Lecture: Extrema der Luftschadstoffe Ozonextrema

Noelia Otero

April-2020

Information & data

-You can find the information and the data at

poincare.met.fu-berlin.de: /home/otero/Lab_extremes/

- data
- Instructions (Lab_April_2020.pdf)
- scripts

-Alternatively you can find the data: *https://github.com/noeliaof/Lab_extremes*

-We will use Rstudio (or R with some graphical interface, e.g. X11).
https://rstudio.com/products/rstudio/download/

Before starting, some Rstudio notes

-During the exercise we will use the following packages:

- MASS
- stats
- ggplot2
- dplyr
- relaimpo

-To install the packages:

install.packages("name")

Introduction

- The main sources of near-surface O3 pollution include both natural and man-made emissions of volatile organic compounds (VOCs) and nitrogen oxides (NOx). Under ultraviolet radiation, they go through a series of photochemical reactions and produce O3.
- Surface ozone concentrations are strongly dependent on meteorological variables, such as solar radiation fluxes, temperature, cloudiness, or wind speed/direction.
- Major episodes of high concentrations of ozone are associated with slow-moving, high-pressure weather systems that usually bring high temperatures and stagant conditions. Therefore, O3 variability is also controlled by meteorological factors.

Objective

The exercise is divided is two main parts as follows:

Exercise:

1. Data analysis: Examination and visualisation (time series, scatter plots, boxplots, histograms..)
2. Regression analysis to assess the ozone variability and the impacts of the different meteorological variables.

Key questions:

- How is the relationship O3 and the meteorological variables?
- Which is the most significant predictor?
- Which is the best model?
- What are the main seasonal differences?

Getting started

```
load("data/data_year_o3.Rda")
head(data_o3)
```

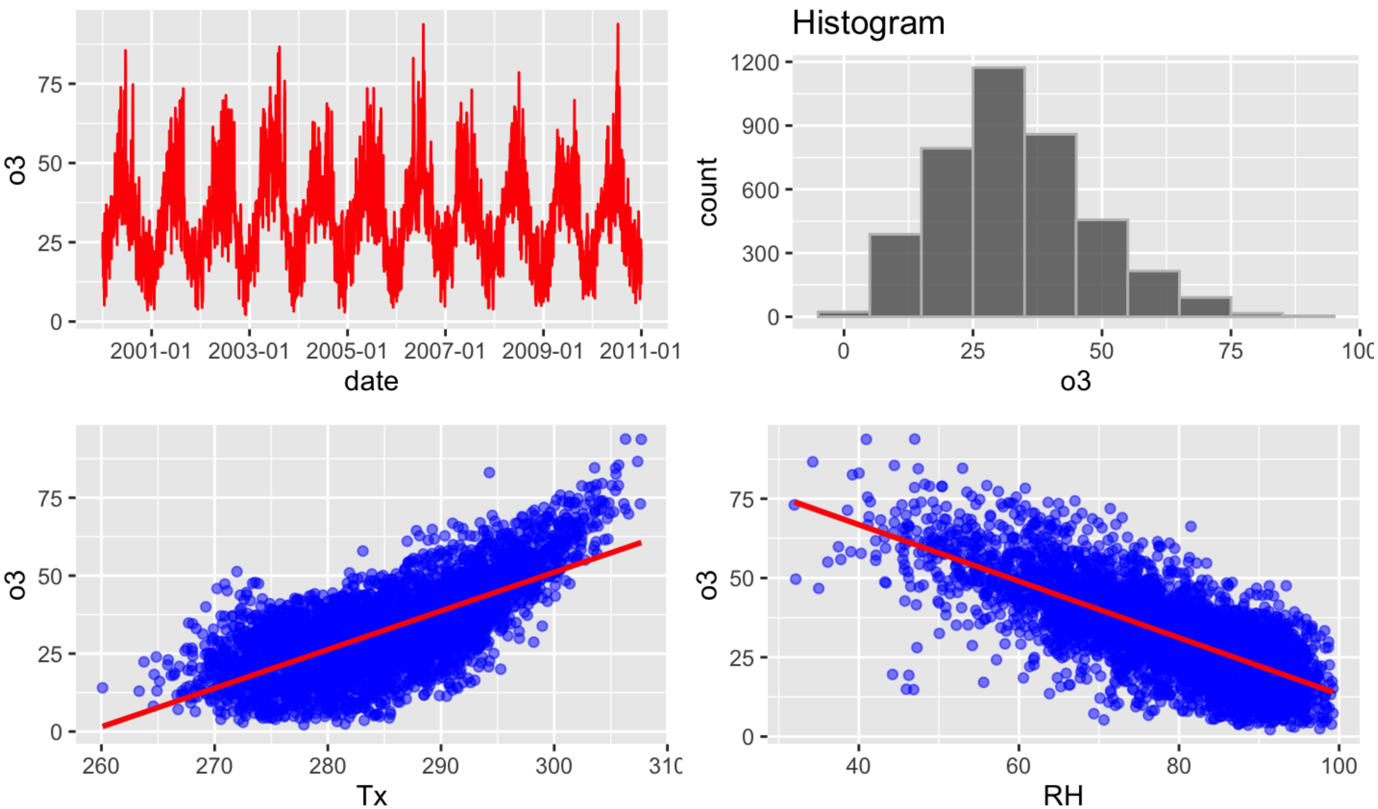
```
##         date       o3      blh       rh      ssrd       tcc     tmax Direction
## 1 2000-01-01 25.73896 370.5228 95.34355 11.083705 0.9210474 277.7885  247.7148
## 2 2000-01-02 25.63040 237.0636 94.27022  1.983333 0.7123587 277.3701  251.1880
## 3 2000-01-03 20.08251 603.9035 85.26027 19.698889 0.9996796 279.8353  231.8117
## 4 2000-01-04 27.85167 515.6228 90.89842 11.294444 0.9987945 278.8301  225.2433
## 5 2000-01-05 27.43265 707.4744 82.16803  4.200000 0.4684357 279.4175  252.5748
## 6 2000-01-06 14.17011 332.4545 84.58965 30.823333 0.3625654 278.6593  207.3831
##         ws
## 1 3.957476
## 2 3.967221
## 3 6.153436
## 4 6.153243
## 5 5.485918
## 6 4.959438
```

