elec_consumption

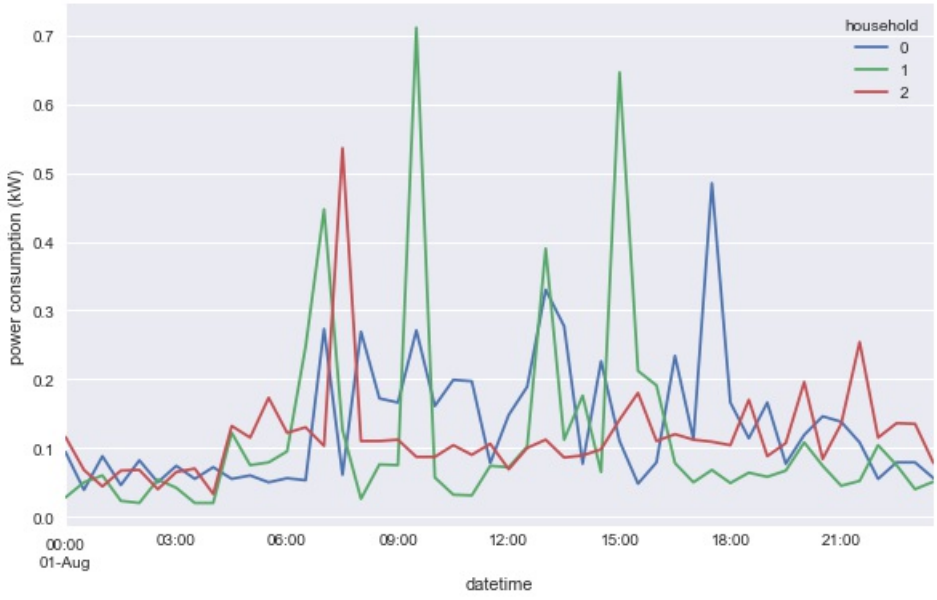
1. Introduction

There are 500 power consumption profiles from 500 households (units) spanning 122 days (Aug 1, 2017 to Nov 30, 2017). Clustering and regression for household power consumption profiles are conducted.

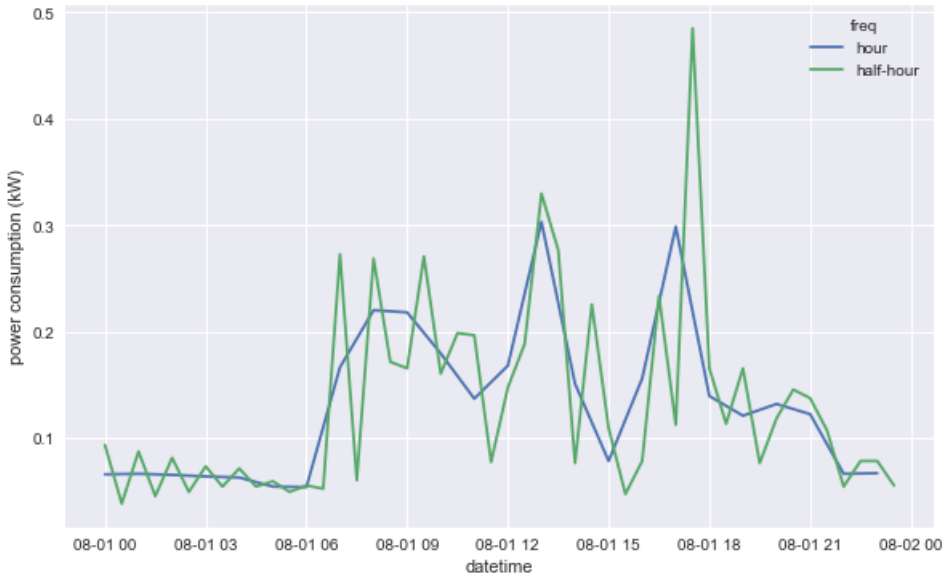
First five observations of first five households are previewed. Physical unit of entries is not mentioned in the data set, and is presumed to be kiloWatt (kW). The physical unit only matters in interpretation because forecasted values are more or less the same.

household	0	1	2	3	4
datetime					
2017-08-01 00:00:00	0.094	0.028	0.116	0.096	0.189
2017-08-01 00:30:00	0.039	0.050	0.068	0.077	0.156
2017-08-01 01:00:00	0.088	0.060	0.044	0.095	0.118
2017-08-01 01:30:00	0.046	0.023	0.067	0.092	0.145
2017-08-01 02:00:00	0.082	0.020	0.068	0.085	0.153

Power consumption profiles of household 0, 1, 2 on Aug 1, 2017 are plotted:

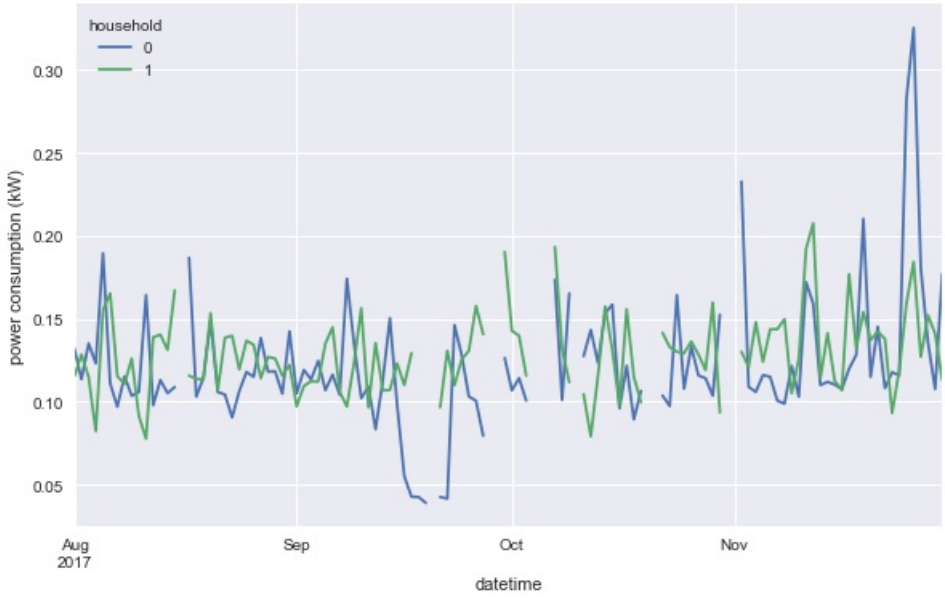


Here is the hourly down-sampled averaged profile of 0 on Aug 1, 2017. It is compared to the original profile.

