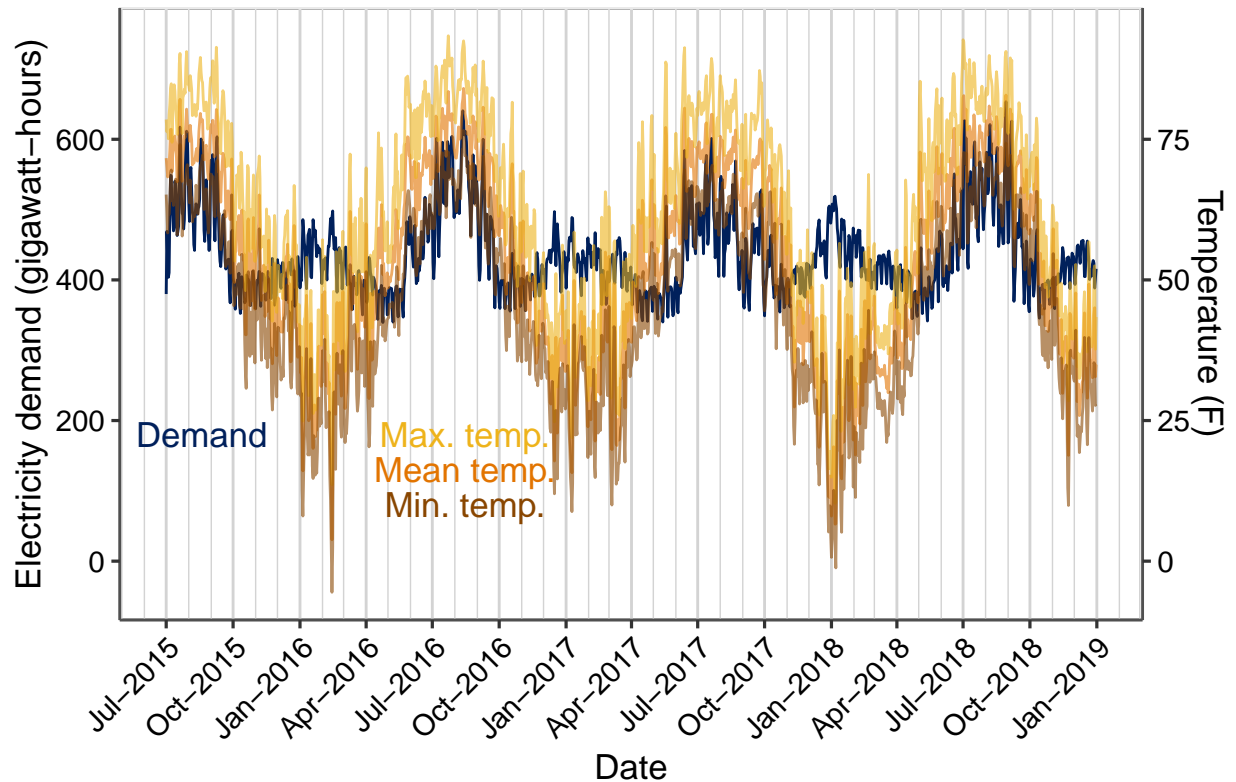
The plot above seems to suggest a variable capturing "absolute average deviation from [roughly] 50F"* may be a more useful variable for predicting daily electricity demand than, say, the daily mean temperature. Therefore, I'll create that variable (using just `tempAvg.F`; logically, there is strong correlation among max./mean/min. temperatures).

*The dataset mean daily-mean temperature is 53.96F, covering 01 Jul 2015 to 31 Dec 2018; I considered using this mean rather than 50F, but the regression model considered further down had almost the same R^2_{adj} using 50F - about 0.923 using the dataset mean and about 0.919 using 50 - and the it's a relatively 'simple' interpretation. It's interesting such a "nice" number (admittedly close to the sample mean) should work so well.

```
##           tempMax.F tempAvg.F tempMin.F
## tempMax.F    1.0000    0.9930    0.9703
## tempAvg.F    0.9930    1.0000    0.9913
## tempMin.F    0.9703    0.9913    1.0000
```

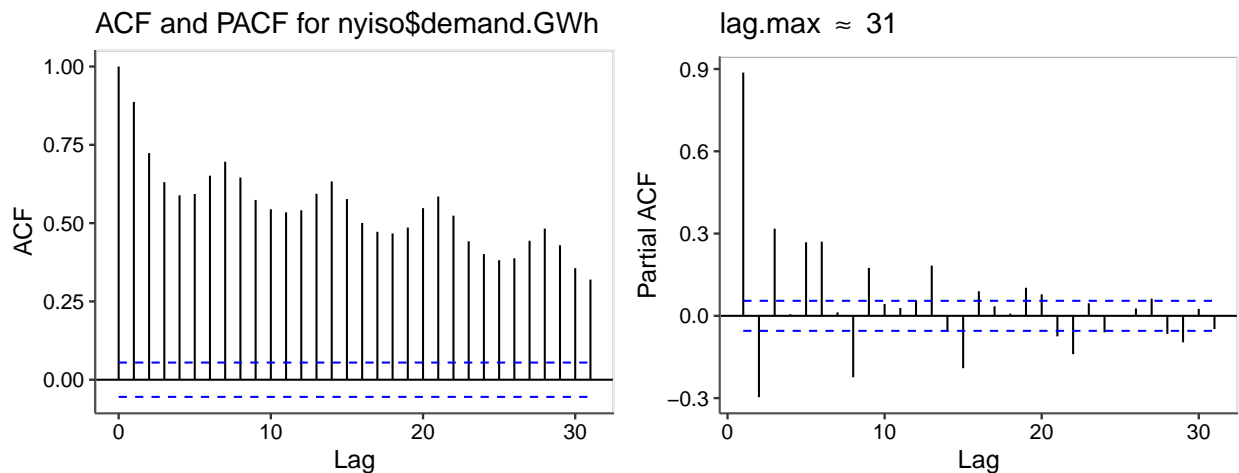NYISO daily electricity demand and |mean temp. dist. from 50F|
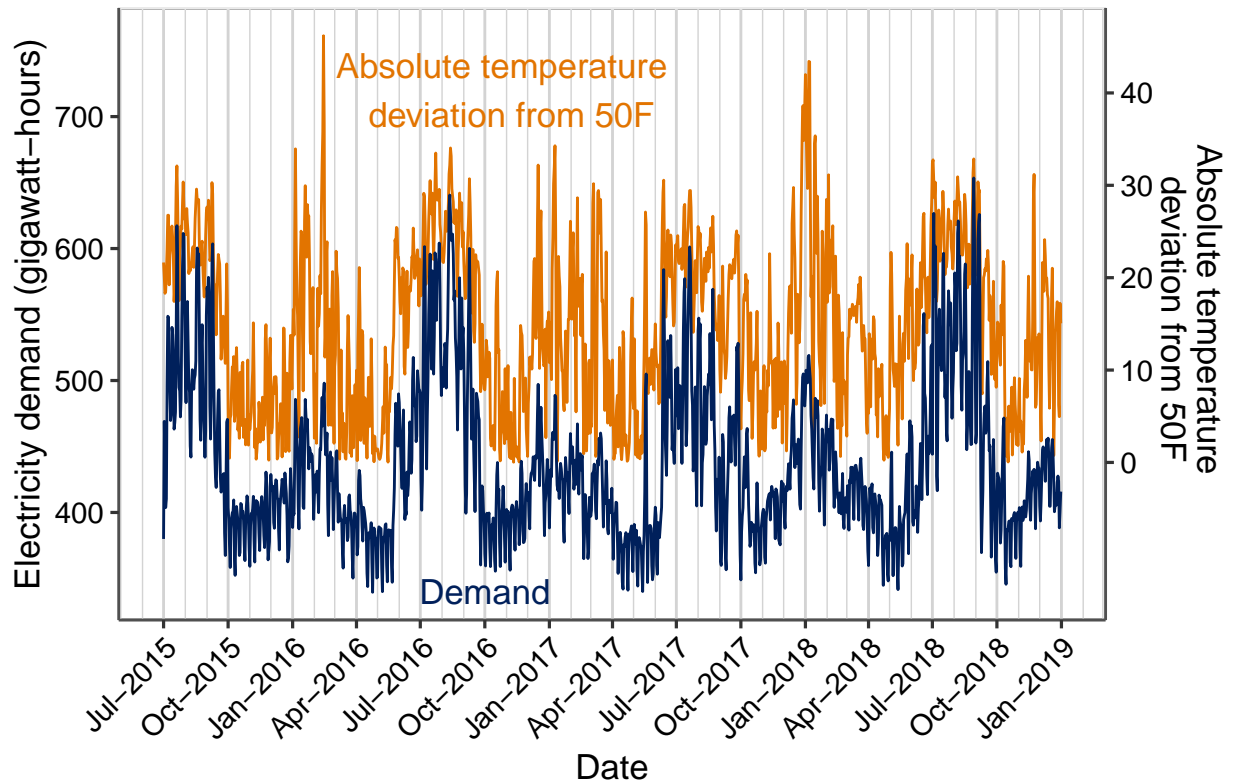
This new variable seems to track relatively well with electricity demand - though I think it would only have good "real world" explanatory power if there really is something connecting "particularly hot or cold" days and electricity usage. An increase in air conditioning activity during hot days seems likely, but for very cold days I'm not sure an uptick in electric heating (either as a primary or supplemental heat source) carries as much plausibility - perhaps there is this in addition to an increase in time spent indoors (and thus increased electrical device use).

This is the first plot so far to make clear there is an unusually low observation in 2018, on November 12 - the previous day was missing quite a few hourly observations, and it may be that there was some issue with the data-recording process in this period. In other words, it seems like the measurement process, rather than the actual value for the metric, was unusual here.

