Climate Simulation, Time-Series, and Uncertainty Quantification

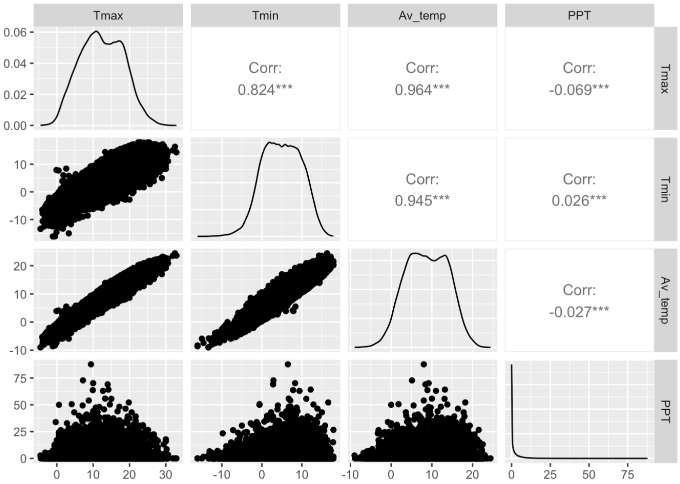
Amber Chang

March, 2024

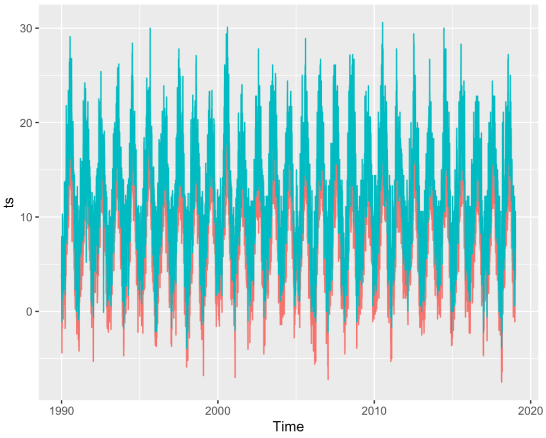
1. Introduction

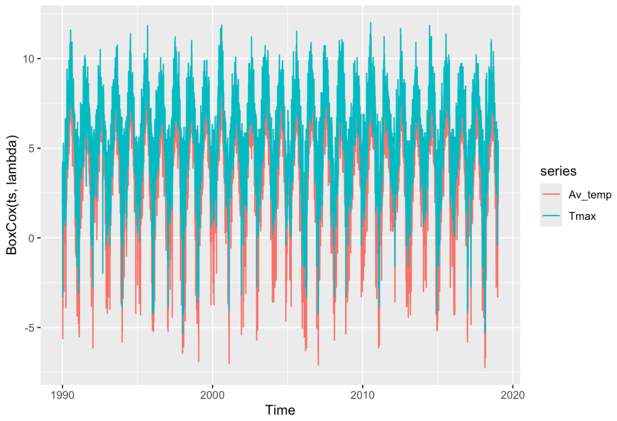
1. In this project, we will be applying AI to earth science data. We will use daily temperature record from the Durham observatory from **1900 to the end of 2019** to predict the daily temperature for 2020. Since this prediction is time-specific, several techniques is explored to construct our temperature prediction model. In construction of our time-series model, various regression methods across frequentist and bayesian would be primarily explored, including ARIMA and MCMC. Specifically, univariate state space model and non-parametric methods would be mainly used for fitting our data for this forecasting problem, **because this allows us to discard our assumption of each model and focus on the prediction itself.** Lastly, in our simulation of our predicted data, uncertainty quantification is taken into account, so instead of just providing a single predicted value, a confidence of predicted interval and variance will be given.

