Beispiele zu LM, GLM, GAM und Bäumen

Friedrich Leisch

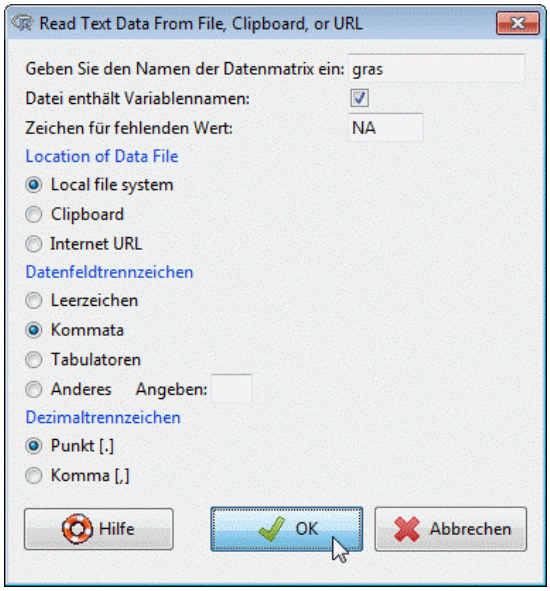
27. November 2013

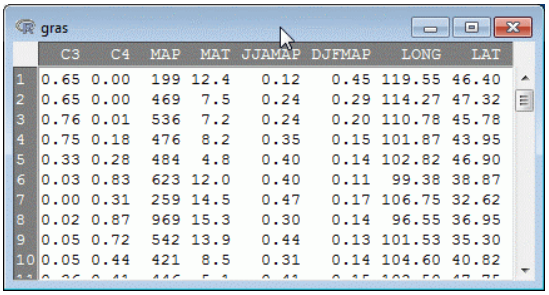
Generalisierte additive Modelle und Bäume können nicht über die Menüs des R Commander gefittet werden, sondern müssen „händisch“ per Kommandozeile angepaßt werden. Die Kommandos unterscheiden sich aber nur minimal von linearen Modellen und GLMs (für die es Menüs gibt).

1 C3-Gräser in Nordamerika

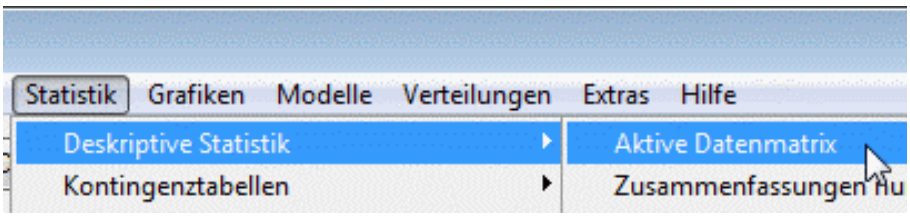
(siehe auch Datei `paruelo.R` in Moodle für mehr Grafiken).

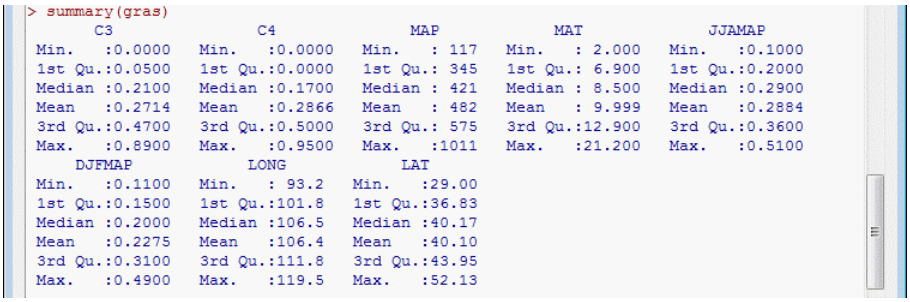
Einlesen der Daten und Summary Statistiken:





	C3	C4	MAP	MAT	JJAMAP	DJFMAP	LONG	LAT
1	0.65	0.00	199	12.4	0.12	0.45	119.55	46.40
2	0.65	0.00	469	7.5	0.24	0.29	114.27	47.32
3	0.76	0.01	536	7.2	0.24	0.20	110.78	45.78
4	0.75	0.18	476	8.2	0.35	0.15	101.87	43.95
5	0.33	0.28	484	4.8	0.40	0.14	102.82	46.90
6	0.03	0.83	623	12.0	0.40	0.11	99.38	38.87
7	0.00	0.31	259	14.5	0.47	0.17	106.75	32.62
8	0.02	0.87	969	15.3	0.30	0.14	96.55	36.95
9	0.05	0.72	542	13.9	0.44	0.13	101.53	35.30
10	0.05	0.44	421	8.5	0.31	0.14	104.60	40.82





```
> summary(gras)
       C3                C4                MAP             MAT            JJAMAP      
 Min.   :0.0000   Min.   :0.0000   Min.   : 117   Min.   : 2.000   Min.   :0.1000  
 1st Qu.:0.0500   1st Qu.:0.0000   1st Qu.: 345   1st Qu.: 6.900   1st Qu.:0.2000  
 Median :0.2100   Median :0.1700   Median : 421   Median : 8.500   Median :0.2900  
 Mean   :0.2714   Mean   :0.2866   Mean   : 482   Mean   : 9.999   Mean   :0.2884  
 3rd Qu.:0.4700   3rd Qu.:0.5000   3rd Qu.: 575   3rd Qu.:12.900   3rd Qu.:0.3600  
 Max.   :0.8900   Max.   :0.9500   Max.   :1011   Max.   :21.200   Max.   :0.5100  
     DJFMAP            LONG            LAT       
 Min.   :0.1100   Min.   : 93.2   Min.   :29.00  
 1st Qu.:0.1500   1st Qu.:101.8   1st Qu.:36.83  
 Median :0.2000   Median :106.5   Median :40.17  
 Mean   :0.2275   Mean   :106.4   Mean   :40.10  
 3rd Qu.:0.3100   3rd Qu.:111.8   3rd Qu.:43.95  
 Max.   :0.4900   Max.   :119.5   Max.   :52.13  
```

Das Einlesen des Datensatzes und die Zusammenfassung der Statistiken kann auch direkt im Skriptfenster des R Commanders erfolgen. (Arbeitsverzeichnis von R enthält hier Kopie der Daten, sonst vollen Pfad zur Datei einfügen). Die entsprechenden Befehle dazu sind:

```
> gras <- read.csv("paruelo.csv")
> summary(gras)
```