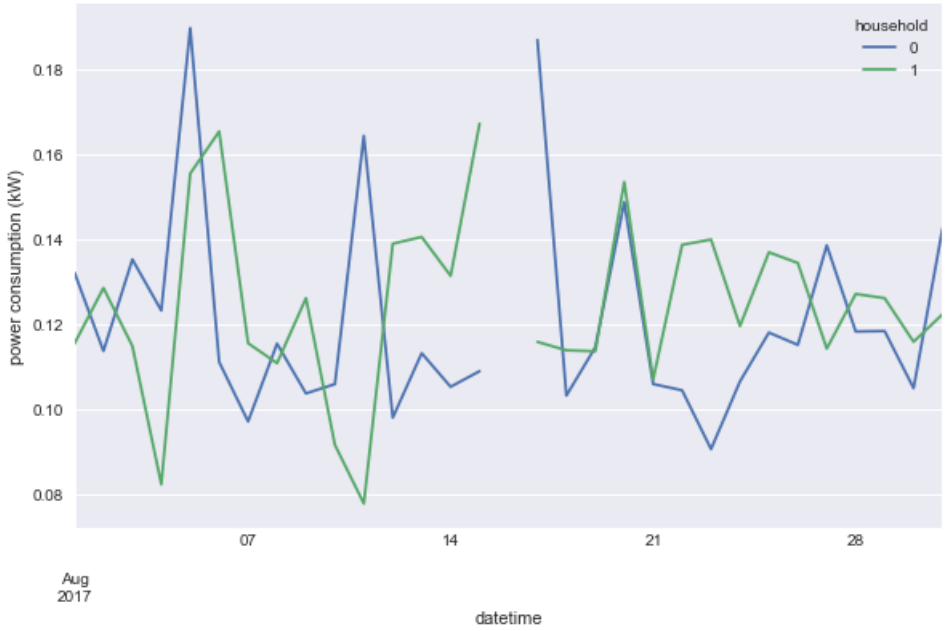


2. Daily Down-Sampled Averaged Profiles

Here are daily down-sampled averaged profiles of households 0 and 1 for the whole period. There are some days when there is no entries. Such missing entries are discussed in the following section.



Entries in Aug, 2017 are zoomed in.

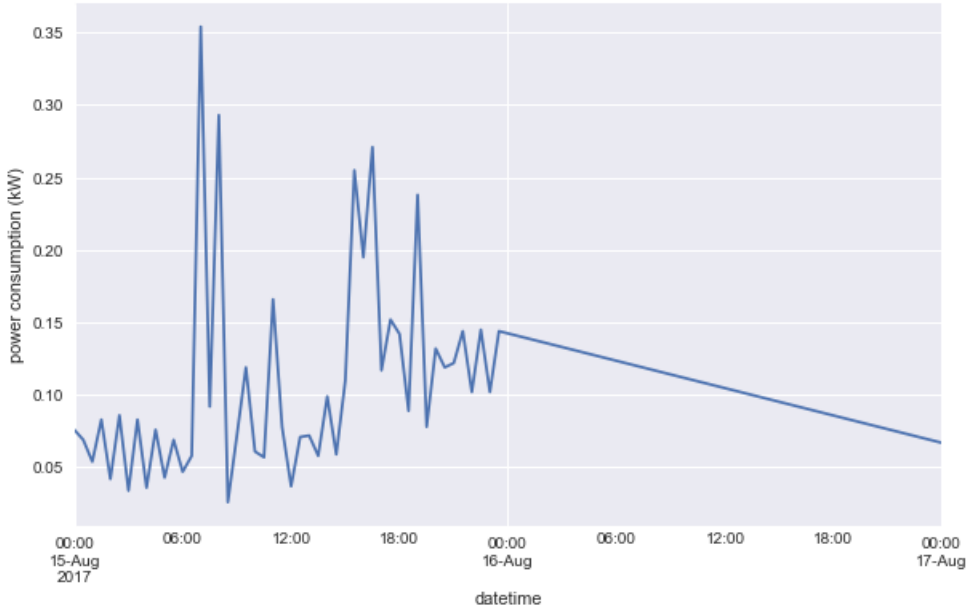


3. Handle Missing Entries

The distribution of missing entries can be summarised as:

- For households 162, 428 and 432, two entries in different sets of dates are missed".
- For all the other households, 48 entries in different sets of dates are missed.
- There are only 186 time points when data is complete.

Linear interpolation is not very useful. For example, how entries Aug 16, 2017 for household 0 are filled can be plotted. Such interpolation can distored both clustering and regression results.



Clustering

1. Introduction

There are at least two challenges in this task:

1. Too many missing days distribute differently among units.
2. All series are non-stationary because of non-business-days.

Why not k-Means

There are dedicated implementation of k-means clustering for time series, like `tslearn.clustering.TimeSeriesKMeans`. However, missing entries must be filled with meaningfull values first, or the algorithm does not work. As shown before, there are only 186 time index when entries are complete. Clearly, neither to fill with 0 nor to interpolate is a good choice in terms of time series.

Monte Carlo simulations can be used, but the resulted effect is not known. The argument is that such common clustering algorithms rely on calculation of cluster centres, which involve multiple units at the same time. Nevertheless, the joint set of missing days from those units might be large compared to that from pair-wise units, so simulated entries exist in more time index, distorting cluster centres.

Moreover, most methods to validate cluster results are based on original data, so they cannot be used, either. Probably the only option left is to validate based on pair-wise similarity measures, like Euclidean distance and Pearson correlation.

For these reasons, it is a good idea to use agglomerative hierarchical clustering with single linkage, which is discussed in the following section.

2. Agglomerative Hierarchical (Single Linkage)

Agglomerative hierarchical clustering (AHC) using single linkage relies on some distance matrix resulted from original data only. Entries in such distance matrix measure the similarity of some unit pair according to a pre-defined criterion. For similarity between time series, Pearson correlation, instead of Euclidean distance, is usually used.

To compute pair-wise correlation, data points with NaNs are removed. Resulted distance matrix has two NaN entries, (463, 485) and (485, 463). They are set as 0 for now.

When number of clusters is set to 4, for example, differences in sizes -- 497, 1, 1, 1 respectively -- are significant. The reason is that units in last three clusters do not have strong correlation with any other unit. That is, they are outliers. Though their distribution among clusters does not have a huge impact, they prevent the algorithm to process units in the first big cluster.

A method based on weighted graphs is proposed in the following section to handle this issue.