NYISO daily electricity demand and monthly economic activity in
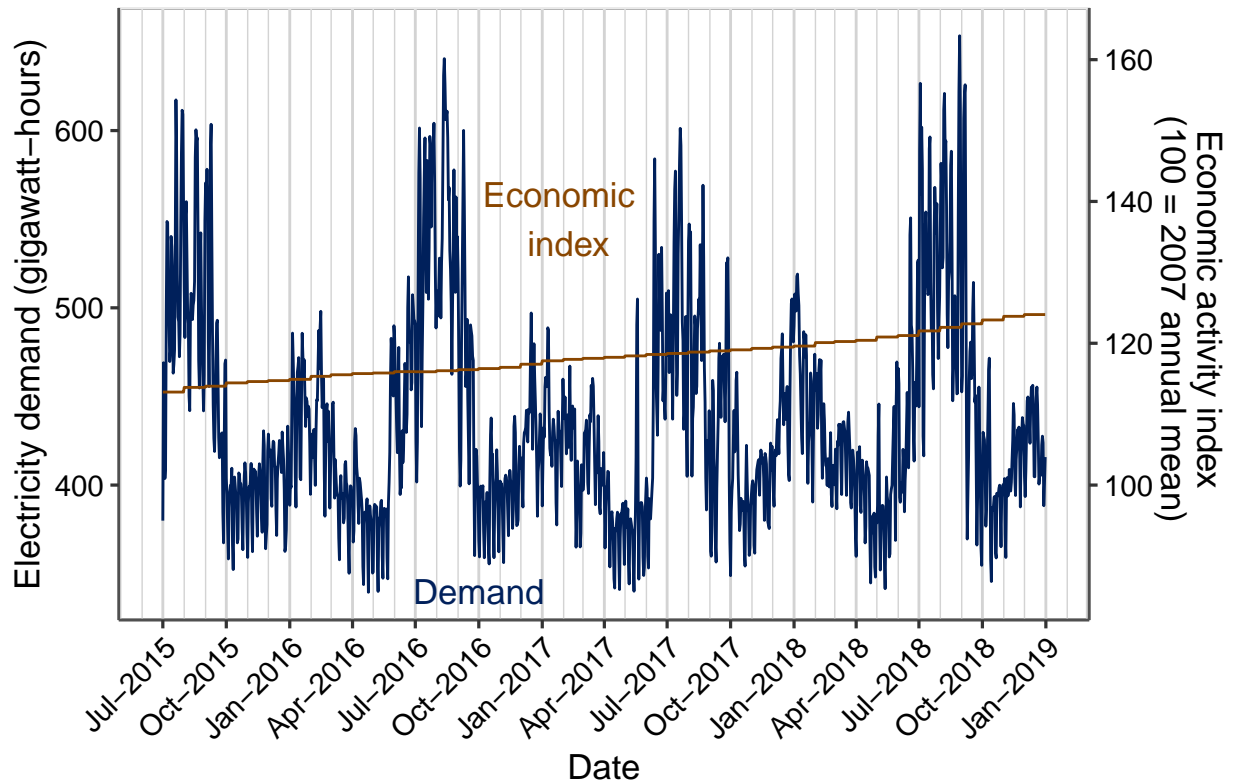
I doubt the monthly economic activity index will provide much prediction utility - it's not granular enough to be of much use for daily data, and it appears to be monotonic - I checked this via `plot(nyiso$econIndex.NY, type="l")`.

ACF and PACF for nyiso$demand.GWh — lag.max=366
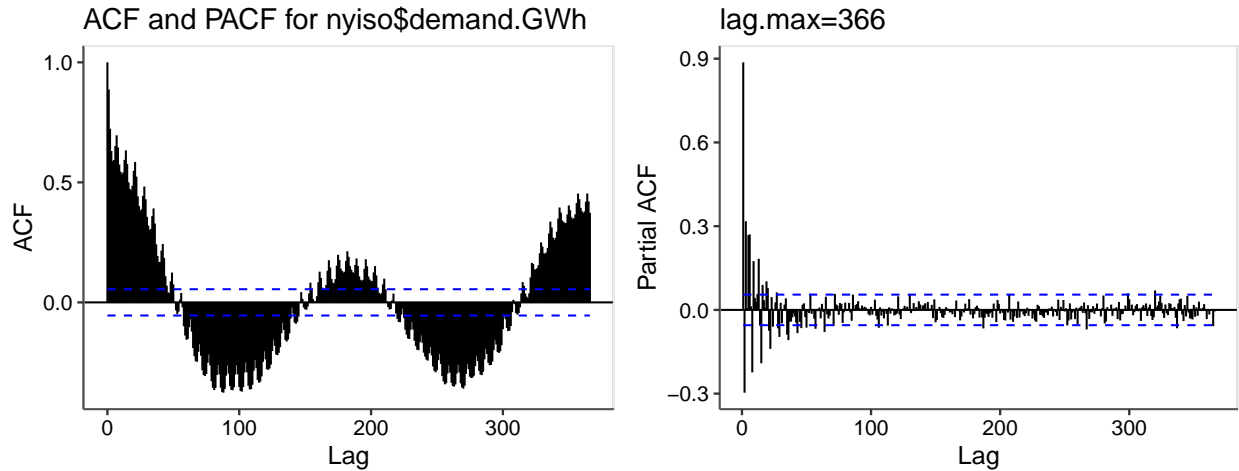
The two sets of ACF and PCF plots clearly indicate the NYISO daily electricity demand data don't reflect a stationary process - this was generally established in the prior plots, but this set of plots also reflect it is not adequately summarized by a "pure" autoregressive or moving average process. We wouldn't expect that given the seasonal pattern in the data.

The ACF 'peaks' in 7-unit increments in the first plot set would seem to indicate a weekly component, and the second plot set might reflect semi-annual and annual cycle peaks given the ACF behavior around (roughly) lags 183 and 365.

NYISO daily electricity demand and aligned 'day ahead' demand
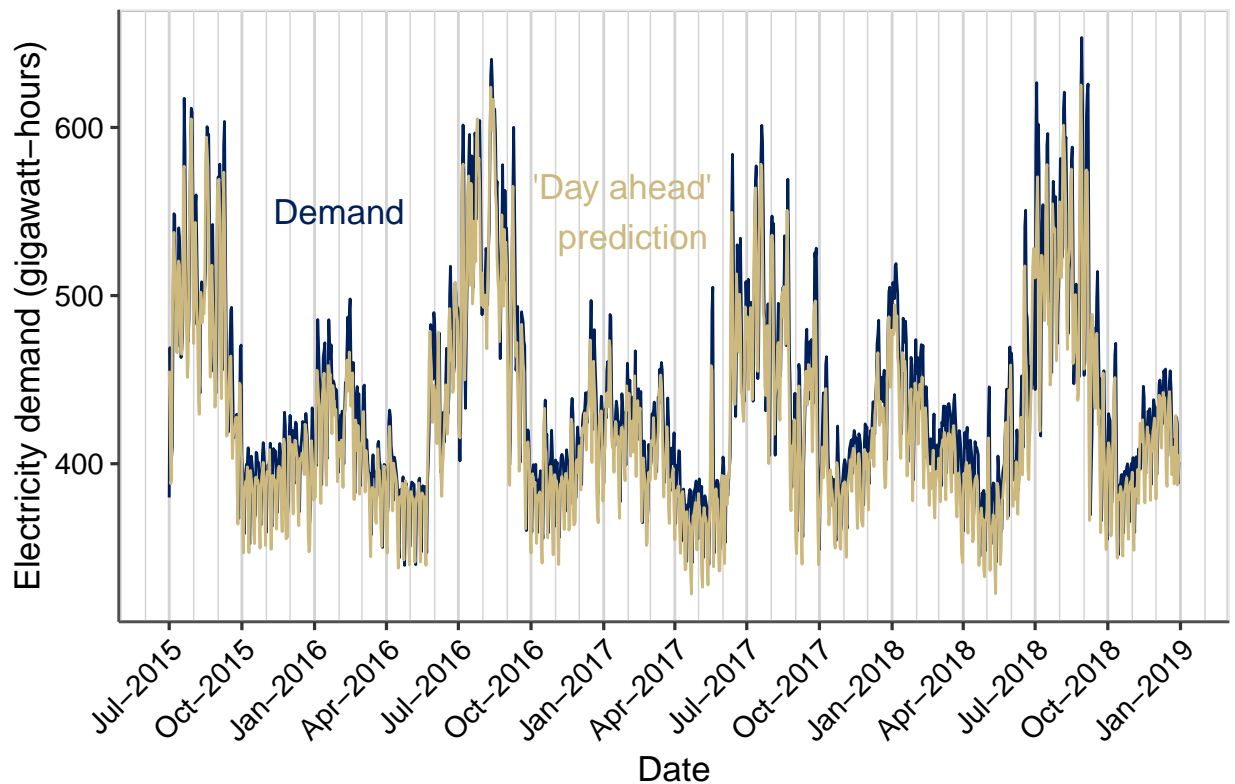
The 'day ahead demand' data from the original dataset aligns pretty well with the actual electricity demand for each day.
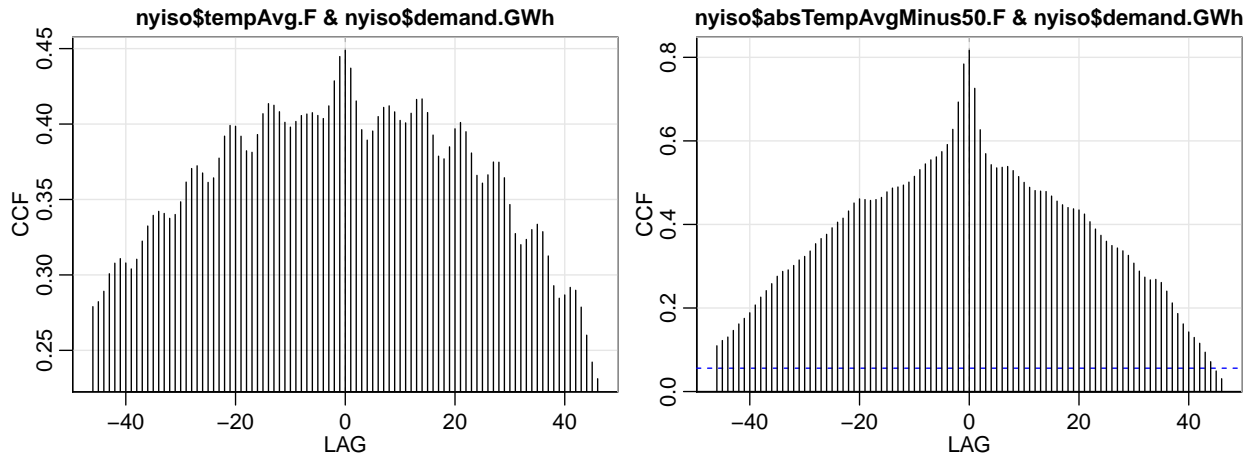

While the newly created "temp. distance from 50F" variable seems to be a much better predictor of the next day or two's electricity demand (the CCF values at right are almost twice that of the left plot), since the process isn't yet stationary and the distribution is so symmetrical about zero there doesn't seem to be much to make of this yet.

