Imperial College London
Department of Mechanical Engineering
MECH60017 Statistics A Coursework

Statistical Modelling, and Analysis on Data Collected Regarding Air-Pollution In a Town In Europe

Oluwaseyi Ayodeji Adeniyi

Word Count:

Date: 07/03/2024

CID: 01858040

TOC

Introduction

Daily data collected over 150 days by a team studying outdoor air-pollution, and consequently the air quality in a town in Europe is to be used to make inferences on the variability of Carbon Monoxide (CO) concentration, with respect to Temperature (T) , and Relative Humidity (%) of the surrounding environment. The aim of this study is to model the CO concentration as a function of the more easily obtained measurements of T, and RH. Additionally, the inclusion of a time index as a covariate is to be investigated, and remarks concerning the quality of data collected, as well as critiques of data analysis methods used will follow.

Exploratory Data Analysis

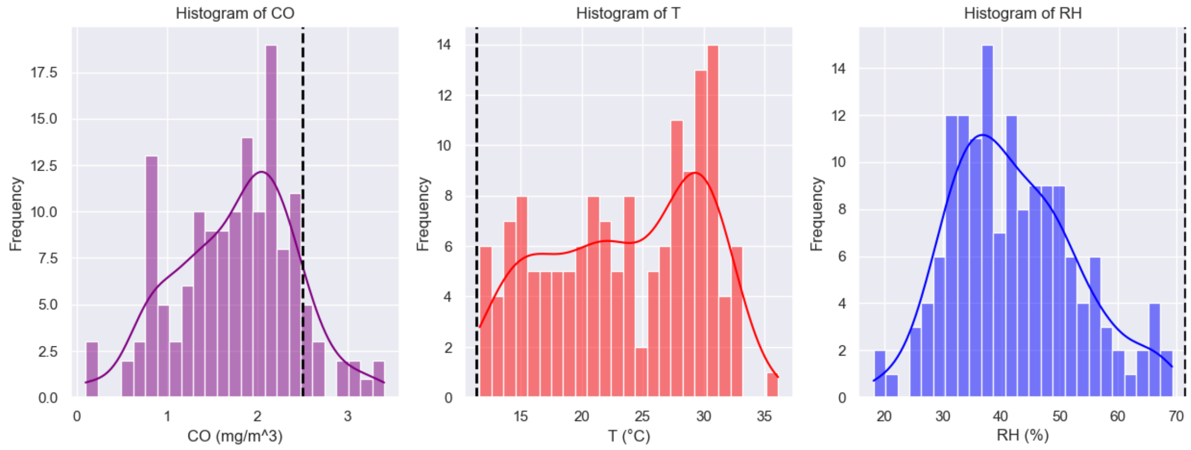


Figure XX: Histograms showing XX (L – R)

From the samples collected, some preliminary analysis was undertaken on the distributions shown by the 3 quantities of interest; see Fig. XX above. The raw data was firstly collated into a set of histograms, bin size 25, with kernel density lines shown to aid distribution inspection. This data is matched against expected values^{A} for the quantities of interest depicted using the black dashed line; see Appendix A.

The leftmost plot, shows Carbon Monoxide (CO) concentration in mg/mm^{3} over the course of the study. It is slightly negatively skewed and asymmetrical. As for the plot in the centre, showing variations in Temperature (°C), which is similarly negatively skewed but much more asymmetrical as the left tail shows there was a large range of temperatures experienced during the 150 day time frame. The rightmost plot is a histogram of the distribution of Relative Humidity. From its kernel density line estimate, it can be noticed that it is positively skewed and although asymmetrical, it is the least asymmetrical of the 3 plots. In comparison to the expected values, none of the plots perfectly match expectations from prior research. The CO concentration being the closest exhibits its peak at circa 2 mg/mm^{3} while the expected value was 2.5 mg/mm^{3}. According to the EEE[1], and another source, the annual mean temperature is between 2 - 17 °C. This was overshot greatly by the measurements which exhibit a peak frequency of 13 in the 26 – 28 °C range. Almost double the expected average. As for the RH, the humidity levels were expected to range from 60.5 % – 82.5%[2], but the population mean lies much lower at circa 37%. From the onset, these discrepancies suggest that there could be some fault in the methods, and/or apparatus used for measurement.

Scatter Plots were configured to gain even better insight into possible relationships within the dataset. The plots are for CO concentration against Temperature, CO concentration against RH, and Temperature against RH. Which yielded correlation coefficients of -0.2363, 0.1288, and -0.6659 respectively; see Fig. XX.?