3. Max Spanning Tree on Distance Matrix

AHC using single linkage is an optimisation problem in essence. To run max spanning tree on distance matrix yields better result, actually. Another advantage of this approach is that the complete graph (with lightgrey-coloured edges in the following figure), which represesnts the distance matrix, provides a more intuitive way to choose clusters.

Demo with Units 70 ~ 99

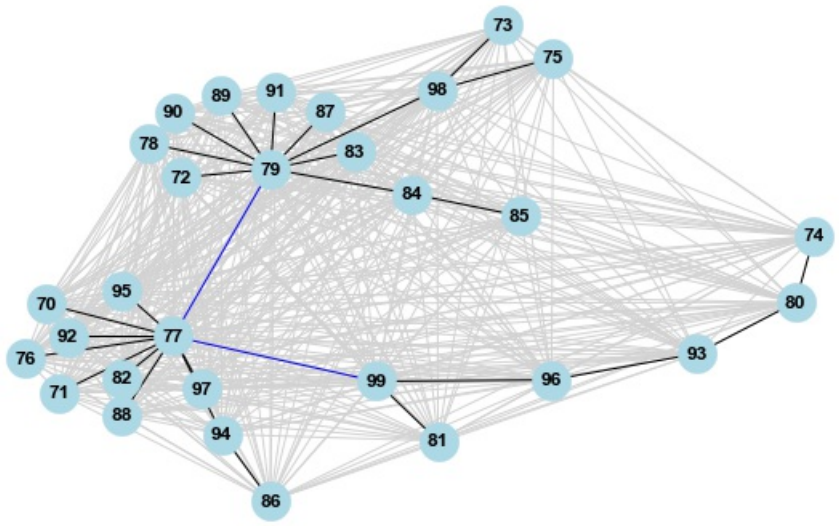
Units 70 ~ 99 are considered in the following example, resulting a clearer figure.

The min spanning tree of the complete graph is highlighted by black edges. The following table highlights two units

with high degrees: 77 and 79. They can be seen as representative units for two clusters. Besides, based on the graph figure, it seems to be a good idea to have another cluster, represented by unit 99.

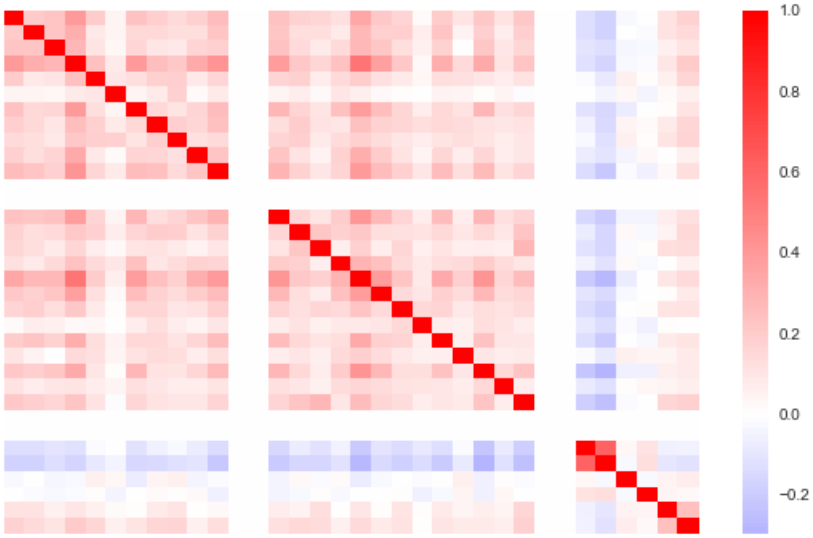
	77	79	98	99	93	94	84	96	80	92
degree	11	10	3	3	2	2	2	2	2	1

Weakest links between such units (two blue edges) are found. If they are removed, the tree becomes a forest with three components, which correspond to three clusters.



The resulted cluster with household 77 being representative unit, for example, has another 10 members: 71, 76, 77, 82, 86, 88, 92, 94, 95, and 97.

The distance matrix can be sorted according to clusters. Here is a heatmap for the previous example. Correlations between members within clusters are shown by three diagonal block matrices, whose entries seem to have higher values. In constrast, off-diagonal block matrices represent inter-cluster correlations, which tend to be low and even negative. So the clustering result is satisfying.



To determine the number of clusters is not a trivial task. With distance matrix and max spanning tree visualised, it is relatively easier to try different numbers and validate results using heatmaps.

4. Effect of Different Resolutions

All the units are to be clustered in this section. Hourly and daily down-sampled profiles are used respectively in first two parts, and results are compared.

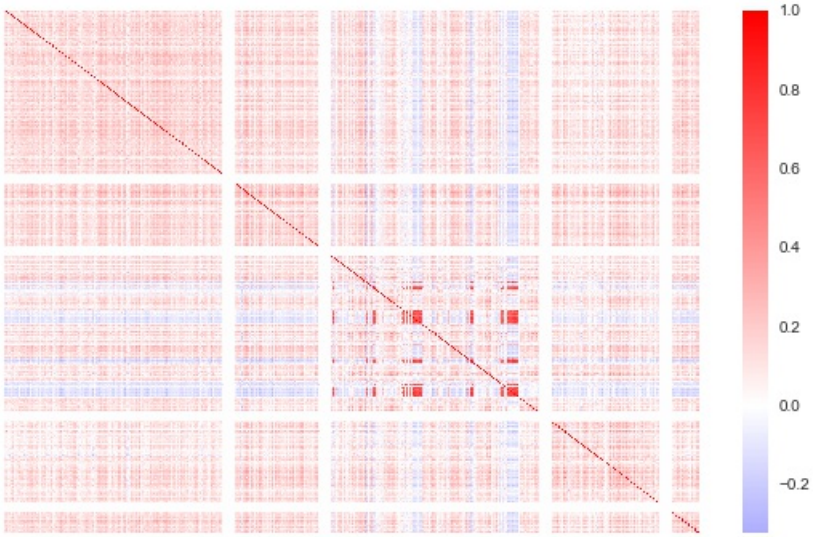
Using Hourly Profiles

Here are nodes with high degrees, which are seen as potential representative units:

	3	79	77	54	197	68	48	222	223	457
degree	100	46	35	24	16	12	12	10	9	9

Will choose 3, 79, 77, 54, and 197 as representative households. It results in 5 clusters.

Here is a heatmap for sorted correlation matrix after clustering. Results are satisfying.

