Imperial College London

Department of Mechanical Engineering

Statistics Coursework Report

Author

Anthony Eden

Supervisor

Ioanna Papatsouma

February 2022

CID: 01853219

Table of Contents

1 Exploratory Data Analysis

Figure 1 indicates that wavelengths in the population are approximately concentrated around two points, 1510.795 nm and 1510.89 nm.

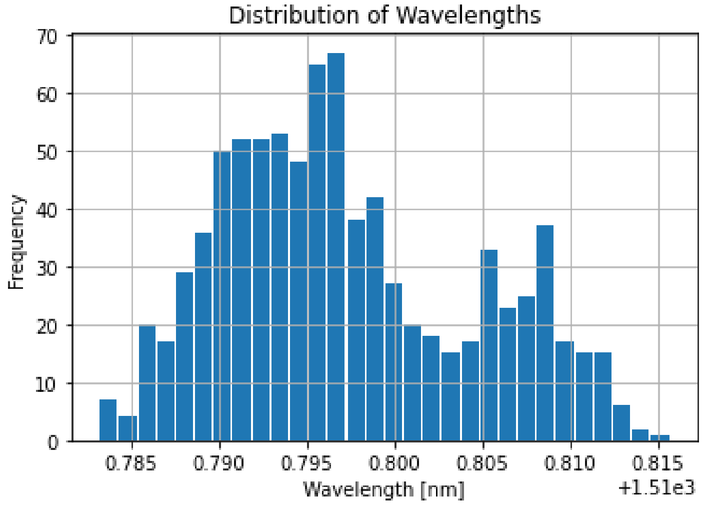

Figure 1: A histogram showing the distribution of wavelengths in the sample.

Figure 2 confirms these two points of interest, as they both lie within the 25th and 75th quantiles respectively.

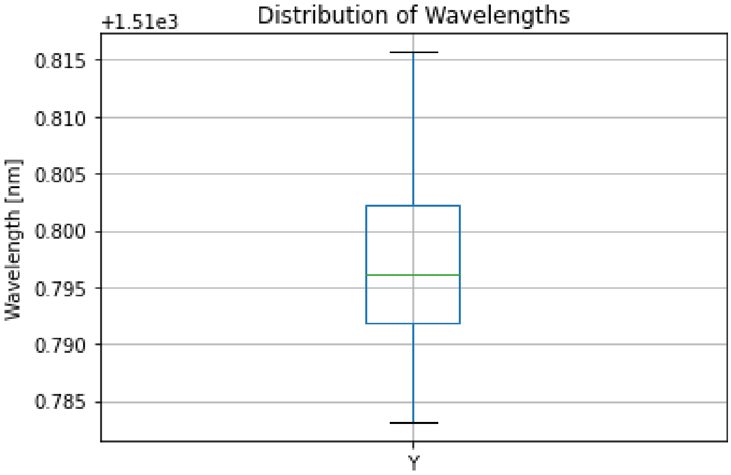

Figure 2: A boxplot showing the distribution of wavelengths in the sample.

The median wavelength, as shown in Table 1, is 1510.796 nm (3 d.p), coinciding with the first concentration peak. This suggests that it is a better reflective metric of the majority of the population than the mean or trimmed mean. The similarities in the mean and trimmed mean to 3 d.p. reflect the absence of outliers in the population as shown in Figure 2.

Table 1: Summary Statistics of The Population (to 5 d.p.)

Mean	Trimmed Mean	Median	Standard Deviation	Interquartile Range
1510.79723	1510.796 90	1510.796 08	0.007 13	0.010 50

2 Modelling

Figure 3 shows the updated scatter plot with linear and quadratic regression used to attempt to fit the data. As expected, both seem to be inappropriate for this data set, given visually there is clearly a more sinusoidal relationship between time and wavelength. The restriction in curvature for both models suggest that the relationship between wavelength and time is not captured appropriately.