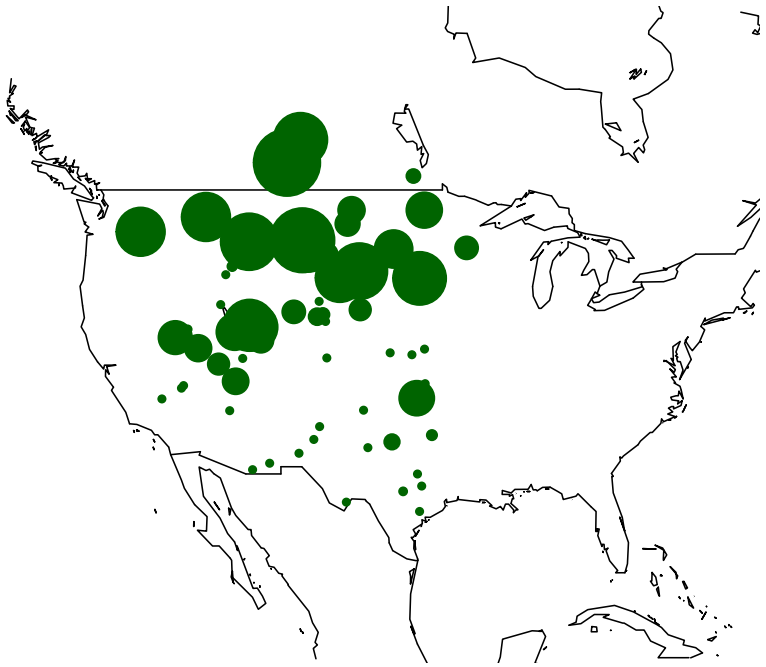
```
       C3                C4                MAP             MAT            JJAMAP      
 Min.   :0.0000   Min.   :0.0000   Min.   : 117   Min.   : 2.000   Min.   :0.1000  
 1st Qu.:0.0500   1st Qu.:0.0000   1st Qu.: 345   1st Qu.: 6.900   1st Qu.:0.2000  
 Median :0.2100   Median :0.1700   Median : 421   Median : 8.500   Median :0.2900  
 Mean   :0.2714   Mean   :0.2866   Mean   : 482   Mean   : 9.999   Mean   :0.2884  
 3rd Qu.:0.4700   3rd Qu.:0.5000   3rd Qu.: 575   3rd Qu.:12.900   3rd Qu.:0.3600  
 Max.   :0.8900   Max.   :0.9500   Max.   :1011   Max.   :21.200   Max.   :0.5100  
     DJFMAP            LONG            LAT       
 Min.   :0.1100   Min.   : 93.2   Min.   :29.00  
 1st Qu.:0.1500   1st Qu.:101.8   1st Qu.:36.83  
 Median :0.2000   Median :106.5   Median :40.17  
 Mean   :0.2275   Mean   :106.4   Mean   :40.10  
 3rd Qu.:0.3100   3rd Qu.:111.8   3rd Qu.:43.95  
 Max.   :0.4900   Max.   :119.5   Max.   :52.13  
```

Bedeutung der Variablen:

```
* C3 - relative abundance of C3 grasses
* C4 - relative abundance of C4 grasses
* MAP - mean annual precipitation (mm)
* MAT - mean annual temperature (oC)
* JJAMAP - proportion of MAP that fell in June, July and August
* DJFMAP - proportion of MAP that fell in December, January and February
* LONG - longitude in centesimal degrees
* LAT - latitude in centesimal degrees
```

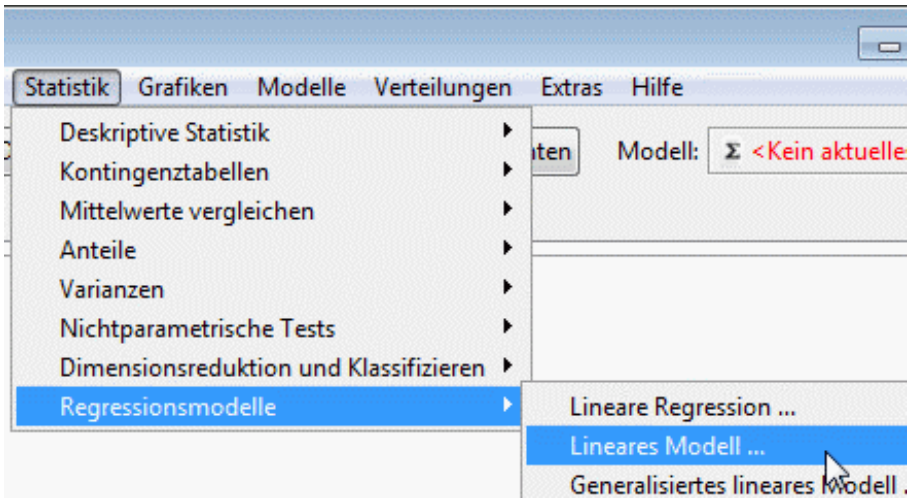
Eine Karte von Nordamerika mit den Beobachtungsstellen, Größe der Punkte entspricht Vorkommen von C3-Gräsern:

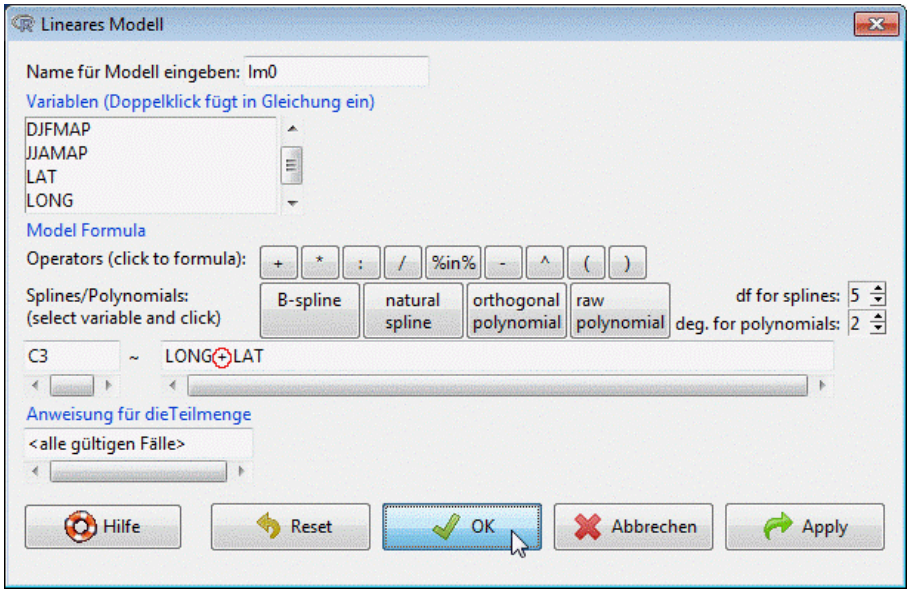
```
> library("maps")
> map("world", xlim = c(-130, -70), ylim = c(20,60))
> points(-gras$LONG, gras$LAT, cex=pmax(1, 10*gras$C3), pch=20, col="darkgreen")
```



1.1 Lineare Modelle

Lineares Modell nur mit Haupteffekten für Längen- und Breitengrad:

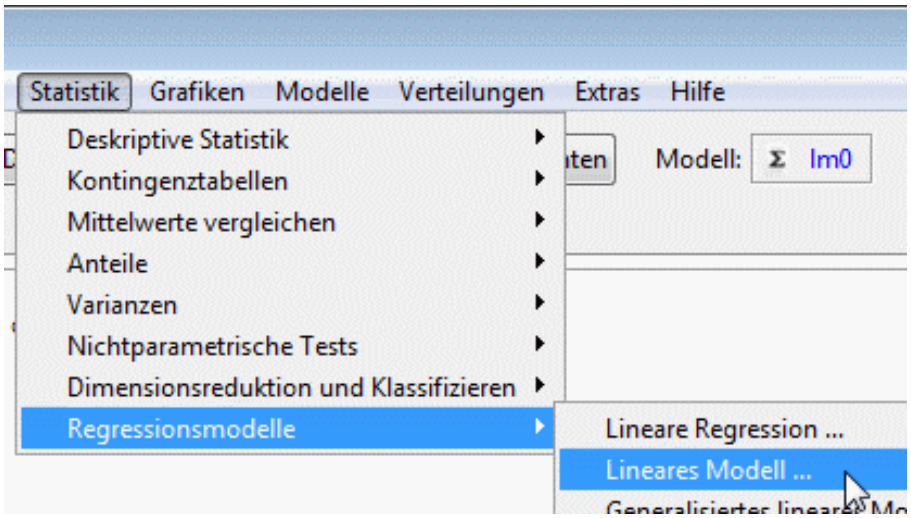


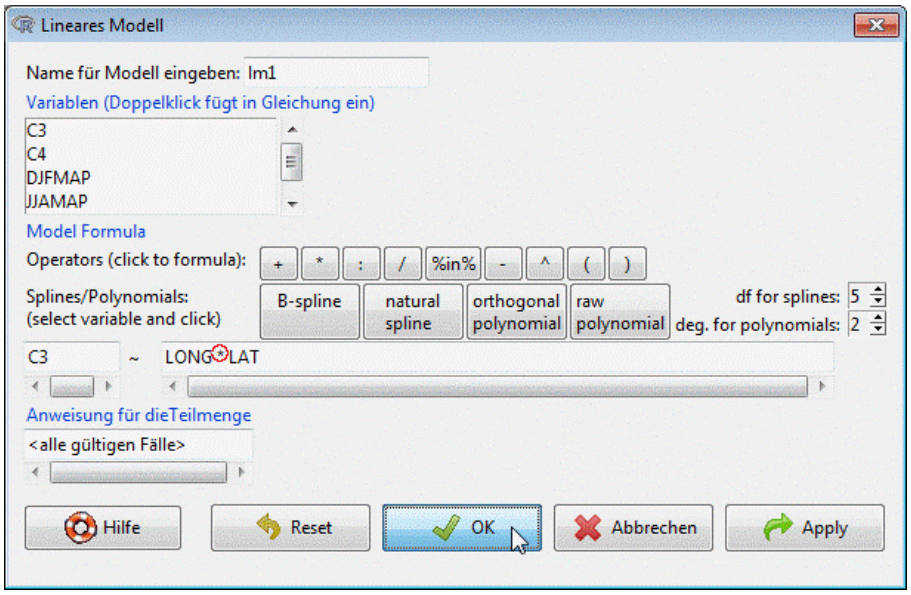


```
> lm0 <- lm(C3~LONG+LAT, data=gras)
> summary(lm0)

Call:
lm(formula = C3 ~ LONG + LAT, data = gras)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41150 -0.15666 -0.00401  0.14823  0.40703

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.9504806  0.4094130  -2.322   0.0232 *
LONG        -0.0009366  0.0036287  -0.258   0.7971
LAT          0.0329518  0.0044035   7.483 1.63e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1972 on 70 degrees of freedom
Multiple R-squared:  0.4454,        Adjusted R-squared:  0.4295
F-statistic: 28.11 on 2 and 70 DF,  p-value: 1.096e-09
```

Interpretation: Längengrad hat keinen signifikanten Einfluß auf das Vorkommen der C3-Gräser, Breitengrad ist hoch signifikant (positiver Koeffizient, daher je nördlicher desto mehr).

Lineares Modell mit Haupteffekten für Längen- und Breitengrad sowie Interaktion:






```
> lm1 <- lm(C3~LONG*LAT, data=gras)
> summary(lm1)

Call:
lm(formula = C3 ~ LONG * LAT, data = gras)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39563 -0.14722 -0.01491  0.11837  0.40268

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.7518079  2.9399294   2.297   0.0247 *
LONG        -0.0752581  0.0283285  -2.657   0.0098 **
LAT         -0.1618176  0.0737967  -2.193   0.0317 *
LONG:LAT     0.0018773  0.0007101   2.644   0.0101 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1893 on 69 degrees of freedom
Multiple R-squared:  0.4964,        Adjusted R-squared:  0.4745
F-statistic: 22.67 on 3 and 69 DF,  p-value: 2.525e-10
```

Interpretation: alle Terme signifikant, im Nordwesten ist C3 am häufigsten (Länge und Breite simultan groß, Koeffizient der Interaktion ist positiv). Die negativen Koffizienten der Haupteffekte sind primär relativ zum Nordwesten zu sehen, der durch die Multiplikation von Längen- und Breitengrad in der Interaktion überproportional starkes Gewicht bekommt. Im nichtlinearen GAM sieht man aber, daß multiplikative Interaktion hier nicht optimal ist.

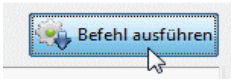
1.2 Generalisierte additive Modelle

Zunächst muss das Paket `mgcv` geladen werden. Das generalisierte additive Modell kann mit Hilfe des R Commanders erzeugt werden, indem erst ein lineares Modell wie in 1.1 erzeugt wird. `lm` muss

dann durch `gam` ersetzt werden und alle Variablen, für die nichtlineare Terme geschätzt werden sollen, müssen in Klammern gepackt und mit einem `s` versehen werden.

Das Skriptfenster sollte dann wie in der Grafik unten aussehen. Entweder wird nach jeder Zeile auf `Befehl ausführen` gedrückt oder die Zeilen werden als Region markiert und einmal `Befehl ausführen` drücken genügt.

