Forecasting High Frequency Data With Complex Multi-layer Seasonality Using Covariates

Shubhabrata Das*
Indian Institute of Management Bangalore, Bangalore, India – shubho@iimb.ernet.in

Akshay Kumar Singh
Indian Institute of Management Bangalore, Bangalore, India – akshay.singh@iimb.ernet.in

Abstract

Increasingly in many contexts like energy load or demand, data is recorded after every minute or every five minutes. Such data exhibit not only annual and weekly seasonality but also within day and within hour seasonality which are often entangled with each other. Various weather covariates also have significant linear and at times nonlinear (quadratic) impact on the dependent variable. Forecasting in such context is methodologically challenging, as ARIMA type models are not directly suitable for data with large seasonal periods or multiple-level seasonality, and relatively newer approaches like TBATS have difficulties in accommodating covariates. In addition to that, there are computational challenges to implement advanced models on such high frequency data. In this work, we first adopt traditional time series and regression based models for daily average data. Subsequently the within hour and within day seasonality is analyzed thoroughly and the estimated seasonal components are integrated with the projected daily average as the final forecast. The methodology is illustrated with energy load data from New York. Results show differential Sunday as well Saturday effect while difference across the different weekdays is not significant. Hourly and within hour seasonality vary in more complicated way and incorporating their estimates improves the forecast substantially.

Keywords: time series; regression; decomposition; big data analytics.

1. Introduction

Forecasting high frequency time series having complex multiple seasonality, like that of electricity load forecast, or mobile network usage, are drawing attention of both the academicians and practitioners (Gould et al. 2008, Au et al. 2011, Livera et al. 2011, Soares et al. 2006, Taylor 2003, among others). On one hand, there is a huge commercial interest due to the ability to capture data at rapid succession and the possible benefit that can be derived from successful forecast that enables judicious resource planning. On the academic front, the task is challenging since existing methods are either not well suited for dealing long seasonal pattern or seasonality at multiple level such as within hour, within day, within week, within year (as in the case of Auto Regressive Integrated Moving Average or ARIMA), or automatic accommodation of critical independent variables like temperature, as in the case of TBATS of Livera et al. (2011). In addition, implementing the time series and regression model for such high frequency data (recorded every minute or at five minute interval) require high computational power. In this work, we consider electricity load as the dependent variable for the purpose of illustration and reference; however the work has wider applicability with forecasting electricity demand, mobile usage etc.

2. Proposed Methodology: An Overview

In this work, we propose a a four-stage approach. In the first stage, daily average load is regressed on all relevant independent variables that may include weather related variables and dummy variables that may account for different days of the week, or weekday, or other structural differences. Usual model selection procedures may be adopted to arrive at appropriate list of covariates which have significant impact. Certain nonlinear transformation of variables may also be considered; in particular, energy load is often found to have quadratic relationship with temperature variables. Whenever the weather related variables are available at higher frequency (e.g. hourly), the corresponding daily averages are considered. Model selection of independent variables are carried out and using change point analysis we also determine whether the different days of the week have significant impact, or whether significant difference exists between weekdays and weekend, or between Saturday and Sunday.

In the second stage traditional pure time series methods such as ARIMA, TBATS may be applied on the residual of the regression from the first stage. Ideally the first two stages should be combined as this would lead to overall better fitting and prediction. However, this is feasible only for ARIMA (for example using xreg in R), but not so with TBATS.
For the third stage, we come up with a methodology for estimating and analyzing the seasonality at higher frequency level (i.e. within hour, within day). These seasonality are estimated using either additive or multiplicative model.
In the final stage, the estimates from the first three stages are combined. The separation between 'high' and 'low' frequency (i.e. separating at what frequency analysis should be done at the first two stages vis-a-vis the third stage) can be dictated by the computational limitation of the user.
For illustrative case study, we take the case of energy load data from New York that is downloaded from http://www.nyiso.com/public/markets_operations/market_data/load_data/index.jsp The data used is at five minute frequency interval from February 1, 2005 till December 31, 2014, with the data from 2014 used essentially as hold-out sample. The information about four weather related variables (maximum and minimum dry and wet bulb temperature) are available only from September 5, 2008 and hence for better part of the study we have considered only that latter time frame.