Getting started

```
summary(data_o3)
```

```
##       date                  o3               blh                rh       
##  Min.   :2000-01-01   Min.   : 2.155   Min.   :  46.52   Min.   :31.94  
##  1st Qu.:2002-10-01   1st Qu.:22.902   1st Qu.: 449.58   1st Qu.:70.31  
##  Median :2005-07-01   Median :32.098   Median : 661.31   Median :79.57  
##  Mean   :2005-07-01   Mean   :33.066   Mean   : 670.45   Mean   :77.88  
##  3rd Qu.:2008-03-31   3rd Qu.:41.693   3rd Qu.: 876.16   3rd Qu.:87.13  
##  Max.   :2010-12-31   Max.   :93.823   Max.   :1635.96   Max.   :99.24  
##       ssrd              tcc              tmax         Direction       
##  Min.   :  0.2878   Min.   :0.0000   Min.   :260.1   Min.   :  0.2008  
##  1st Qu.: 32.6486   1st Qu.:0.4155   1st Qu.:279.1   1st Qu.:141.3467  
##  Median : 98.8338   Median :0.6554   Median :285.6   Median :228.6122  
##  Mean   :112.2667   Mean   :0.6089   Mean   :285.5   Mean   :207.6878  
##  3rd Qu.:179.6789   3rd Qu.:0.8466   3rd Qu.:291.7   3rd Qu.:270.1296  
##  Max.   :313.5396   Max.   :1.0000   Max.   :307.7   Max.   :359.9870  
##        ws         
##  Min.   : 0.5442  
##  1st Qu.: 2.6418  
##  Median : 3.6527  
##  Mean   : 4.0167  
##  3rd Qu.: 5.0847  
##  Max.   :11.1696  
```