ISONE data
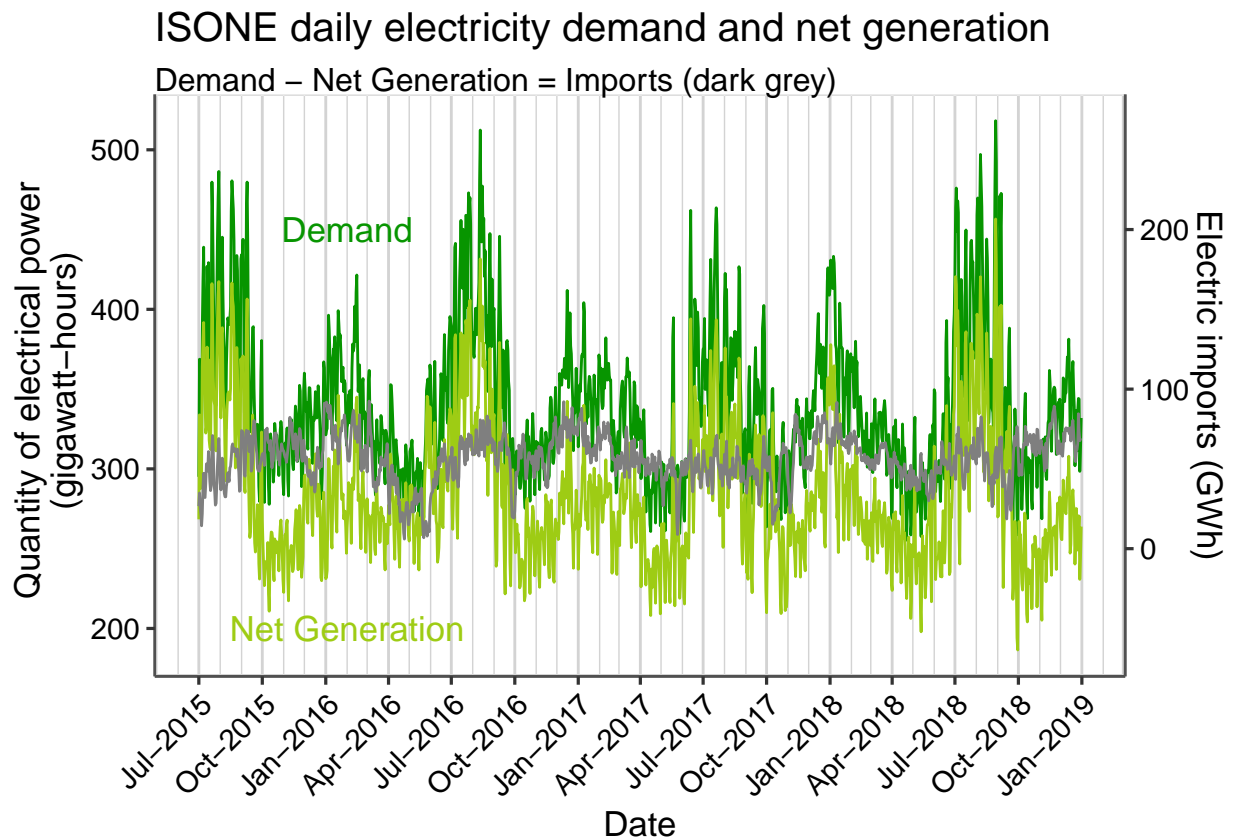
Interestingly, the ISONE demand and net generation data follows a similar general pattern as the NYISO data (largest peaks in summer months with a secondary peak in winter months), though daily demand is roughly 100 GWh less in the New England states.
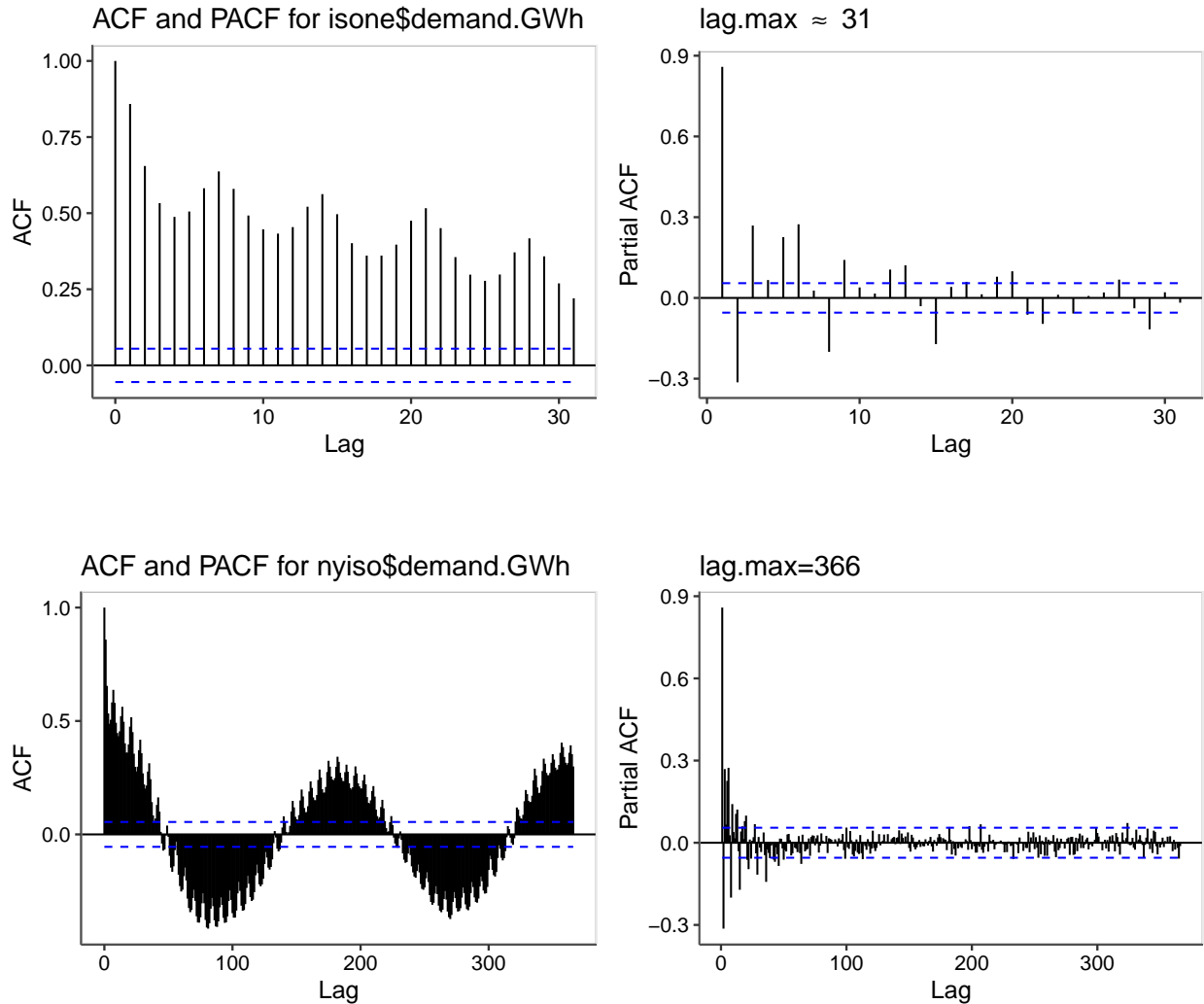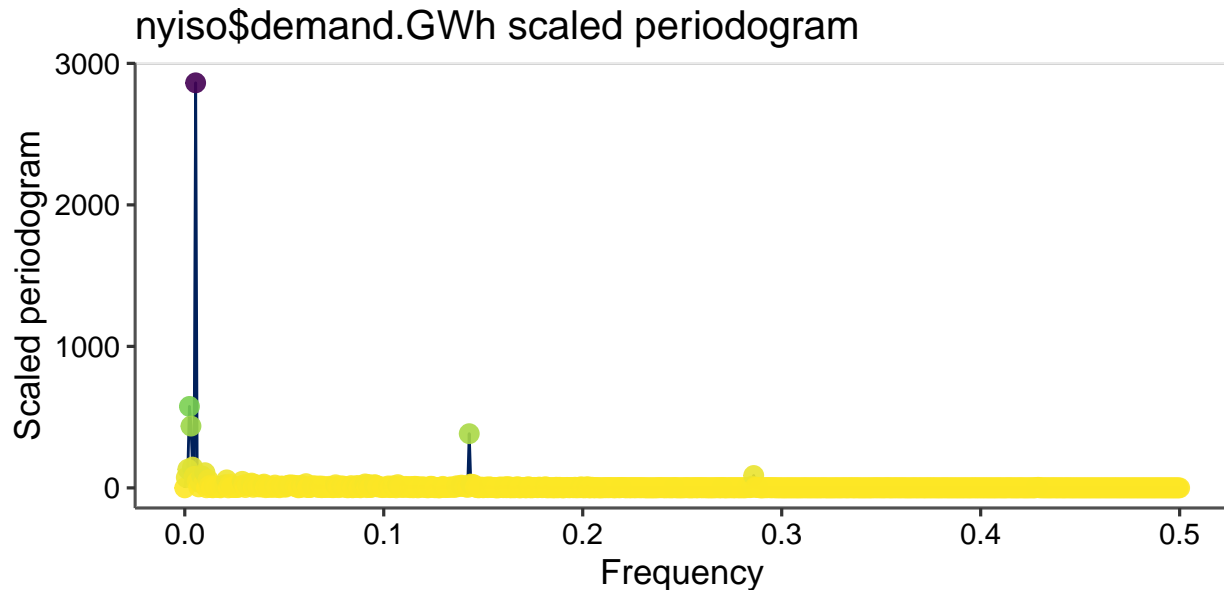

