Midterm Peer Review

Project 12 – Household Electricity Consumption

The stated goal of the project is to model the **total global daily** power consumption in a household. The report does not explicitly clarify what "global" means in this context.

Data Preparation: The original dataset has a one-minute sampling rate. Dates with at least one missing record are dropped entirely. Then, the data is aggregated to get the total daily power consumption and dates corresponding to outliers are also dropped. However, **dropping dates disrupts the continuity of the time series**. As time series analysis relies on the temporal ordering of observations, this could potentially affect discovery of the underlying autocorrelation structure and seasonal patterns. Lastly, the time series is subsampled for computational reasons.

EDA: The description "distribution of the **daily average**" for the time plot is misleading as each data point represents the **daily total**. The authors could've explained their visual interpretation of variance stationarity. The interpretation of the month-level boxplots is logical. The ACF and PACF are interpreted but the distinction between the two is not made clear.

At this point, the **analysis suddenly switches to total monthly** power consumption. This is **not the initially stated goal** and the report **never explains why this is done**. This also reduces the **no. of data points to 47**, which is less than the recommended minimum of 100 in the project instructions.

Trend: **OLS** polynomial regression is used to fit a trend but the **assumption of iid errors is never verified**. So, the conclusion of "no trend" is not very reliable. It would've been more appropriate to carry out regression with ARMA errors which accounts for serial autocorrelation between errors that is typical of time series data. In fact, even the STL decomposition done later shows a trend component, but this is not related back to the conclusions drawn here.

Spectral Analysis: The x-axis of the periodogram **does not indicate the units of frequency**. Bandwidth is displayed but not explained. The dominant period is estimated visually – a scientific report is expected to include the precise calculated value.

ACF/PACF: The lags (x-axes) are decimal valued – the analysis doesn't address this.

STL Decomposition: The seasonal component is interpreted correctly but the trend is not interpreted at all. The authors could've included a note that for very short time series, STL decomposition should be interpreted with caution.

SARIMA: 1st order differencing is used for monthly data based on visual analysis of the daily data time plot. AIC based model selection – it is only stated that different SAR/SMA pairs are tried and 1/1 achieves the lowest AIC among all but the AIC tables for each (for different AR/MA pairs) are never shown. The final model SARMA((0,0,4) x $(1,1,1)_{12}$) is not too complex.

Model Diagnostics: The fitted model summary is **not interpreted – all coefficients except ma4 seem to be insignificant** at the 5% level based on their Fischer standard errors. Causality/Invertibility: The MA roots seem to be just inside or on the unit circle (the reader has to calculate the magnitude), but the authors conclude that all roots fall outside the unit circle. Residual analysis is adequate.

Forecast: The authors could've used a held-out set and compared predictions to true values instead of concluding that the model is good based on visually observed patterns in the plotted forecast.

References are not cited in-line which makes it difficult for the reader to connect the content to the sources.

Project 08 – Solar Wind

The goal of the project is to model and forecast X-ray flux data related to solar wind. Although the WHAT is stated, the WHY is missing.

Data Processing: Missing values are imputed with linear interpolation, but the report does not mention how many values were missing. The motivation for using the log transformation is not clear.

Smoothing: Rolling mean and iterative rolling mean are used to reduce noise but the concepts are not explained. The rolling mean is later used to detrend the series.

There are two aligned time series in the dataset - ‘flux1’ and ‘flux2’. It is unclear how these are different and why ‘flux2’ is chosen for further analysis. Further, the report doesn’t explain how the training set is constructed (I had to look through the code for this).

ACF: Good, rigorous interpretation of the plot. Seasonality is inferred from the oscillatory pattern.

ARIMA/SARIMA: Formal model definition in the form of equations is missing.
* The 1st order difference is used to “induce stationarity” but the report doesn't previously discuss the stationarity of the time series.
* There seems to be an issue with the implementation – in forecast::auto.arima(max.p=50, max.q=50, seasonal=TRUE, stepwise=FALSE) with the default seasonal parameters (max.P=2, max.Q=2), the maximum value of p+q+P+Q for the models tried is given by max.order (5 by default). So, it is likely that the full parameter space intended by the authors is never explored.
* The authors also don't state what criterion is used to choose the “best model” (ARIMA(2,1,2)).
* Insightful analysis of the fitted model summary (coefficients, mean error). MAPE and MASE are used to interpret the forecast accuracy – the report could’ve been more self-contained if definitions of these metrics were included.

ARIMA Forecast:
* The ARIMA forecast shown is for log(flux2) and not flux2 as the title states.
* The limitations of the model in capturing the complex patterns in the data are correctly identified.

Spectral Analysis: The x-axis of the periodogram **does not indicate the units of frequency**. Bandwidth is displayed but not explained. It is not explained how the dominant period is calculated.

STL Decomposition: Report does not include any explanation or formal definition.
* A larger window could’ve been used to get smoother trend and seasonal components.
* Separate ARIMA models are fit to each of the components – the authors do not comment on how too many parameters could potentially lead to overfitting/poor generalization.
* The order of the models fitted is never shown.
* The forecast still doesn’t seem to be capturing the patterns in the data well. “The forecasted values also show a general trend of increasing” -- this doesn’t seem to be true from the plot.

Model Diagnostics: Adequate check for iid normal errors assumption.

References are not cited in-line which makes it difficult for the reader to connect the content to the sources.