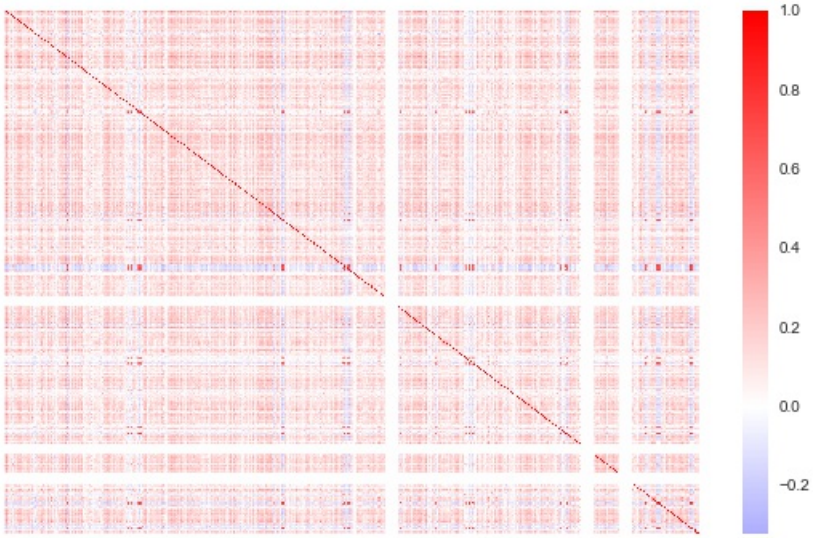


Using Daily Profiles

Here are units with high degrees, and household 80, 74, 155, 249 are chosen as potential representative units. So number of clusters is 4.

	80	74	155	249	388	68	456	465	478	222
degree	31	17	10	10	9	9	8	8	7	7

Here is a heatmap for sorted correlation matrix after clustering.



Compare Results

Units are usually labelled by integers in clustering result, and integer values are irrelevant. Some metrics can be used to quantify difference between two clustering results. Three symmetric metrics are used here, because the true label is unknown.

Adjusted Rand index is 1 when two clustering results are exactly the same, and being -1 corresponds to the biggest

difference possible. There are another two indices. Normalized mutual information (NMI) is often used in the literature, while adjusted mutual information (AMI) was proposed more recently and is normalized against chance. (1 is best; negative is bad)

	index	value	when results are different	when results are similar
rand	adjusted Rand index	-0.015766	close to -1	close to 1
AMI	adjusted mutual information	0.016228	being negative	close to 1
NMI	normalized mutual information	0.026368	being negative	close to 1

So, two clustering results are quite different. There are multiple reasons behind:

- Lots of variation is lost when profiles are down-sampled in daily resolution.
- Outliers (spikes) may distort two results in different manners.
- The fact that missing entries come in batchs has significant impact on clustering with hourly profiles.

How to handle and analyse such issues is discussed in the next section.

5. Future Work

- Multiple results with different number of clusters should be investigated.
- **Structured clustering** is to cluster based on parameters in statistical models. It is especially handy to cluster time series.
- Pearson coefficients are based on original profiles for now. Scaling (amplitude) and translation (offset) variances most likely dominate clustering results. Preprocessing can be applied to reduce such effect.
- Sum of inter-cluster correlations should be minimised at the same time, the problem can be formulated as an MILP, which, however, can be NP-hard.
- Outlier can be detected and removed, reducing their impact on squared terms in some commands.

6. Results

Clusters based on hourly profiles:

```
3: [1, 3, 8, 10, 16, 25, 26, 29, 32, 36, 40, 41, 42, 45, 49, 50, 55, 59, 61, 65, 67, 71, 72, 73,
76, 78, 82, 84, 86, 87, 92, 94, 95, 96, 99, 106, 110, 111, 113, 129, 134, 137, 138, 139, 141, 143
, 146, 147, 155, 156, 157, 159, 161, 164, 169, 177, 184, 189, 193, 198, 199, 203, 205, 206, 209,
211, 213, 217, 220, 223, 224, 225, 226, 228, 230, 232, 233, 234, 235, 238, 241, 243, 246, 255, 2
56, 257, 259, 262, 264, 266, 268, 270, 278, 279, 280, 283, 285, 290, 295, 297, 298, 299, 301, 30
5, 306, 307, 308, 312, 313, 315, 316, 318, 322, 323, 325, 327, 330, 331, 332, 334, 335, 336, 337
, 340, 341, 342, 348, 349, 351, 361, 363, 364, 369, 373, 375, 379, 380, 382, 399, 400, 406, 408,
410, 411, 414, 417, 418, 421, 427, 428, 429, 431, 435, 437, 446, 448, 452, 454, 459, 460, 467, 4
79, 480, 481, 483, 488, 489, 490, 492, 499]

79: [0, 9, 12, 21, 27, 38, 46, 53, 63, 79, 83, 85, 90, 91, 112, 131, 132, 133, 136, 140, 144, 148
, 152, 162, 163, 165, 178, 179, 204, 221, 227, 236, 239, 244, 245, 252, 265, 272, 274, 286, 293,
296, 300, 304, 314, 321, 329, 338, 343, 347, 352, 360, 368, 377, 381, 396, 403, 422, 430, 441, 4
45, 447, 469, 470, 494]

77: [4, 5, 6, 13, 14, 17, 19, 20, 23, 24, 30, 34, 35, 39, 43, 52, 57, 58, 60, 62, 64, 66, 68, 70,
74, 75, 77, 80, 93, 97, 98, 100, 101, 102, 103, 104, 105, 107, 108, 109, 119, 120, 121, 123, 125
, 130, 142, 145, 151, 154, 158, 166, 167, 168, 170, 171, 172, 173, 174, 176, 180, 181, 182, 183,
185, 186, 187, 188, 191, 192, 194, 195, 202, 210, 212, 214, 215, 216, 218, 219, 222, 229, 231, 2
37, 247, 248, 249, 250, 251, 253, 254, 263, 269, 273, 284, 287, 292, 294, 309, 310, 317, 328, 33
3, 354, 355, 358, 362, 367, 371, 372, 374, 376, 378, 384, 385, 390, 392, 393, 397, 402, 413, 415
, 419, 420, 424, 425, 426, 433, 434, 436, 438, 439, 440, 442, 443, 444, 450, 451, 453, 455, 456,
457, 458, 461, 462, 463, 464, 465, 471, 473, 474, 478, 482, 484, 485, 491, 493, 495, 496, 497, 4
98]

54: [2, 7, 11, 15, 31, 33, 37, 44, 47, 48, 54, 56, 69, 81, 89, 114, 115, 116, 117, 122, 124, 126,
127, 128, 149, 153, 160, 175, 190, 196, 200, 207, 240, 258, 260, 267, 271, 276, 277, 281, 282, 2
88, 289, 302, 303, 311, 320, 324, 326, 339, 344, 345, 346, 350, 353, 356, 357, 359, 365, 366, 37
0, 383, 386, 387, 388, 389, 391, 394, 395, 398, 401, 404, 407, 409, 412, 416, 423, 432, 449, 466
, 468, 472, 475, 477]

197: [18, 22, 28, 51, 88, 118, 135, 150, 197, 201, 208, 242, 261, 275, 291, 319, 405, 476, 486,
487]
```

