ENVSCI 203/ EARTHSCI 263: 'Modelling of Environmental Systems'

Air Pollution Modelling

Assignment #3

Introduction

This lab is designed to provide you with some experience of working with empirical models. You will use linear regression to derive an empirical model of air pollution (read the refresher at the end of the handout *before starting* if you are not familiar with how linear regression works, as well as the relevant material on Cecil). The model is then evaluated at a second site on the basis of independent data. This lab will give you experience in building and then evaluating a simple empirical model, and relate to material covered in Lectures 10 and 16. **The lab write-up is due in Week 43 – that is the week starting on Monday 17th of October– by noon on the day that your practical class is held.**

The System

We are concerned with a model of air pollution in streetscapes as conceptualised by Johnson *et al.* (1976). Johnson *et al.* were interested in air pollution in built-up city streets. The relevant pages from the book in which this model is described are appended to this handout – you don't need to go and track down the original! We will consider carbon monoxide (CO) emissions from the vehicles as the pollutant. Therefore, we require a model that will allow us to predict the concentration of CO as a function of the significant variables.

The Model

The model assumes that emissions are released within a street and that the air recirculates about a primary vortex within the street; pollutants only escape via the top of the canyon (Fig. 1).

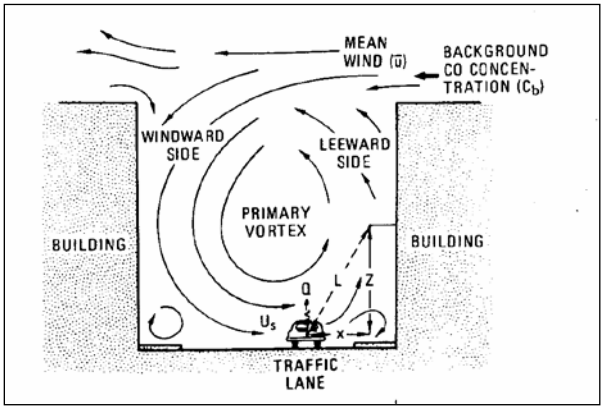

Figure 1 Schematic view of air circulation in an urban 'canyon' (from Johnson *et al.*, 1976).

Johnson *et al.* (1976) modelled the street as a canyon: closed sides with the only opening for exchange of air and pollution above the building that make up the sides. They collected a set of data that consisted of hourly values of CO concentration and other parameters they considered may be important. From the data set they concluded that during periods when the wind blows more or less normal to the street the CO concentration is related to the traffic flow rate, the average vehicle speed, and the wind speed at the roof top. From their results, Johnson *et al.* proposed the following relationship:

$$C_p = a\left[\frac{N\,S^{-0.75}}{u+0.5}\right] + b \qquad \text{Eq. 1}$$

where:

C_P is the predicted concentration of CO (ppm)

N is the traffic flow (vehicles.h^{-1})

S is the average vehicle speed (miles.h^{-1})

u is the wind speed at the rooftop (m.s^{-1})

For convenience, we will call the term in the square brackets X. This simplifies the model to:

$$C_p = aX + b \qquad \text{Eq. 2}$$

Note that X has no theoretical basis. Johnson *et al.* (1976) proposed this expression (it is Eq. 29 in their book chapter) because of all the different ones they tried they found this

gave the best fit to the data; hence, this is an empirical model. The variables *a* and *b* are the regression parameters, the values of which are to be found by a linear regression of the observations. We are modelling CO concentration as a linear function of *X*.

The Data

The emphasis in this lab is on the use of existing data to calibrate (or tune) and then evaluate/test an empirical model. You will calibrate the model for one site and then test it at another. The data supplied for the purpose consists of 10-minute averages for two sites in Hamilton, New Zealand (in file *SiteData.xls*). Your main task is to:

1. Use the data from the Vercoe site to calibrate the model. You will also evaluate the model's performance at this site.
2. Evaluate the performance of the model at the Te Rapa site (validate the model for use at a independent site).

For the purpose of modelling, we will assume that the vehicle speed is constant throughout the period of observation and that the average vehicle speed for the Vercoe site is 40 miles.h^{-1}.

Your Tasks

Part A: Observations Based on the Raw Data from Vercoe

First, construct, on separate graphs, time series of the observed concentrations, the wind speed and the traffic flow rates for the periods of data for the **Vercoe site** and comment on what you observe.

