# MA30091 Applied Statistics 2023/24

# Practical 1: Defoliation in Spruce Trees

## Background

Forest health monitoring schemes were set up across Europe in the 1980s in response to concern about air pollution causing a deterioration of forest health, and have continued since then. Trees which are not healthy lose their foliage; how many leaves or needles are lost depends on how badly trees are affected. Coniferous trees only recover slowly from bad health, since they carry needles from 7 years of growth.

Recent threats to forest health are climatic extremes likely to be due to global climate change, increased ground ozone levels and nitrogen deposition. Changes in weather patterns have resulted in severe droughts and hence may also have caused a deterioration of forest health. Here we will analyse data from Baden-Württemberg, Germany. In this area spruce, a coniferous tree, is the most common tree species and we will be concentrating on this species. The forests in this area are, compared to the UK, very large and extend for example in the case of the Black Forest over about 200 kilometres.

Forest scientists are interested in which of the explanatory variables drive tree defoliation, which of the variables have the strongest effect and what type of effect the variables have. They would also like to predict defoliation of spruce, given certain values of the explanatory variables.

## Data

Data on percentage tree defoliation from a monitoring survey carried out annually in Baden-Württemberg, Germany are available. The response variable defoliation, other site specific explanatory variables (e.g. soil type) and tree specific variables (e.g. age) are recorded yearly. Here we will analyse the data collected in 1994 on spruce trees.

The response variable recorded in the survey is the percentage of defoliation in the tree crown (i.e. the top of the tree) estimated by eye in 5% classes for each individual tree using binoculars. At each grid location 24 trees are systematically selected. Here we will analyse the aggregated response over a maximum of 24 sample trees at each grid location. This response is provided on the scale of a proportion (**ratio**). The aggregation was carried out because the forests are heavily managed and at any grid location trees will almost always have the same age and will have been planted in a regular pattern, hence there are generally no big differences between trees in terms of height and dominance.

As explanatory variables we consider mean age, site-specific and weather characteristics. Note that all these variables are mean or modes over the respective observed values at the 24 trees sampled at each grid location. In addition we consider summary statistics on precipitation and weather during the preceding 2.5 years of the sample survey in 1994 at each grid location. These variables are derived from the predictions of a mathematical weather model. Explanatory variables on pollution levels are not available.

The data have been cleaned, and there should not be any errors.

The data may be downloaded from **https://moodle.bath.ac.uk/mod/folder/view.php?id=1306030**

## Description of variables

|  |  |
| --- | --- |
| ratio | mean proportion defoliation of the trees sampled at the grid location. |
| tmin, tmedian, tmax | minimum, median and maximum daily mean temperature (°C x 100) for the preceding 2.5 years (1/1/1992 – 30/6/1994) at the grid location |
| pmedian, pmax | minimum, median and maximum daily precipitation in mm\*100 for the preceding 2.5 years (1/1/1992 – 30/6/1994) at the grid location |
| daysZeroP | number of days with zero precipitation for the preceding 2.5 years at grid location |
| daysT20 | number of days with mean daily temperature above 20°C for the preceding 2.5 years at grid location |
| daysBoth | number of days with zero precipitation and mean daily temperature above 20°C for the preceding 2.5 years at grid location |
| area | different growth areas characterise different growth conditions |
| slopedir | mean direction of the slope of the trees sampled at the grid location |
| relief | topology (plateau, valley, etc.) at the grid location |
| soiltxt | what type of texture the soil has at the grid location |
| soiltype | soil type at the grid location |
| altitude | mean altitude in metres of the sampled trees at the grid location |
| age | mean age of the sampled trees at the grid location |
| slopegrad | mean of slope gradient at the grid location |
| nobs | number of spruce trees sampled at location (between 1 and 24) |

## Instructions

Your task is to carry out a data analysis to determine which of the explanatory variables drive tree defoliation. Which of the variables have the strongest effect and what type of effect do they have?

In **Week 1** you should concentrate on carrying out an initial data analysis of the data to familiarise yourself with the data and describe general patterns.

* Identify the response variable and key explanatory variables.
* Determine the type of variables (eg categorical, continuous).
* Produce simple numeric and graphical summaries.

In **Week 3** you will work on formulating a sensible model, checking the fit and interpreting results:

* What type of analysis is required (eg descriptive, prediction, hypothesis testing)?
* Choose an appropriate linear or generalised linear model to address this question.
* What types of bias may be present and how can your model address these?
* How well does the model fit?
* What are assumptions behind the model and are they valid?
* Answer the research question.