The ISONE ACF and PACF plots look very, very similar to those for NYISO data.
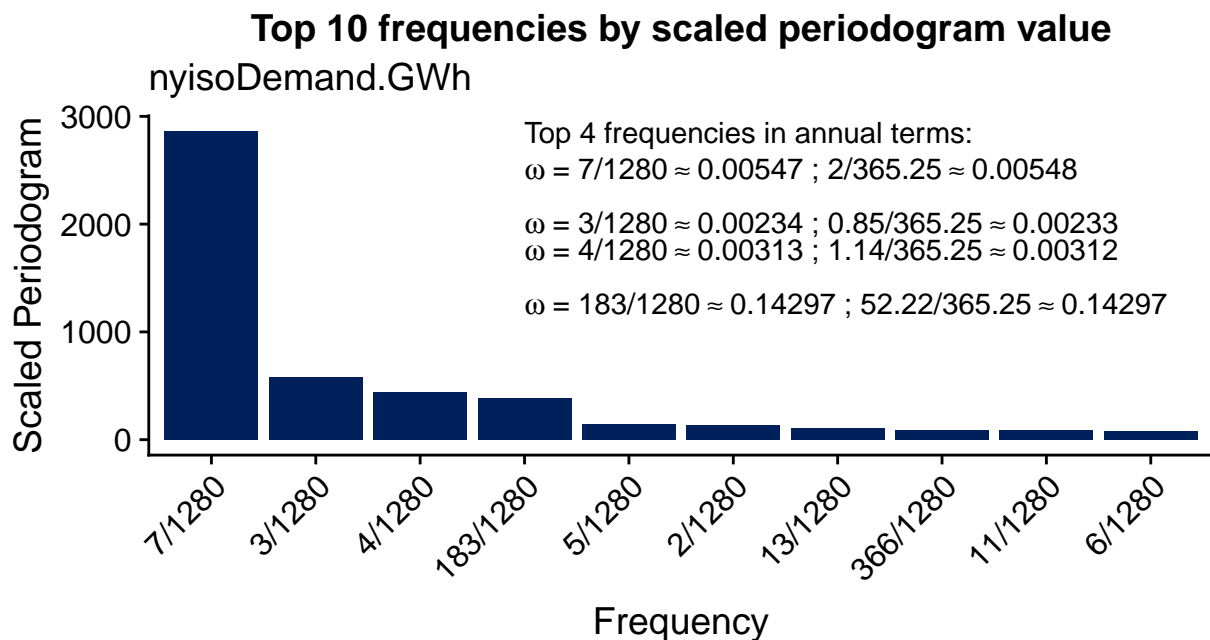
NYISO analysis

Because the series has such a clear waveform pattern, it seems reasonable to start by trying to identify key underlying frequencies.

nyiso$demand.GWh scaled periodogram

There are clearly some frequencies that are much more prominent than the rest, especially at low frequencies. The next plot more clearly illustrates that these align with once- or twice-yearly frequencies, which should correspond with the summer/winter peaks seen in the EDA plots. The highest-value frequency that also stands out seems to align (very roughly) with a twice-monthly frequency, which will also be considered.

Top 10 frequencies by scaled periodogram value

nyisoDemand.GWh

Top 4 frequencies in annual terms:
$\omega = 7/1280 \approx 0.00547$; $2/365.25 \approx 0.00548$

$\omega = 3/1280 \approx 0.00234$; $0.85/365.25 \approx 0.00233$
$\omega = 4/1280 \approx 0.00313$; $1.14/365.25 \approx 0.00312$

$\omega = 183/1280 \approx 0.14297$; $52.22/365.25 \approx 0.14297$

**Regression modelling - frequencies, temperatures, weekdays, months (plus ARMA errors)**

The following model is a bit of a "kitchen sink" approach. It's worth noting that adding the 'absolute temperature deviance from 50F' variable moved the R^2_{adj} of the 'penultimate' model from 0.6552 to 0.8196.

The corresponding AIC change was from 12761.71 to 11933.98, respectively. From there, including the 'previous day's demand' variable (per Dr. Frey's suggestion) improved the R^2_{adj} from 0.8196 to 0.9193 (and AIC from 11933.98 to 10895.85).

The 'once per 2 months' frequency (183/1280 above) didn't have statistically significant sine or cosine terms using either 52 (closest integer) or 60 ('2 months') out of 365.25. The quadratic time indicator was marginally statistically significant (p-value 0.056), so I kept it in the model.

It wasn't possible to fit all but one month indicator (or even all but five) in the model due to singularities, which seems odd but perhaps this is somehow connected to the sine/cosine frequency variables. I chose to not include month indicator variables - selectively including some months improved the model fit, but it seemed overly arbitrary.

```
##
## Call:
## lm(formula = demand.GWh ~ dayIndex + dayIndex2.ctr + prevDayDemand +
##     isMon + isTue + isWed + isThu + isFri + isSat + sin.1xYr +
##     cos.1xYr + sin.2xYr + cos.2xYr + absTempAvgMinus50.F, data = nyiso2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.145  -9.150   0.341   9.198  81.607
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.226e+02  6.136e+00  19.972  < 2e-16
## dayIndex           -3.449e-03  1.308e-03  -2.637 0.008466
## dayIndex2.ctr       7.816e-06  4.079e-06   1.916 0.055542
## prevDayDemand       6.032e-01  1.552e-02  38.864  < 2e-16
## isMon               4.006e+01  1.791e+00  22.370  < 2e-16
## isTue               3.186e+01  1.809e+00  17.617  < 2e-16
## isWed               2.452e+01  1.851e+00  13.250  < 2e-16
## isThu               2.344e+01  1.843e+00  12.717  < 2e-16
## isFri               2.057e+01  1.839e+00  11.183  < 2e-16
## isSat              -5.127e+00  1.831e+00  -2.800 0.005184
## sin.1xYr            4.175e+00  7.820e-01   5.339 1.10e-07
## cos.1xYr            3.978e+00  7.664e-01   5.191 2.43e-07
## sin.2xYr            2.681e+00  9.347e-01   2.868 0.004199
## cos.2xYr            2.691e+00  7.866e-01   3.421 0.000643
## absTempAvgMinus50.F 2.187e+00  8.484e-02  25.776  < 2e-16
##
## Residual standard error: 17.01 on 1264 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.9202, Adjusted R-squared:  0.9193
## F-statistic:  1042 on 14 and 1264 DF,  p-value: < 2.2e-16
```

