Chapter 6
Load Data: Preparation, Analysis and Feature Generation

Chapter 5 introduced the general definition of a forecast and many of the concepts and categorisation of the different types and features of time series forecasts. To develop an appropriate model requires identifying genuine patterns and relationships in the time series data. This requires a detailed investigation and analysis of the data, since selecting the correct input features is, arguably, at least as important as selecting the most appropriate forecast model. This data analysis and feature generation is the main focus of this chapter. However, prior to this it is important to understand whether the data is of sufficient quality to allow the training of a good forecast model. The next section begins by considering important features of high quality data and potential preprocessing which may be required. This is followed by methods for analysing the load data and identifying features which may be useful inputs to a forecast model.

6.1 Preparation and Pre-processing

Before training the models, the data must be assessed and cleaned, otherwise the forecasts will be trained on flawed data and the outputs will be inaccurate, misleading, or meaningless. As the machine learning mantra succinctly puts it 'garbage in, garbage out'. Unlike contrived data often used in textbook examples to demonstrate techniques, real data is messy, has little-to-no formatting, and is rarely error free.

Some of the most common data issues are:

1. **Missing values**—Due to errors in communications or faults in the monitoring equipment, recorded data is rarely complete.
2. **Extreme values**—These can be excessively large, or excessively small values. This is particularly tricky to determine, especially for highly volatile data like the low voltage energy demand analysed in this book. Some extreme values can be identified since they are outside the parameters of the system, e.g. exceeding the circuit breaker limits of a house and therefore technically not possible (unless something is wrong with the circuit breaker of course!). However, for the majority

S. Haben et al., *Core Concepts and Methods in Load Forecasting*,
https://doi.org/10.1007/978-3-031-27852-5_6

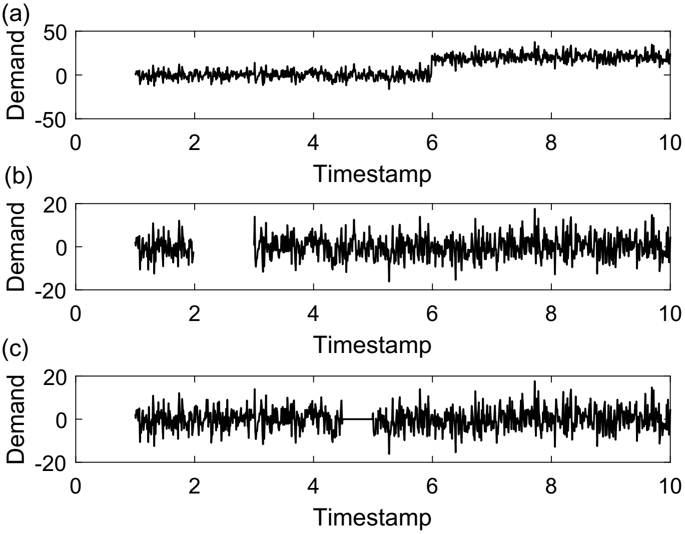

Fig. 6.1 Examples of time series plots with **a** a change in the level of demand, **b** with missing data, and **c** with a section of constant values

of the time, identification is tricky since it is often not possible to confirm whether a value is valid or has been recorded incorrectly.

3. **Anomalous values**—Depending on the application, some values are clearly incorrect. For example on an LV network feeder with no generation all the demand should be positive, with no values smaller than zero. Hence negative demand is clearly impossible and, on feeders with large demand, recorded values of zero should be treated with suspicion.

Examples of some time series with possible anomalous and erroneous data is shown in Fig. 6.1 for difference situations. Plot (a) shows a demand level increase at a particular point in the time series. This could be a fix to the monitoring equipment, or could be a genuine change caused by, say, the occupants installing a new high demand appliance (like a heat pump or electric vehicle) which causes an overall increase in the baseline demand (also referred to as concept shift, Sect. 13.6.3). Plot (b) shows an example of when data is missing from the time series data. Sometimes data is missing for only occasional points, or like in this example it can be over a long period of time. The latter can occur when monitoring equipment experiences a fault. Finally plot (c) shows a section of constant values. Again these can be caused by sensor equipment faults, however they are much harder to detect, especially if they only occur over short periods, since they may not be obvious from basic analysis or simple time series plots.

The impact of these erroneous values can have a detrimental effect on the quality of the models and the accuracy of the corresponding forecasts. For this reason it is important that their effect is mitigated or removed which is typically done by deleting them from the data set. However, removing values that are not erroneous could reduce the accuracy of the forecasts. This is especially true when the aim is to accurately forecast extreme values such as peaks. Further, when producing probabilistic forecasts it could mean that the tails of the distribution are not properly calibrated.

The topic of data pre-processing is a complex area in its own right and much of it is beyond the scope of this book. Some extra references are included to more advanced techniques in Appendix D.1. For this book, it is sufficient to concentrate on some simple but common methods for identifying and cleaning up time series data.

6.1.1 Outlier Identification

Missing values are obviously easy to identify and anomalous values will also be simple to check for. For that reason the focus is on identifying outliers which in our case will be those values which are unusually large or small. Visual methods are a common way to determine which points are outliers, but this can be considered a little subjective and hence could lead to biases. For stationary time series a more systematic identification is to identify those points which are further from the central value of the data than would be expected given the spread of the sample of the data. One of the most common ways of doing this is to identify those points which are a few sample standard deviations from the sample mean (see Sect. 3.5). Recall for normally distributed data, 95% of values are within 2 standard deviations of the mean, and 99% of values are withing 3 standard deviations. However, often the assumption of data following a normal distribution is not valid and hence the true distribution of the data is unknown. Despite this, the standard deviation approach can often be a useful measure to understanding which points may be outliers but care should be taken if the data is skewed and not symmetrical. For data which is not normally distributed a more robust, but less commonly used way, is to consider how many interquartile ranges the data is from the sample median (Sect. 3.2). More generally estimating which quantiles the data lies in can also identify outliers. The quantiles are often more robust to outliers than the standard deviation approach and therefore can be preferable for defining thresholds in the data.

An illustration of using two standard deviations from the mean to detect large values is shown in Fig. 6.2. In this relatively contrived example, three points are clearly outliers and these have been successfully identified by using the standard deviation criteria.