Clusters based on daily profiles:

```
80: [1, 2, 3, 8, 10, 11, 12, 13, 17, 18, 19, 21, 22, 24, 25, 26, 28, 29, 30, 31, 32, 33, 35, 37,
38, 40, 41, 42, 45, 48, 50, 51, 52, 53, 55, 56, 57, 58, 59, 62, 63, 64, 67, 69, 70, 71, 76, 77, 8
0, 83, 84, 86, 88, 89, 90, 91, 94, 97, 98, 100, 104, 106, 107, 108, 109, 110, 112, 113, 114, 115
, 116, 117, 118, 120, 128, 129, 132, 134, 136, 141, 146, 147, 148, 149, 151, 152, 153, 158, 159,
160, 161, 163, 165, 166, 168, 172, 173, 174, 175, 177, 178, 179, 181, 185, 188, 190, 193, 195, 1
97, 198, 199, 200, 201, 202, 203, 205, 206, 209, 212, 217, 218, 221, 223, 225, 227, 228, 229, 23
0, 232, 234, 238, 239, 240, 242, 243, 246, 247, 248, 252, 253, 254, 256, 257, 259, 261, 264, 268
, 269, 271, 272, 274, 275, 276, 277, 279, 281, 282, 283, 284, 286, 288, 292, 293, 294, 295, 297,
298, 299, 300, 301, 302, 304, 306, 307, 308, 311, 312, 313, 314, 315, 316, 317, 319, 324, 325, 3
28, 329, 330, 332, 333, 334, 337, 338, 340, 342, 344, 345, 346, 350, 353, 355, 356, 357, 358, 35
9, 360, 362, 363, 365, 366, 367, 369, 371, 372, 373, 374, 377, 378, 380, 381, 383, 386, 387, 388
, 393, 394, 395, 396, 397, 398, 399, 400, 401, 403, 404, 408, 409, 410, 412, 413, 414, 415, 416,
417, 418, 421, 422, 424, 425, 426, 428, 432, 434, 436, 438, 445, 447, 448, 451, 452, 453, 454, 4
55, 456, 458, 459, 460, 463, 464, 465, 466, 469, 471, 472, 474, 475, 476, 478, 481, 482, 485, 48
6, 487, 488, 489, 491, 492, 493, 494, 498, 499]

74: [0, 4, 5, 9, 14, 15, 20, 23, 27, 39, 43, 46, 60, 61, 65, 66, 68, 72, 74, 75, 78, 79, 81, 82,
85, 93, 95, 99, 101, 103, 105, 111, 121, 122, 123, 124, 125, 126, 127, 130, 131, 133, 135, 139,
140, 143, 144, 150, 156, 164, 167, 169, 170, 176, 180, 182, 189, 191, 194, 196, 204, 207, 208, 2
10, 211, 214, 215, 216, 219, 222, 224, 226, 231, 235, 236, 237, 241, 244, 245, 251, 255, 258, 26
0, 262, 263, 265, 267, 270, 289, 291, 296, 303, 310, 321, 323, 326, 331, 335, 341, 343, 347, 348
, 352, 354, 368, 370, 375, 376, 379, 382, 385, 389, 391, 392, 402, 405, 411, 430, 431, 433, 435,
437, 440, 441, 442, 443, 444, 449, 450, 462, 473, 479, 480, 483, 484, 490, 495, 496, 497]

155: [6, 44, 92, 155, 157, 162, 184, 233, 266, 280, 287, 309, 318, 322, 327, 361, 364, 384, 406,
407]

249: [7, 16, 34, 36, 47, 49, 54, 73, 87, 96, 102, 119, 137, 138, 142, 145, 154, 171, 183, 186, 18
7, 192, 213, 220, 249, 250, 273, 278, 285, 290, 305, 320, 336, 339, 349, 351, 390, 419, 420, 423
, 427, 429, 439, 446, 457, 461, 467, 468, 470, 477]
```

Regression

1. Introduction

Methods for two sets of forecasts, using hourly and daily profiles respectively, are discussed here.

Why not Multivariate Time Series

It is known that there is no relationship between profiles from different households, because their behaviours are not directly correlated. The only regressor considered should be time. Others like temperature may contribute, but are unknown for now.

So it is not necessary to use vector auto regression (VAR). Profiles can be modelled independently.

Why Facebook `Prophet`

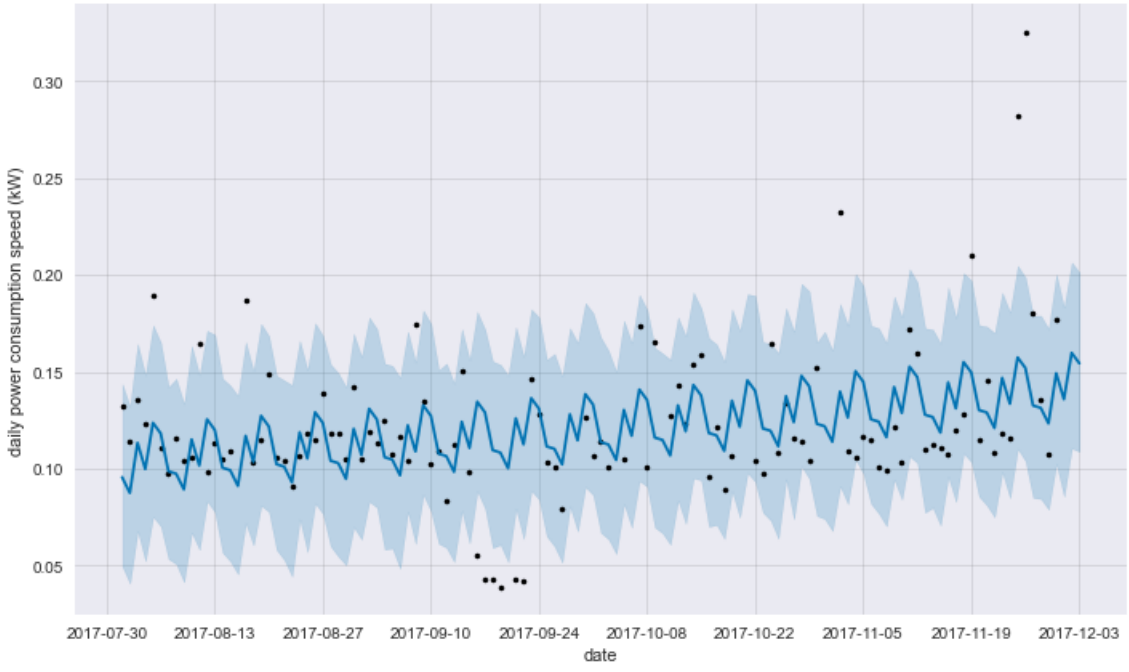
`Prophet` is an open-source package for modelling and forecasting univariate time series by Facebook. Its model structure is similar to that of generalised additive models (GAM), where distinct non-linear terms are integrated in a linear way. `Prophet` works best when the series shows strong and multiple seasonality, and it is robust to missing values and effect from holidays, making it the perfect choice in our case.