ACF & PACF for nyisoDemandGWh.lm$residuals
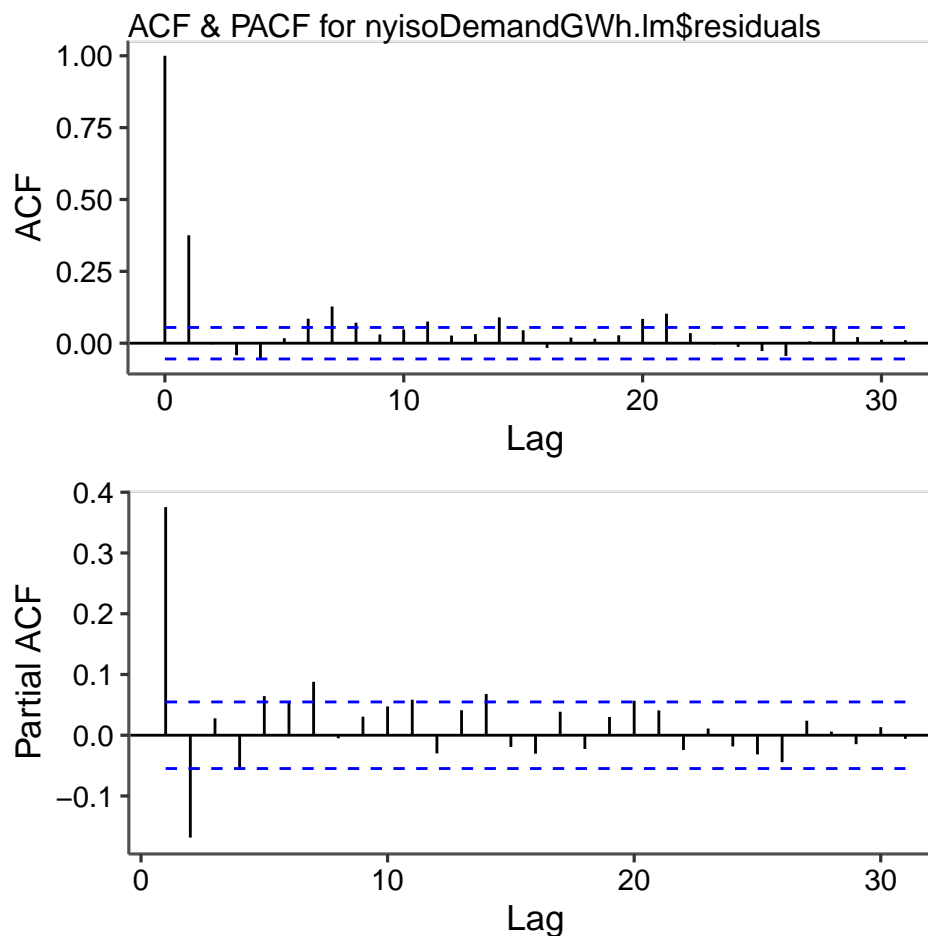
This seems like a prime candidate for fitting an ARMA model to the residuals - ARMA(1,1) may work based on the above.

```
##     row col
## p:5   6   2
```

```
##      q:0   q:1   q:2   q:3   q:4   q:5
## p:0 6.658 6.472 6.474 6.476 6.473 6.475
## p:1 6.506 6.474 6.471 6.472 6.475 6.474
## p:2 6.477 6.475 6.477 6.468 6.472 6.470
## p:3 6.478 6.472 6.473 6.470 6.466 6.469
## p:4 6.476 6.473 6.475 6.469 6.468 6.469
## p:5 6.473 6.462 6.463 6.470 6.463 6.471
```
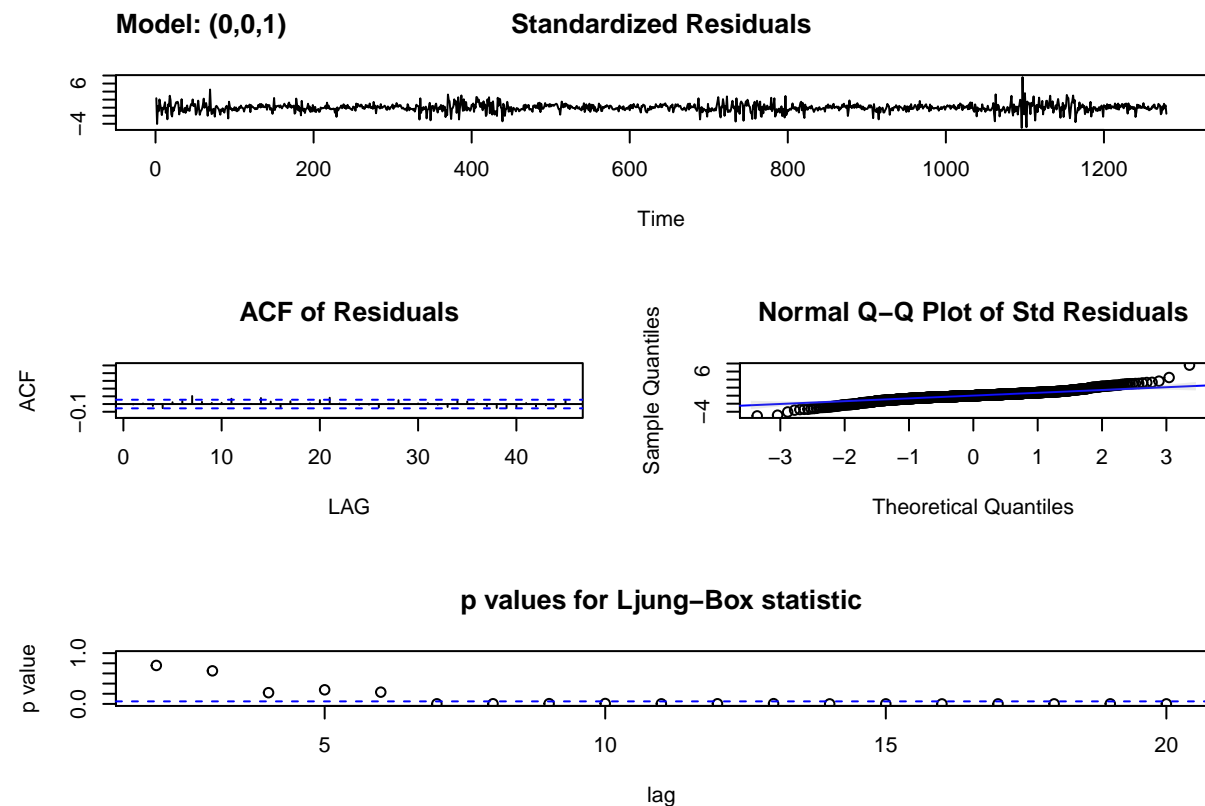
While ARMA(5,1) has the "best" AIC value, ARMA(0,1) a.k.a. MA(1) comes close here.

```
## initial  value 2.828071
## iter   2 value 2.739168
## iter   3 value 2.736675
## iter   4 value 2.736133
## iter   5 value 2.736127
## iter   6 value 2.736127
```

```
## iter   7 value 2.736127
## iter   8 value 2.736127
## iter   8 value 2.736127
## iter   8 value 2.736127
## final  value 2.736127
## converged
## initial  value 2.734792
## iter   2 value 2.734755
## iter   3 value 2.734753
## iter   3 value 2.734753
## iter   3 value 2.734753
## final  value 2.734753
## converged
```

Model: (0,0,1)   **Standardized Residuals**

ACF of Residuals   **Normal Q–Q Plot of Std Residuals**

p values for Ljung–Box statistic


```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##     Q), period = S), xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc,
##     REPORT = 1, reltol = tol))
##
## Coefficients:
##          ma1    xmean
##       0.4567   0.0265
## s.e.  0.0257   0.6273
```

```
##
## sigma^2 estimated as 237.3:  log likelihood = -5312.57,  aic = 10631.14
##
## $degrees_of_freedom
## [1] 1277
##
## $ttable
##        Estimate     SE t.value p.value
## ma1      0.4567 0.0257 17.7971  0.0000
## xmean    0.0265 0.6273  0.0422  0.9663
##
## $AIC
## [1] 6.472451
##
## $AICc
## [1] 6.47403
##
## $BIC
## [1] 5.48051
```

Excepting the worrisome skewedness in the tails of the Normal Q-Q plot (indicating heavy tails, especially at right), the diagnostic plots look quite good. The mean value of the process is very, very plausibly zero (p-value = 0.880), and the MA(1) component is very statistically significant.

Predicting from the model

The 'in-model' data ends at the very end of 2018 (31 Dec 2018) but there's data up to mid-February 2019, so predicting early 2019's demand (January through 14 Feb 2019) and comparing to 'the real thing' is pretty easy. Given the Ljung-Box diagnostic plot above, I suspect the predictions after about a week may stray fairly far from the observed values.
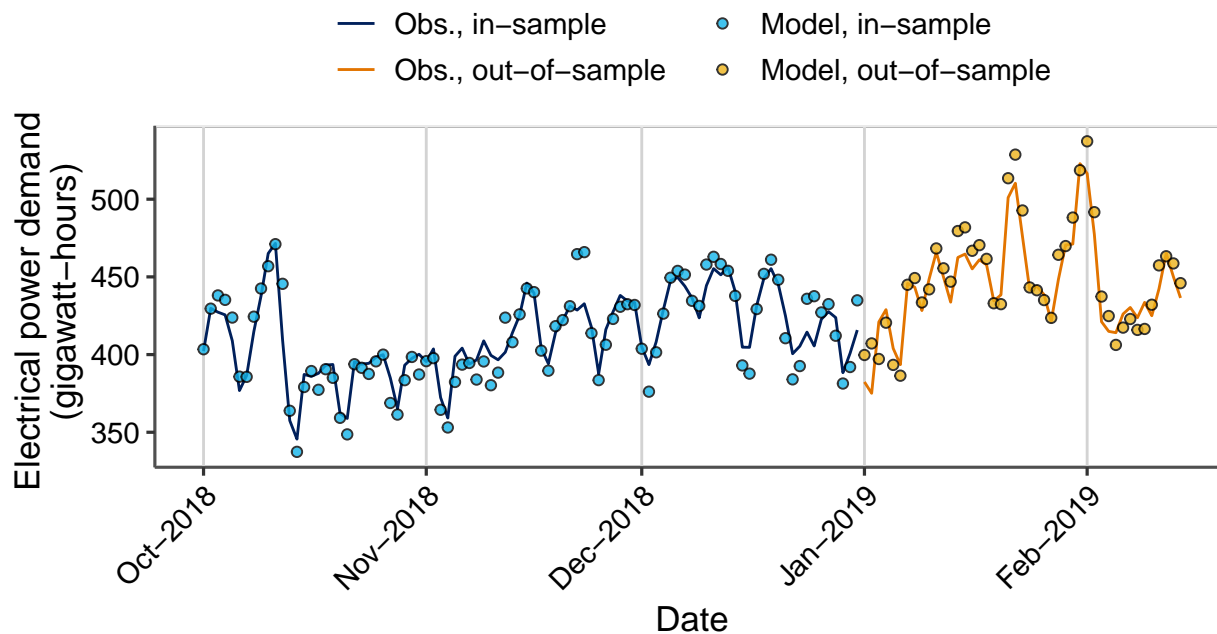
NYISO demand.GWh observed vs model fit