Part B: Calibrating the Empirical Model at Vercoe

We now want to use the observations from the Vercoe site to fit the empirical model of the equation given above. To do this we will find *a* and *b* using the method of least squares estimation[1].

[1] See Appendix B if you don't know what 'least squares' refers to.

1. Construct a scatter plot to show that the observations from the Vercoe site do indeed suggest a linear relationship between *X* and CO concentrations. To do this you need to create a new column for *X*, and calculate values based on Eq. 1, above[2].

2. Fit the model – based on the observations from the Vercoe site, you need to determine the optimum values for *a* and *b* for the street canyon model.

   To do this, right click on the the data you have plotted and select the **Add Trendline** … option under **Chart**. Select the **Linear** option under **Type** and choose to **Display Equation on the Chart** under **Options.** This will show you the values of *a* and *b* and hence give the equation for the model ($y' = aX + b$) – you can now use the model parameter values to create a column of predicted concentrations.

   Make sure you have the dependent and independent variables on the correct axes - think about what is being predicted and what is being used to make that prediction!

3. Assess the quality of the model performance by:
   - Constructing a predicted versus observed scatter plot. It is always useful to include the perfect fit line (predicted = observed, **not** the line of best fit) for such comparisons. It is also helpful to make such scatter plots square in size (i.e., physical length of both axes the same and axes range the same). This allows for a better visualisation of how different the predicted values are from the observed values.
   - Comparing the observed and predicted values of the carbon monoxide concentrations by plotting them as a time series with the predicted values overlaid on the observed values.

4. Assess whether there are any significant departures from the linear model by plotting:
   - observed versus residual (remember, the residual is the predicted – the observed)

[2] Note that to raise a value to a power in Excel you use the ^ symbol; so, for example, $2^{-0.75}$ would be entered as 2^-0.75

- residual as a time-series (this will allow you to determine whether the model error varies systematically over the measurement period)

5. Construct a table of model evaluation metrics (*n*, mean observed, mean predicted, SD observed, SD predicted, RMSE, MAE, RRMSE) for the Vercoe site and comment on the results; details for these metrics are provided in the handout from Lecture 15, and equations are given at the end of this handout (Appendix A).

6. Based on 1 to 4, comment on the performance of the model at the Vercoe data.

PART C: Validation of the model at Te Rapa

An empirical model is only really tested if it is evaluated using a set of data that is independent from the data used to parameterise the model. Your task in this lab is to evaluate the model you constructed last week using independent data and, by doing so, to test the applicability of the model to the second site, Te Rapa.

1. Use the model constructed in Part A (the optimum parameters for the model determined using the Vercoe site) to model carbon monoxide concentrations at the Te Rapa site. *That is use the estimates of* a *and* b *from the Vercoe site to model CO concentration at Te Rapa.* Note that there are missing data in this set – as is common with monitoring data – you will need to decide how to deal with this; the easiest option is just to leave a gap in the graphs.

2. Make an initial assessment of how applicable the model is at Te Rapa site by plotting predicted (modelled) versus observed carbon monoxide concentrations.

3. Compare the observed and predicted values of the carbon monoxide concentrations by plotting both on a time series graph.

4. Construct a table of model evaluation metrics (*n*, mean observed, mean predicted, SD observed, SD predicted, RMSE, MAE, RRMSE) for **BOTH** sites and comment on the

results; details for these metrics are provided in the handout from Lecture 16, and equations are given at the end of the handout.

5. Assess whether there are any significant departures from the linear model by plotting:

- observed (*x*-axis) versus residual (*y*-axis)
- change in residual over time

6. Comment on the overall performance of the model at the Te Rapa site.

PART D: General Discussion of the Model

1. By considering what would happen without the 0.5 constant in the denominator of X, explain why it has been included in the empirical model. If in doubt, try applying the (modified) model to the Te Rapa data to see what happens.

2. Suggest another variable to include in the street canyon model that you think may help to improve its predictive ability. Explain how this variable affects carbon monoxide concentrations.

Your Report

Your lab report must include:

- A brief introduction, with 2-3 appropriate references (½ page is sufficient).
- The main results, including:
    1. The graphs that are specifically asked for, for **BOTH** sites
    2. A table of model evaluation parameters (i.e., MAE, RMSE, etc.) for both sites – you must provide an interpretation of these parameters and **NOT** just the raw tables of data!
    3. Interpretation/discussion of all results and graphs that you include in your report.
- Answers to questions posed above, and comments where appropriate.
- A short discussion of the limitations of this model.