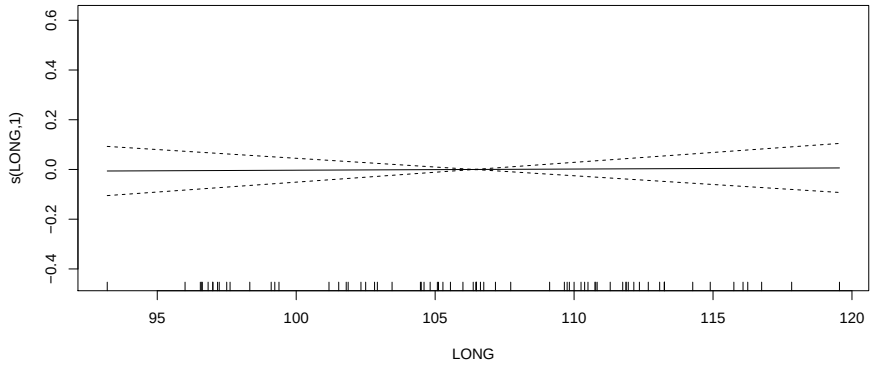
```
library("mgcv")
gam1 <- gam(C3 ~ s(LONG) + s(LAT), data=gras)
summary(gam1)
```

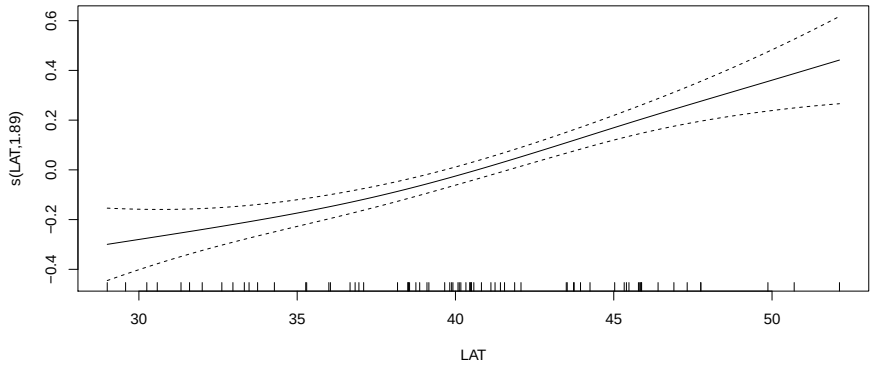



```
Family: gaussian
Link function: identity

Formula:
C3 ~ s(LONG) + s(LAT)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2714     0.0228    11.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df      F  p-value
s(LONG) 1.00   1.00  0.016    0.901
s(LAT)  1.89   2.39 24.313 8.27e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.443   Deviance explained = 46.6%
GCV score = 0.040087  Scale est. = 0.037951  n = 73
```

Wir bekommen nun 2 Tabellen: eine für den linearen Teil des Modells (hier nur Intercept), der genau gleich wie in der Ausgabe von `lm()` zu interpretieren ist. Die zweite Tabelle beschreibt die nichtlinearen Terme, hier `s(LONG)` und `s(LAT)`. Dabei gibt es KEINE REGRESSIONSKOEFFIZIENTEN, für jede Variable wird eine komplette Kurve geschätzt, diese zeichnet man am besten zur Interpretation. Die Tabelle für die nichtlinearen Terme hat trotzdem vier Spalten: die ersten beiden (`edf` und `Ref.df`) messen die Nichtlinearität der geschätzten Kurve über die Freiheitsgrade (estimated degrees of freedom). Ein `edf=1` entspricht einer Geraden, die Variable könnte also in den linearen Teil des Modells übernommen werden (hier bei Längengrad der Fall). Die zweiten beiden Spalten machen einen F-Test, ob der Term überhaupt einen signifikanten Einfluß auf die Zielvariable hat. In unserem Beispiel ist nur der Breitengrad signifikant, und zwar leicht nichtlinear (`edf=1.89`). Stärkere Nichtlinearität heißt nicht unbedingt höhere Signifikanz und umgekehrt.

Plots der nichtlinearen Effekte bekommt man mit `plot(gam1)`, in interaktiven Sessions muß für jede Grafik einmal ENTER gedrückt werden:



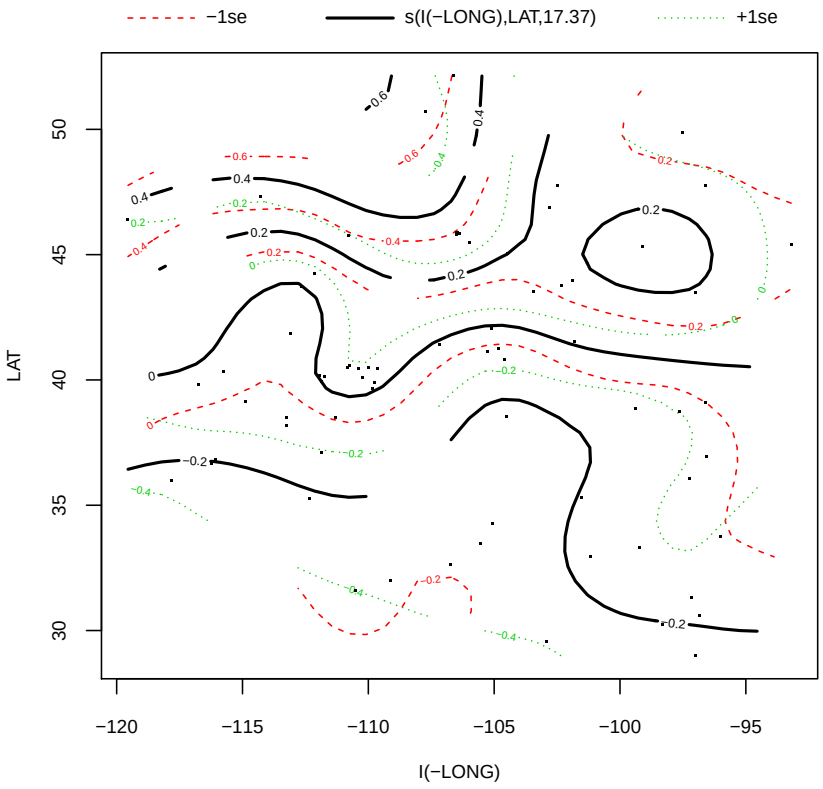


Will man eine Interaktion zwischen zwei Variablen nichtlinear modellieren, packt man beide in denselben glatten Term. Das `I(-LONG)` nimmt den Längengrad mit negativem Vorzeichen, damit dann in den Plots die Westküste der USA (Länge ca -120) auch links ist und die Orientierung stimmt:

```
> gam2 <- gam(C3~s(I(-LONG), LAT), data=gras)
> summary(gam2)

Family: gaussian
Link function: identity

Formula:
C3 ~ s(I(-LONG), LAT)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.27137    0.01977   13.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                  edf Ref.df    F  p-value
s(I(-LONG),LAT) 17.37  22.23 4.74 1.39e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.582   Deviance explained = 68.3%
GCV score = 0.038131  Scale est. = 0.028534  n = 73
```

Wir sehen einen stark signifikanten nichtlinearen Effekt, Interpretation nur über Grafiken:

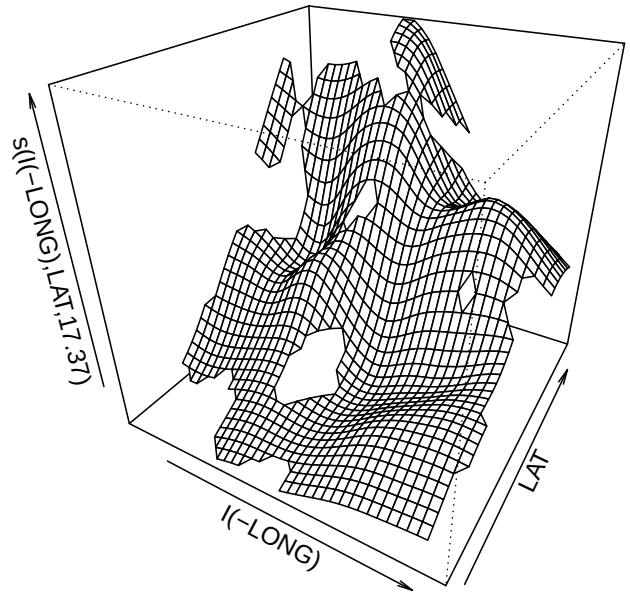
```
> plot(gam2)
```



Im Nordwesten größter positiver Beitrag (plus 0.4 und mehr), im Süden negativer Beitrag (Höhenschichtlinie von -0.2).

