Pruscha, Göttlein:

Regression Analysis for Forest Inventory Data with Time and Space Dependencies

Sonderforschungsbereich 386, Paper 173 (1999)

Regression Analysis for Forest Inventory Data with Time and Space Dependencies

Helmut Pruscha and Axel Göttlein
University of Munich and Technical University of Munich

Summary

In this paper the data of a forest health inventory are analysed. Since 1983 the degree of defoliation (damage), together with various explanatory variables (covariates) concerning stand, site, soil and weather, are recorded by the second of the two authors, in the forest district Rothenbuch (Spessart, Bavaria). The focus is on the space and time dependencies of the data. The mutual relationship of space-time functions on the one side and the set of covariates on the other is worked out. To this end we employ *generalized linear models (GLMs)* for ordinal response variables and employ *semiparametric* estimation approaches and appropriate *residual* methods. It turns out that (i) the contribution of space-time functions is quantitatively comparable with that of the set of covariates, (ii) the data contain much more (timely and spatially) sequential structure than smooth space-time structure, (iii) a fine analysis of the individual sites in the area can be carried out with respect to predictive power of the covariates.

Keywords: Forest inventory data; Cumulative regression model; Generalized linear model; Ordinal residual values; Semiparametric model; Space-time dependencies.

1 Introduction and Data

We are concerned with a set of forest health data in a district of the Spessart (Bavaria). The inventory was carried out over the period of 15 years, 1983-1997, and comprises 80 sites with beech trees. The damage variable Y describes the percentage of leaves lost by the trees. It is gained by visual inspections and is measured on an ordinal scale,

$$0 \approx 0\%,\ 1 \approx 12.5\%,\ 2 \approx 25\%, \ldots,\ 6 \approx 75\% \quad \text{defoliation.}$$

The damage values Y vary drastically over space (Fig. 1a), with a minimum averaged value per observation point of 0.0 and a maximum of 3.40 and –to a less degree– over time (min = 0.15, max = 0.84; Fig. 1b, data are here averaged over the sites). Fig. 2 shows the positions of the 80 beech sites in the 16 km $\times$ 8 km area and combines space and time information: the mean damage values of the three consecutive 5-years periods are coded into a 3 digit number. Many beeches in south-western sites are of good health, different to the beeches in the south and west of the village Rothenbuch (situated at the coordinates $(x, y) = (10, 5)$) some of them showing a drastical increase in their damage values over the 15 years of observation.