The process is much more complicated when the data is not stationary. When there are simple and obvious trends in the time series, but the variance is fixed over time, then the standard deviation can be calculated from the detrended series. The detrended series is simply the difference between the observations and a model fit, i.e. the model residuals (Sect. 5.2). Two examples are shown in Fig. 6.3 for one series with linearly increasing trend and one with a simply single periodic behaviour. Also included are the lines of best fit as well as the lines indicating distances of two standard

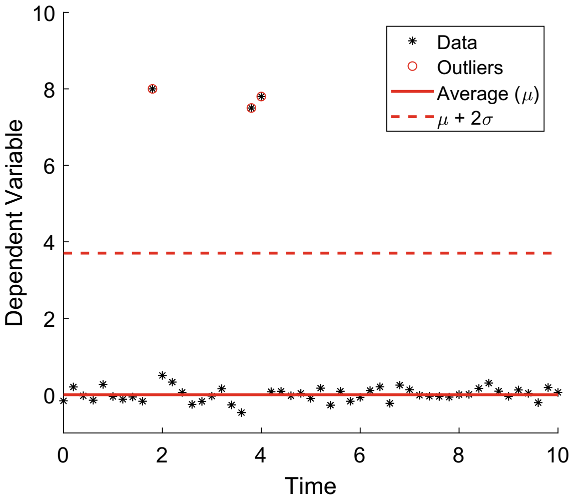

Fig. 6.2 Example of a stationary time series (black markers) with three outliers (red circles). Also included is the average value (solid red line) and the average plus two standard deviations of the points. Any points further than two standard deviations from the mean are labelled as outliers

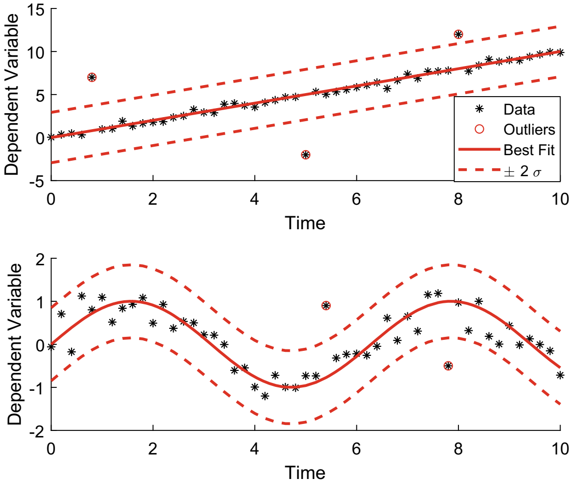

Fig. 6.3 Examples of non-stationary time series (black markers) with outliers (red circles) for a time series with linear trend (top) and one with seasonality (bottom). Also included are the lines of best fit for the data

deviations of the residual series from the line of best fit. Without the detrending it would not be clear that some of these points are outliers since they are within the range of the full data set. A complication with this approach is that the trend fit may be poor because of training on data with outliers/anomalous values, especially if there are a lot of them relative to the full dataset. In that case the extreme values may not register as outliers.

For more general time series it is not possible to easily detrend in order to identify outliers and may not be advisable as it requires making assumption about the underlying process. Hence the forecaster may include some of their biases in the model assumptions and incorrectly label some normal values as outliers. One approach is to develop a forecast model and compare the performance with, and without, the assumed outliers and analyse their effect on the accuracy. There are other more sophisticated models for identifying outliers beyond the scope of this book, some of which can be found in the further reading in Appendix D.1.

6.1.2 *Imputation*

After identifying anomalous and outlier data a decision must be made as to whether to remove the values or not. If the values are known to be incorrect then they should definitely be removed from the data set. Otherwise if it cannot be confirmed whether a value is truly an outlier or incorrect value then it is recommended to keep the value in. One possibility is to run the models both with and without the anomalous values to see if the forecast is sensitive to the changes. If there are only a few anomalous values then their inclusion may have little effect on the overall model accuracy anyway. In some cases there may be a large amount of missing data, in which case it may be impossible to produce any good model with decent accuracy. What constitutes "enough" data depends on the type of data, the application and other design elements such as forecast horizon. For short term forecasts (say one day ahead), sometimes a reasonable benchmark forecast can be produced using only four to five weeks of data (See the Case Study in Chap. 14).

When data is missing, or has been removed due to cleaning actions, there will be gaps in the dataset. This causes a number of issues. Firstly, it makes data handling more complicated as techniques must be considered which ignore missing instances, and secondly, it produces potential biases in the data since certain features may be more prominent in the reduced dataset then they otherwise would be. If the missing data is relatively random then there is very little bias introduced and the forecasts can be trained on the reduced data without concern that the models will be skewed by any biases.

An alternative method for dealing with missing values is to insert or *impute* other values. The process is known as **imputation**. This simplifies analysis and model training on the data and there are several different ways to choose the values to insert:

- **Simple Average**—this maintains the sample mean of the data but ignores any trends or seasonalities. A moving average (over a moving window around the data) can be used to better fit any trends.
- **Last value**—this retains the trends in the data and means that the missing values are filled with recent values. However, this ignores any other patterns which may be in the data. It is worth noting that many data acquisition systems used to collect power data often forward fill at regular intervals if they stop receiving data.
- **Seasonal Average**—if the data has seasonality then this retains those features. For example, if the data has daily seasonality then a missing value at 4pm could be filled by using an average of the values at 4pm from the previous days.

- **Regression**—Regression is considered in Chap. 9 but essentially this means using a weighted average of surrounding values to fill the missing data. This allows more complex relationships to be included to impute missing values.
- **Interpolation**—More generally a curve could be fitted (e.g. a polynomial) to the surrounding values of the missing data point and the value on the curve at the missing point can be used to impute the value (See an example of interpolation in Sect. 9.6).

6.1.3 Normalisation and Transformations

Even after cleaning the data, in it's raw form, the data may not be suitable for using directly within a model. For example, for a linear regression model (Sect. 9.3) there are assumptions that the errors follow a Gaussian distribution. This is unlikely to be always true, especially for smart meter data which in some cases has been shown to follow a lognormal distribution (see Sect. 3.1). Further, notice that unless there are reverse power flows at the meter, say due to solar PV generation, then smart meter demand should always be nonnegative (i.e. positive or zero) which is not true for Gaussian distributed data, but will be for a lognormal distribution. Recall (Sect. 3.1) a random variable z has a lognormal distribution if it has PDF of the form

$$f(z) = \frac{1}{z\sqrt{2\pi}\sigma} \exp\left(\frac{-(\ln(z) - \mu)^2}{\sigma^2}\right). \tag{6.29}$$

By definition this just means that the transformed variable e.g.

$$x = (\ln(z) - \mu)/\sigma \tag{6.30}$$

has a Gaussian/Normal distribution. In other words if data is nonnegative, has one long positively skewed tail, applying a lognormal transformation may produce normally distributed data. This in turn may be easier to manipulate and utilise. After training a model on this data an inverse transform can be applied to the forecasts to obtain physically representative data. An example of lognormally distributed data ($\mu = 0$, $\sigma = 0.5$) is shown in the histogram in Fig. 6.4 (left), together with the same data but log-transformed (right). Notice the lognormal transformed data is now symmetric and bell-shaped as expected and now allows negative values.