In 3d:

```
> plot(gam2, pers=TRUE)
```



1.3 Regressionsbaum

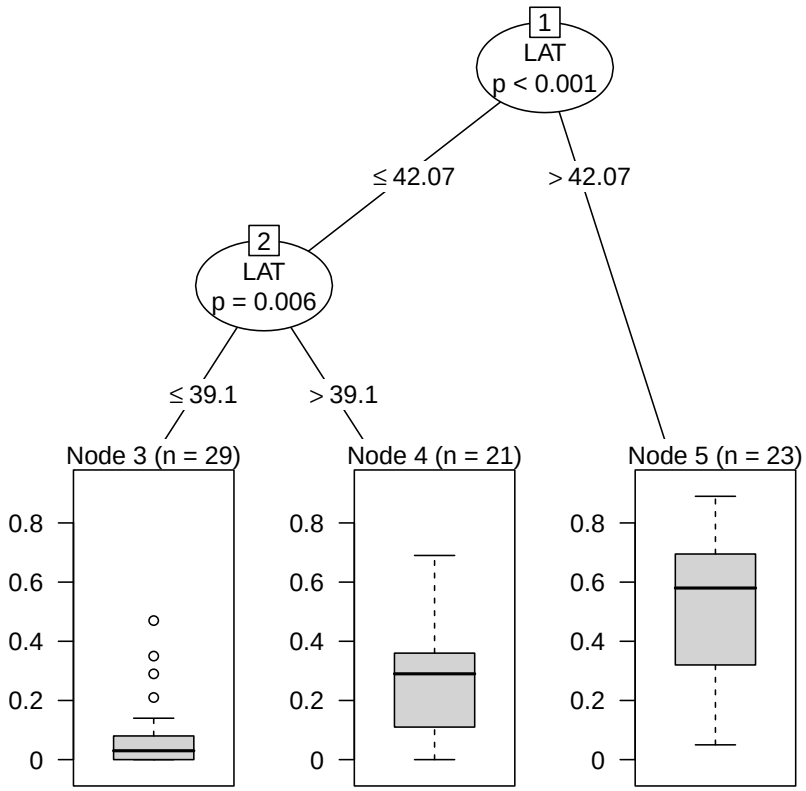
Paket `party` von CRAN installieren, dann:

```
> library("party", quietly=TRUE)
> baum1 <- ctree(C3~LONG+LAT, data=gras)
> baum1

         Conditional inference tree with 3 terminal nodes

Response:  C3
Inputs:  LONG, LAT
Number of observations:  73

1) LAT <= 42.07; criterion = 1, statistic = 32.03
  2) LAT <= 39.1; criterion = 0.994, statistic = 8.786
    3)*  weights = 29
  2) LAT > 39.1
    4)*  weights = 21
1) LAT > 42.07
  5)*  weights = 23
```

Bäume lassen nach Konstruktion alle möglichen Interaktionen zwischen allen Variablen zu: kommen zwei Variablen im selben Zweig des Baumes vor (= Pfad von Wurzel zu Blatt) so wird eine

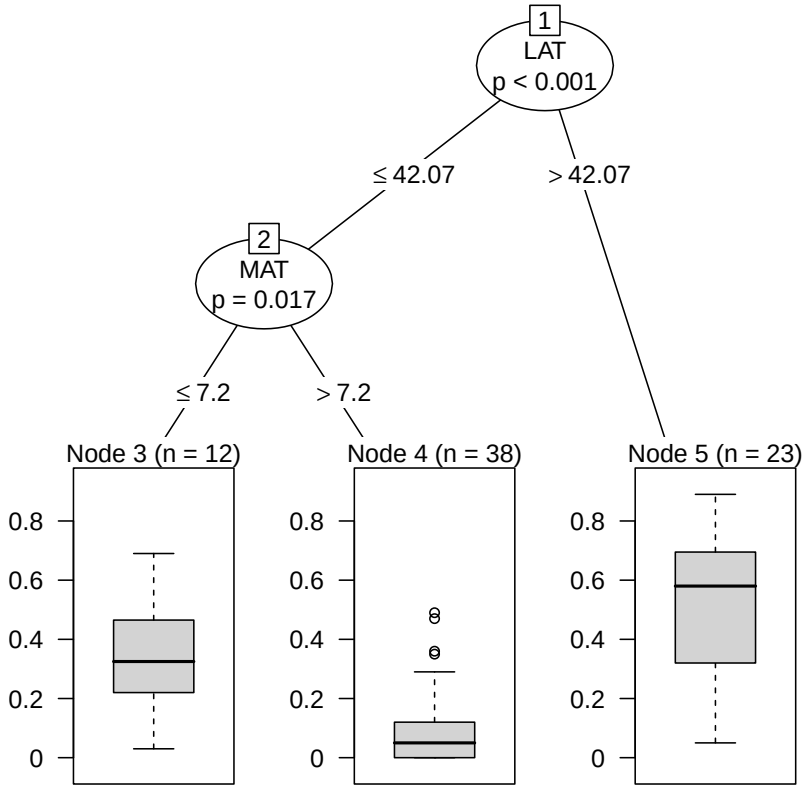
Interaktion geschätzt (Blätter haben simultane Bedingungen an alle Variablen im Zweig). Die Modellformel für Bäume enthält daher nur durch + getrennte Variablennamen. Dieser einfache Baum trennt nur in 3 Streifen: südlich vom 39. Breitengrad, zwischen 39 und 42, und nördlich des 42. Breitengrades. In den Endknoten (Blättern) des Baumes finden sich jeweils Boxplots der Verteilung von C3 in dieser Teilmenge.

```
> plot(baum1)
```



Läßt man alle Variablen zu, wird südlich des 42. Breitengrades nach mittlerer Temperatur statt Breite getrennt, das ist wegen Küste bzw Inland mit Grenze zu Mexiko der bessere Prädiktor:

```
> baum2 <- ctree(C3~MAP+MAT+JJAMAP+DJFMAP+LONG+LAT, data=gras)
> plot(baum2)
```



Die p-Werte in den Verzweigungsknoten des Baumes kommen von einem statistischen Test, ob sich die beiden durch die Teilung resultierenden Gruppen signifikant in der Zielvariable unterscheiden. Das `p < 0.001` in Knoten 1 ensteht also folgendemaßen: Der Datensatz wird in „südlich“ und „nördlich“ (des 42. Breitengrades) geteilt, die entsprechenden Gruppengrößen sind $n = 50 = 12 + 38$ und $n = 23$. Für diese beiden Gruppen von Beobachtungen wird nun getestet, ob sich die Mittelwerte von `C3` signifikant unterscheiden (was sie tun, sonst gäbe es den Knoten im Baum nicht). Im Knoten 2 werden die 50 südlichen Beobachtungen nach Temperatur geteilt, getestet wird wieder auf Unterschied Mittelwert `C3` in den beiden Gruppen mit 12 und 38 Beobachtungen.