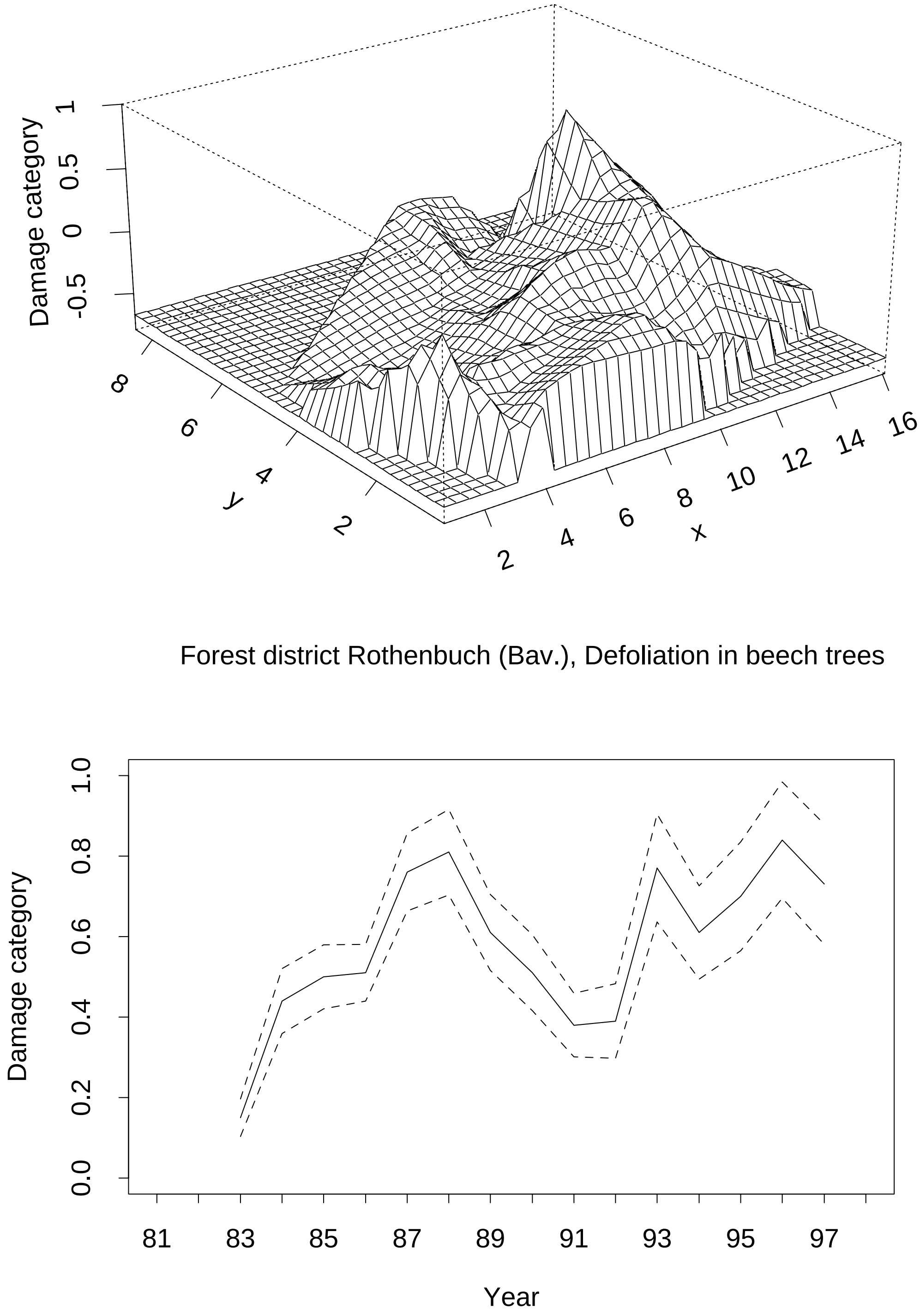


Figure 1: **a)** above: Spatial distribution of damages in beech trees over the area of 80 sites under observation (averaged over the 15 years). The plot was produced by a very close local polynomial fitting (`S+, loess(x,y,span=0.1)`) and is presented with a z-axis centered at zero. **b)** below: Timely development of damages in beech trees over the 15 years of observation. The mean values (averaged over the 80 sites) are plotted, together with a band of $1\times$ standard-error of the mean.

