

02424 Assignment 1: Dioxin emission

This is the first of three mandatory assignments for the course 02424. Submission must contain only one collected file in portable document format (pdf); other formats will not be accepted.

Background

Dioxin is a shorthand name given to a family of 75 different chemical compounds with a similar chemical structure and a set of biological effects though their potency varies widely. Dioxins are organochlorines or compounds formed when a chlorine molecule binds with a carbon molecule during combustion or in some type of industrial production process. Dioxins cause cancer in humans and animals. Animal tests also show interference with reproduction, development and immune system function at low doses. This raises substantial concern about current human exposure levels.

The toxicity of a dioxin compound will vary depending on the number and position within the molecule of the chlorine atoms. One of the most well known (and most toxic) forms of dioxin is 2,3,7,8-tetrachlorodibenzo-p-dioxin, or TCDD28 (the 'Seveso' dioxin).

Municipal solid waste (MSW) and medical waste incinerators are the most significant sources of dioxin in the environment collectively accounting for roughly 85-87 % of all known dioxin emission.

The purpose of this project

The purpose of this assignment is to find a model for the variation of measured dioxin emission at a number of Danish municipal solid waste incinerator plants. It is of particular interest to investigate whether the operating conditions influence the dioxin emission.

In order to create the data needed for setting up such a model, a large number of experiments have been conducted at a number of Danish MSW incinerator plants. During these experiments gas samples have been collected and the dioxin emission estimated in the samples. Likewise a large number of possible explanatory variables have been measured.

The experiments

Care must be taken in planning the experiments to ensure that useful models and reliable conclusions can be formulated. Furthermore, dioxin measurements are difficult to obtain and the analyses are expensive. In order to ensure reliable conclusions and to obtain maximum information in the data, statistical designed experiments have been used.

The experiments have been conducted at three Danish MSW plants. For one of the plants the experiment was repeated at a later time. The layout of an experiment is shown in Figure 1.

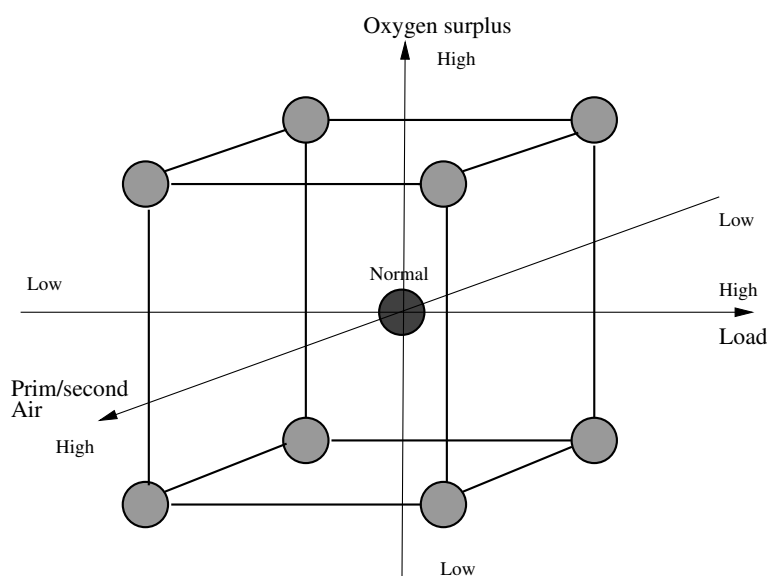


Figure 1: Experimental setup corresponding to a single visit at a MSW plant.

Dependent variable

The dependent variable is the concentration of dioxin in the combustion air. More precisely the concentration is measured as the total amount of dioxin in ng pr. m^3 . In the data file the dependent variable is called DIOX. **You should consider transforming the dependent variable in the analysis.**

Explanatory variables

The total set of explanatory variables is most conveniently divided into the so-called block effects, active and passive variables.

Block effects

The experiments are conducted under some conditions which we need to take into account in the modelling. These conditions are often termed **block effects**. In the list of block effects the name of the effects are shown in brackets. The block effects are:

- MSW Plant (PLANT). The experiments have been conducted at three Danish MSW plans named RENO_N, KARA, and RENO_S.
- Time (TIME). For the plant RENO_N, the experiment was repeated at a later time point (1,2).
- Laboratory (LAB). Two laboratories have been used for the analysis, one in Denmark (KK) and one in US (USA). It is very difficult (and expensive) to measure the amount of dioxin, hence the data is assumed to be encumbered with considerable measurement noise.