2 Heuschrecke Tetrix Subulata (Säbel-Dornschrecke)

(siehe auch Datei `ghuepfer.R` in Moodle für mehr Grafiken und Analyse der Psophus Stridulus). Die Daten können wieder direkt oder über das Menü des R Commander eingelesen werden, danach empfiehlt sich wie immer eine Kontrolle der numerischen Zusammenfassung der Daten (siehe 1).

```
> ghuepfer <- read.csv("ghuepfer.csv")
> summary(ghuepfer)
```

```
       X                Y              Stadt             SAWald
 Min.   :4288273   Min.   :5244262   Min.   :0.00000   Min.   :0.0000
 1st Qu.:4403528   1st Qu.:5348734   1st Qu.:0.01144   1st Qu.:0.1825
 Median :4455359   Median :5420756   Median :0.02961   Median :0.3026
 Mean   :4459217   Mean   :5422686   Mean   :0.05027   Mean   :0.3439
 3rd Qu.:4514920   3rd Qu.:5499912   3rd Qu.:0.05916   3rd Qu.:0.4700
 Max.   :4632062   Max.   :5600352   Max.   :0.82090   Max.   :1.0000
     Acker             Wiesen         TETRIX.SUBULATA  PSOPHUS.STRIDULUS
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.04403   1st Qu.:0.04333   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.25831   Median :0.10794   Median :0.0000   Median :0.0000
 Mean   :0.29753   Mean   :0.17123   Mean   :0.3788   Mean   :0.1074
 3rd Qu.:0.49590   3rd Qu.:0.22532   3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :0.98425   Max.   :0.89501   Max.   :1.0000   Max.   :1.0000
    AnzArten
 Min.   : 0.00
 1st Qu.:11.00
 Median :16.00
 Mean   :15.71
 3rd Qu.:20.00
 Max.   :41.00
```

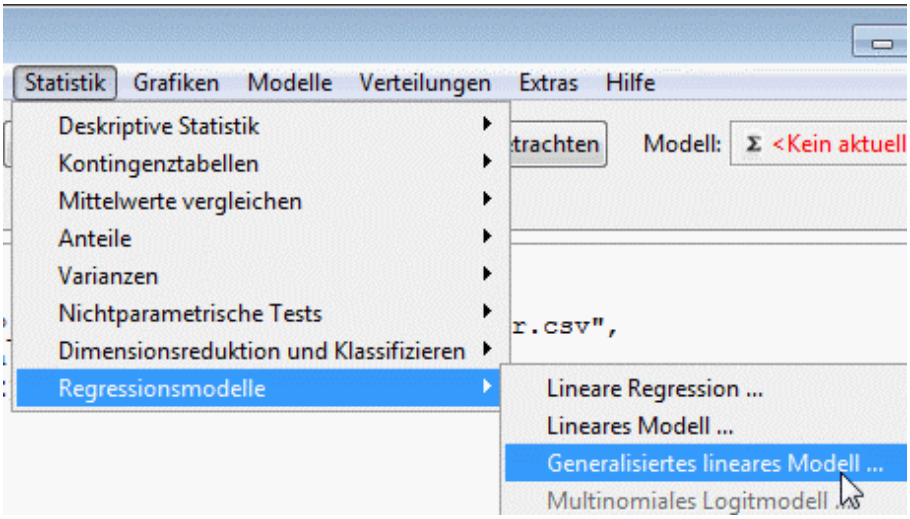
Bedeutung der Variablen:

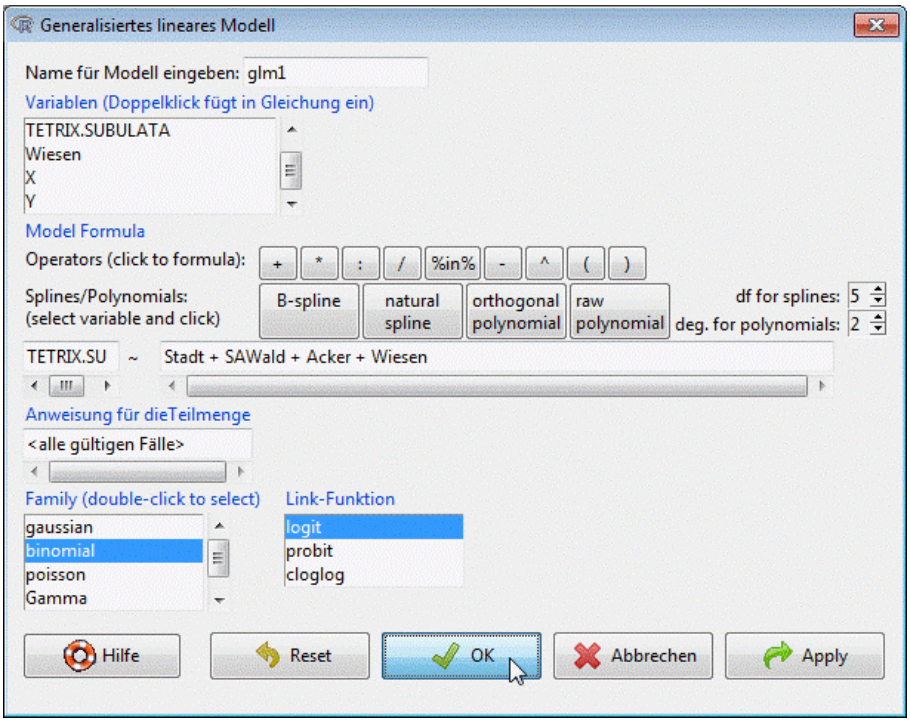
* X, Y: Position des Quadrates in Bayern
* Stadt: Prozent Bodennutzung Stadt
* SAWald: Prozent Bodennutzung Wald (Summe aller Waldarten)
* Acker: Prozent Bodennutzung Acker
* Wiesen: Prozent Bodennutzung Wiesen
* TETRIX.SUBULATA: Vorkommen ja/nein
* PSOPHUS.STRIDULUS: Vorkommen ja/nein

Im folgenden steht „Wald“ immer für `SAWald`.

