Introduction

In this group project we will be conducting an investigation into the use of Vector Autoregression (VAR) and Long Short-Term Memory (LSTM) models for forecasting air pollution in Beijing.

We'll be interested in analysing the computational performance of our chosen models, how these models will generalise to a larger dataset and what methods we could use to assist this. Our dataset consists of 8 variables which have been measured hourly over the course of 5 years. Initially, we'll look at a VAR model and take this as our baseline. The VAR model makes intuitive sense in the context of the problem and is a mathematically accessible model. It looks at the previous few days data to try and predict what each variable will be the next day using a linear combination of the previous values. We'll then look at two deep learning models, a univariate LSTM model and a multivariate LSTM model. These are a type of Recurrent Neural Network (RNN) and allow for prediction using sequential data, accounting for long-term dependencies. Whilst univariate relies solely on pollution levels, multivariate will incorporate additional weather variables, introducing complexity and allowing us to interpret if additional features are necessary for LSTM networks.

Our main areas of interest with regards to assessing the computational performance will be the time it takes the models to train, the amount of memory the model uses, and the complexity of the model. These will be key when we come to think about how the work we've carried out can be applied to a much larger amount of data.

Furthermore, we look at how using distributed computing technology such as may increase the performance of our model by allowing for faster data processing.