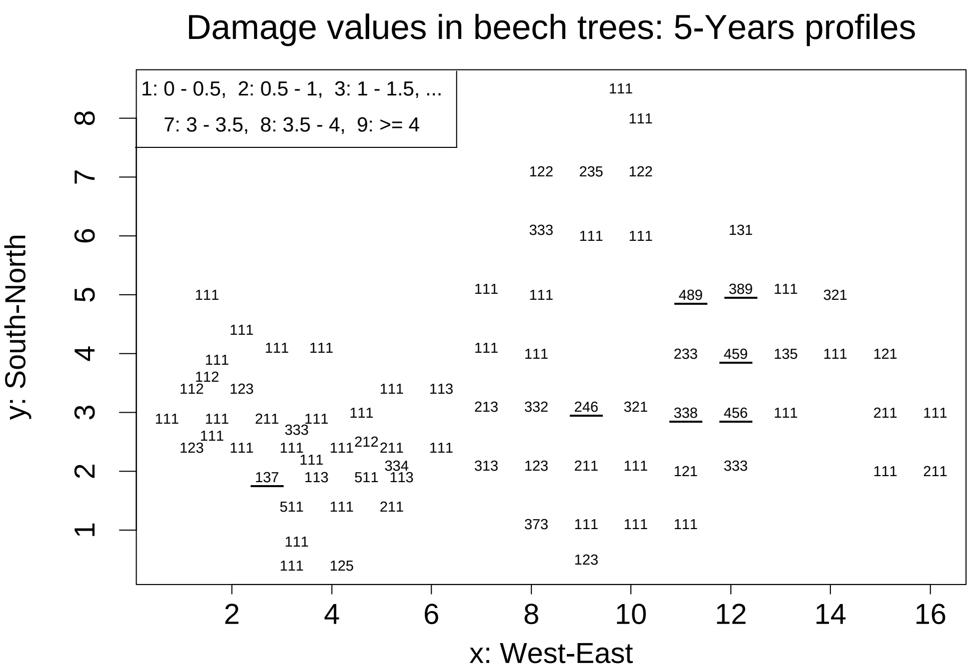


Figure 2: The damage profile for each of the 80 sites, coded by a 3 digit number, giving average values for the three consecutive 5-years periods. The code numbers of sites with strongly increasing damages are underlined.

The damage category for the year t at the site with west-east and south-north coordinates (x, y) will be denoted by $Y_{t,xy}$. In addition to the *response* $Y_{t,xy}$, a vector $Z_{t,xy}$ of 14 *covariates* was recorded, consisting of variables concerning the

- stand: age, age^2, canopy, fructification
- site: elevation, part of slope, steepness, number of tree species
- soil: humus layer, pH-values (H_2O in 1 and 16 cm depth), fertilizing, moisture
- climate: dryness

Some of the covariates vary over time, some over the sites, some vary over time and sites.

It is evident that the actual observation $Y_{t,xy}$ may depend on both, the covariates $Z_{t,xy}$ and the space-time coordinates. The time-dependencies of the response variable Y_t can be modelled in two different ways: by introducing as a regressor variable

(i) the last-year-value Y_{t-1} at the site

(ii) a smooth *time function* $s(t)$, which is fitted to the data.

The approach (i) leads to a time series model with discrete response variable. Those time series models are usually defined on a *generalized linear model* (GLM) basis and were already investigated by Kaufmann (1987), Zeger and Qaqish (1988), Pruscha (1993), Li (1994). Similarly, the space dependencies can be accounted for by regressing on

(i) the neighbourhood values Y_t^* of the site

(ii) a smooth *space function* $a(x, y)$, which is fitted to the data.

The aim of the present study is

1. to compare approach (i) with approach (ii) (and then to dismiss the regressor variables Y_{t-1}, Y_t^* of approach (i), since we are mainly interested in the smooth space-time structure of the data)

2. to separate, assess and visualize the contributions of the 14 covariates Z and of the space-time information t, xy (with respect to the damage category of trees)

3. to characterize individual sites by their different levels and different developments of damages over the 15 years period, before and after adjustment for the covariates (can changes in the damage values be traced back to the covariates?)

Point 3 seems to be especially useful from the forestry point of view.

The statistical model we will underly our data is a *semiparametric* regression model based on GLMs. The technique we will employ consists of a residual method for ordinal response variables introduced earlier (Pruscha, 1994).

Applications of GLM-type regression models to forest inventories can be found in Kublin (1987), Göttlein & Pruscha (1992), Pritscher et al. (1995), with the latter two emphazising the time series and the spatial aspects, resp., of the data.

2 Regression Models

We are faced with a response variable Y which assumes the distinct ordered values $0, 1, \ldots, q$ (actually we have $q = 6$).

Our data set will be analysed by semiparametric regression models on the basis of GLMs taking into regard the ordinal scale of response variable Y (*cumulative* regression models, see McCullagh, 1980).

To be precise, we start with some notation. If p_j denotes the probability for the damage category j, i.e. for the event $Y = j$, $j = 0, \ldots, q$, we introduce *cumulative probabilities* by

$$p_{(j)} = p_0 + \ldots + p_j, \qquad \text{i.e.} \quad p_0 = p_{(0)} \leq p_{(1)} \leq \ldots \leq p_{(q)} = 1.$$

The statistical models will comprise the following quantities:

- covariate vector $Z_{t,xy}$ for the current year and site

- last year value $Y_{t-1,xy}$ for the site
- neighbourhood value $Y_{t,(xy)^*}$ for the current year, averaged over all sites being within the range of 1 km of the point (x,y)
- smooth space function $a(x,y)$ and time function $s(t)$.