Active and passive variables

The explanatory variables are considered as being either active or passive. The active variables are those varied according to the experimental plan, while the passive variables are all other measured variables which might influence the dependent variable.

Active variables:

- Oxygen surplus in gas (OXYGEN)
- Plant load (LOAD)
- Air distribution (primary/secondary) (PRSEK)

In order to obtain the optimal amount of information each measurement corresponds to one of the corners in a cubus, as shown in Figure 1. The order of experiments within the cubus is randomised. Even though the experiment is planned as described above it is often rather difficult in practice to obtain the desired values of the active variables, Therefore is is more reasonable to use the actual measured values related to the design. Those values are:

- Measured oxygen surplus is called (O2). In the data file a value of the oxygen corrected by the mean is given by (O2COR).

- The normalized measured plant load is called (NEFFEKT). It is defined as the difference between the actual effect and the mean effect divided by the mean effect. A value of $\text{NEFFEKT} = -0.15$ means that the load is 15 % less than the design load.
- The ratio between the primary air and the secondary air is called (QRAT).

Passive variables:

- Gas flow (QROEG) (m^3/h)
- Combustion chamber temperature (TOVN) ($^{\circ}C$)
- Gas temperature (TROEG) ($^{\circ}C$)
- Pressure in the chamber (POVN)
- CO_2 (CO2) (ppm)
- CO (CO) (ppm)
- SO_2 (SO2) (mg/m^3)
- HCl (HCL) (mg/m^3)
- H_2O (H2O) (%)

CO_2 , CO, SO_2 , HCl and H_2O are measured in the gas.

The observations are numbered by a unique number as represented by OBSERV.

How to get the data?

The data can be found in File sharing on CampusNet in a file called `dioxin.csv`. The first row of the file contains the name of the variables referred to in the text.

The data can be imported into R with the following command:

```
dat <- read.table("dioxin.csv", sep=',', head=TRUE)
```

Goal of the project

The goal of this project is to find a model for the variations observed in the reported dioxin emission. For the analysis use $\alpha = 5\%$. It is important that you

1. Read the introduction carefully
2. Specify the models and the underlying assumptions
3. Explain how you reduce the models
4. Check the underlying assumptions (residuals)
5. Explain your results (remember to take all transformations you may have done into account)

Particularly we want you to answer the following questions/do the following analysis:

1. Start with some preliminary/explorative analysis of the data by making some plots.
2. Set up a simple additive model only with the active and the block variables. Reduce the model if possible.
3. Set up a similar model as before but now use the measured values of the active variables along with the block variables. Reduce the model if possible.
4. Using the model with the measured active variables, predict the dioxin emission in the first visit to the RENO_N plant, analysed in the KK laboratory with $O2COR = 0.5$, $NEFFEKT = -0.01$ and $QRAT = 0.5$ (you should disregard values of variables that you have removed from the model). Give a 95 % prediction interval as well.
5. Does the dioxin emission depend on the operating conditions ($O2COR$, $NEFFEKT$ or $QRAT$)? If 'yes', how can the dioxin emission be reduced by changing these conditions.
6. Do you see any differences between the considered MSW plants? What about the two laboratories?
7. Set up a final model, this time with the passive variables as well. You may want to consider including higher order terms in the model to find the model that gives the most complete description of the variation in

the dioxin emission. Use residual analysis to validate the model and check if some observations are particularly influential. Give estimates of the parameters in the model and their uncertainties.

8. Make a brief abstract for your “grandmother”, i.e. summarize your final model in plain words (no more than 250 words).
9. (Possibly difficult) It has been shown that the precision in the KK laboratory is better than in the USA laboratory. Having this in mind, re-estimate your final model, using likelihood estimates of the weight between the measurements in the two labs. Also consider the uncertainty of the estimated weight.¹

¹A better estimate for the weights is the REML estimate (which we will cover later on in the course), and you may compare with the `gls` function in the `nlme` package using the `weights` argument.