Visualisation



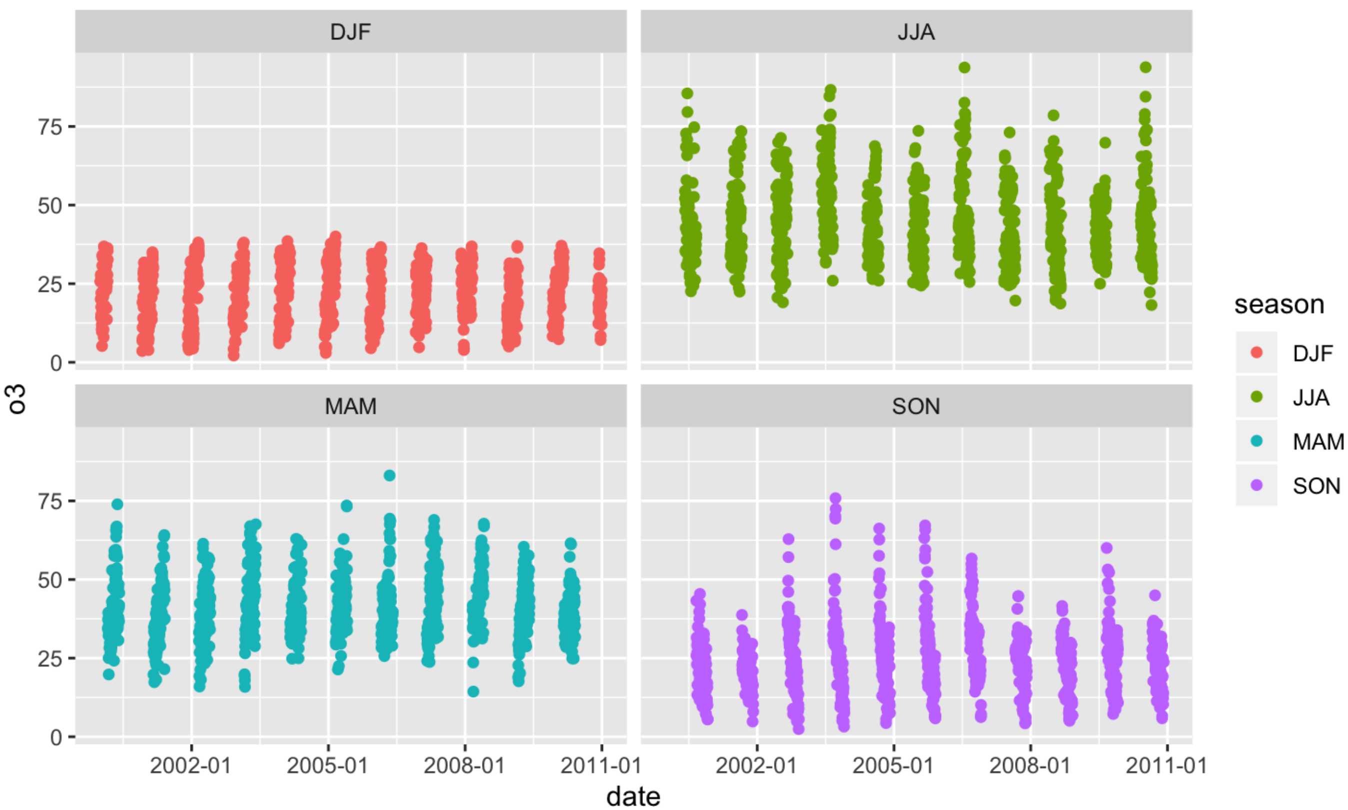
Visualisation-Seasonal cycle

- Ozone season usually ranges from April to September.
- Here, we will split the data into seasons to visualise the effect of the seasonal cycle.
- Then, we will restrict the analysis to spring and summer.

```
library(dplyr)
data_o3 <- data_o3%>%
    mutate(season=ifelse( format(date, "%m")>="01" & format(date, "%m") <="02" |format(date, "%m")=="12", "DJF",
                       ifelse( format(date, "%m")>="03" &  format(date, "%m")<="05", "MAM",
                              ifelse( format(date, "%m")>="06" &  format(date, "%m")<="08", "JJA",
                                     ifelse(format(date, "%m")>="09" &  format(date, "%m")<="11","SON",NA)))))
```