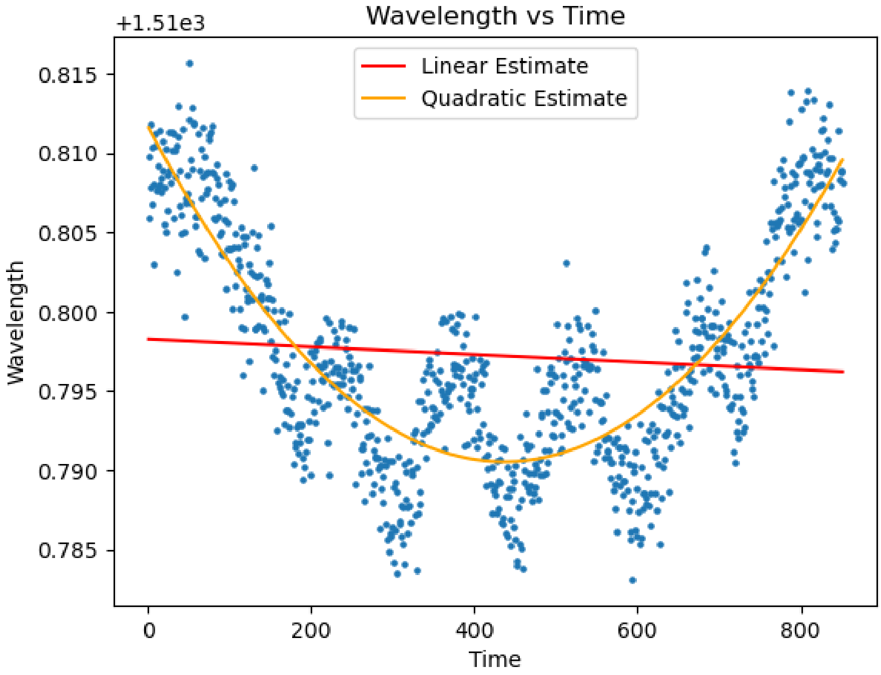

Figure 3: A linear and quadratic regression model used to fit wavelength variation with time.

It is clear that a higher order polynomial is required to fit the data. To find the optimal order for this, AIC values were compared using Equation 1. The AIC rewards models for having a good fit to the data, as encapsulated by the second term, to ensure models do not underfit the data. It also penalises models for having a greater number of parameters, encapsulated by the first parameter, indicating the propensity of such models to overfit the data. With this in mind, models with the lowest AIC values are favourable.

$$AIC = 2 \cdot q - 2 \cdot l(\hat{\beta}, \hat{\sigma}^2) \tag{1}$$

After identifying the region of minimal AIC values (see Figure 8 in Appendix) table 2 shows that the lowest AIC value is found at the 26th order estimate.

Table 2: AIC values for varying polynomial orders 'k'

k	AIC
24	-7799.1598075729
25	-7797.2212570251
26	-7800.8772466149
27	-7799.9489705192
28	-7798.1772572019

The identification of the maximum log likelihood in Equation 1 is based on the assumption that residual errors are normally distributed. To confirm this, Figure 4 shows a Q-Q plot of the data against the quantiles of a normal distribution. Deducing graphically, the data seems to indeed approximately follow a normal distribution.

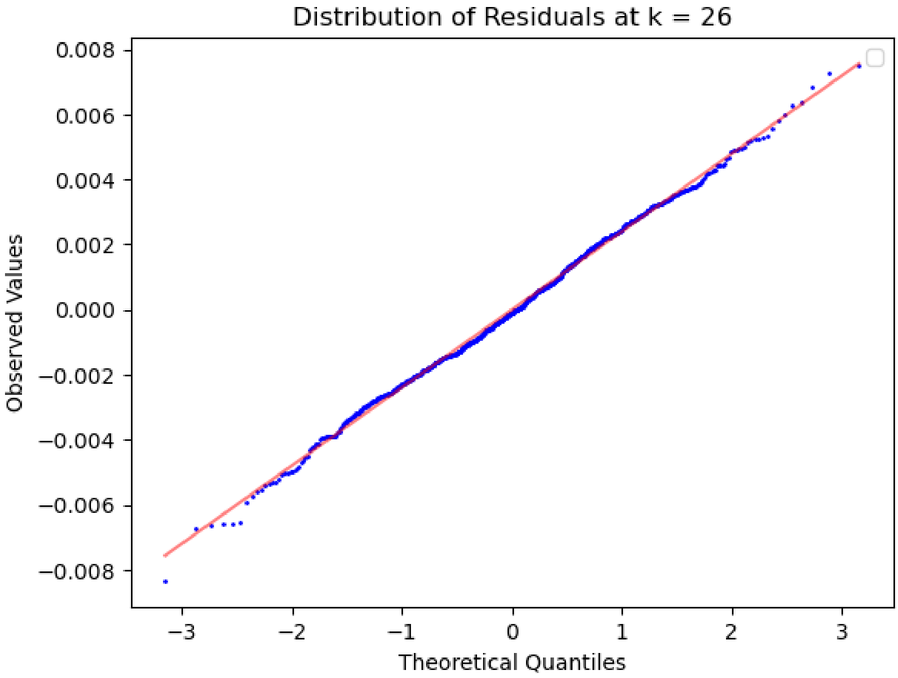

Figure 4: A Q-Q plot of the residual errors against the quantiles of a normal distribution.

Figure 5 confirms these residuals approximating well to a normal distribution.