2. Exploratory Data Analysis

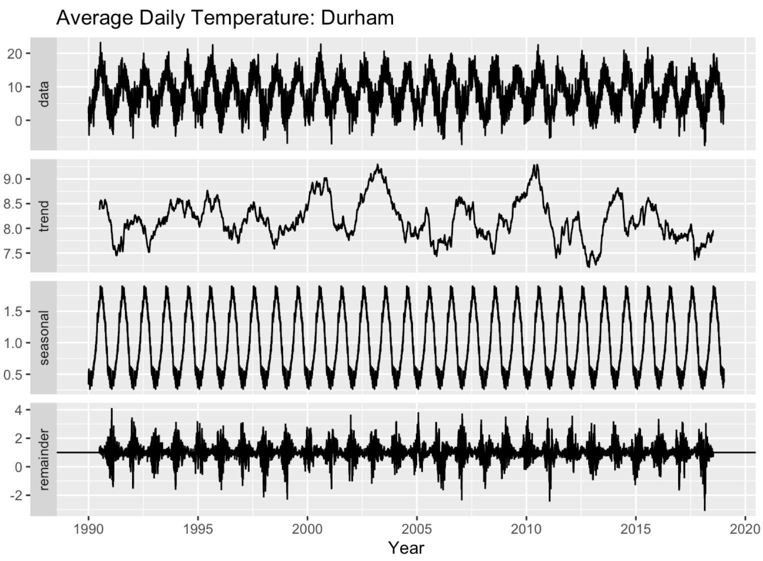


標題





Before Box−Cox Transformation



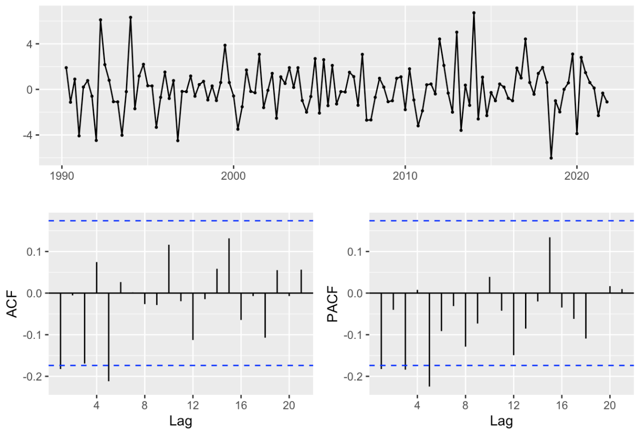
Classical decomposition of Time Series Data

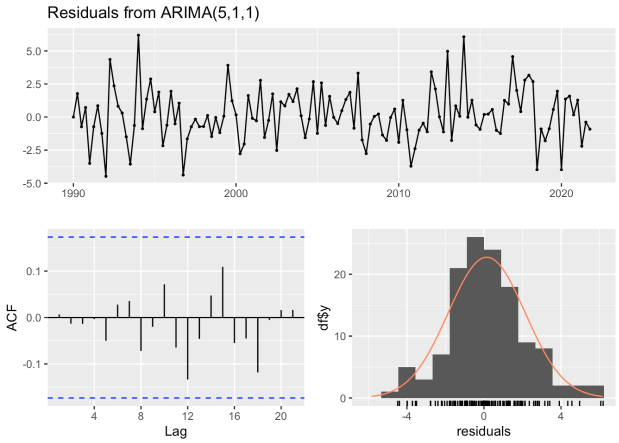
This illustrates a classical multiplicative decomposition of the average daily temperature in Durham form 1990 to 2019.

2. Approaches Tried

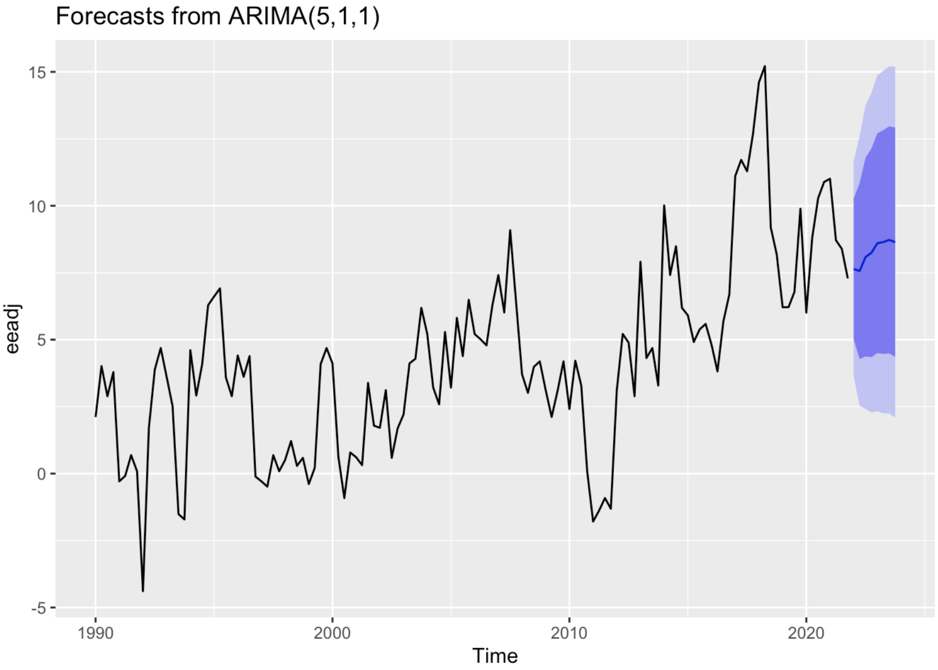
1. ARIMA: : Frequentist methods

One approaches is to use ARIMA. If we combine differencing with autoregression and a moving average model, we obtain a non-seasonal ARIMA model, an acronym for AutoRegressive Integrated Moving Average (in this context, "integration" is the reverse of differencing).