1 Jul 2015 through 14 Feb 2019

No model estimate for 1 Jul 2015 (no 'previous day demand')
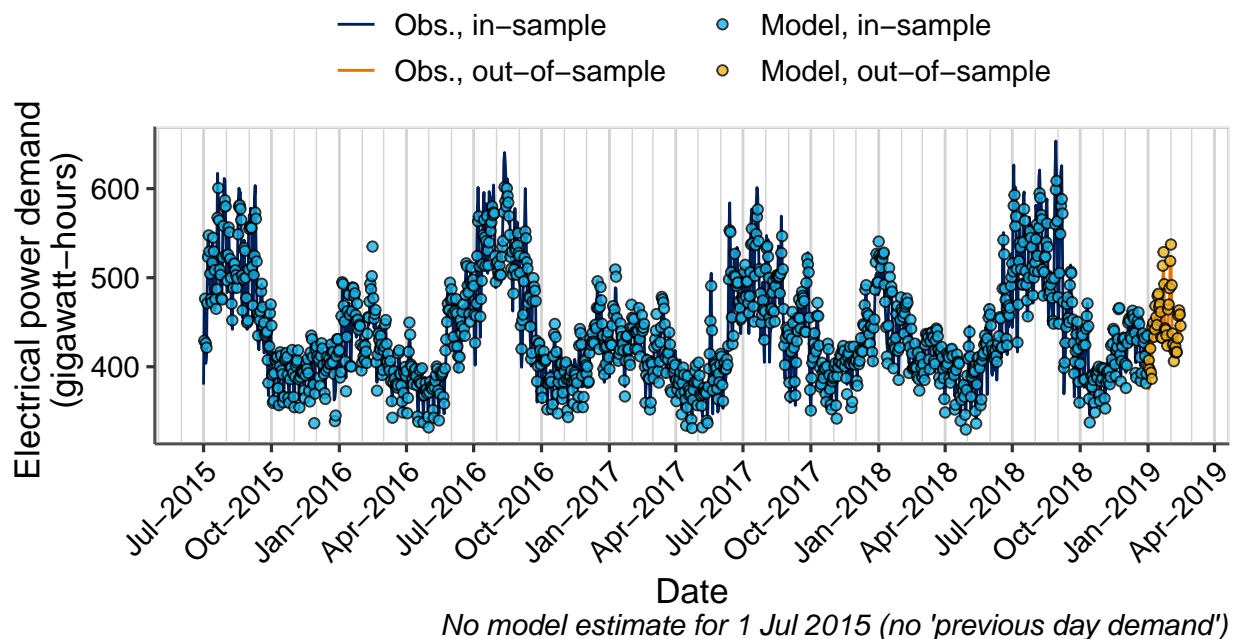
NYISO demand.GWh observed vs model fit

1 Oct 2018 through 14 Feb 2019

Not bad! It doesn't look as close as the in-dataset "day ahead demand" variable (when aligned with the appropriate day), but the model seems reasonable. Perhaps I was overly concerned about the model's residual-estimating performance 1+ weeks out.

Let's compare the root-mean-square error (RMSE) for the model vs. 'aligned day-ahead demand' and take a quick look at where the model misses most:

```
## # A tibble: 4 x 3
##   source                 where          RMSE
##   <chr>                  <chr>         <dbl>
## 1 model                  in-sample      16.9
## 2 alignedDayAheadDemand  in-sample      32.0
## 3 model                  out-of-sample  12.2
## 4 alignedDayAheadDemand  out-of-sample  27.6
```

Very interesting - the model has much lower RMSE both in- and out-of-sample vs. the 'aligned day-ahead demand' metric in the dataset. 'Out-of-sample' RMSE may be lower in both cases since there are much fewer observations (and thus less 'misses' to accumulate) - though I'd think dividng by the number of cases would adjust for this. Perhaps normalized RMSE (dividing the above results by the sample-appropriate mean of `demand.GWh`) would be better, but for now I'll stick with the above as a relative comparison.

```
##             date prevDayDemand   weekday demand.GWh modEst modelResid
## 1   07 Jul 2018          578.1  Saturday      427.7  509.8      -82.1
## 2   02 Jul 2018          444.1    Monday      588.6  507.0       81.6
## 3   01 Jul 2018          526.9    Sunday      444.1  515.9      -71.8
## 4   08 Sep 2015          491.9   Tuesday      589.1  524.8       64.3
## 5   28 May 2018          409.3    Monday      368.8  431.7      -62.9
## 6   28 Aug 2018          550.4   Tuesday      621.8  560.0       61.8
## 7   04 Sep 2018          537.7   Tuesday      607.0  549.0       58.0
## 8   11 Aug 2016          582.4  Thursday      631.1  573.3       57.8
## 9   22 Jul 2016          533.9    Friday      596.7  539.6       57.1
## 10  06 Aug 2018          528.3    Monday      611.7  555.9       55.8
```

Unsurprisingly, two of the model's worst 'misses' (largest absolute residuals) occur on 1-2 July 2018, days which had considerable missingness (and thus imputation). I checked values of the other covariates for these cases, but there didn't seem to be anything extraordinary - I believe the model was simply overly-reliant on the `prevDayDemand` values.

ISONE analysis

Since the ISONE `demand.GWh` data was so similar (in form if not magnitude) to the NYISO data, I'd expect very similar results for the spectral analysis. Unfortunately there isn't temperature data available now, so overall model performance will likely suffer accordingly.

Spectral analysis
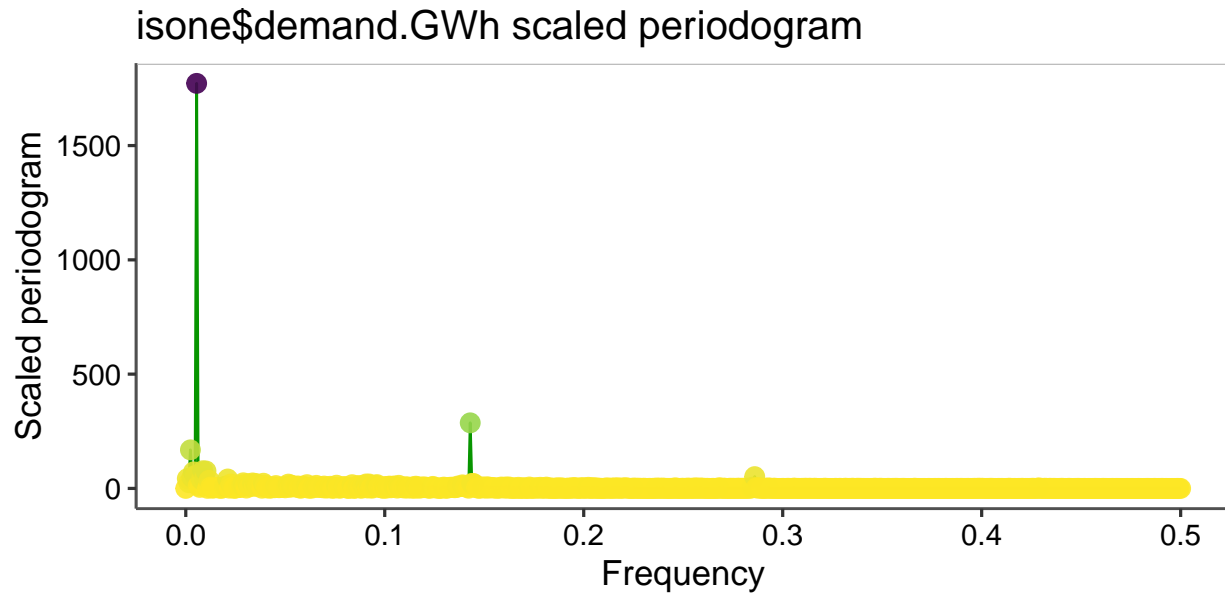
Interestingly the very-top frequencies are similar to those of NYISO demand, but there is less uncertainty for the very low-frequency value(s).
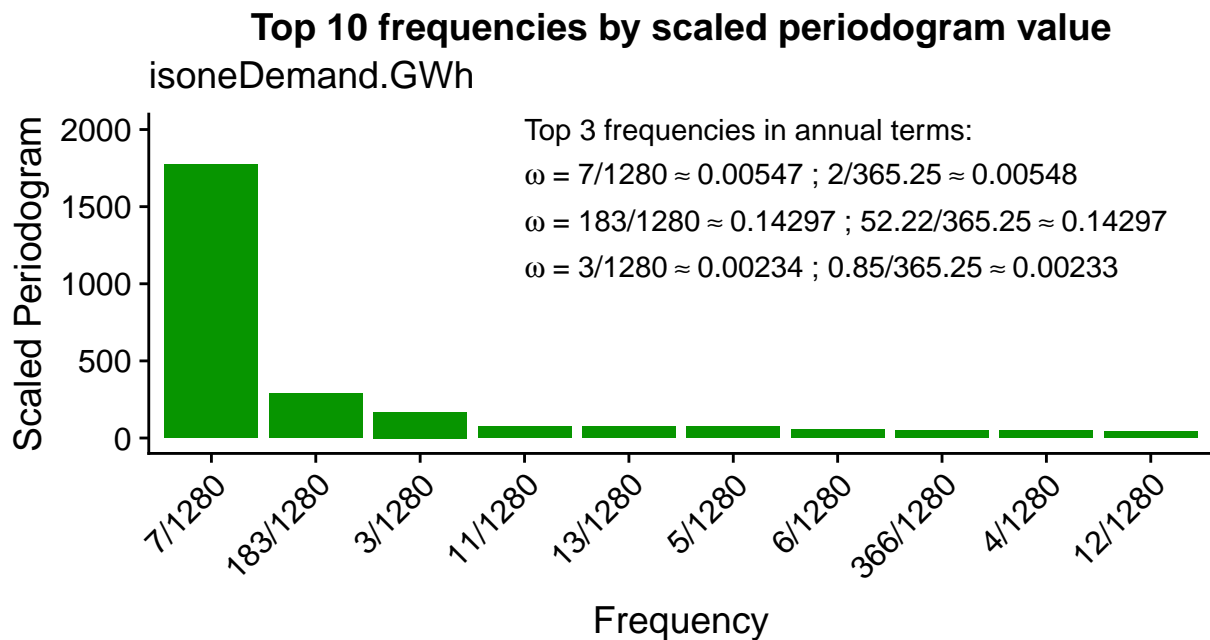
Some quick takeaways:

- Once more, the 'once-per-year' is by far the most prominent frequency detected, though there are fewer 'accompanying' frequencies around this compared to the NYISO demand data.

- The (approximately) 'once every two months' frequency appears again, though once more it is quite low relative to the 'top' frequency.

- NYISO and ISONE demand share 9 of the same 10 'top' frequencies, only differing in the 'right tail' of the set (NYISO has 2/1280 as its #6 frequency, ISONE has 12/1280 as its #10 frequency). The order and relative magnitude of the shared top frequencies differ, though 7/1280 is far and away tops for both series, 3/1280 is top-3 for both, and 183/1280 is top-4 for both.

Regression modelling - frequencies, temperatures, weekdays, months (plus ARMA errors)

This will adhere as closely as possible to the NYISO model, though it lacks temperature data and the retained fundamental frequencies (sine/cosine covariates) may differ based on statistical significance. The quadratic time indicator has a fairly high p-value (about 0.3) and so was dropped.