Another common transformation applied to raw data is **normalisation**, where the data is scaled prior to visualisation and/or modelling development. This can have two main advantages. Firstly, consider a situation where more than one variables is being modelled, but they have very different ranges of values, say one is bound between 0 and 10 and another is between 0 and 1000. In this case, it can be very difficult to visualise or understand the relationships between them due to the extreme difference in their relative variations. Scaling these values can better highlight these relationships.

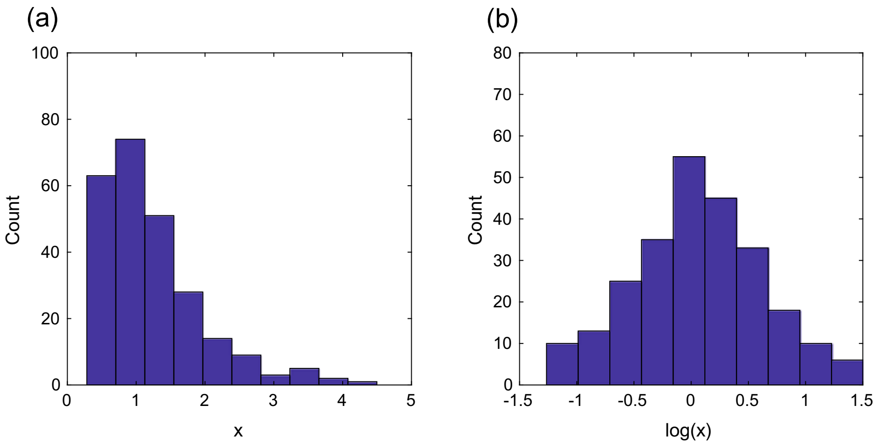

Fig. 6.4 Example of the distribution of a sample of data **a** with a lognormal distribution and then **b** the same data but transformed using the log function

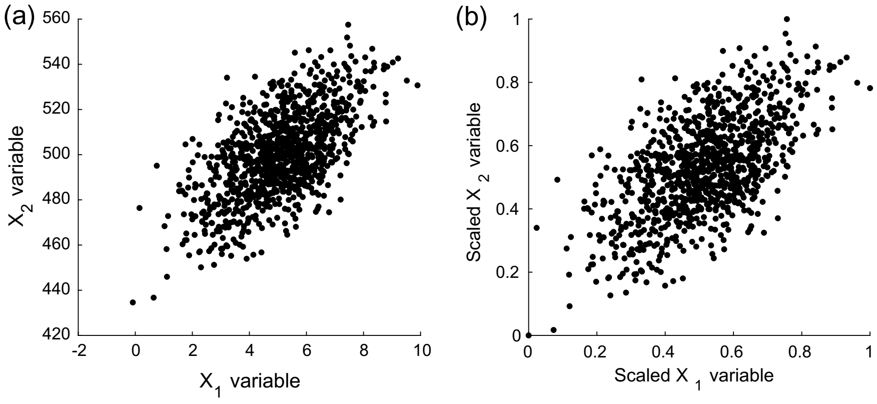

Fig. 6.5 Examples of two variables from a bivariate distribution (**a**) original variables (**b**) the same data but each variable has been scaled to be between 0 and 1

The second major reason for scaling is to help with training the parameters of a model (see Chap. 8 on training). The scale of the data may effect the scale of the parameters. By normalising the data, this restricts the range of the data, and reduces the search space for the optimal parameters. A simple two variable example is shown in Fig. 6.5 where the scatter plot of the original variables is shown in (a) and the rescaled variables are in (b). The scaling has been performed so that each variable is between 0 and 1. In the original data the spread of X_1 is between -2 and 10, in contrast the X_2 variable is between 400 and 500, hence has bigger magnitude and spread. If we were training a linear model, e.g. $y = aX_1 + bX_2$ on this data we would see that b would be relatively small compared to a otherwise the responses to changes in the dependent variables (within their respective ranges) would produce wildly different responses y. It also means that the coefficients therefore have different search ranges. In contrast the normalised variables would reduce the search space for the linear coefficients a, b. Note that the relationships between the variables will still be preserved by the normalisation albeit scaled.

As will be shown below, there are several ways to normalise the data, but an important requirement is that the data can be rescaled back to its original size. One of the most common forms is the **Min-Max Scaler** which transforms the variables into the interval [0, 1]. Assuming that the time series has a maximum and minimum value given by $x_{\max}$ and $x_{\min}$ respectively, then the scaled version of any data point x is given by

$$\hat{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{6.31}$$

and thus the series transforms to one which has the maximum value of one, and the minimum of zero. Note if you have an extreme outlier then $x_{\max}$ and $x_{\min}$ may produce an unsuitable normalisation. In this case it may be worth cleaning/preprocessing the data before performing the scaling (see Sect. 6.1.1). To recover the original values you simply rearrange the calculation:

$$x = x_{\min} + \hat{x}(x_{\max} - x_{\min}). \tag{6.32}$$

Another common methods is the **Standard Scaler** which subtracts the mean of the series, μ, and divides by the standard deviation, σ,

$$\hat{x} = \frac{x - \mu}{\sigma}. \tag{6.33}$$

The new series has mean zero and a unit variance and the values are no longer constrained within [0, 1]. Note that the mean may be strongly effected by outliers so as with the Min-max scaler it may be worth removing them (Sect. 6.1.1) before proceeding with the normalisation.

To avoid the effect of outliers there are other normalisations such as the **Robust Scaler** which uses the median x_{50}, and the interquartile range $x_{75} - x_{25}$ which is the difference between the 25th and 75th percentile (See Sect. 3.2 for further details on percentiles/quantiles). The scaled variables are given by

$$\hat{x} = \frac{x - x_{50}}{x_{75} - x_{25}}. \tag{6.34}$$

This has median zero but note that there may still be outliers in the transformed series.