2.1 Lineares Logitmodell (Binomial-GLM)

Generalisiertes lineares Modell mit Binomial-Verteilung für die Auftretenswahscheinlichkeit der Säbel-Dornschrecke:





```
> glm1 <- glm(TETRIX.SUBULATA~Stadt+SAWald+Acker+Wiesen, data=ghuepfer,
+             family=binomial)
> summary(glm1)

Call:
glm(formula = TETRIX.SUBULATA ~ Stadt + SAWald + Acker + Wiesen,
    family = binomial, data = ghuepfer)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6651  -0.9666  -0.8469   1.3070   1.6496

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.32300    0.32283   1.001 0.317055
Stadt        0.94478    0.75013   1.259 0.207853
SAWald      -1.38710    0.40012  -3.467 0.000527 ***
Acker       -1.35568    0.38263  -3.543 0.000396 ***
Wiesen       0.02524    0.42690   0.059 0.952848
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2557.1  on 1926  degrees of freedom
Residual deviance: 2507.9  on 1922  degrees of freedom
AIC: 2517.9

Number of Fisher Scoring iterations: 4
```

Interpretation: Art kommt im Wald und auf Äckern seltener vor, wegen positivem Intercept ist die grundsätzliche Auftretenswahrscheinlichkeit nicht klein (bei dieser Art ca 38%). Nur Wald und Acker haben scheinbar (siehe nichtlineare Modelle weiter unten) einen signifikanten Einfluß, dieser ist negativ, die Art kommt in Wäldern und Äckern seltener vor als im Rest von Bayern.

Das Chancenverhältnis für feste Werte aller erklärenden Variablen ist definiert als

$$\frac{\mathrm{P}(y = 1 \,|\, \text{Stadt, Wald, Acker, Wiesen})}{\mathrm{P}(y = 0 \,|\, \text{Stadt, Wald, Acker, Wiesen})}$$

Zwischen einem Quadrat mit gar keinem Wald (`SAWald = 0`) und einem reinen Waldgebiet (`SAWald = 1 = 100%`) verändert sich das Chancenverhältnis die Heuschrecke zu beobachten auf rund ein Viertel:

```
> betaWald <- coef(glm1)["SAWald"]

> betaWald

[1] -1.3871

> exp(betaWald)

[1] 0.2497987
```

Dabei wird angenommen, daß sich in den anderen Variablen nichts ändert (was bei Anteilen an Bodennutzung natürlich schwer geht). Erhöht sich der Waldanteil um 10% verändert sich das Chancenverhältnis auf

```
> exp(betaWald * 0.1)

[1] 0.8704804
```

die Chance die Heuschrecke anzutreffen sinkt also um rund 13%.

Für ein konkretes Quadrat mit z.B. keinem Stadtanteil (0), Hälfte Wald (0.5), ein Viertel Äcker (0.25) und 5% Wiesen (0.05) ergibt sich die Wahrscheinlichkeit, die Säbel-Dornschrecke anzutreffen mit

```
> ## Beta: Koeffizienten des Modells
> beta <- coef(glm1)
> beta

(Intercept)       Stadt      SAWald       Acker      Wiesen
 0.32299745  0.94478251 -1.38709996 -1.35567678  0.02524298

> ## Eta: linearer Praediktor
> eta1 <- beta[1] + 0*beta[2] + 0.5*beta[3] + 0.25*beta[4] + 0.05*beta[5]

> ## Wahrscheinlichkeit
> wkt1 <- exp(eta1)/(1+exp(eta1))
> wkt1

[1] 0.3299946
```

die vom Modell vorhergesagte Wahrscheinlichkeit die Säbel-Dornschrecke anzutreffen ist also rund ein Drittel.

Steigt der Waldanteil von 50% auf 60% und alles andere bleibt gleich (0% Stadt, 25% Acker, 5% Wiesen) sinkt die Wahrscheinlichkeit von 0.33 auf rund 0.30:

```
> eta2 <- beta[1] + 0*beta[2] + 0.6*beta[3] + 0.25*beta[4] + 0.05*beta[5]

> wkt2 <- exp(eta2)/(1+exp(eta2))
> wkt2

[1] 0.3000794
```

Die Chancenverhältisse für die beiden Quadrate ergeben sich wegen $\mathrm{P}(y = 0) = 1 - \mathrm{P}(y = 1)$ als

```
> cv1 <- wkt1/(1-wkt1)
> cv1

[1] 0.4925252

> cv2 <- wkt2/(1-wkt2)
> cv2

[1] 0.4287336
```

Im ersten Quadrat ist die Chance die Säbel-Dornschrecke anzutreffen mit einem Drittel nur rund halb so groß, wie sie nicht anzutreffen (zwei Drittel), daher ist das Chancenverhältnis `cv1` ungefähr 0.5. Will man nun `cv1` und `cv2` vergleichen, so ist (mit ein wenig Übung) der Quotient leichter zu interpretieren, weil beide Terme ja selber bereits Quotienten sind:

```
> cv2/cv1

[1] 0.8704804

> exp(betaWald*0.1)

[1] 0.8704804
```

Der Ausdruck *„e hoch Regressionskoeffizient“* im linearen Logitmodell beschreibt also, wie sich die Chancenverhältnisse ändern, wenn sich nur diese eine Variable ändert. Negative Koeffizienten heißt die Wahrscheinlichkeit sinkt, positive daß die Wahrscheinlichkeit steigt.

Die manuellen Berechnungen der Wahrscheinlichkeiten wie oben sind in R natürlich nicht notwendig, dafür gibt es die Funktion `predict()`. Diese kann sowohl den linearen Prädiktor wie auch die Wahrscheinlichkeiten berechnen.

2.2 Generalisierte additive Modelle

Das generalisierte additive Modell kann wieder mit Hilfe des R Commanders erzeugt werden, indem erst ein generalisiertes lineares Modell wie in **??** erzeugt, `glm` durch `gam` ersetzt wird und jede erklärende Variable in Klammern gepackt und mit einem `s` versehen wird. Das GAM ergibt sich also durch:

```
> gam1 <-  gam(TETRIX.SUBULATA~s(Stadt)+s(SAWald)+s(Acker)+s(Wiesen),
+              data=ghuepfer, family=binomial)
> summary(gam1)

Family: binomial
Link function: logit

Formula:
TETRIX.SUBULATA ~ s(Stadt) + s(SAWald) + s(Acker) + s(Wiesen)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.51248    0.04802  -10.67   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
            edf Ref.df Chi.sq  p-value
s(Stadt)  2.918  3.658 21.437 0.000218 ***
s(SAWald) 1.000  1.001  8.668 0.003245 **
s(Acker)  1.001  1.002 10.511 0.001192 **
s(Wiesen) 1.981  2.494  2.955 0.303437
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.037   Deviance explained = 3.09%
UBRE score = 0.29416  Scale est. = 1         n = 1927
```