I retained the time indicator variable despite its high p-value (about 0.23), as it seems fundamental to a time series regression model.

The coefficients are understandably different than for the NYISO demand model, since here we cannot condition on absolute temperature deviation from 50F. It seems somewhat that the indicator variable `isSat` is negative, indicating that (adjusting for all other variables in the model) a typical Saturday would be expected to have lower electricity demand than for a typical Sunday (the reference level for weekday) - perhaps New England just really, really enjoys watching the Patriots play.

```
##
## Call:
## lm(formula = demand.GWh ~ dayIndex + prevDayDemand + isMon +
##     isTue + isWed + isThu + isFri + isSat + sin.1xYr + cos.1xYr +
##     sin.2xYr + cos.2xYr, data = isone2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -102.679   -9.857   -0.701    9.406   81.640
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   63.356545   5.901680  10.735  < 2e-16
## dayIndex      -0.001703   0.001432  -1.189  0.23462
## prevDayDemand  0.780671   0.017405  44.853  < 2e-16
## isMon         33.618195   1.964814  17.110  < 2e-16
## isTue         23.234412   1.989960  11.676  < 2e-16
## isWed         14.871468   2.029681   7.327 4.18e-13
## isThu         12.691243   2.025103   6.267 5.04e-10
## isFri         10.017198   2.016522   4.968 7.71e-07
## isSat         -9.762667   2.000351  -4.880 1.19e-06
## sin.1xYr       3.260117   0.795626   4.098 4.44e-05
## cos.1xYr       1.970908   0.756287   2.606  0.00927
## sin.2xYr       6.740207   0.963827   6.993 4.34e-12
## cos.2xYr       6.401827   0.846048   7.567 7.33e-14
##
## Residual standard error: 18.74 on 1266 degrees of freedom
```

```
##    (1 observation deleted due to missingness)
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.836
## F-statistic: 543.9 on 12 and 1266 DF,  p-value: < 2.2e-16
```
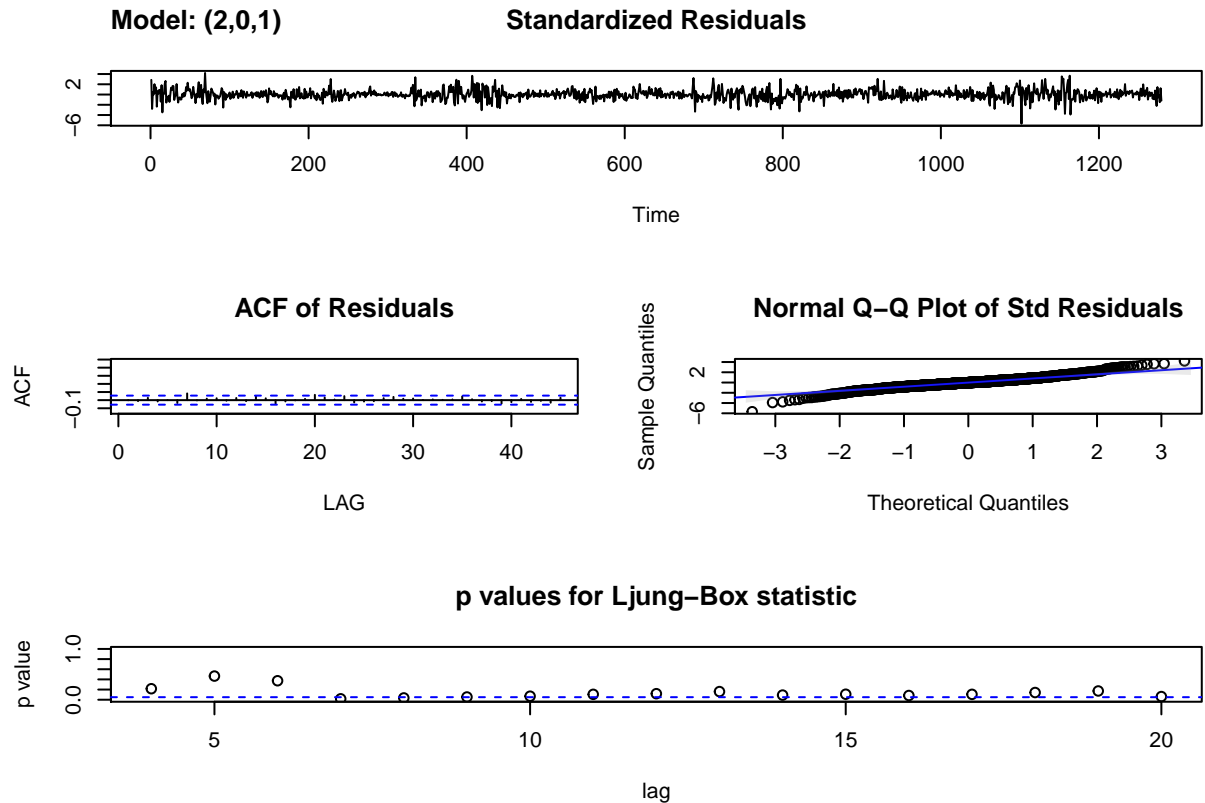

ACF & PACF for isoneDemandGWh.lm$residuals
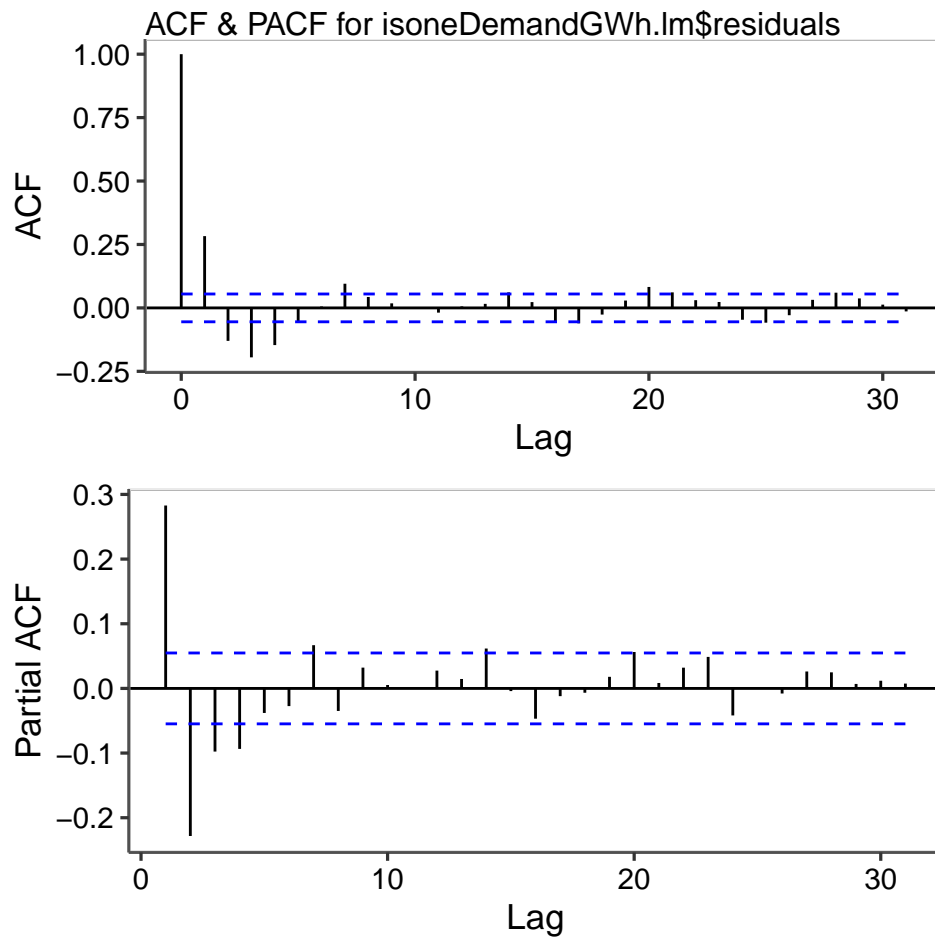
While ARMA(1,1) looked like a possible candidate for NYISO demand model residuals (and MA(1) was the final selection), here a slightly more complicated model might be called for.

```
##     row col
## p:5   6   6
```

```
##       q:0   q:1   q:2   q:3   q:4   q:5
## p:0 6.853 6.739 6.737 6.719 6.704 6.705
## p:1 6.771 6.739 6.715 6.708 6.705 6.706
## p:2 6.717 6.701 6.703 6.701 6.705 6.707
## p:3 6.709 6.703 6.701 6.702 6.702 6.703
## p:4 6.702 6.702 6.702 6.701 6.702 6.707
## p:5 6.702 6.704 6.703 6.702 6.701 6.699
```

As with NYISO model residuals, a fairly complicated process (ARMA(5,5)) is 'best', but a much simpler alternative (here, ARMA(2,1)) is quite close.

```
## initial  value 2.921949
## iter   2 value 2.880762
## iter   3 value 2.855088
## iter   4 value 2.854137
## iter   5 value 2.853859
## iter   6 value 2.851780
## iter   7 value 2.849019
## iter   8 value 2.846970
## iter   9 value 2.844159
## iter  10 value 2.844153
## iter  11 value 2.844106
## iter  12 value 2.844062
## iter  13 value 2.844020
## iter  14 value 2.843954
## iter  15 value 2.843935
## iter  16 value 2.843933
## iter  17 value 2.843932
## iter  17 value 2.843932
## iter  17 value 2.843932
## final  value 2.843932
## converged
## initial  value 2.847930
## iter   2 value 2.847899
## iter   3 value 2.847837
## iter   4 value 2.847722
## iter   5 value 2.847649
## iter   6 value 2.847633
## iter   7 value 2.847632
## iter   8 value 2.847632
## iter   8 value 2.847632
## iter   8 value 2.847632
## final  value 2.847632
## converged
```