[1] European Environment Agency

[2] CoolWeather.net

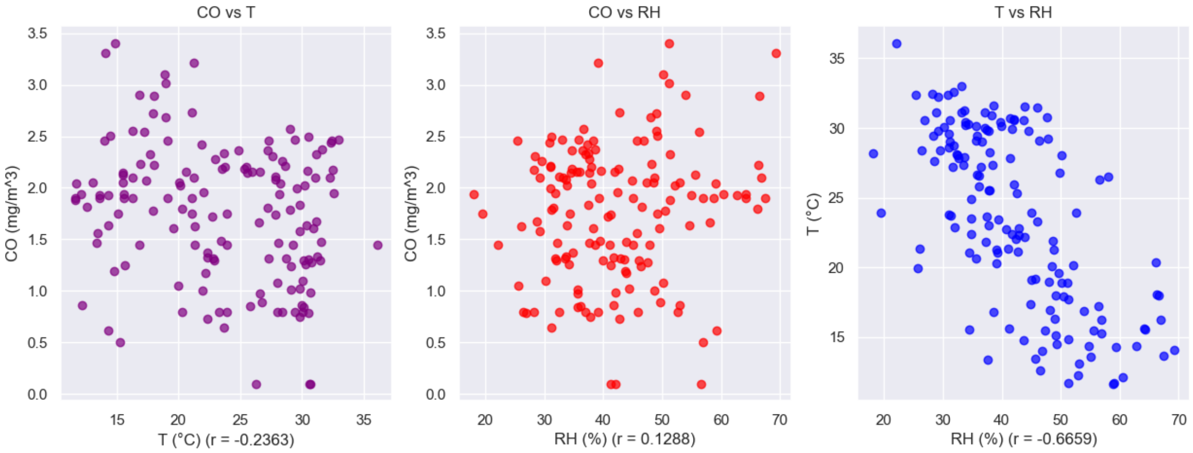


Figure XX: Scatter *Plots of XX (L - R)*

Modelling

The scatter diagrams in Fig. XX had linear regression models fit over i, and ii the resulting coefficients of determination are 0.0558, and 0.0166 respectively. This implied that the model suggests the variation in Carbon Monoxide Concentration was only about 5.6% due to changes in Temperature , and 1.7% due to changes in the Relative Humidity of the surrounding Atmosphere. According to textbook heuristic, since both values of r^2 obtained are below 0.5 i.e., 50%, the correlation must be weak. Another consideration would be that the relationship is non-linear, but from the scatter chart, no higher order polynomial is immediately obvious, and therefore the inference that the relationships are weak must be true.

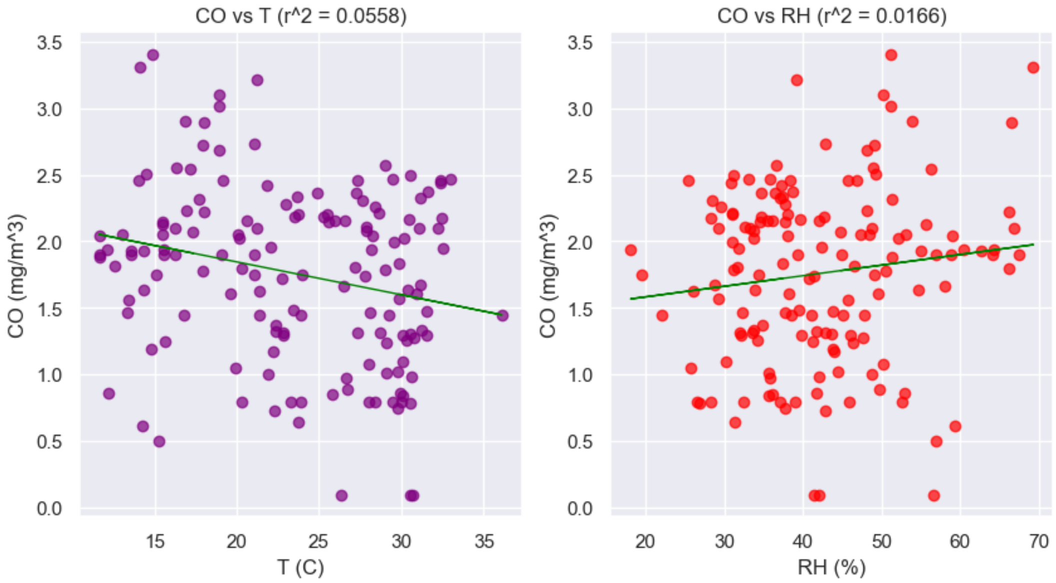


Figure XX: Scatter Plots with Linear Regression Fit onto them

As is evident from Fig. XX above, the lack of any trends in the plotted points, even when a best fit line is drawn shows that the linear regression model is not a great fit for the data sample available. Further investigation is needed, and given access to the entire population the next logical thing is to apply a linear model to a scatter of CO against T and RH making use of a polynomial regression instead . This resulted in a slight improvement in coefficient of determination but much to minimal to recharacterize the coefficient as moderate; increased r^2 from .

Changing from linear to a polynomial regression however, shows instant improvements to the calculated coefficient of determination. The outcome of increasing the order of the polynomial regression from linear to quartic is shown in the form of a bar chart in Fig. XX below. Here we can see that a quartic regression yields a coefficient corresponding to > 20% variance which while still weak, is a step in the right direction. Increasing the order of the polynomial continues this relationship until n = 10 is reached where there is a steep drop off to circa r^2 = -1.5 and at n = 11 the curve immediately bounces back and continues on its upward trend; see Fig. XX.