Nichtlineare Terme sind also nur für Stadt und Wiesen notwendig, der Beitrag von Wiese ist aber nicht signifikant. Wir passen daher ein kleineres Modell an, in dem nur Stadt einen nichtlinearen Einfluß auf die Zielgröße hat:

```
> gam1a <-  gam(TETRIX.SUBULATA~s(Stadt)+SAWald+Acker,
+               data=ghuepfer, family=binomial)
> summary(gam1a)

Family: binomial
Link function: logit

Formula:
TETRIX.SUBULATA ~ s(Stadt) + SAWald + Acker

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3486     0.1569   2.223   0.0262 *
SAWald       -1.2787     0.2916  -4.385 1.16e-05 ***
Acker        -1.4139     0.2360  -5.992 2.07e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
           edf Ref.df Chi.sq  p-value
s(Stadt) 2.981  3.736  22.94 0.000118 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.0355   Deviance explained = 2.92%
UBRE score = 0.29448  Scale est. = 1         n = 1927
```

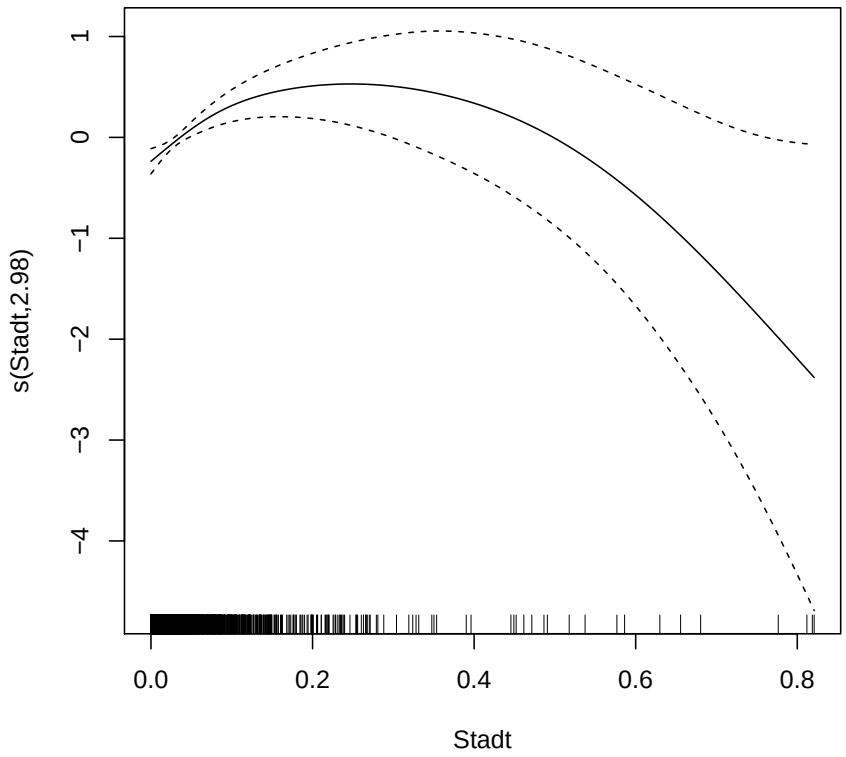
Die Regressionskoeffizienten für die linearen Einflußgrößen Wald und Acker sind sehr ähnlich zum linearen Logitmodell, die Interpretation ist identisch: 10% mehr Wald ändert Chancenverhältnis um

```
> coef(gam1a)["SAWald"]

   SAWald
-1.278697

> exp(0.1 * coef(gam1a)["SAWald"])

  SAWald
0.879968
```

(die 0.1 sind wieder die 10% Änderung des Waldanteils).

Der Einfluß von Stadt ist am besten in einer Grafik zu sehen:

```
> plot(gam1a)
```



Bis zu einem Stadtanteil von 30% steigt die Auftretenswahrscheinlichkeit, danach sinkt er stark (es gibt aber da dann natürlich nur sehr wenig Beobachtungen). Anhand der Grafik kann man auch sehen, warum der Einfluß von Stadt im linearen Modell nicht signifikant war: Wenn man die Kurve durch eine Gerade approximiert, so ist diese annähernd horizontal, und es scheint keinen Zusammenhang zwischen Stadt und Auftretenswahrscheinlichkeit zu geben.

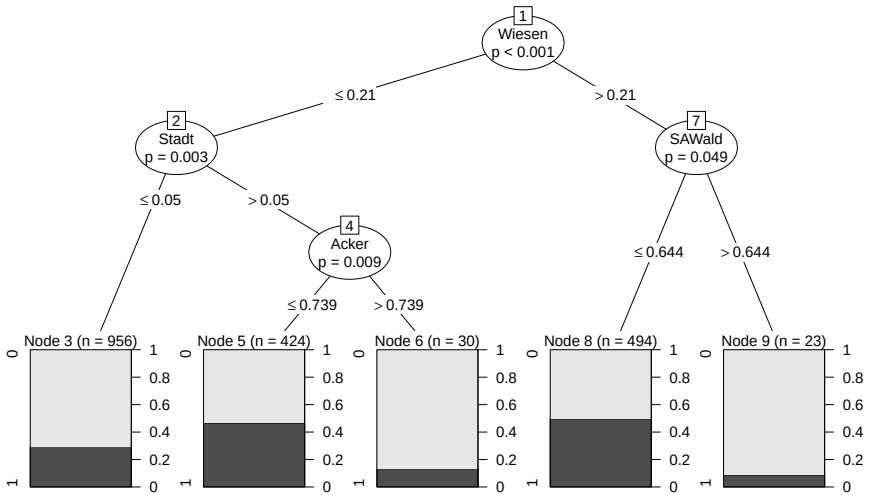
2.3 Klassifikationsbaum

```
> baum1 <-  ctree(factor(TETRIX.SUBULATA)~Stadt+SAWald+Acker+Wiesen,
+                 data=ghuepfer)
```

Das `factor()` in der Formel sagt `ctree()`, daß die binäre Größe `TETRIX.SUBULATA` als kategorisch angesehen werden soll (Auftreten ja/nein) und nicht als die beiden Zahlen 0 und 1. Für metrische Zielgrößen würde wie bei den Gräsern ein Regressionsbaum angelegt.

Interpretation wieder über die Grafik:

```
> plot(baum1)
```



Zuerst wird unterschieden, ob der Wiesenanteil kleiner oder größer als 21% ist. Bei mehr als 21% Wiesen wird nach Wald geteilt. Insgesamt 494 Quadrate im Datensatz haben einen Wiesenanteil über 21% und Waldanteil kleiner als 64%, diese sind im Endknoten 8. In diesen 494 Quadraten kommt die Säbel-Dornschrecke mit einer Wahrscheinlichkeit von ungefähr 0.5 vor (dunkler grauer Balken). Bei Wiesenanteil über 21% und Waldanteil über 64% sinkt die Auftretenswahrscheinlichkeit massiv. Das ist konform mit den anderen Modellen: Diese Art kommt im Wald seltener vor.