Model: (2,0,1) Standardized Residuals

ACF of Residuals

Normal Q–Q Plot of Std Residuals

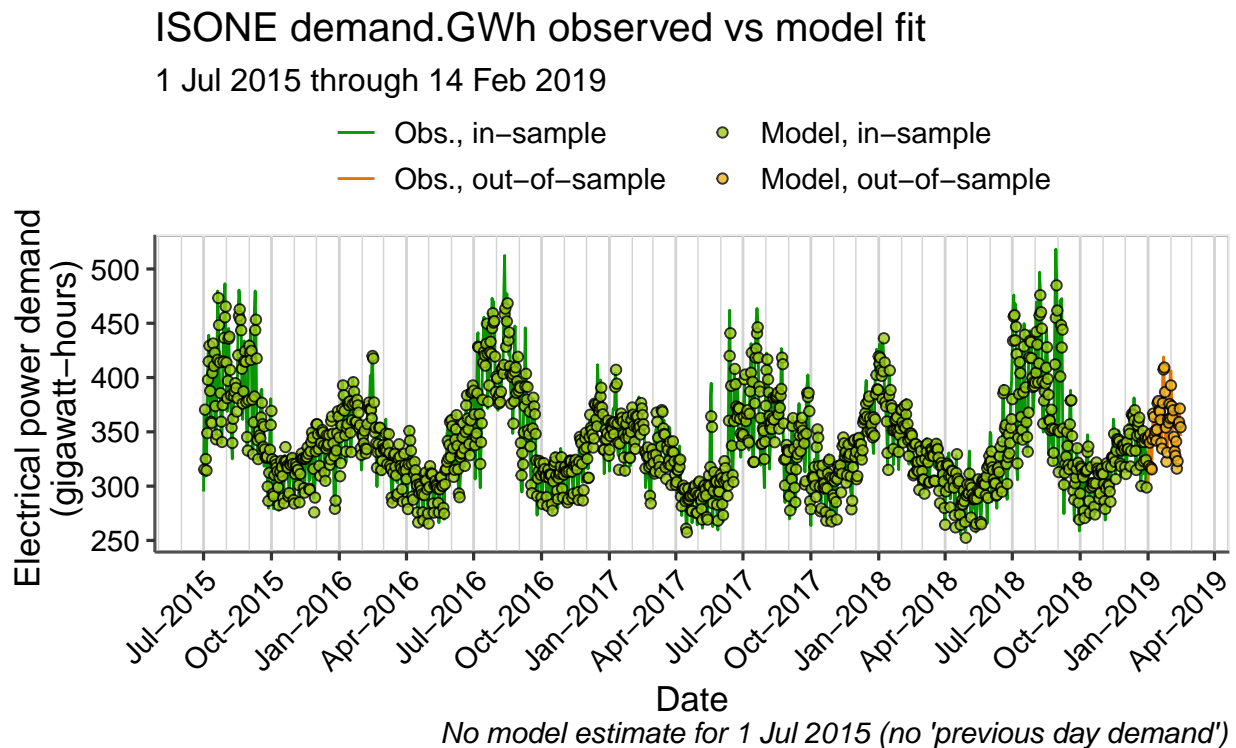## p values for Ljung–Box statistic


```
## $fit
##
## Call:
## stats::arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D,
##      Q), period = S), xreg = xmean, include.mean = FALSE, optim.control = list(trace = trc,
##      REPORT = 1, reltol = tol))
##
## Coefficients:
##           ar1       ar2       ma1    xmean
##        0.8287   -0.3665   -0.5138   0.0052
## s.e.   0.0696    0.0275    0.0729   0.4363
##
## sigma^2 estimated as 297.4:  log likelihood = -5456.94,  aic = 10923.89
##
## $degrees_of_freedom
## [1] 1275
##
## $ttable
##        Estimate      SE  t.value p.value
## ar1      0.8287  0.0696  11.8983  0.0000
## ar2     -0.3665  0.0275 -13.3220  0.0000
## ma1     -0.5138  0.0729  -7.0487  0.0000
## xmean    0.0052  0.4363   0.0119  0.9905
##
## $AIC
## [1] 6.701319
```

```
##
## $AICc
## [1] 6.70292
##
## $BIC
## [1] 5.717438
```

Once more, with feeling: the ISONE model's diagnostics are similar to those of NYISO data, though here the left tail of the Q-Q norm plot seems heavier and the model (ARMA(2,1)).

Predicting from the model

The 'in-model' data ends at the very end of 2018 (31 Dec 2018) but there's data up to mid-February 2019, so predicting early 2019's demand (January through 14 Feb 2019) and comparing to 'the real thing' is pretty easy. Given the Ljung-Box diagnostic plot above, I suspect the predictions after about a week may stray fairly far from the observed values.

ISONE demand.GWh observed vs model fit
1 Jul 2015 through 14 Feb 2019

No model estimate for 1 Jul 2015 (no 'previous day demand')

ISONE demand.GWh observed vs model fit
1 Oct 2018 through 14 Feb 2019
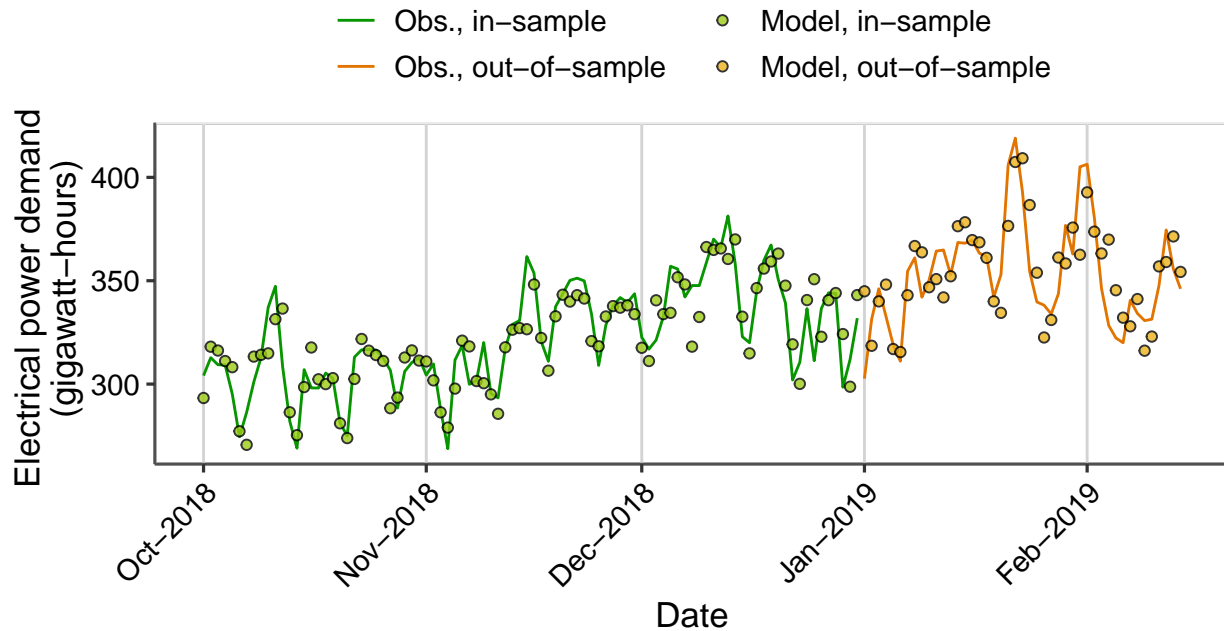
The model seems to fit the data fairly well, and predict reasonably well, but it seems to predict less well more quickly than does the NYISO demand model.

```
## # A tibble: 4 x 3
##   source                 where         RMSE
##   <chr>                  <chr>         <dbl>
## 1 model                  in-sample      18.6
## 2 alignedDayAheadDemand  in-sample      25.2
## 3 model                  out-of-sample  17.4
## 4 alignedDayAheadDemand  out-of-sample  21.7
```

The model seems to outperform the 'aligned day-ahead demand' from the ISONE dataset, as did the NYISO model previously.

```
##           date prevDayDemand  weekday demand.GWh modEst modelResid
## 1  07 Jul 2018        460.7 Saturday      318.9  421.6     -102.7
## 2  31 Aug 2018        492.3   Friday      376.5  461.7      -85.3
## 3  08 Sep 2015        374.8  Tuesday      464.2  382.6       81.6
## 4  04 Sep 2018        399.2  Tuesday      471.1  401.0       70.1
## 5  11 Aug 2016        416.8 Thursday      480.2  411.7       68.5
## 6  22 Jul 2016        394.1   Friday      458.9  392.8       66.1
## 7  14 Jul 2017        400.3   Friday      330.6  396.1      -65.6
## 8  29 Sep 2017        384.7   Friday      304.0  369.1      -65.0
## 9  12 Jun 2017        335.7   Monday      425.1  360.2       64.8
## 10 18 May 2017        319.5 Thursday      382.4  317.6       64.8
```

The ISONE demand model has larger absolute residuals in its 'top 10 misses' set compared to the NYISO data, with 7 Jul 2018 a relatively large 'miss'. It seems one more that `prevDayDemand` is largely involved here.

24

Conclusion

This was an enjoyable exploration of modeling time series data, and it seems that with just a few variables and a healthy dose of data formatting and R one can model 'well-behaved' electricity demand data without too much trouble.

Ancillary analysis considerations: Spectral analysis of `netGen.GWh`

Unfortunately I ran out of time to conduct more extensive analyses, but some ideas for extending this analysis are:

- Including more recent data and observing the results, including perhaps truly predicting future observations and seeing how the model fares

- Bringing in temperature data for the ISONE dataset and making a more direct comparison between NYISO and ISONE models

- Analyzing `netGen.GWh`

- Analyzing hourly-level data (which might require more granular temperature data)

- Finding more useful/less arbitrary covariates (e.g. the 'degree heating days' idea Dr. Frey mentioned)

- Checking how energy demand compares vs. the model on certain specific dates (e.g. major holidays or events such as the Super Bowl)