Now consider how to fit a model to scaled data in a simple example. Consider the linear model

$$y = ax + b, \tag{6.35}$$

for variables x, y and coefficients $a, b \in \mathbb{R}$. Suppose we scale the variables x, y to $\hat{x}$, and $\hat{y}$ respectively. The parameters $\hat{a}, \hat{b}$ are found for the scaled model

$$\hat{y} = \hat{a}\hat{x} + \hat{b}, \tag{6.36}$$

and by substituting the rescaling (e.g. Eq. (6.32)) you can estimate the original simple model. This case is much simpler than others as the relationships are all linear. The transformation may not be possible for other more complicated cases.

6.1.4 Other Pre-processing

As well as dealing with missing and anomalous data there is also some standard pre-processing procedures which also should be considered.

Firstly, many time series forecasting methods depend on the values being spaced equally in time. This is often the case as much monitoring equipment is calibrated to record at uniform intervals. However, if the data is not at uniform intervals the data can be estimated at these time steps by interpolating to the time steps of interest. This obviously creates extra errors (in this case errors from the estimation) but they will be smaller the higher the resolution of the original data, and the smaller the volatility of the data. Unfortunately, there is few techniques and packages for dealing directly with time series data which is not recorded at uniformly spaced intervals. However, since most energy monitoring is either high resolution or is designed for regular intervals the rest of the book will assume the data is equally spaced in time with negligible interpolation errors.

Another common issue is the fact that different input data is defined at different temporal resolutions. For example the load data may be recorded at half hourly intervals but the corresponding temperature data may be at hourly resolution. If the temperature variables are important explanatory input variables for a load forecast then it may be worth resolving both data sets to the common resolution of hourly data.[1] For energy data (kWh) this essentially means summing the data over the two half hours, whereas for average Power (kW) data this would require averaging over the two half hour points. The latter (Power) is the more common representation of load data.

6.2 Feature Selection and Engineering

One of the most important factors for creating an accurate forecast is choosing the most appropriate features to include in the model. In many cases the features are more important than the forecast model used. One option is to include all the

[1] Alternatively the temperature data could be interpolated to half hourly.

available features and fit a model which penalises the number of parameters used in the model, such as information criteria and regularisation techniques (Sects. 8.2.2 and 8.2.4). These methods can be used to select variables and create a parsimonious model. In this section several methods will be considered for identifying potentially important relationships between the dependent and independent variables.

As will be discussed in Sect. 8.1 the aim for choosing features is to achieve a bias-variance trade-off. In other words, to try and include all the important features that describe the data, but not too many that the model will end up overfitting. If the number of potential features to include in the model is large then it may be worth considering the information criteria and regularisation techniques mentioned above to reduce the features to the most important ones.

6.2.1 Domain Knowledge

Some variables can be automatically selected based on the domain of interest. For example, if trying to forecast ice-cream sales it would be reasonable to suspect that the outside temperature is a strong determining factor. Similarly when considering residential households energy usage it would be sensible to assume that the demand would be related to time of day and day of the week due to typical behavioural patterns (However, note it is also easy to find households, such as shift workers, who probably won't neatly fit this assumption). In each application there are some strong candidates which, if available, could be included as features in the forecast model. At the very least, further investigation should be applied, e.g. using the visualisation techniques presented in Sect. 6.2.2.

If the obvious candidates are not readily available or cannot be measured, then **proxy values** could also be considered. These are values which are closely correlated to the value considered. For example, the weather data may not be available in the exact location of interest but may be available from an adjacent town. Another real life example of proxy values is where scientists use ice core and tree ring data as a proxy for the past climate.

6.2.2 Visual Analysis

Visualisation is essential for better understanding the data and determining potential features to include in your models. They can also be used to confirm (or deny) relationships that the forecaster anticipates would be useful, or discover entirely new relationships.

The most obvious visualisation for time series data is to simply plot the data against time, unsurprisingly this is called a **time series plot**. Many examples have already been shown including Fig. 5.1. These examples show some simple features

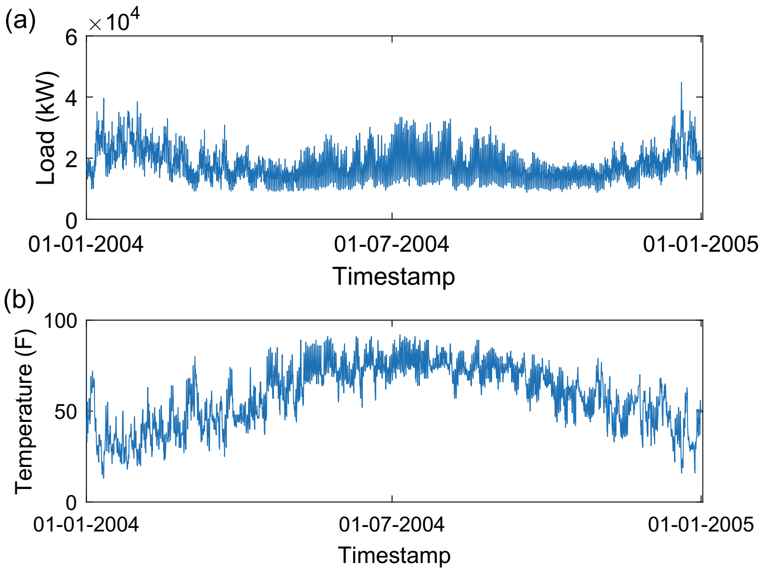

Fig. 6.6 Example of hourly time series of **a** electricity demand in kW, and **b** temperature in Fahrenheit for the whole year of 2004 in the GEFCOM 2012 data set [1]

which are readily identified from a plot including whether the data is stationary (does the distribution of data change in time), and any seasonal or linear trends.

Figure 6.6 shows an example of real, hourly, electricity demand (for one zone for an American Utility) using one year of the Global Energy Forecasting Competition 2012 data (GEFCOM 2012)[2] and the temperature from a nearby weather station (for the data and more details on GEFCOM 2012 see [1]). Instantly evident is the annual seasonalities in both the demand and the temperature values. The demand has high values at the start, middle and end of the year, likely due to the use of increased heating in the winter periods and increased use of air-conditioning in the summer months. It is clear then that the temperature and the demand time series are correlated with each other.

The relationship between temperature and demand is more easily visualised through a **scatter plot** which plots one variable against the other, with each point corresponding to each hourly period. This is shown in Fig. 6.7. In this form the relationship between the variables is much clearer and other characteristics are evident. For example, it can now be seen that, for temperatures less than 50 °F, the demand increases as the temperature decreases but at a much slower rate than the increase in demands with increases in temperatures above 50 °F. Using the extra information which the scatter plot has revealed, the shape of the relationship can be used to develop more accurate forecast models.