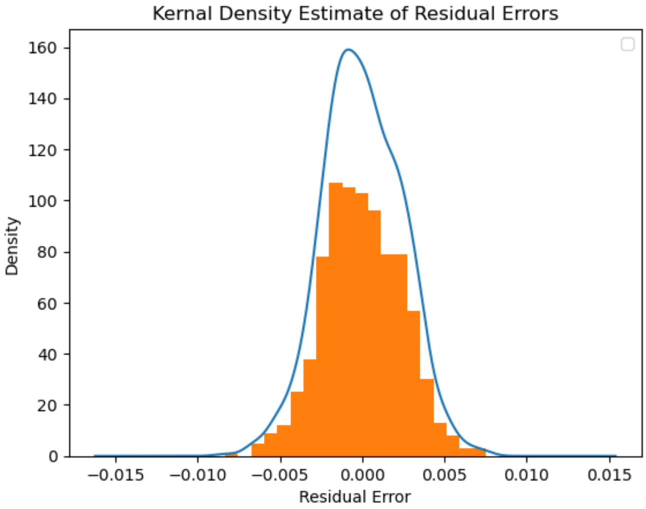

Figure 5: A kernel density estimate plot of the residual errors.

A fitted model of order 26 is plotted on a subset of the data sampled every 10 time indices in Figure 6. Note that the x-axis of the data has been normalised according to the z-distribution. This not only reduces instability effects that large values can have on the coefficients of a model, but also allows for the comparison of other predictor variables in future if needed (e.g. wavelength vs temperature at each time index).

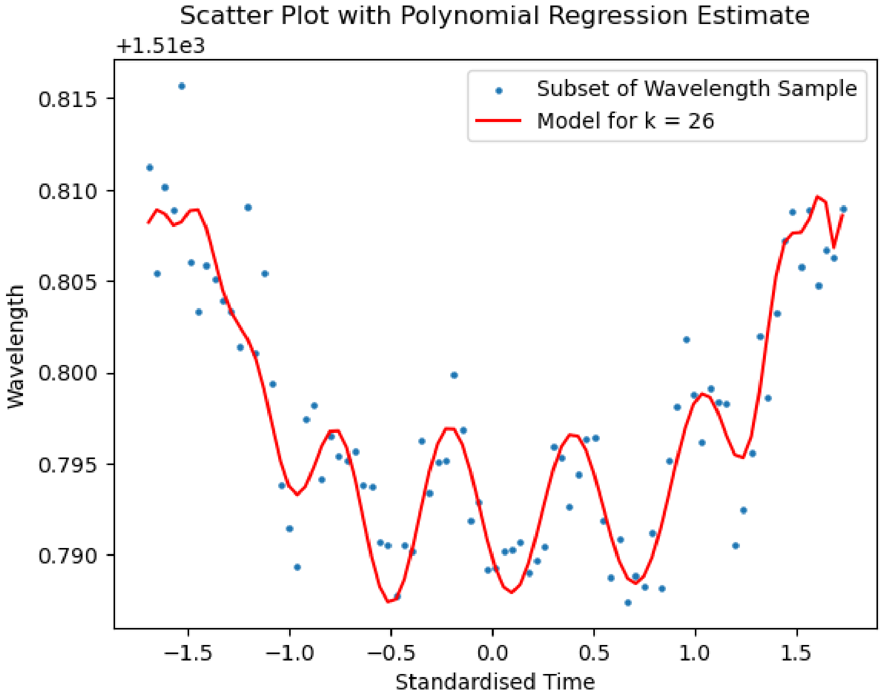

Figure 6: A scatter plot of wavelength against time using a subset of the initial dataset, fitted with a 26th order polynomial model.

3 Bootstrapping

A 95% confidence interval was constructed from point-wise boostrapped samples of various sizes at each time index. This allows a confidence band to be constructed as shown in Figure 7. This Figure also indicates well the positive correlation between boostrapped sample size and convergence between the true and bootstrapped confidence bands.