3. Modeling and Analysis of Average daily Load.

Overall to model the daily average load, the following class of models are tried:

- Multiple Regression incorporating trend, day effect, effect of temperature variables
- TBATS (Livera et al. (2011) with weekly and/or annual seasonality
- Seasonal (yearly) ARIMA
- Multiple regression followed by TBATS on residuals from regression
- Multiple regression followed by Seasonal ARIMA on residuals from regression

First we look at the daily, weekly and monthly average load to see if there is significant trend with time. The summary of the four regression is enclosed below:

Table 1: Summary – Long Term Linear Trend

Average frequency	$\hat{\beta}$	Std. Error ($\hat{\beta}$)	p-value	R^2
daily	-0.0319	0.0152	0.0359	0.12%
weekly	-0.2130	0.2355	0.3663	0.16%
monthly	-0.9515	2.0122	0.6372	0.19%
yearly	-14.3069	10.3256	0.2033	19.35%

Thus, we observe a very slight negative linear trend; taken in isolation, slope is significant *only* for daily (average) load. However, as seen subsequently, significance improves with consideration of other explanatory variables.
Next we identify the appropriate set of independent variables that ought to be used as regressors. Towards this, we consider the following list of Independent variables: Day no. (to account for possible daily linear trend) (DNo); four (daily) temperature variables: the maximum and minimum DRY bulb temperature ($t_{\max}, t_{\min}$), and the maximum and minimum WET bulb temperature ($b_{\max}, b_{\min}$); dummy variable for weekday ($D_{weekday}$); six dummy variables for Monday, Tuesday, ..., Saturday ($D_M, D_T, D_W, D_{Th}, D_F, D_{Sa}$).
We consider the following eight linear models (with R^2 values from the fit indicated within parenthesis):

(A) $\bar{DL} = \beta_0 + \beta_1 DNo \quad (R^2 = 0.05\%)$

(B) $\bar{DL} = \beta_0 + \beta_1 DNo + \beta_2 t_{\max} + \beta_3 t_{\min} + \beta_4 b_{\max} + \beta_5 b_{\min} \quad (R^2 = 31.31\%)$

(C) $\bar{DL} = \beta_0 + \beta_1 DNo + \beta_2 D_{weekday} \quad (R^2 = 11.30\%)$

(D) $\bar{DL} = \beta_0 + \beta_1 DNo + \beta_2 t_{\max} + \beta_3 t_{\min} + \beta_4 b_{\max} + \beta_5 b_{\min} + \beta_6 D_{weekday} \quad (R^2 = 42.48\%)$

(E) $\bar{DL} = \beta_0 + \beta_1 DNo + \beta_2 D_{weekday} + \beta_3 D_{Sa} \quad (R^2 = 11.39\%)$

(F) $\bar{DL} = \beta_0 + \beta_1 DNo + \beta_2 t_{\max} + \beta_3 t_{\min} + \beta_4 b_{\max} + \beta_5 b_{\min} + \beta_6 D_{weekday} + \beta_7 D_{Sa} \quad (R^2 = 42.57\%)$

(G) $\bar{DL} = \beta_0 + \beta_1 DNo + \beta_2 D_M + \beta_3 D_T + \beta_4 D_W + \beta_5 D_{Th} + \beta_6 D_F + \beta_7 D_{Sa} \quad (R^2 = 11.55\%)$

(H) $\bar{DL} = \beta_0 + \beta_1 DNo + +\beta_2 t_{\max} + \beta_3 t_{\min} + \beta_4 b_{\max} + \beta_5 b_{\min} + \beta_6 D_M + \beta_7 D_T + \beta_8 D_W + \beta_9 D_{Th} + \beta_{10} D_F + \beta_{11} D_{Sa} \quad (R^2 = 42.70\%)$

Among the above, we select the best model in the framework of subset model vis-a-vis full model (change point model); selected comparison is included below:

Table 2: Compare between Four Linear Models:
P-values in Full vs. Subset Model comparison

	# parameter (p)	6	7	8	12
R-square	Model	B	D	F	H
31.31%	B		0.00%	0.00%	0.00%
42.48%	D			5.15%	11.94%
42.58%	F				29.10%
42.70%	H				

Thus model D is selected, although model F is only marginally worse off. This suggests that all the four weather variable have significant impact; the trend is small but significant. There is significant difference between a weekday and weekend. Difference between different weekdays is not at all significant; however difference between Saturday and Sunday is borderline (not significant at 5% level, but significant at 5.25%). Since the R^2 was only about 42.5%, we explore the possibility of quadratic relationship between temperature variables and average load. For the sake of brevity, we skip the details of model selection at this stage and instead report only the final regression model selected which has ($R^2 = 84.37\%$):

$$\bar{DL} = \beta_0 + \beta_1 DNo + \beta_2 t_{\max} + \beta_3 t_{\min} + \beta_4 b_{\max} + \beta_5 b_{\min} + \beta_6 t^2_{\max} + \beta_7 t^2_{\min} + \beta_8 b^2_{\min} + \beta_9 D_{weekday} + \beta_{10} D_{Sa}.$$

We refer to this model as F2-, as this is essentially the quadratic version of Model F; this is because the quadratic terms for the temperature variables are incorporated; except the same for that of the maximum wet bulb term, as it is not significant. The parameter estimates, their standard errors and significance are listed below.

Table 3: Parameter estimates and Significance in Model (F2-)

	(Intercept)	day	weekday	sat	max_temp	min_temp	max_bulb	min_bulb	$\max^2_{temp}$	$\min^2_{temp}$	$\min^2_{bulb}$
Estimate	9553.22	-0.09	712.25	91.07	-113.53	-59.87	16.31	-44.52	0.94	0.82	0.51
Std. Error	104.03	0.01	22.36	28.85	7.71	14.55	2.85	12.50	0.06	0.13	0.12
Sig.	0.00%	0.00%	0.00%	0.16%	0.00%	0.00%	0.00%	0.04%	0.00%	0.00%	0.00%

It establishes that both Saturday and Sunday are significantly different from weekdays and from each other, while there is no significant differences among the weekdays. It also establishes that the maximum as well as minimum dry bulb temperature and the minimum wet bulb temperature have quadratic impact on the daily average load while the maximum wet bulb temperature has linear impact.

TBATS is run with with different levels of seasonality only weekly or only annual seasonality, or both on the daily average load. The parameters selected optimally by the package is given below. The last row corresponds to the direct modeling of load at the five minute gap and hence it is not directly comparable with the rest. Thus among the TBATS models, AIC criterion points towards using trigonometric seasonal model of two layers of periodicity (weekly as well yearly), no Box-Cox transformation, a very mild level of damping trend parameter, and both autoregressive and moving average order for errors being five.

Table 4: Parameters Selected by TBATS for different implementations of seasonality

Seasonality period	BATS / TBATS	Box-Cox par. ω	ARMA order (p, q)	Damping par. ϕ	# harmonics in Trig. Seasonality $k_1, \ldots$	Error sd. σ	AIC
weekly	BATS	0.002	(5,5)	1	·	0.120	37306
yearly	TBATS	0	(5,3)	0.997	6	0.054	37292
weekly & yearly	TBATS	0	(5,5)	0.999	3,16	0.053	37283
within hour, daily, wkly, yrly	TBATS	0.306	(0,0)	0.955	1,7,6,6	0.052	10855512

Within the (seasonal) ARIMA class, an extensive search lead to two choices for (p, d, q): (5,1,5) and (3,1,3) by the AIC and BIC criterion respectively; the fit from either model are almost identical.