[2] Available from http://blog.drhongtao.com/2016/07/gefcom2012-load-forecasting-data.html.

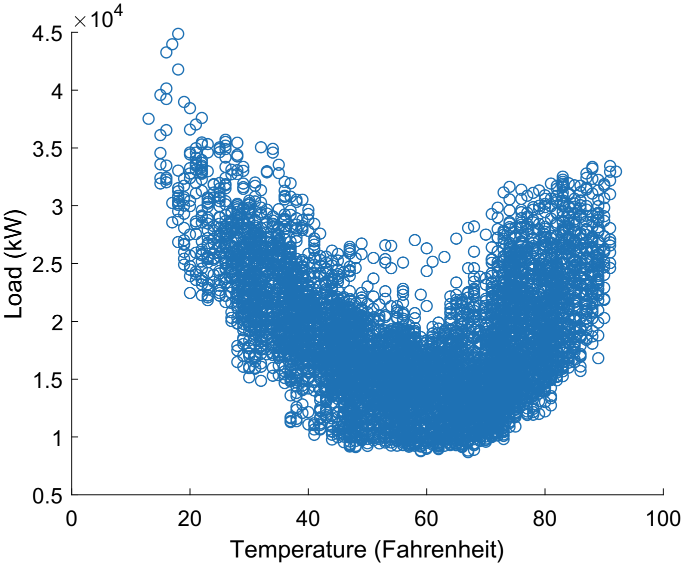

Fig. 6.7 Scatter plot of the hourly load versus the hourly temperature for 2004 in the GEFCOM 2012 data [1]

If there are several variables then it can be cumbersome to produce scatter plots for all the different relationships between the different variables. A more concise representation is a **pair plot** (also known as a scatter plot matrix). This is just a matrix of scatter plots which compares the relationship between each pair of variables. An example of a scatter plot is shown for simulated data in Fig. 6.8 for three variables. Row $k \in \{1, 2, 3\}$, column $j \in \{1, 2, 3\}$ represents the scatter plot for kth variable against jth variable. Notice that row j and column k displays the same information, just reflected, since the plots are reflected across the diagonal of the scatter matrix. Often, as in this example, the kth diagonal contains a histogram of the kth variable. In other words it shows an estimate of the marginal distribution of this variable (see Sect. 3.3 for the definition of a marginal distribution).

It is worth bearing in mind throughout this section and Sect. 6.2.4 that although variables are correlated this does not mean there is a causal link between them (the adage "correlation doesn't imply causation") but this also doesn't mean that features are not useful for the purposes of forecasting, this is known as *Granger causality*.

6.2.3 Assessing Linear Relationships

The scatter plots in Sect. 6.2.2 can suggest different relationships between variables. The Pearson correlation $Corr(X, Y)$ for two random variables X and Y is defined as

$$Corr(X, Y) = \frac{\mathbb{E}(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))}{\sigma_X \sigma_Y}, \tag{6.37}$$

and is a useful measure of the linear correlation between them. Here σ_X and σ_Y are the standard deviations of X and Y respectively, and $\mathbb{E}(X)$, $\mathbb{E}(Y)$ are the corresponding expected values (see Sect. 3.3) for more details). The value ranges from -1 to 1. A value of 1 means the values are perfectly linearly align and positively correlated, i.e. the increase in X will correspond to a linear increase in Y. For a value of -1 the values are again perfectly linearly aligned but this time negatively correlated, so the increase in one variable will simultaneously correspond to a linear decrease in the other variable, and vice versa. Values inbetween indicate less correlation, with $Corr(X, Y) = 0$ indicating no correlation at all.

The lines of best fit and the correlation coefficients are shown for each pair of variables in Fig. 6.8. In this case it is clear that variables $X1$ and $X2$ have very little correlation (0.05), as do variables $X1$ and $X3$ (0.01). In contrast variables $X2$ and $X3$ are strongly positively correlated (0.85).

If several predictors are highly correlated with each other then there can be difficulty in understanding their individual effects on the dependent variable. It may mean that the model is splitting the importance of each variable (e.g. via its trained coefficient) in a way which would be very different if only one of the variables was included in the forecast model. The importance of a variable could be underestimated when included in a model with a variable for which it is highly correlated as its influence may appear minimal. In theses cases it may be worthwhile testing

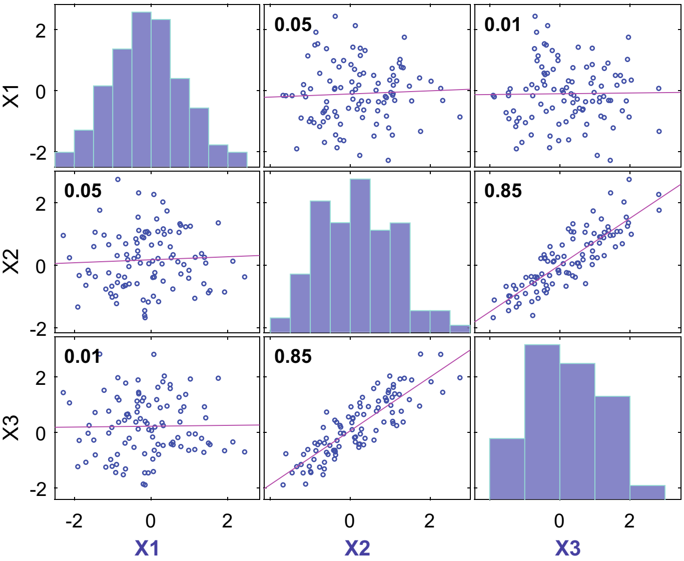

Fig. 6.8 Scatter plot matrix example for three variables. Also included is the line of best fit and their pearson correlation coefficient (see Sect. 6.2.3)

models with different combinations of the correlated inputs to see their effect on the forecast accuracy. The effect of collinearity is discussed further in Sect. 13.6.1.