References / Useful Reading

Dirks, K.N., Johns, M.D.,Hay, J.E. and Sturman, A.P. 2003. A semi-empirical model for predicting the effect of changes in traffic flow patterns on carbon monoxide concentrations. *Atmospheric Environment* **37**: 2719-2724 [on Cecil].

Johnson, W.B., Skllarew, R.C. and Turner, D.B. (1976). Urban air quality simulation modelling. In: Stern, A.G. (ed.) *Air Pollution, Volume 1* (3rd ed). Academic Press. New York, p. 503 – 562 [model is discussed on p. 529-31 and is attached to this handout].

Appendix A – Formulae for Model Evaluation

1. Mean absolute error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|(P_i - O_i)\right|$$

where: N is the no. of pairs of data points for comparison, P_i is the i^{th} predicted value, and O_i is the i^{th} observed value

This is the mean of the absolute value of the residuals. To compute the absolute value use the ABS function in Excel. You need to compute the residual for each pair of observations and predictions, calculate the absolute value of each and find the average of those values. The MAE tells you the average distance (i.e., the residual) from a prediction to an observation.

2. Root mean squared error

$$RMSE = \sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2\right]}$$

This is the root-mean squared error. You need to compute the residual for each pair of observations and predictions, square each of those values, find the average of those squared values and then square-root it. The RMSE tells you the average squared distance from a prediction to an observation – the use of the square means that larger errors are up-weighted relative to smaller ones.

3. Relative root mean squared error (RMSE)

$$RRMSE = \frac{RMSE}{\overline{O}}$$

This is the relative RMSE. It is used to compare model performance at sites where the values may differ greatly. To make this clearer, imagine two sites A and B. At A the average observation is 10000 and at B it is 10, the error in prediction is 10% at both.

This means the MAE at A will be around 1000 but at B only 1 and so comparing a model's performance via the MAE across multiple sites is not a fair test. The RRMSE corrects for this.

Remember that:

- $\sum$ denotes 'sum of', so $\sum_{i=1}^{N} x_i$ means 'sum all of the *x* values from the 1st to the *N*th'.
- | *x* | denotes the absolute value of *x* (the absolute value is the distance from zero), so |-6| = 6 and |6| = 6 (strictly, the absolute value of *x* is $\sqrt{x^2}$)

Appendix B – What Does 'Least Squares' Mean?

Relationships between Variables: Linear Regression

Often we are interested in the relationship between variables, or in trying to predict the value of one variable (*y*) based on the value of another variable (*x*). In these cases we can use regression. First, some terminology: the independent variable is usually called *x* and the dependent variable is called *y*. This is because we depend on the values of *x* to predict *y*.

The simplest possible relationship between our two variables is a straight line. Although there are methods for making predictions when the relationship is nonlinear, these methods are not considered here. Given that the relationship is linear, the prediction problem becomes one of finding the straight line that best fits the data. Since the terms 'regression' and 'prediction' are almost synonymous, this line is called the regression line. The mathematical form of the regression line (see Fig. A1), predicting *y* (the dependent) from *x* (the independent) is:

$$y' = mx + c \qquad \text{(Eq. A1)}$$

where: *x* is the variable represented on the abscissa (*x*-axis), *m* is the slope of the line, *c* is the point at which the lines intercepts the *y*-axis, and *y'* consists of the predicted values of *y* for the various values of *x*.

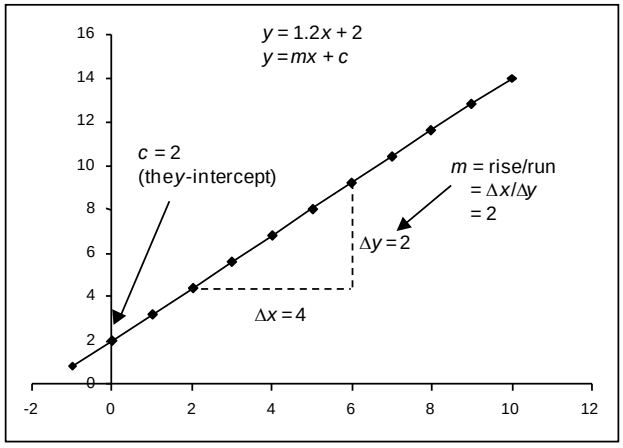

Figure A1 Schematic overview of the linear regression model ($y' = mx + c$).