Table 5: Summary of Forecast evaluation in predicting daily average

Sl. No.	Method	Model period 2008-13 MSE	MAD	MAPE	Hold out period 2014 MSE	MAD	MAPE
1	Regression (F2-)	101151	245.16	4.01%	341279	398.17	6.63%
2	TBATS with only weekly seas.	547286	571.2	9.28%	3031240	1482.99	26.49%
3	TBATS with only yrly seas.	127477	260.35	4.12%	499144	513.82	8.51%
4	TBATS with weekly & annual seas.	79752	181.48	2.82%	405261	424.49	6.90%
5	Regression, then TBATS on reg. res.	57631	171.51	2.79%	801432	774.51	13.63%
6	Regression, then yrly seas. ARIMA (3,1,3) on reg. res.	91680	201.03	3.28%	835482	783.67	13.79%
7	yrly seas. Arima (4,0,2) with xreg	69081	169.96	2.73%	387038	370.25	6.06%

Table 5 provides the summary of fit and prediction of the competing methods to forecast daily average load. In the last method, the first two stages are integrated, as permitted by ARIMA using xreg option in R. ARIMA order (4,0,2) is found to be best on BIC criterion among $d = 0$; but an all exhaustive optimal search of optimal order of the ARMA runs into computational obstacles (including when $d = 1$. At any rate this model yields better result, (while order (5,0,5), selected by AIC criterion yields almost identical result on both model and holdout period. On the model period the performance is best when Regressions is followed TBATS on regression residuals. However, on the hold-out period, the the performance of regression (F2-) as well as (simple) TBATS with weekly and annual seasonality (Sl no 1 & 4) is better, but ARIMA with xreg performs the best among all competing models.

4. Analyzing within hour and within day fluctuation – Estimating Seasonality at higher frequency

4.1 Block or Within Hour Seasonality. In order to analyze if within hour variation is significant, we construct the deviations from corresponding average hourly loads for all the observations recorded at five-minute interval. There are 12 such residuals correspond to the 12 (five minute) block from every hour of data. These residuals collated from all hours and days are now segregated by the block number. We note that all other factors (like day, hour of the day, temp, etc.) are absent (removed) in the residuals obtained. (Each bucket has 86904 residuals.) Now we analyze the distribution, summary statistics of these residuals across the 12 blocks. For example, an ANOVA establishes significant difference across the twelve blocks. The summary is enclosed below:

Table 6a: ANOVA: Compare mean deviation between 12 five-minute blocks

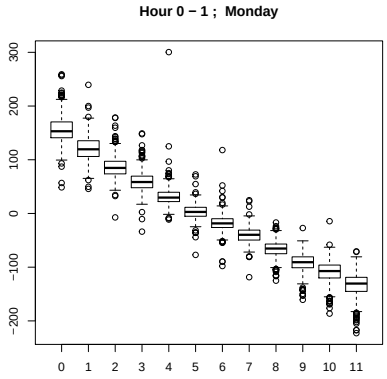
	Sum of Squares	df	Mean Square	F	p-value
due to block	1725864.101	11	156896.7	24.8706	3.22603E-52
within/ error	6578754458	1042836	6308.523		

Table 6b: ANOVA: Compare mean deviation from hourly average between 12 five-minute blocks

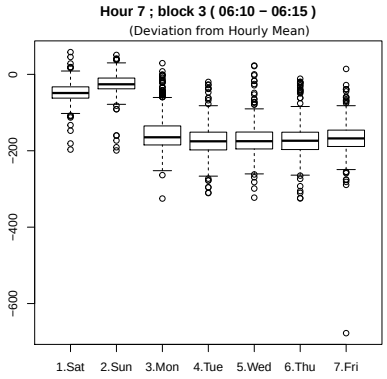
Block No	mean_dev	Std-dev	SE	p-value
1	-0.031	134.705	0.457	94.67%
2	-0.863	108.313	0.367	**1.88%**
3	-2.799	76.424	0.259	**0.00%**
4	-0.850	54.141	0.184	**0.00%**
5	-1.183	33.514	0.114	**0.00%**
6	0.713	20.081	0.068	**0.00%**
7	0.371	23.222	0.079	**0.00%**
8	0.044	39.757	0.135	74.43%
9	-0.366	59.806	0.203	*7.13%*
10	1.645	79.011	0.268	**0.00%**
11	1.948	98.133	0.333	**0.00%**
12	1.370	118.141	0.401	**0.06%**

Thus prima-facie, we note a strongly significant, but small seasonal impact at the block level (within hour). The individual block means can be taken as seasonal estimate of the blocks. However, we do not recommend that as further analysis reveals more prominent pattern. Towards this, we segregate the above analysis by further subdividing the twelve buckets based on (i) different hours within a day; and (ii) day (Mon/Tues/ . . . /Sun) of the week.