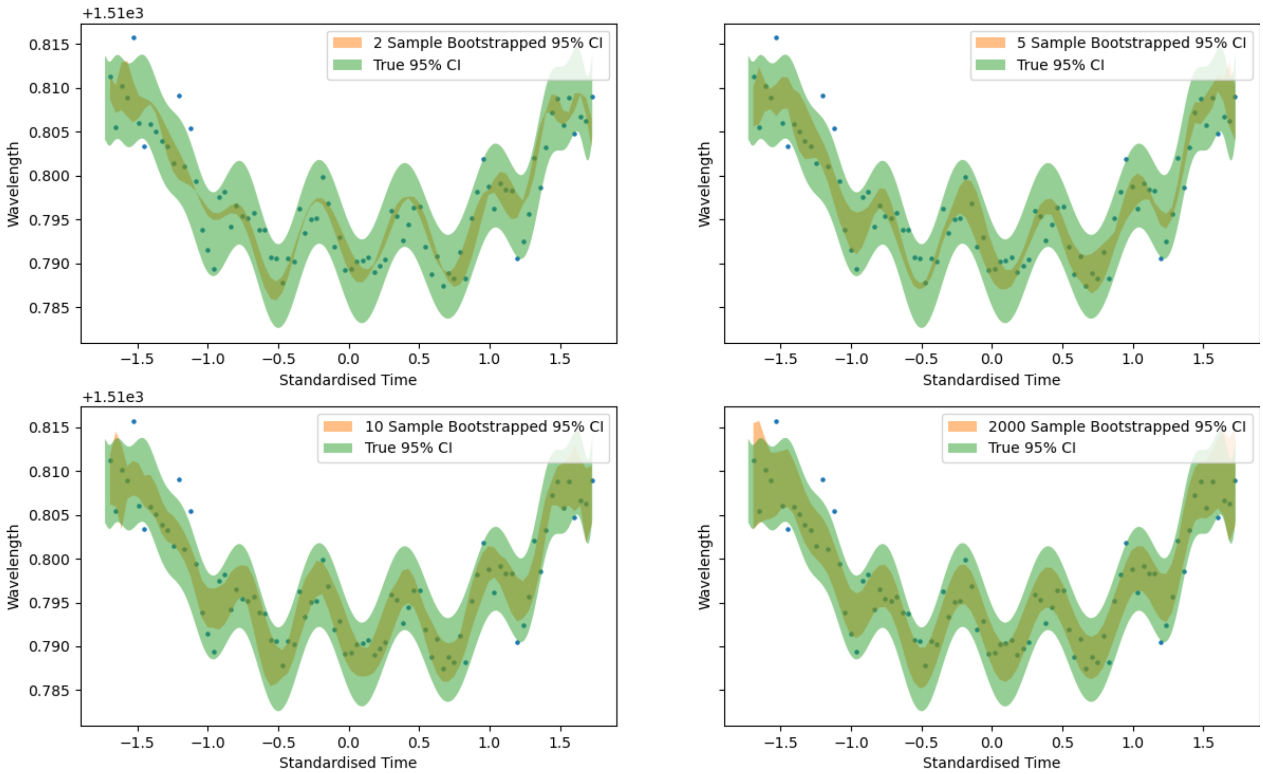

Figure 7: Confidence band plots of various point-wise bootstrapped sample sizes.

The true confidence band has been caluclated using the mean squared error of the residuals of the original dataset (see Appendix Equation 2). Figure 4 has demonstrated that the residual values can be assumed to approximately follow a normal distribution, and hence the error values were used in tandem with the z-distribution to construct this band.

The wider band of the true 95% CI makes intuitive sense - as this band has been derived from the entire 851 sample set, of lower residual variance, it implies that the true prediction model is less certain than the bootstrapped samples suggest, with greater context (i.e. more data points) to support this.

4 Appendix

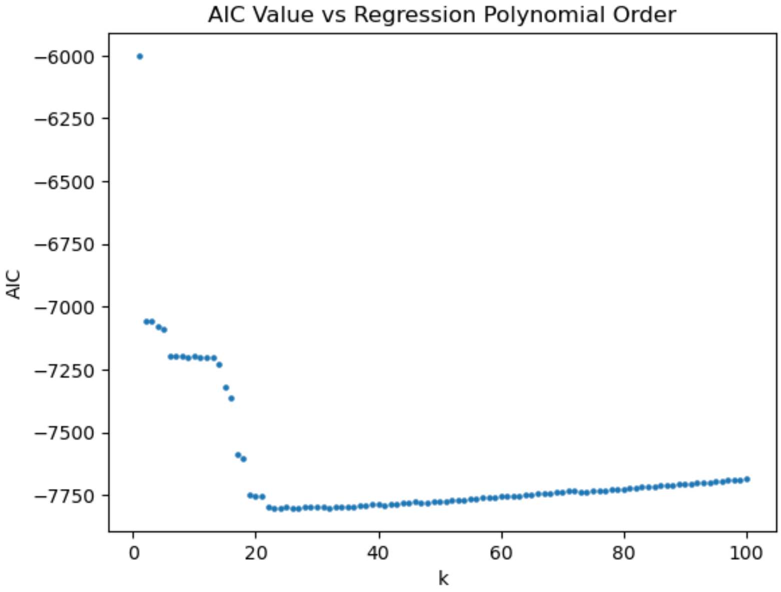

Figure 8: A plot of AIC values against polynomial order, used to find the region of interest for optimal polynomial order.

$$StandardError_{regression} = RMSE\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{2}$$

Where $RMSE$ is the root mean squared error, n is the number of observations, x_0 is the value of x at which the predicted value is being estimated, $\bar{x}$ is the mean of the x values, and $\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sum of squares of the deviations of the x values from their mean[1].

[1] Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). "Applied Linear Statistical Models" (5th ed.). McGraw-Hill/Irwin.