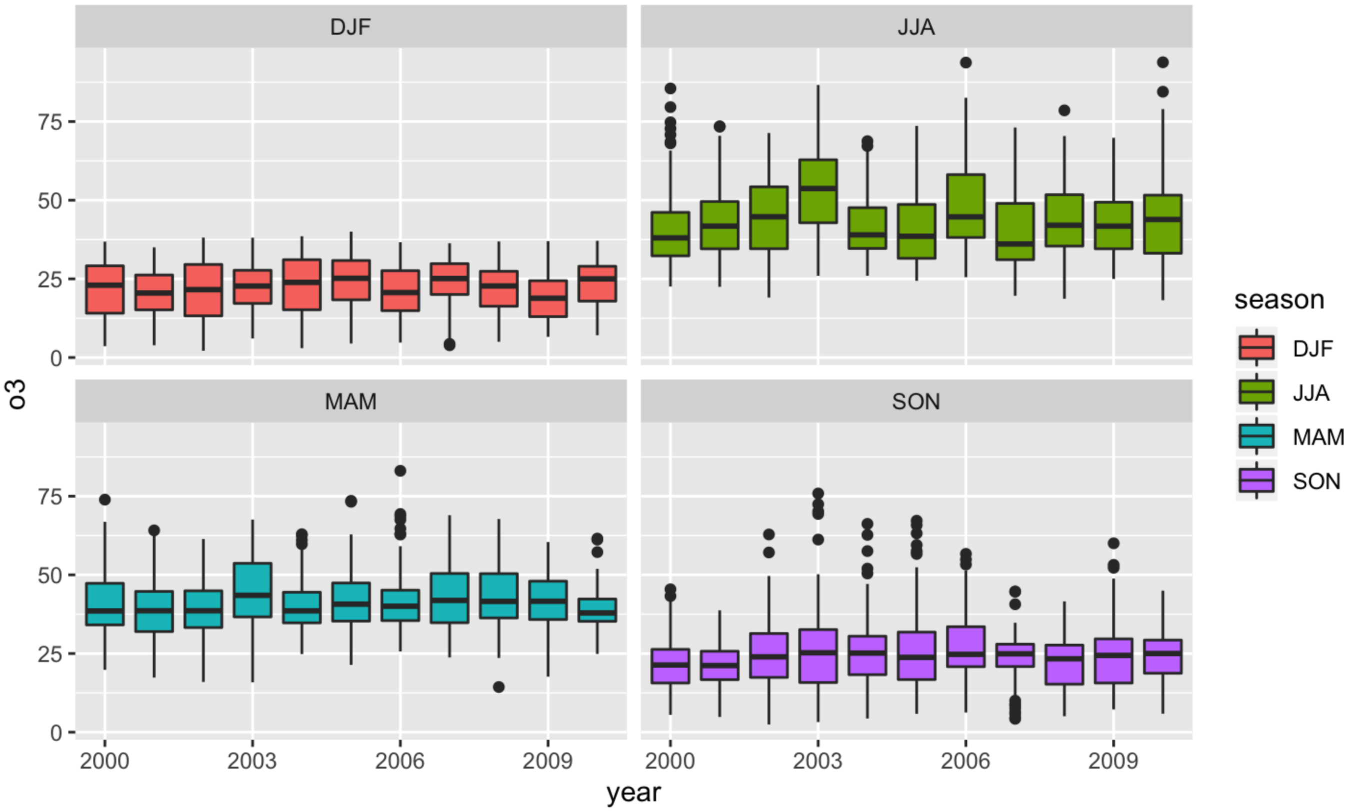
Visualisation-Seasonal cycle

```
ggplot2::ggplot(data_o3, aes(x=date, y=o3, color=season)) +
      geom_point() +
      scale_x_date(date_breaks = "3 years", date_labels = "%Y-%m") + facet_wrap(~season, ncol=2)
```