The Pearson correlation is limited to linear relationships and therefore is not useful for measuring the potential of nonlinear models for the relationships between two variables. For example, it is clear that the relationship between load and temperature is not linear in Fig. 6.7. A common way to test the descriptive quality of a model between variables is the so-called **coefficient of determination** or R^2 value (R squared), which describes how much a model (for example a line) describes the data, and is defined by:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{k=1}^{N} r_k^2}{\sum_{k=1}^{N}(Y_k - \bar{Y})^2}, \tag{6.38}$$

where $\bar{Y} = (1/N)(\sum_{k=1}^{N} Y_k)$ is the mean of the observations, r_k are the residuals between the model and the observations, $SS_{tot} = \sum_{k=1}^{N}(Y_k - \bar{Y})^2$ is the total sum of squares difference between the observations and the mean, and $SS_{res} = \sum_{k=1}^{N} r_k^2$ is the sum of square residuals. R^2 typically takes values between zero and one (but can take negative values when the model is worse than the mean estimate), with the best case $R^2 = 1$ since the model would perfectly fit the observations and hence $SS_{res} = 0$. Explanatory variables with larger R^2 value can be considered more important for describing the dependent variables than those with smaller values. The value of the coefficient of determination can be interpreted as how much of the variation in the dependent variable is captured by the model, so for example, an $R^2 = 0.75$ indicates 75% of the variation is explained by the model. Care must be taken when comparing different models. Better fits (and thus larger R^2 values can usually be achieved by increasing the number of parameters, hence models with different numbers of inputs cannot be compared with the traditional R^2 values. Instead an adjusted R-squared is often used which still compares how much of the variation is captured by the models but also controls for their complexity (the different numbers of parameters). The adjusted coefficient of determination is defined as

$$AdjR^2 = 1 - (1 - R^2)\frac{N-1}{N-p-1}, \tag{6.39}$$

where p is the number of independent variables in the model (excluding any constant term in the model). The adjusted R^2 value is always less than or equal to the R^2 value, Suppose a new parameter is added to the model, then the adjusted coefficient of determination increases if the improvement in R^2 is more than would be expected by chance. Note, that a good R-squared may be achieved by simply overfitting the data but this doesn't mean the forecasts will be accurate (Recall Sect. 8.1.2), and the adjusted R^2 helps to mitigate against this.

6.2.4 Temporal Correlation Analysis

The previous sections have compared at least one variable against another variable. However, in time series, the future values frequently depend on their historical values. The **autocorrelation** function (ACF) can be used to assess these temporal interdependencies. The autocorrelation simply calculates the correlation between the time series and a shifted (or lagged) version of itself. For a time series $(L_1, L_2, \ldots, L_N)^T$ the ACF at lag k is defined as

$$\rho(k) = \frac{1}{N\sigma^2} \sum_{i=1}^{N-k} (L_i - \mu)(L_{i+k} - \mu) = \frac{R(k)}{R(0)} = \frac{R(k)}{\sigma^2}, \tag{6.40}$$

where μ is the sample mean and σ^2 is the sample variance for the time series (see Sect. 3.5 for more details). The important lags can be found by examining an autocorrelation plot, which plots the autocorrelation as a function of the number of lags. The autocorrelation plot for the GEFCOM 2014 hourly demand data is shown in Fig. 6.9. It is clear that there is strong daily seasonalities in the data given the cyclical nature of the ACF and the bigger peaks at lags of multiples of 24 h.

A drawback of the autocorrelation function is that values at shorter lags contribute to the value of the autocorrelation function at longer lags. The partial autocorrelation

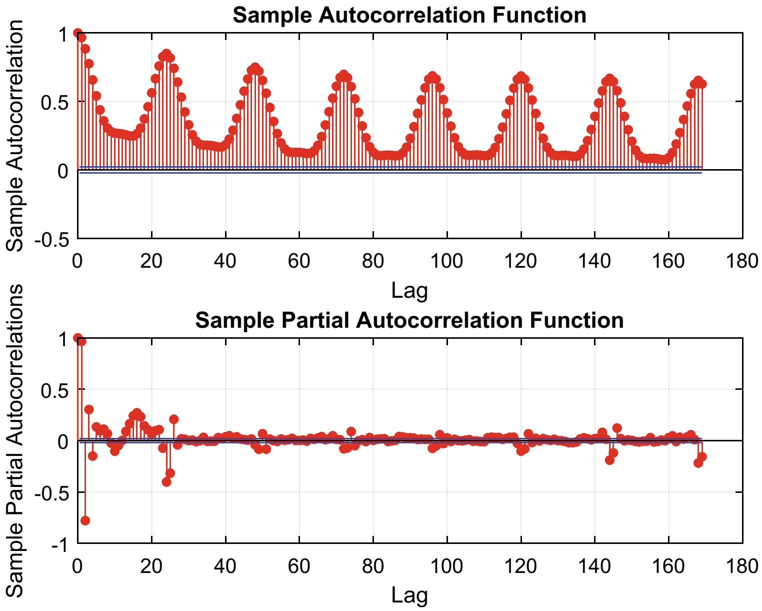

Fig. 6.9 Example of autocorrelation (top) and partial autocorrelation (bottom) for the hourly load data from GEFCOM 2014 [1]

function (PACF) at lag k can reduce this effect by removing the effects of the lags at $1, \ldots, k-1$ (see Sect. 3.5 for the formal definition). In Fig. 6.9 the partial autocorrelation for the GEFCOM load series is also shown in the bottom plot. Notice now, that the sizes of the PACF values are much lower at the daily lags of 48, 72, ... but the value increases slightly at lag 168, which corresponds to a full week and indicates weekly correlations in the time series. This means that the lag at hour 24 is still significant, as is the weekly lag, but other daily lags i.e. at two, three, four days previous etc. are perhaps less significant since the correlation shown in the PACF is much weaker.

Large autocorrelations or partial autocorrelations at particular lags can inform which historical data to include in a forecast model. For example, if, as above, there is some relatively strong correlations at weekly lags it is worth considering including the data from previous weeks from the same time period of day, in the forecast model. Historical inputs from the same time series to a model are often called **autoregressive components**. As will be shown in Sect. 9.4, the ACF and PACF plots play an important role in the parameter selection of ARIMA models.

As shown in Sect. 3.5 we can also consider the **cross-correlation** between two separate time series. This not only shows the correlation between them, but also the correlation between lags of the two series. For example, although heating demand may be driven by cold temperature, there may be a delayed response (especially in homes with good insulation) and hence it may be important to include the temperature as input together with the values from a few time steps previously. Consideration of the cross-correlation (Sect. 3.5) can be used like the autocorrelation plots to identify significant lags to include within your model. An example between the temperature and load data for the GEFCOM data is shown in Fig. 6.10. The x-axis shows the lags (up to a 24 h either way) between the series. Notice they are allowed to be negative here since either series can be lagged, the negative lag means the correlation is between lagged (historical) values of the temperature against the load without lag. First notice the cross correlation is negative but not too strong. in fact it would be much more negative if comparisons had been made between the series in Winter, and similarly been positive in Summer, due to the heating versus air-conditioning patterns as shown in Fig. 6.7. However, since the cross correlation measure linear correlation the coefficient values are much weaker due to conflating the positive and negative correlations. Another thing to notice is that the values are not symmetric around zero lag. This is because there is likely a delay between the effect of temperature on the overall load.

