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 and seeing how our models would work on a large scale.

1 What we're doing

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.

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 all the variables from the previous few days of data and uses a linear combination of these variables to try and predict what each of these will be the next day. This model will give us a good baseline in order to get a feel for how long a simple model takes to perform its task and the computational complexity involved.

We'll then look at two deep learning models, a univariate Long short-term memory (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. These are more complex models than what we looked at initially and allows us to assess how the computational performance changes as the models complexity increases.

Furthermore, we look at how using distributed computing technology such as PySpark may increase the performance of our model by allowing for faster data processing to gain an insight into how our models would perform at a much larger scale.

2 Investigating Computational Performance

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, such as datasets that look at pollution and its related variables over the course of decades.