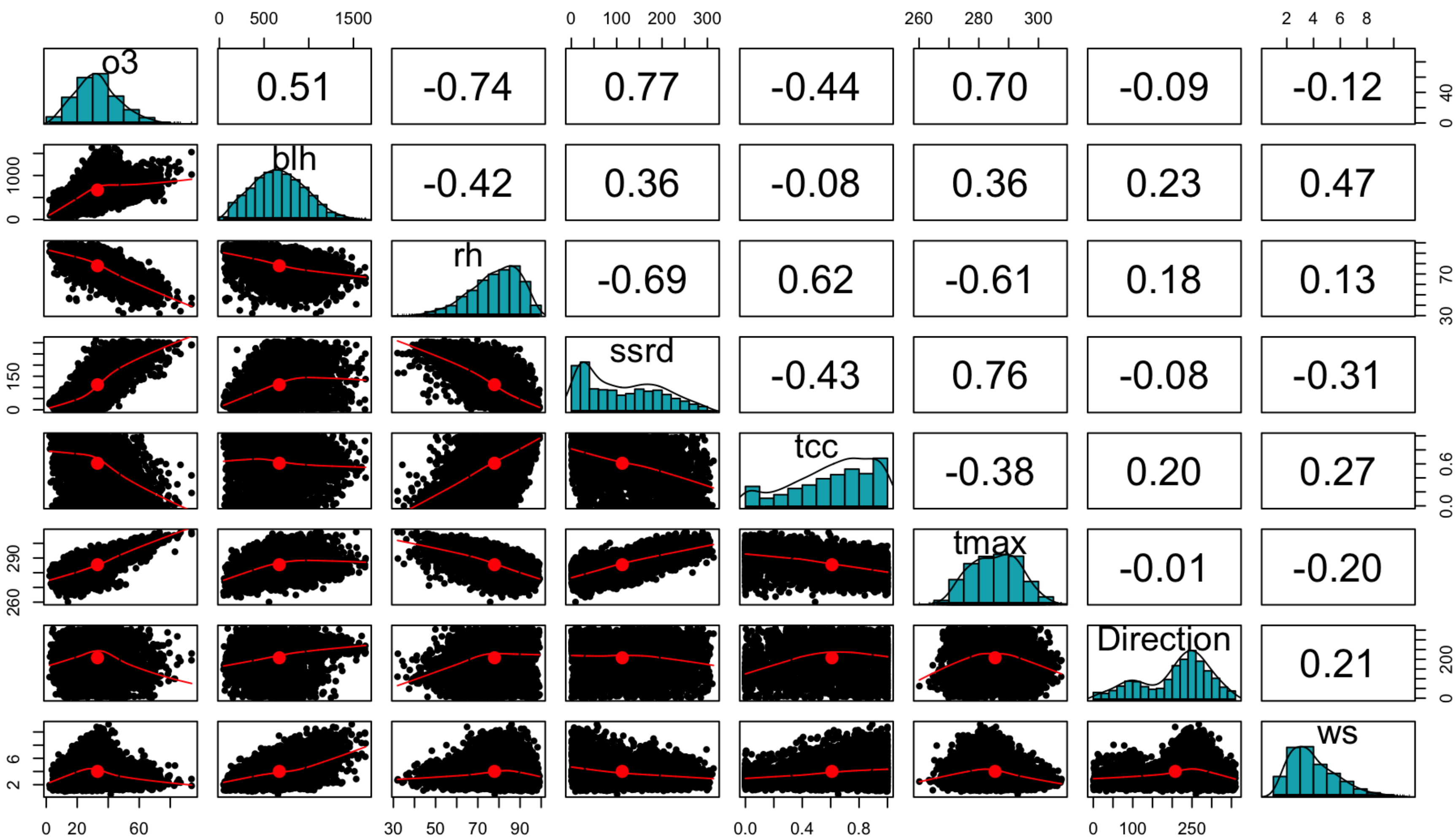


Visualisation-Seasonal cycle

```
ggplot2::ggplot(data_o3, aes(x=format(date,"%Y"), y=o3, fill=season)) +
  scale_x_discrete(breaks=seq(2000,2010,3))+ xlab("year")+
      geom_boxplot() + facet_wrap(~season, ncol=2)
```



Visualisation-Correlations



Regression analysis

The simple linear model can be written as:

$$\hat{y} = a + \beta x$$

where a is the intercept and β is the slope.

Let's start modelling the relationship between ozone and the meteorological variables. Since the ozone season usually ranges between April and September, we will focus on spring and summer.

```
m1  <- lm(o3~tmax,data=data_o3,na.action=na.omit)
```

```
##
## Call:
## lm(formula = o3 ~ tmax, data = data_o3, na.action = na.omit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.687  -7.498  -0.146   7.436  39.064
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -320.98253    5.64745  -56.84   <2e-16 ***
## tmax           1.24030    0.01978   62.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.3 on 4016 degrees of freedom
## Multiple R-squared:  0.4948, Adjusted R-squared:  0.4947
## F-statistic:  3934 on 1 and 4016 DF,  p-value: < 2.2e-16
```

Multiple regression analysis

Now, we can fit a new model by adding more variables. We are interesting in building a model that better explain the O3 variabily. Then, we need to examine which variables give us the best model. Ultimately, we want to see which variable is the main “driver” (i.e. explaining the larger proportion of O3 variability)

```
m2  <- lm(o3~tmax+rh,data=data_jja,na.action=na.omit)
```

summary(m2) # see the summary of the model

```
m3  <- lm(o3~tmax+rh+ssrd,data=data_jja,na.action=na.omit)
```

