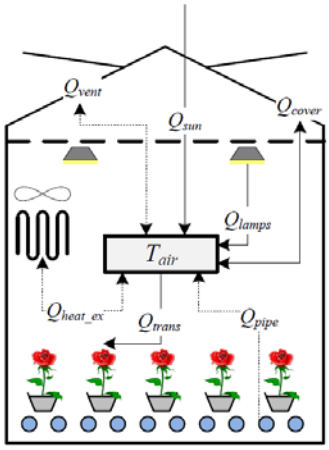
Practical Pattern Recognition, Biosystems Engineering, Friday Week 1

Temperature prediction inside a greenhouse

Control over climate variables such as temperature, carbon-dioxide concentration, and humidity, is an important aspect in horticulture. There are all sorts of technology available that help the grower control the climate in a greenhouse, for example screens that block sunlight, or windows that open mechanically, CO_2 injection, and heaters. However, greenhouse climate is a very complex system, and it is not easy to control, even with advanced technology. In this practical we focus on the temperature inside the greenhouse. Figure 1 shows the various energy fluxes that influence temperature.



Energy fluxes ($W\,m^{-2}$) (1)

- Q_{sun} = Radiation from sun
- Q_{cover} = Cover losses
- Q_{trans} = Crop transpiration
- Q_{lamps} = Artificial lighting
- Q_{vent} = Ventilation
- Q_{he} = OPAC heat exchanger
- Q_{pipe} = Pipe rail heating

See Van Beveren et al. (2013).

Figure 1. *Various energy fluxes that influence the air temperature inside the greenhouse.*

A lot of sunshine (external variable) may increase the temperature too much, and an appropriate control action is to open a window, or close a light screen. How much you want to close it, depends on the amount of sunlight, but also on the state of other variables, for example on whether a heater is on or not. A first step in helping the grower controlling the climate is making a predictive model. This may help the grower to select appropriate control actions. The variable to be predicted is the temperature inside the greenhouse. Many variables are measured. Some variables can be controlled by the grower himself. Some variables are external, which means they may be regarded as disturbances, or uncontrollable. See Table 1 for an overview.

Variable	Name	Dimension	Type of variable
Temperature inside	Temp	°C	Output
Relative humidity inside	H	%	External
CO2 inside	CO2	Ppm	External
Wind speed outside	Vwind	m/s	External
Temperature outside	Tout	%	External
CO2 outside	CO2out	Ppm	External
CO2 dosage	CO2dos	Kg/ha/hr	Control
Artificial lighting	Lamp	%	Control
Opening of light screen	LS	%	Control
Opening of energy screen	ES	%	Control
Windows opening leeward	WL	%	Control
Windows opening windward	WW	%	Control
Temperature of pipe to heat exchanger	Tp	°C	Control

Table 1. Types and names of measured variables inside a rose greenhouse.

Note: relative humidity inside and CO2 inside are actually variables that may be controlled by the grower, but for now we regard them as external.

We make a linear regression model for temperature inside, as a function of the external variables and control variables, and investigate the prediction accuracy.

Instructions for exercises:

Please keep in mind the following attention points:

- Clear English
- Clear explanation of your answers
- No thesis format, i.e., no motivational introduction and no general discussion or general conclusions.

Exercises

1. Read the data file Data2011.csv in RStudio. This data set contains hourly measurements during 100 days in 2011, so in total 2400 measurements. Call this set 'GHdata'. Make a data inspection with the command 'pairs'. In terms of temperature (T) correlations, what do you observe?
2. Make a multiple linear regression model with all variables. Inspect leverage points and residuals with the command 'plot'. Are the conditions for linear regression met?
3. Perform an Anova with the command 'summary'. Almost every variable is significant. Does this imply they are also contributing to the model in terms of prediction? Write down the formula describing the model
4. Remove all non-significant ($p>0.05$) variables, e.g., with the command structure 'GHdata[c(x,x,x)]<-list(NULL)'. Some coefficient values have changed. Do you observe a relation between change in coefficient size, and significance of the corresponding variable? Does this surprise you? Give a statistical explanation using equation 3.14 from the book.
5. Remove all non-significant variables again, and repeat this until all variables are significant.
   a. Write down the model. Which variable selection method does this algorithm resemble most? What is the difference between these methods?
   b. Explain what goes wrong with this method regarding prediction accuracy (hint:use R^2adjusted values).
6. Start again by reloading the dataset. Use the command "rm(list=ls())" to remove all previous objects and data.

a. This time we take another approach. Remove in each step the least likely variable, until R^2_{adj} starts decreasing. Only remove variables with a p-value higher than 0.05.

b. Write down the model, and compare it with the one obtained in 5.

7. Pipe temperature seems not very significant, however it surely has influence on air temperature. What could be a cause of the lack of evidence in our dataset? (Hint: think of the role of the grower)
8. a. What are the assumptions behind this model regarding causality?
   b. what are the assumptions behind this model regarding dynamics? Discuss whether all assumptions are realistic.
9. See if you can improve the model prediction accuracy by model extension: transform y, introduce interaction terms (start with model obtained in 6.). Which practical problem do you encounter? Write down your best model in terms of R^2_{adj}. Compare this model with the one obtained in 6. Discuss your observations.
10. Describe how you think this model may help the grower in obtaining desired temperature, regarding financial costs, and technological feasibility.

Warning model

Finally, we would like to give the grower a warning in case our model predicts that the temperature inside the greenhouse will be high ($T>22\ ^oC$). For this, we make a logistic regression model.

11. Read the data file Data2011_2.csv. This dataset is the same as Data2011.csv, but extended with the columns Hot ("yes" or "no"), and Type ("Test" or "Training"), to indicate whether the temperature is high or normal (hot/not hot), and whether this data should be used for testing or training the model. Use the following code to create a confusion matrix cold/not cold predictions based on the temperature outside, and sun radiation:

```
glm.fit=glm(Hot~Tout+Rsun,GHdata[Type=="Train",],family=binomial)
glm.probs=predict(glm.fit,GHdata[which(Type=="Test"),],type="response")
glm.pred=rep("no",length(which(Type=="Test")))
```

```
glm.pred[glm.probs>x]="yes"
Table1=table(glm.pred,Hot[which(Type=="Test")])
```

12. What does **x** represent? Insert the value corresponding to a threshold of $p=0.5$.
13. Run the code. What is the corresponding true positive rate, and what is the false positive rate? Show your calculations.
14. Make formulas for true and false positive rates in terms of the table, by filling in the question marks:

```
FPR=Table[?,?]/(Table[?,?]+Table[?,?])
TPR= Table[?,?]/(Table[?,?]+Table[?,?])
```

15. Generate a warning model ROC curve by varying **x**. You may use the code

```
iter=100;
FPR = rep(NA, iter); TPR = rep(NA, iter)
for (i in 1:iter) {
  p=1/iter*i  # vary p from 0.01 to 1
  glm.pred=rep("no",length(which(Type=="Test")))
  glm.pred[glm.probs>p]="yes"
  Table1=table(glm.pred,Hot[which(Type=="Test")])
  FPR[i]=Table[?,?]/(Table[?,?]+Table[?,?])
  TPR[i]= Table[?,?]/(Table[?,?]+Table[?,?])
}
par(mfrow=c(1,1));
plot(FPR,TPR, ylim=c(0,1), ylab = "True positive rate", xlab="False positive rate")
```

16. What would you say about the reliability of these models when looking at the ROC curve? What could in your opinion improve reliability?