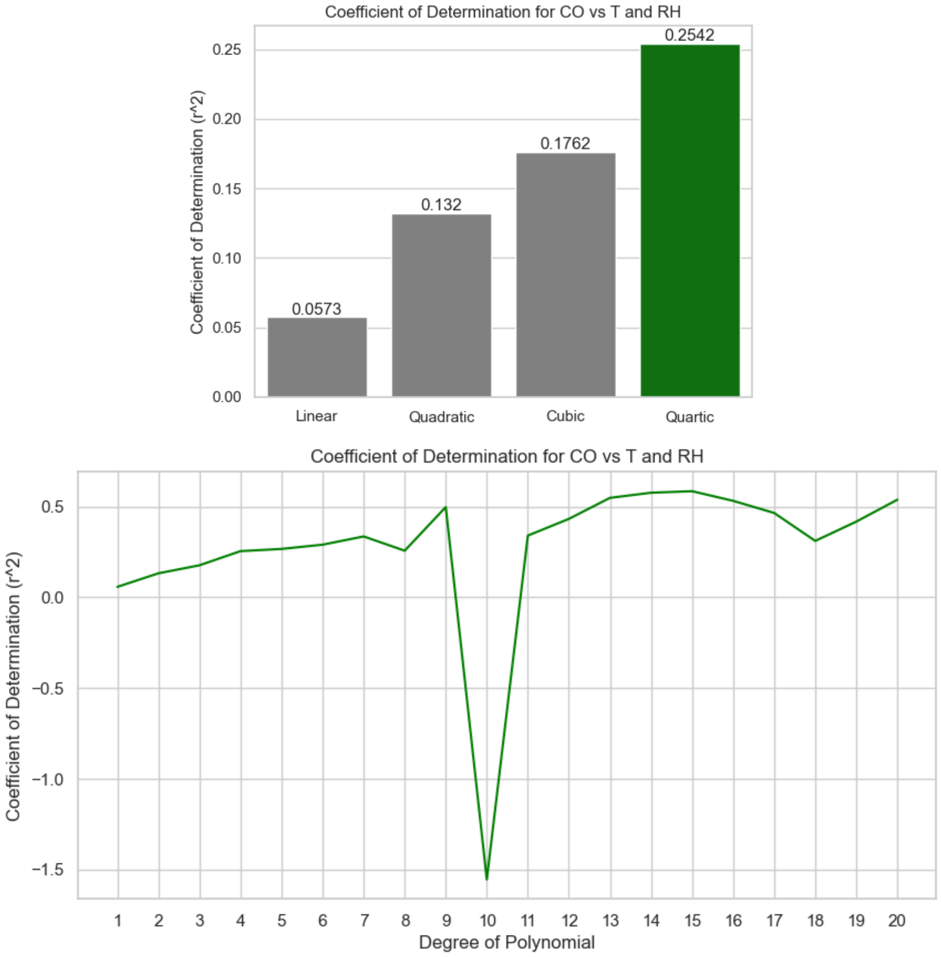
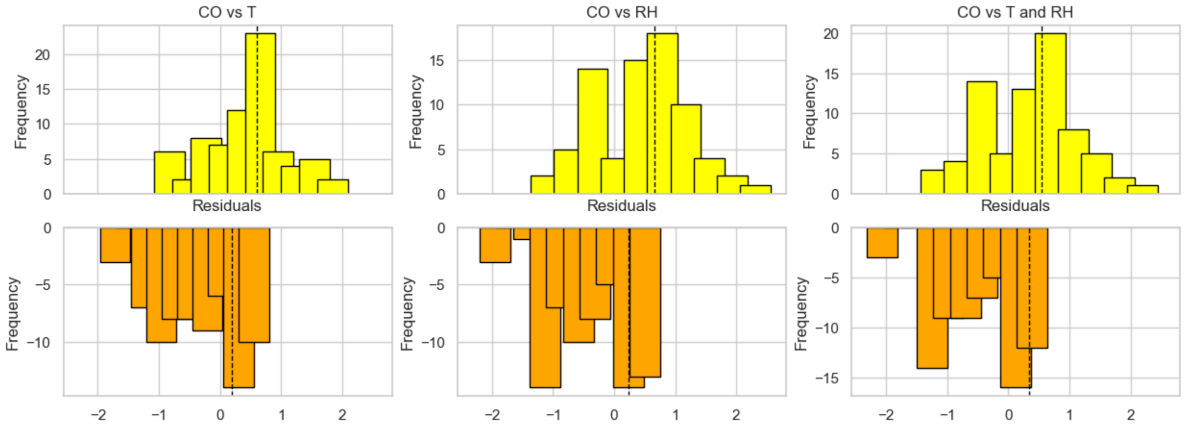


Figure XX: Coefficient of Determination as n increases

Prediction

Furthering the investigation into the relationships present, we employ the 2-fold cross-validation method, in which we use one half of the sample to instantiate a model, and the other half of the sample to test said model, recording the residuals that occur for each prediction. The 2-fold cross-validation method was



performed on all 3 models namely, CO vs T, CO vs RH, and CO vs T and RH.

Including Time as a Feature

Appendix A

References