summary(m3) # see the summary of the model

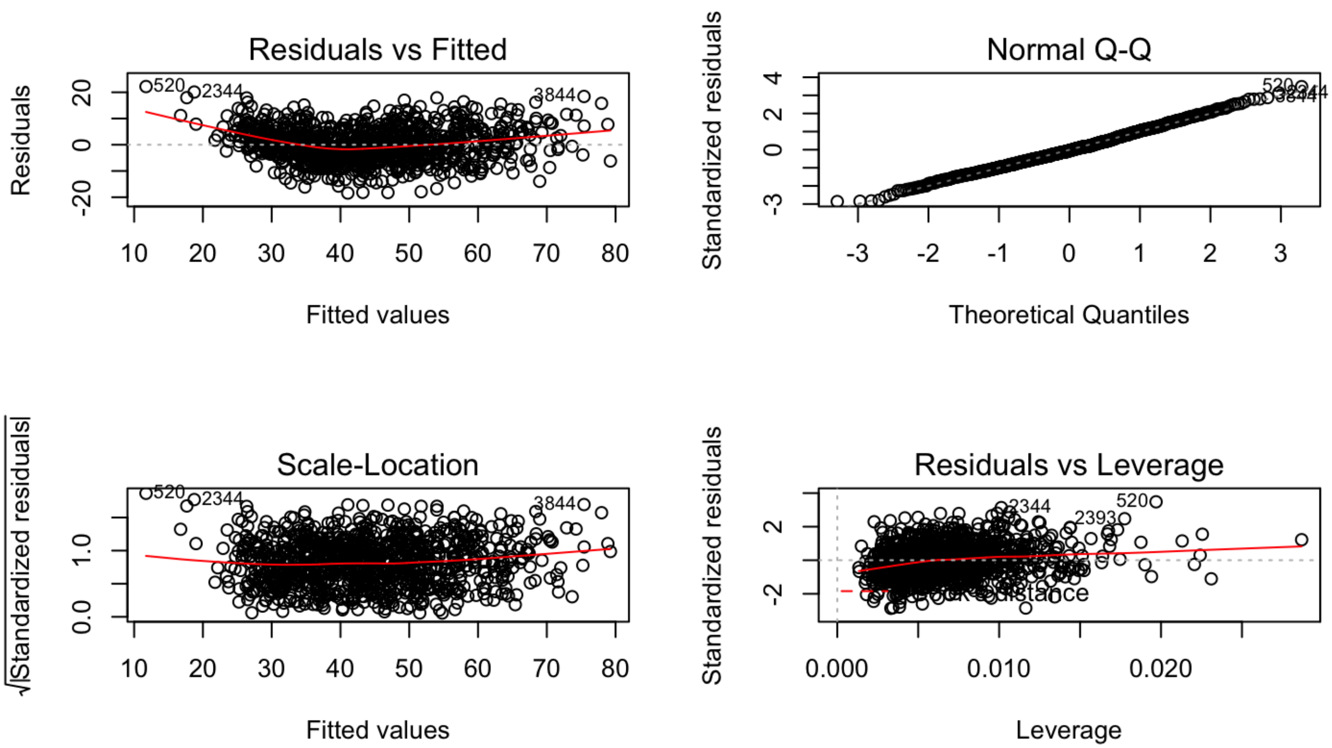
Multiple regression analysis

Fit a full model:

```
mfull  <- lm(o3~tmax+rh+ssrd+tcc+ws+Direction,data=data_jja,na.action=na.omit)
summary(mfull)
```

```
##
## Call:
## lm(formula = o3 ~ tmax + rh + ssrd + tcc + ws + Direction, data = data_jja,
##     na.action = na.omit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4031  -4.3544  -0.3266   4.2860  22.1882
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.222e+02  2.158e+01 -19.568  < 2e-16 ***
## tmax         1.646e+00  6.866e-02  23.978  < 2e-16 ***
## rh          -2.917e-01  3.086e-02  -9.452  < 2e-16 ***
## ssrd         2.609e-02  4.537e-03   5.750 1.18e-08 ***
## tcc         -2.978e-01  1.081e+00  -0.275  0.78306
## ws          -1.096e+00  1.588e-01  -6.904 8.93e-12 ***
## Direction    6.744e-03  2.568e-03   2.627  0.00875 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.438 on 1005 degrees of freedom
## Multiple R-squared:  0.7547, Adjusted R-squared:  0.7533
## F-statistic: 515.4 on 6 and 1005 DF,  p-value: < 2.2e-16
```

Multiple regression analysis-model check



Model selection - Stepwise regression

The stepwise regression (or stepwise selection) consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error.

In R, stepAIC() (MASS package), choose the best model by AIC(Akaike Information Criterion (AIC). It has an option named direction, which can take the following values: i) “both” (for stepwise regression, both forward and backward selection); “backward” (for backward selection) and “forward” (for forward selection). It return the best final model.

$$AIC = 2K - 2logLik$$

where loglik is the log-likelihood (how well the model fits the data) and K is the number of the parameters.