For illustration, we include one sample Box-plot from either analysis in Figures 1 and 2. On the left, we a Boxplot of deviations for these 12 blocks as observed between midnight and 1 am on Mondays. The set of boxplots on the right compare the seasonal impact of the third block (06:10 am) across the seven days of the week. These analysis (details omitted for the the sake of brevity) establish that within hour fluctuation or seasonality at the block level is substantial, dependent on hour of the day as well as day of week, although the difference across different weekdays is typically not significant. All the above analysis leads to seasonal estimates of within hour fluctuation that depend on hour of the day, as well as day of the week; the detailed list of this seasonality is not reported here, but used in the final stage of forecasting. For example, the seasonal estimate for the 6:10-6:15 block on a Monday turns out to be -153.35, while the same for a Saturday is -47.44. We may note that either of them is of much greater (absolute) magnitude than -2.799, as noted in Table 6b.



(a) Figure 1



(b) Figure 2

4.2 Hourly or Within day seasonality. Average hourly load time series, plotted separately for the different hours of the day, show interesting gradual pattern. Twenty four separate forecasting would not be efficient as it would not capture the linkages. Instead following the same path as in capturing within hour variation, now we consider the deviation of average hourly load from the average daily load throughout the period and analyze the pattern of these hourly seasonality. The statistical comparison of the hourly seasonality across the different days of the week are carried out to determine significance of it. In our context with NY load, given large quantum of data, this was not critical; however this could be used to simply models, in the case of smaller data sets. The plot for the first six hours of the day (midnight to 6 am) is enclosed to provide an idea of the hourly seasonality.

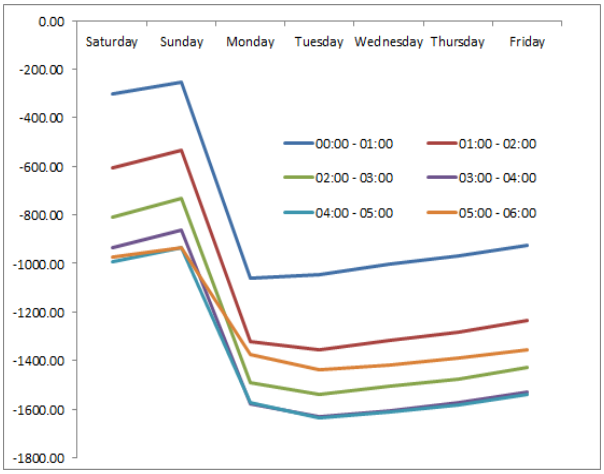


Figure 3: Seasonal effect of hour across days – midnight to 6 am

5. Combining Stages.

In the first two stages, we end up with projected daily load $\hat{\bar{L}}_{day}$, while the hourly and within-hour seasonality $\hat{S}_{hour}$ and $\hat{S}_{block}$ are estimated in the third stage. We now combine all in additive framework (since the seasonal estimates are defined in that setup) to project the load at the block level:

$$\hat{L} = \hat{\bar{L}}_{day} + \hat{S}_{hour} + \hat{S}_{block}$$

Thus, e.g. to predict load for March 3 (Tuesday), 2015 at 6:15 am, one should use:

- $\hat{\bar{L}}_{day}$: projected average daily load for March 3, 2015, based on either Regression model or TBATS (or ARIMA / ARIMA with regression error)
- $\hat{S}_{hour}$: the 7th hour effect for a weekday
- $\hat{S}_{block}$: the block effect for the fourth block of the 7th hour of a Tuesday