2. Demo: Daily Profile of Household 0

Though profiles in the same cluster tend to show similar patterns, it is relatively easy to build different statistical model for different households. Daily profiles of household 0 is modelled by Facebook `Prophet` and seasonal ARIMA in this section.

Facebook `Prophet`

Here is resulted Line plot from the model. Point forecasts and confidence intervals are plotted.

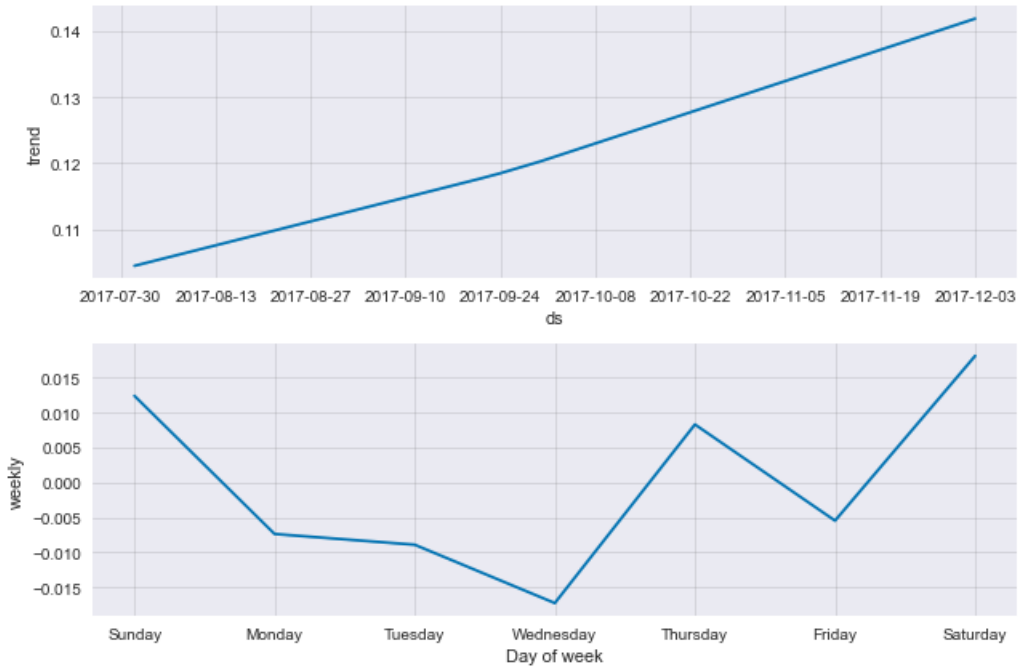


Forcasts and confidence intervals for future 3 days:

	yhat	yhat_lower	yhat_upper
date			
2017-12-01	0.135803	0.085666	0.183257
2017-12-02	0.159741	0.110511	0.206479
2017-12-03	0.154350	0.108852	0.201318

According to the previous figure, there is a trending and a weekly seasonal components, which are shown in following two figures respectively. The trending component increased from 0.104 to 0.141 in 122 days. That is, the average power consumption increased by 34.93 % in merely 4 months. Such trending effect must be taken into account.

Surprisingly, there are variations in power consumption over one week. Household 0 tends to consume more during weekends and Thursday. The difference between values on Saturday and on Wednesday amounts to 28.98 % of the mean power consumption, so this effect is essential as well.

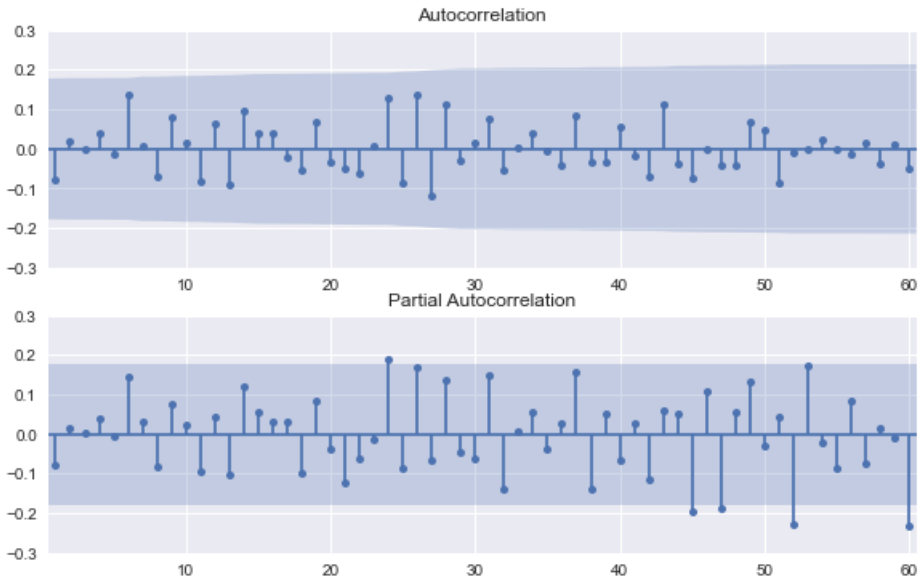


Seasonal ARIMA

It is impossible to plot ACF and PACF for original series because of missing entries. As discussed before, they are adjacent so it is hard to fill with meaningful values. Previous result provides a starting point. To have a weekly season is a good idea, and it should be "integrated", in order to remove the long-term trend.