6.2.5 *Basic Functions as Features*

The feature selection methods above have largely required manual investigation and visual analysis. This can be quite time consuming and impractical for forecasting large numbers of time series. A more automatic way to train a model is to build it from individual components, which will be called *basis functions*. This is particularly

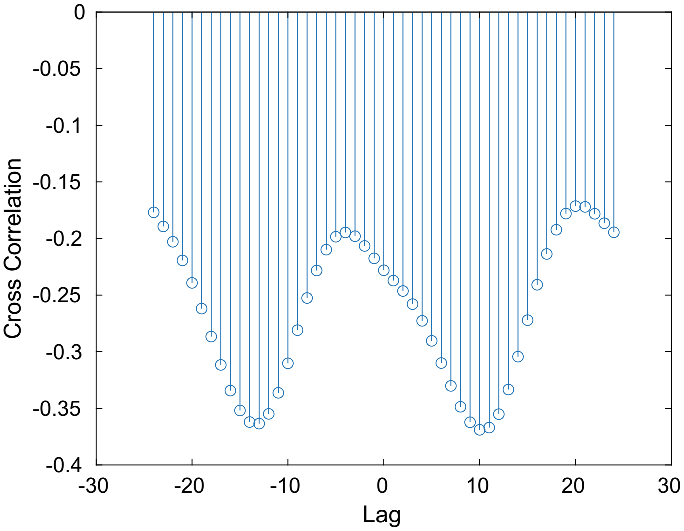

Fig. 6.10 Cross correlation coefficients between hourly temperature and hourly load for 2004 in the GEFCOM 2012 data [1]

useful for time series exhibiting periodic behaviour as is the case with energy demand time series. In other words a time series (or segment of a time series) Y_t can be written as a linear combination of simple functions/vectors $\phi_k(t)$,

$$Y_t = \sum_{k=1}^{\infty} \alpha_k \phi_k(t). \tag{6.41}$$

Where α_k are the coefficients that must be found. The most common example of this is the Fourier Series which has periodic basis functions of the form $\sin\left(\frac{2\pi kt}{K}\right)$ and $\cos\left(\frac{2\pi kt}{K}\right)$ for $t \in [-K, K]$.

Using basis functions as features means that each time series can be trained whilst reducing the development of bespoke features for each time series. Instead of the infinite sum in Eq. (6.41), in practice a finite sum is chosen, however as mentioned in the previous section, this could cause overfitting of the time series to the training data and thus must be carefully chosen. Methods for choosing an appropriate number of terms, are given in Sect. 8.2.2. A special case of basis features, using splines is described for generalised additive models in Sect. 9.6.

6.2.6 Common Features in Load Forecasting

Load is comprised of several different types of consumers, from residential, to small-to-medium enterprises (also called SMEs) such as hairdressers, small shops, etc. There is also larger commercial consumers (larger supermarkets, schools, etc.), and finally industrial consumers which usual comprise of larger demands such as steel production and other heavy industries. Forecasts may be required for the demand of these individual customers or for the load at the substations, or over larger areas and therefore consist of aggregations of these individuals consumers and other connections. These other connections may be anything from street-lighting, but also distributed generation (e.g. solar farms or wind turbines). Therefore there is no simple set of features which model all types of demand. However, there may be some features which are common, or at the very least, worth testing to identify whether they would make useful explanatory variables. This section will discuss some common ones.

First, it is worth mentioning that even the same type of consumer (domestic, SME, etc.) may still have very different demand behaviours from each other with very different drivers. For example, a house that uses electric heating will likely have electricity demand driven by temperature, in contrast one that is gas heated may have electricity demand which has little-to-no influence from the weather. Secondly, the demand of aggregations of consumers will likely becomes more regular the larger the aggregation although the main drivers may be less clear, or alternatively some features could become more pronounced. As shown in both Fig. 1.2 (Sect. 1.2) and Fig. 2.5 (Sect. 2.3), the weekly regularity is improved with larger aggregations.

We have already discussed weather quite a lot in this Sect. 6.2 but as in the gas versus electric heating example mentioned above it may not be obvious if weather will have an effect on the demand without further investigation. Further, as discussed in Sect. 6.2.4 it is also worth checking cross-correlation and the lagged variables as there may be delays in the effect. If a relationship is observed it is unlikely to be completely linear, as illustrated in Fig. 6.7. In this plot of temperature versus demand, demand increases below 50 °F demand due to heating but there is also an increase for temperatures above 60° due to air conditioning. In this case it may be worth modelling the relationship as a piecewise linear model, a polynomial, splines, or other basis functions (Sect. 6.2.5).

In addition to temperature, other weather variables can also be important for demand forecasts:

1. Wind Speed: higher wind speeds can increase the effect of *experienced* temperature, e.g. colder temperatures will feel colder the faster the wind speed. This may have a knock on effect that the heating is turned on sooner. This variable is known as **wind chill** and is engineered by combining the temperature and wind speed.
2. Humidity: Similarly higher humidity's can increase the felt temperature. One way to describe this is the **humidity index**, a combination of humidity and temperature.

3. Solar Radiance and visibility: the sun radiates on the earth and increases the temperature of the earth. If there is a lot of cloud then less radiation reaches the surface and the temperature will be lower than on a clear day.