It would require simple adjustment to implement multiplicative model of seasonality in this framework. The hourly and within hour seasonal estimates would be estimated differently (based on ratio, as opposed to difference), leading to eventual forecast at block frequency: $\hat{L} = \hat{\bar{L}}_{day} \times \hat{S}_{hour} \times \hat{S}_{block}$.
Table 7 summarizes the performance of the key methods in forecasting/fitting at five minute interval on training period as well as on the holdout period:

Table 7: Summary of Forecast evaluation in predicting load at 5 minute interval

	Model period 2008-13			Hold out period 2014		
Method	MSE	MAD	MAPE	MSE	MAD	MAPE
Multiple Regression (Reg.) (F2-) w/o seas. Adj.	868423	786.61	13.65%	1046664	818.77	14.51%
Block & hrly seas. adj. on Reg. (F2-)	835554	776.72	13.68%	413143	443.22	7.48%
TBATS (with weekly and annual seas.) w/o seas. Adj	847024	780.46	13.45%	1110650	850.26	14.86%
Block & hrly seas. adj. on TBATS (weekly and yrly seas.)	177870	293.63	4.71%	529839	511.84	8.52%
Block & hrly seas. adj. on [Reg. (F2-), then TBATS on reg. res.]	879514	791.28	13.93%	405752	420.71	7.04%
Block & hrly seas. adj. on [Reg. (F2-), then Seas. (yrly) ARIMA (3,1,3) on reg. res.]	826026	776.40	13.68%	1142428	884.87	15.38%
Block & hrly seas. adj. on [yrly seas. Arima (4,0,2) with xreg]	803427	770.17	13.57%	1122191	875.95	15.31%
TBATS with within hour, daily, weekly, yrly seas.	487	16.71	0.28%	691829	640.74	10.77%

Table 7 shows the utility of proposed seasonality estimation within day and integrating it with daily average forecast based on TBATS, or regression. The TBATS model with all four levels of seasonality is not only computer intensive, it runs into the possible danger from over-fitting, as although it provides an superlative fit during the model period, its performance is much worse (even compared to the competing methods) on the holdout period. While the holdout period of one year may appear to be too long, for certain long-term policy decisions, this is appropriate.

6. Summary and Concluding Comments.
In this work we propose a stepwise modeling approach of high frequency data of complex seasonality that integrates different existing methods with estimating seasonality at very high frequency. The work illustrates an additive form of seasonality, but multiplicative form of seasonality is also indicated. The differentiating level of frequency at which the two stages of computation are to be carried out may be dictated by the available computing power. However the presented analysis establishes that a reasonably efficient forecasting method may be arrived at in stepwise modeling.
The work will be extended including numerical computation involving multiplicative form of seasonality. Also, a seasonal ARIMA model may work better with average weekly data (shorter period of seasonality, 52 as opposed to 365) and subsequently (within week) seasonality due to day of the week may be estimated along the same lines as that of hourly (and within hour) seasonality and incorporated in the final forecast. Also accuracy of the proposed stepwise method will be investigated with greater details including situation where the model may be recursively adapted with additional bulk of data points.

References.
Au, S.T., Ma, G.Q., & Yeung, S.N. (2011) Automatic Forecast of Double Seasonal Time Series with Applications on Mobility Network Traffic Prediction http://www.research.att.com/techdocs/TD_100381.pdf
Gould, P.G., Koehler, A.B., Ord, J.K., Snyder, R.D., Hyndman, R.J., & Araghi, F.V. (2008). Forecasting time series with multiple seasonal patterns. *European Journal of Operational Research* 191, 207 – 222.
Livera, A.M.D., Hyndman, R.J., & Snyder R.D. (2011). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing *Journal of the American Statistical Association*, 106 (496), 1513 – 1527.
Soares, L.J., & Souza, L.R. (2006). Forecasting electricity demand using generalized long memory. *International Journal of Forecasting* 22, 17 – 28.
Taylor, J.W. (2003). Short-term Electricity Demand Forecasting Using Double Seasonal Exponential Smoothing *The Journal of the Operational Research Society*, 54(8), 799 – 805.