Stepwise regression

```
m.null <- lm(o3~1, data=data_jja)
m.f  <- stepAIC(m.null, direction="forward", scope=list(lower=m.null, upper=mfull))
```

```
## Start:  AIC=5186.49
## o3 ~ 1
## 
##              Df Sum of Sq    RSS    AIC
## + tmax        1    117085  52770 4005.5
## + rh          1     86136  83718 4472.5
## + tcc         1     52427 117427 4814.9
## + ssrd        1     50844 119011 4828.5
## + ws          1     30673 139181 4986.9
## + Direction   1     18184 151671 5073.9
## <none>                    169855 5186.5
## 
## Step:  AIC=4005.46
## o3 ~ tmax
## 
##              Df Sum of Sq   RSS    AIC
## + rh          1    7334.0 45436 3856.0
## + ssrd        1    5160.0 47610 3903.3
## + tcc         1    2141.9 50628 3965.5
## + ws          1    1158.3 51612 3985.0
## <none>                    52770 4005.5
## + Direction   1      46.9 52723 4006.6
## 
## Step:  AIC=3856.03
## o3 ~ tmax + rh
## 
##              Df Sum of Sq   RSS    AIC
## + ws          1   2136.41 43300 3809.3
## + ssrd        1   1649.97 43786 3820.6
## + Direction   1     96.61 45339 3855.9
```

Setpwise regression

```
m.b  <- stepAIC(mfull, direction="backward")
```

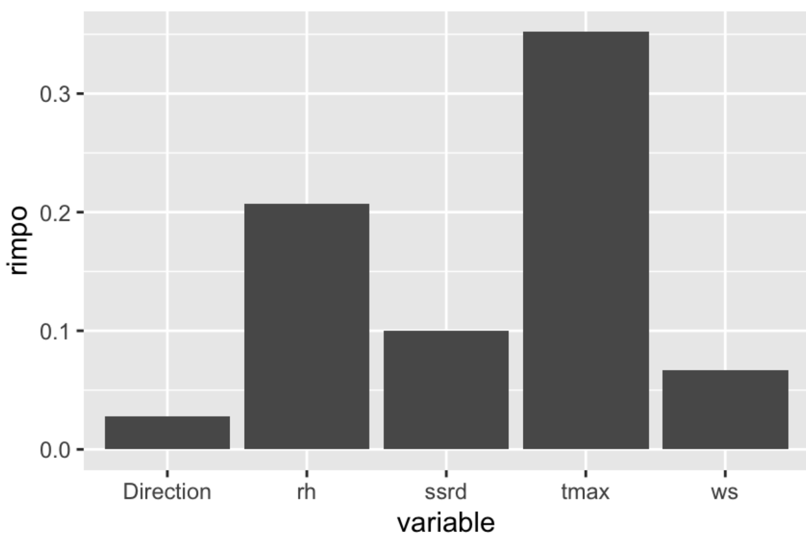
```
## Start:  AIC=3776.22
## o3 ~ tmax + rh + ssrd + tcc + ws + Direction
##
##             Df Sum of Sq   RSS    AIC
## - tcc        1       3.1 41663 3774.3
## <none>                   41660 3776.2
## - Direction  1     286.0 41946 3781.1
## - ssrd       1    1370.5 43030 3807.0
## - ws         1    1976.0 43636 3821.1
## - rh         1    3703.4 45363 3860.4
## - tmax       1   23832.0 65492 4232.0
##
## Step:  AIC=3774.29
## o3 ~ tmax + rh + ssrd + ws + Direction
##
##             Df Sum of Sq   RSS    AIC
## <none>                   41663 3774.3
## - Direction  1     285.4 41948 3779.2
## - ssrd       1    1375.2 43038 3805.2
## - ws         1    2000.1 43663 3819.7
## - rh         1    4697.0 46360 3880.4
## - tmax       1   24393.3 66056 4238.7
```

Variable importance

We want to identify which predictor has a large contribution to the total explained deviance. We can now calculate the relative importance of each predictor.