For example, an integrated SARIMA (with AR1, MA1 and weekly seasonal AR1) can be built. The seasonal AR1 term (value 0.1603, std error 0.145, pvalue 0.269) is a bit controversial, but stays due to the previous finding. The normality and heteroskedasticity assumptions are both rejected, so this is not good model.

Good news is that there is nothing left according to ACF and PACF. This example is actually the final model selected after a manual model selection.



Pseudo Out-of-Sample Validation

There are systematic ways for validation. It is challenging to conduct a cross validation for time series models, especially when there is long-term trend. Pseudo out-of-sample validation is used instead. Here is a description from `statsmodels`:

"A common use case is to cross-validate forecasting methods by performing h-step-ahead forecasts recursively using the following process"

- Fit model parameters on a training sample
- Produce h-step-ahead forecasts from the end of that sample
- Compare forecasts against test dataset to compute error rate
- Expand the sample to include the next observation, and repeat

MSE of 3-step forecast using Facebook Prophet is 0.00356, and its root is 48.83 % of mean power consumption. So this is not a good model.

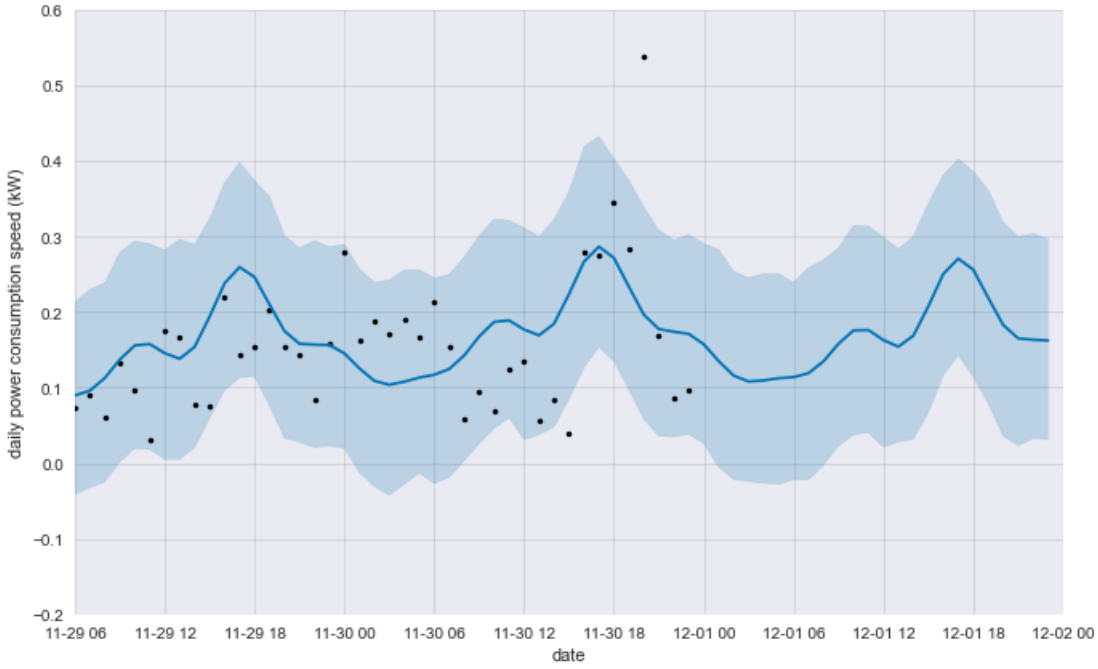
MSE of 3-step forecast using seasonal ARIMA is 0.00370, which is 3.94 % higher than MSE using Prophet, and its root is 49.78 % of mean power consumption. Despite of the increment, this model structure will be used to model all other daily profiles.

---

3. Demo: Hourly Profile of Household 0

`Prophet` can even be used to model time series with hourly (sub-daily) resolution.

Here is a line plot for in-sample and out-of-sample forecasts and confidence intervals.

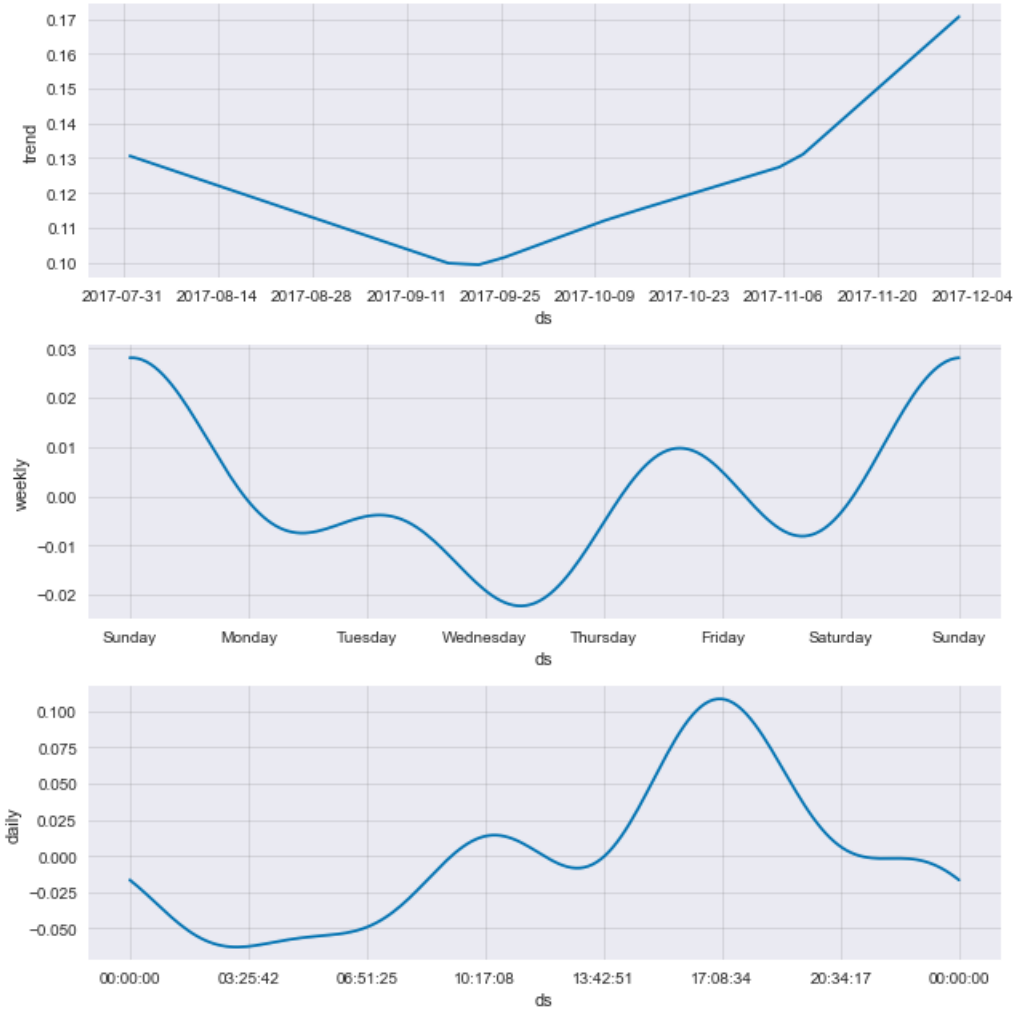


Out-of-sample forecasts for future 4 hours are:

	yhat	yhat_lower	yhat_upper
datetime			
2017-12-01 00:00:00	0.157450	0.026090	0.290846
2017-12-01 01:00:00	0.135131	-0.005102	0.282784
2017-12-01 02:00:00	0.115922	-0.021668	0.253831
2017-12-01 03:00:00	0.108154	-0.024035	0.245774

Multiple Seasonality

Here are plots for one trending and two seasonality components.



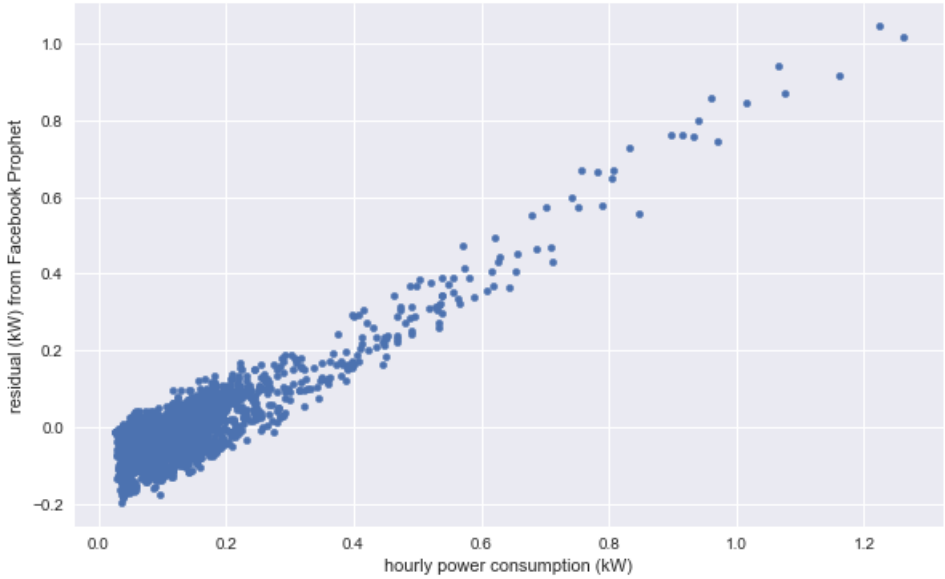
Unfortunately, it is impossible to apply seasonal ARIMA, because of multiple seasonality.

Specification Tests

The assumption that residuals from Prophet model is rejected.

According to White's test for Heteroscedasticity, variance of residuals does vary with power consumptions.

As suggested by the previous test, variance of residuals increases with respect to power consumptions. That is, when power consumption is high, the in-sample forcast is expected to have a high residual. The same goes for out-of-sample forecasts.



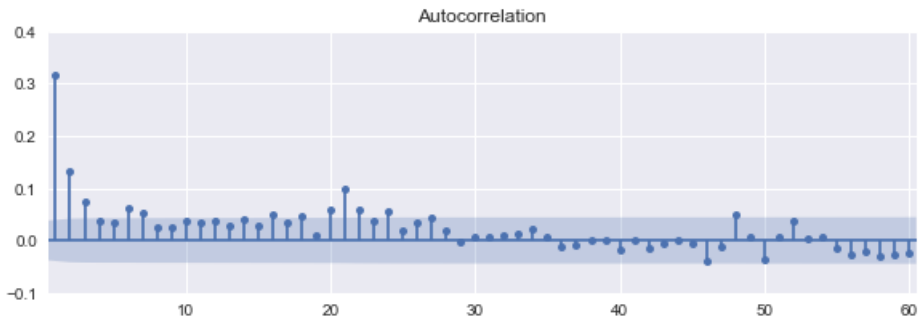
In spite of the previous finding, when residuals are compared to corresponding values, they can be 4 times larger. So

it is hard to say that more precise forecasts can be made when power consumption is low (usually during nights).

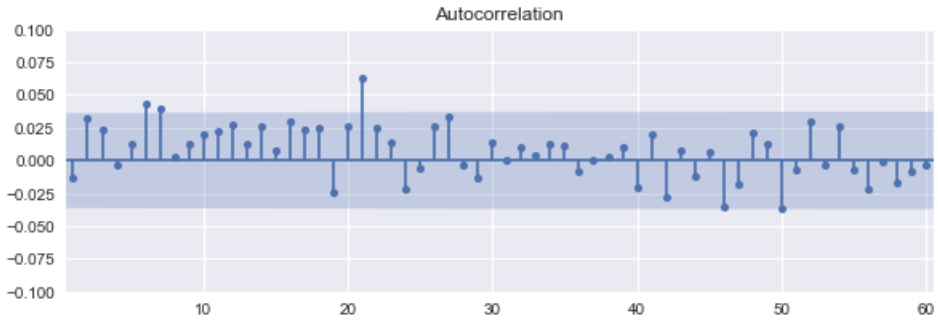
Autocorrelation Tests and ARIMA Model

All the Ljung-Box tests for lags 1, 2, 3, 4, 5, 24, and 168 reject the assumption that the residual is not correlated.

The test is supported by ACF, when missing values are dropped.



Such information can be modelled by a SARIMA with AR1, AR1 (24), MA1 (24). Independence assumptions in lags 1, 2, 3, 4, and 5 are not rejected, so the model has been improved.



However, the model is still not satisfying, because it does not pass normality test and heteroskedasticity test.

4. Future Work

- To combine multiple series belonging to different units is not a good idea, through there is a simple way to do it (estimating same model over multiple time series, `crossvalidated`).
- Apply more flexible methods and tune parameters using pseudo out-of-sample validation.

5. Results

Two sets of forecasts for 500 households are made:

- 3-step forecasts based on daily profile, in a comma-separated CSV file
- 4-step forecasts based on hourly profile, in another comma-separated CSV file