So, how do we calculate the values of *m* and *c*? Again, the formulae might look tricky but in reality they are not too bad. First, we need to calculate the slope of the regression line (*m*):

$$m = \frac{\sum (x - \bar{x})^2 (y - \bar{y})^2}{\sum (x - \bar{x})^2} \qquad \text{Eq. A2}$$

This equation for *m* (the slope of the regression line) is clearly related to the equations for calculating *r* – the correlation coefficient. In fact, *m* is equal to the ratio of the co-variation of *x* and *y* to the variation in *x*.

Now, we can calculate *c* (the *y*-intercept) as:

$$c = \bar{y} - m\bar{x} \qquad \text{Eq. A3}$$

Having calculated *m* and *c* we are in a position to calculate an expected value of *y* (i.e. *y*') for every value of *x*!

The resulting line is that which comes close to the points as possible. As an example, consider the following data:

x	***y***	***y'***	***y-y'***	***(y-y')²***
2	5	4.8	0.2	0.04
3	6	6.6	-0.6	0.36
4	9	8.4	0.6	0.36
5	10	10.2	-0.2	0.04

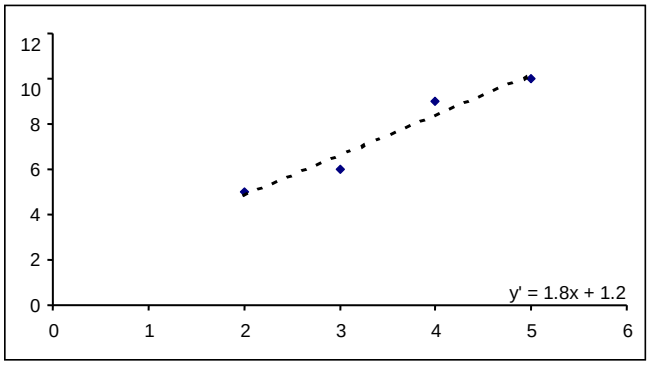

Figure A2 Simple example of the calculation of the linear regression model.

The four (*x,y*) pairs found in the first two columns of the table are indicated in the plot as the four points. The line in the plot is the best fitting straight line; it has a slope (*m*) of 1.8 and a Y-intercept (*c*) of 1.2. The first value of *y*' is 4.8. This was computed as: (1.8).(2) + 1.2 = 4.8. Previously I said that the regression line is the best fitting straight line through the data. More technically, the regression line minimizes the sum of the squared

differences between *y* and *y*'. The third column of the table shows these differences and the fourth column shows the squared differences. The sum of these squared differences (0.04 + 0.36 + 0.04 + 0.36 = .80) is smaller than it would be for any other straight line through the data. Since the sum of squared deviations is minimized, this criterion for the best fit is called the 'least squares criterion'. Notice that the sum of the differences (.2 -.6 + .6 -.2) is zero.

The $y - y'$ part of the analysis above is also called the *residual* (i.e. this is the difference between the observed and the predicted values of *y* for each observation). As described above, the residual is what we want to try and minimise when performing linear regression. If the regression we've used is a good model of the data then the residual values will all be relatively quite small. Because the technique we have used to calculate the regression minimises the square of the residuals it is sometimes called the 'ordinary least squares' method.

The residual values can be used in another way: by comparing the variance in the residuals to the variance in the dependent variable (*y*) we come up with a simple measure of how well the variance in *y* is explained by the regression line. Obviously the more variance in *y* that is explained the better the regression! This measure is usually called the coefficient of determination, or r^2 (because it is the square of the correlation coefficient, *r*), and is simply the ratio of the variance in the residuals to the variance in *y*:

$$r^2 = \frac{\text{variance in residuals}}{\text{variance in } y} \qquad \text{Eq. A3}$$

If all the regression points lie exactly on the line then the variance in the residuals and y will be identical and $r^2 = 1$. The greater the scatter of y around the regression line then the closer to zero r^2 will be. r^2 is a simple, intuitive measure of the success of a regression in terms of the amount of variance it explains.