Daily and weekly periodicity is often common in electricity loads. Residential and commercial demands are driven by human behaviour and needs and hence follow daily and weekly patterns. However, of course there are exceptions such as doctors and nurses who may have different daily patterns in their behaviour due to the different shifts they work. Including autoregressive effects at daily and weekly lags (e.g. a lag of 24 and 168 respectively for hourly timeseries) is one way to include the periodicity in the model. However, this doesn't model the general day of the week effects. To include an effect for "Monday", "Tuesday", etc. means including the effect of seven categorical variables into a model of electricity demand which is a continuous variables. This requires adding an update, or change, to the demand which is different depending on the day. A common way to do this is via so-called **dummy variables**. Dummy variables can take values of zero or one depending on the falsity or truth of the presence of a variable. For example, say $W(k)$ is a variable which is one if the time step k occurs on a day which is a weekday (Monday, Tuesday,..., Friday), and zero if not, i.e.

$$W(k) = \begin{cases} 1, & \text{if time step k occurs on weekday} \\ 0, & \text{otherwise} \end{cases}.$$

This is called the dummy variable for a weekday. If instead the model requires the effect of each weekday to be represented, it is required to include a corresponding dummy variable for each day. In this case, define the dummy variables $D_j(k)$, for $j = 1, \ldots, 7$ to be

$$D_j(k) = \begin{cases} 1, & \text{if time step k occurs on day } j \text{ of the week} \\ 0, & \text{otherwise} \end{cases}.$$

It is often not desirable to define N dummy variables to represent the entire set of N possible values. For example, there is only seven days of the week, which means the seventh day is simply defined as not being any of the other six days. Only six variables are needed since they can be added as a correction to the default seventh day. Since one variable can be modelled by the other six if all seven variables are included in the modelling this creates colinear variables (Sect. 13.6.1) which can cause issues with the training of parameters in the final model. This is also known as the **dummy variable trap**. Note that, the term "dummy variables" is often used in statistical modelling whereas in machine learning, it is often called **one-hot encoding**.

We've already seen in Sects. 5.1 and 6.1.1 there is often long term changes in the demand time series signal. This could be large scale annual seasonality or linear trends for example. Demand series is often changing, there is new technologies, or more efficient versions of the same appliances, or there may be changes in which customers are connected to the network (a hairdressers turning into a convenience store). If these differences are clearly observed in the time series then it may be worth including them as explicit variables within your model. A linear trend in the series could be included by simply including the time step indicator in the model. In a simple linear model (Sect. 9.3) a trend is included by adding a term such as bt where t is the time step and $b \in \mathbb{R}$ is a parameter to be trained.

Another common feature of demand time series is annual seasonal trend. Energy demand often increases in the winter due to increased demand, and in the summer may be at its lowest level since the temperature may be warm enough so no heating is required (in hotter counties there is often an increase in demand due to air conditioning appliances). These patterns represent periodic patterns and hence should be included in the models. One option would be to simply add the day or time period of the year, by, for example generating a large number of dummy variables. However, in this case there wouldn't be many historical examples to train the parameters and the model may not generalise very well. It is often preferable to use periodic variables which can estimate the seasonality with fewer parameters. One very basic example is to use basis functions (Sect. 6.2.5). Trigonometric functions such as $\sin at$ and $\cos bt$ are one option. The parameters $a, b \in \mathbb{R}$ can be chosen (or preferably trained) to ensure that the period is appropriately chosen to match the pattern within the signal. Multiple trigonometric functions (with different periods) can be chosen to improve the fit and generate complicated patterns, see Fig. 6.11. There are more complicated choices as well such as wavelet functions, which are not explored here. We give an example of using trigonometric functions within a linear model in the Case Study in Sect. 14.2.

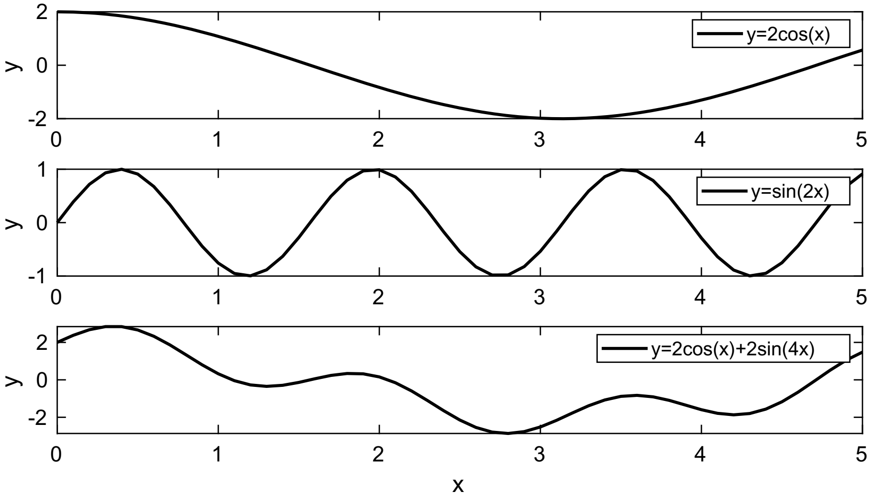

Fig. 6.11 Demonstration of how to model seasonal patterns with trigonometric functions. Two seasonal functions (top and middle) have been combined to generate a more complicated seasonal pattern (bottom)

6.3 Questions

For the questions which require using real demand data, try using some of the data as listed in Appendix D.4.

1. Consider different methods for imputing missing values. Select a demand time series. Simulate missing values by removing them from the time series (save them for comparison later). Now consider filling them in using some of the techniques given in Sect. 6.1.2. Comparing to the real values. What methods seems to perform the best? Why do you think this may be?
2. Select a demand time series, and select four weeks worth of data. Calculate the sample mean and standard deviation. Which values are more than two standard deviations from the mean? What about three standard deviations? Do these values look unrealistic or too large? Now calculate the median and interquartile range. How many interquartile ranges from the median are the largest values?
3. Take a time series with average kW or kWh values at half hourly resolution. Convert the data into hourly by averaging the data over each pair of consecutive half hours. Reconvert the data back to half hourly by linear interpolation. What is the difference in the reconstructed half hourly data compared to the original data? Where is the error largest, why is this? Try using a higher order polynomial (e.g. cubic) for the interpolation. Is this more accurate? How does it compare using household smart meter data versus system level data (e.g. GEFCOM 2014)?
4. Other than temperature what may be some other important weather variables which may affect the electricity demand within a home, or more general for the national demand? What about non-weather data, what else would be good indicators of demand?
5. Consider the London Smart meter data.[3] Plot a scatter plot of the different weather values against demand. Which variables have the strongest relationship with demand? Are the weather variables related to each other? What is the correlation between them? Which has the largest correlation value.
6. Take a demand time series. Plot the autocorrelation and partial autocorrelation of the series. At what lags is the correlation strongest? Plot a scatter plot of the demand against lags of the demand series. Include a lag of one time step, and also the lags which gave the largest autocorrelation values. Are the relationships linear? If they are linear calculate the adjusted coefficients of determination. Which lags give the biggest values?

[3] https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london?select=weather_hourly_darksky.csv.

Reference

1. T. Hong, P. Pinson, S. Fan, Global energy forecasting competition 2012. Int. J. Forecast. **30**(2), 357–363 (2014)