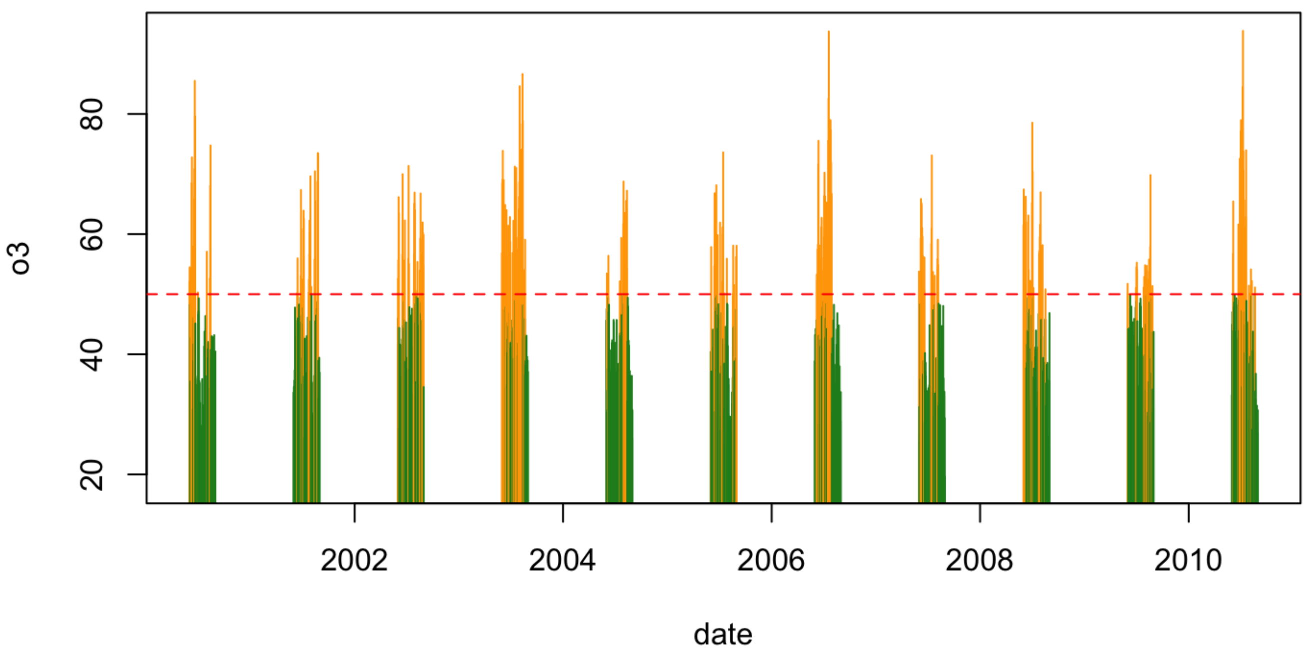
```
library(relaimpo)
relImportance <- calc.relimp(m.f, type="lmg")
```

```
# plots
ggplot2::ggplot(df.rimpo, aes(x=variable, y=rimpo))+ geom_bar(stat = "identity")
```



Ozone exceedances

```
plot(o3~date, data=data_jja, type="h", col=o3_50)
abline(h=ths, lty=2, col="red")
```



Ozone exceedances-Logistic regression

We use logistic regression (LR) to model the probability of ozone exceedances over a threshold. Occurrences of threshold exceedance can take values of 0 (not exceeded) or 1 (exceeded), so the associated distribution for probabilities of these exceedances is the binomial distribution.

In R, it can be done with GLM and it is similar than the MLR case, but with another distribution.

```
fitglm_tx <- glm(o3~tmax,data=data_jja,family="binomial")
exp(coef(fitglm_tx))
```

```
##  (Intercept)         tmax
## 1.393375e-20 1.197077e+00
```

```
summary(fitglm_tx)
```

```
##
## Call:
## glm(formula = o3 ~ tmax, family = "binomial", data = data_jja)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6489   0.0290   0.0400   0.0513   0.1112
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -45.7200    77.0747  -0.593    0.553
## tmax          0.1799     0.2647   0.680    0.497
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.838  on 1011  degrees of freedom
```

Ozone exceedances-Multiple Logistic regression

We can add more predictors:

```
fitglm <- glm(o3~tmax+rh+ssrd+blh+Direction+ws,data=data_jja,family="binomial")
```

summary(fitglm)

You can now also apply model selection:

-modelglm <- stepAIC(fitglm,direction=“both”)

and get the predictions:

-pred <- predict(modelglm,type=“response”)

key-questions

We have used regression analysis to examine ozone variability and the influence of meteorological variables.

- Which model fits better?
- What are the main meteorological drivers of ozone?
- What are the main differences between spring and summer?
- When the number of exceedances is greater?

References

- Statistical Methods in the Atmospheric Sciences, Daniel Wilks
- Otero, N., Sillmann, J., Schnell, J. L., Rust, H. W., and Butler, T.: Synoptic and meteorological drivers of extreme ozone concentrations over Europe, Environmental Research Letters, 11, 24 005, doi:10.1088/1748-9326/11/2/024005 (ref. therein)
- Camalier L, Cox W and Dolwick P 2007 The effects of meteorology on ozone in urban areas and their use in assessing ozone trends Atmos. Environ. 41 7127–37 (ref. therein)