By looking at ACF and PACF plots above, we have decided to specify ARIMA(5,1,1) model. The residual plot has also shown as a normal distribution, which corresponds with our expectation of the model.



2. Monte Carlo Marco Chain Posterior Sampling: Bayesian methods

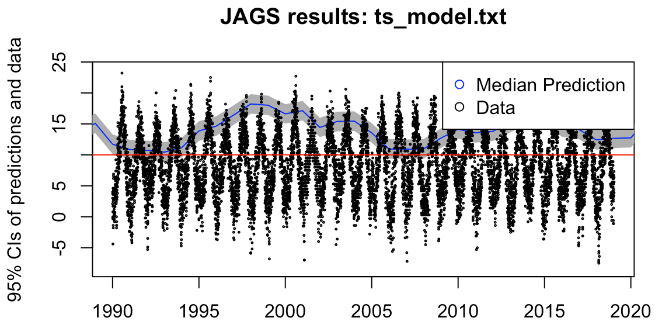
Another approaches is to use **Bayesia**n time series models to perform forecasting. This means we fit a model, and use those posterior distributions to forecast as a secondary step. We assume a prior and likelihood for our parameters, which are usually specified as a gaussian distribution, and update them with our dataset to gain a posterior probability. This approach is generally more evidence-based than prior probability sampling, since we incrementally updated our model by taking into account of both our subjective assumption of the parameters of interest as well as our observation data.

Technically, by using a more streamlined approach is to do this within the JAGS code itself. We can take advantage of the fact that JAGS allows us to include NAs in the response variable (but never in the predictors). We used the same Durham's **daily average temperature** dataset, and the univariate state-space model described above to forecast three time steps into the future. This will write each of the models with the same **univariate state-space** form.

To efficiently estimate our parameters, **some important arguments** for our Jags model have to be considered. Running multiple MCMC chains (e.g., 3 or more) helps assess convergence and improve the robustness of the results. A burn-in of at least 5000 iterations ensures that the MCMC chains have converged to the target distribution before sampling. Thinning rate is 10, total number of retained samples per chain:

$$\text{Total retained samples} = \frac{n_{\text{iter}} - n_{\text{burnin}}}{\text{thinning rate}} = \frac{5000}{10} = 500$$

In terms of assessing the predictive power for MCMC, c**onvergence and mixing** is a necessary criteria to do assessment. RMSE will be assessed to check the accuracy of our model.



3. Limitation of the approach

As with most prediction interval calculations, ARIMA-based intervals tend to be too narrow. This occurs because only the variation in the errors has been accounted for. There is also variation in the parameter estimates, and in the model order, that has not been included in the calculation. In addition, the calculation assumes that the historical patterns that have been modelled will continue into the forecast period.

Assumption of Stationarity: ARIMA models assume that the historical patterns observed in the data will continue into the forecast period. However, this assumption might not always hold true, especially if the underlying processes change over time. Ignoring potential shifts in the data-generating process can lead to inaccurate prediction intervals.

In contrast, Jags can provide a more flexible approach to tackle aforementioned issues associated with ARIMA. For instance, the probability of errors and quantity of parameters are inherently given to estimate.

However, **Current Jags (MCMC) do not support bi-modal distribution, a mixture of two normal distribution. This may lead to a very biased sampling process.** We observe that this MCMC process tend to be overestimating. In addition, the computation cost is extremely high. Even though this could be tackled by thinning and burning, the computational time is still higher than ARIMA.

4. Further Improvement

If I had more time and resources, I would have tried out more hyper parameters of ARIMA model and would have used bootstrap method to simulate the model of parameters in order to be more certain about the range of the Durham Daily temperature according to the Central Limit Theorem, as the number of simulation gets larger.

In terms of MCMC model, rather than specifying likelihood in normal distribution, bi-modal likelihood can be considered by using GMM to simulate with two of its means and two of its variances. In addition, convergence and mixing process could have been done more thoroughly to check accuracy of our predicted values in order to avoid overestimation or underestimation.