The number of years will be denoted by T (here $T = 15$), the number of beech sites under observation by N (here $N = 80$), the set of the coordinates (x,y) of the N beech sites by the symbol $\mathbb{D}$. Then the *conditional* cumulative probabilities

$$p_{t,xy(j)} = \mathbb{P}(Y_{t,xy} \leq j \,|\, \mathcal{H}_{t,xy}),$$

given the information $\mathcal{H}_{t,xy}$ up to time t for site (x,y), is modelled in the form

$$p_{t,xy(j)} = F(\eta_{t,xy(j)}), \quad t = 1,2,\ldots,T, \quad (x,y) \in \mathbb{D}, \quad j = 0,\ldots,q, \tag{1}$$

where $F : \mathbb{R} \to \mathbb{R}$ is an appropriate cumulative distribution function and η is the *regression term* of the model. We take the logistic distribution function, i.e.

$$F(s) = \frac{1}{1+e^{-s}}, \; s \in \mathbb{R},$$

arriving at a cumulative *logistic* model. As regression term we will choose –according to the two approaches mentioned above– a parametric and a semiparametric equation, resp., namely

$$\begin{array}{lrcl} (i) & \eta_{t,xy(j)} & = & \kappa_{(j)} + \beta^\top \cdot Z_{t,xy} + \gamma \cdot Y_{t-1,xy} + \delta \cdot Y_{t,(xy)^*} \\ (ii) & \eta_{t,xy(j)} & = & \kappa_{(j)} + \beta^\top \cdot Z_{t,xy} + a(x,y) + s(t), \end{array}$$

where $\kappa_{(j)}$ are *treshold* constants and $\beta \in \mathbb{R}^p$ comprises the unknown regression coefficients for the covariates (here $p = 14$).

The overall performance of the model (1) with regression term η can be quantified by

- the negative log-likelihood $\text{negLLH} = -\sum_{t,xy} \log\, p_{t,xy,Y_{t,xy}}(\hat{\theta})$
- the percentage of expected correct Monte Carlo predictions $\text{ECMCP\%} = 100 \cdot \sum_{t,xy} p_{t,xy,Y_{t,xy}}(\hat{\theta})/(N \cdot T)$.

Here, $p_{t,xy,j} = p_{t,xy(j)} - p_{t,xy(j-1)} = \mathbb{P}(Y_{t,xy} = j|\mathcal{H}_{t,xy})$ denotes the probability of the occurrence of category j, at year t and site (x,y). The estimator $\hat{\theta}$ of the unknown parameters $\theta = (\kappa, \beta, a, s)$ of model (1) with semiparametric regression term (ii) are gained by the `Splus` procedure

```
gam(Y ~ Z + loess(x,y,span=0.2) + loess(t,span=0.5),quasi(link=log)),
```

see Venables & Ripley (1997), adopted to cumulative logistic regression (`loess` is a nonparametric local polynomial fitting procedure, involving as neighbouring points `span` % of the data points).

Regression Term		negLLH	ECMCP%
1	covariates Z	958.30	55.71 %
2	covariates $Z + a(x,y) + s(t)$	895.21	58.48 %
3	covariates $Z + Y_{t-1,xy} + Y_{t,(xy)^*}$	619.23	70.52 %
4	$a(x,y) + s(t)$	997.70	53.11 %
5	$Y_{t-1,xy} + Y_{t,(xy)^*}$	680.13	68.25 %

Starting with regression term 1 in the table, the increase of goodness-of-fit is relatively small (although statistically significant) when applying the regression term 2 with smooth space-time functions and much greater when applying the term 3 with last year- and neighbourhood-values. This finding is supported when starting with an empty regressor set and then applying the space-time functions alone (i.e. regressor term 4) or the adjacent values $Y_{t-1,xy}$, $Y_{t,(xy)^*}$ alone (i.e. regressor term 5 in the above table).

As an obvious first result we can state, that the forest inventory data contain much more (timely and spatially) sequential structure than smooth space-time structure.

As a second result we note that the contribution of the space and time functions $a(x,y)$ and $s(t)$ is quantitatively comparable with that of the regressor set Z; compare regressor terms 1 and 4 of the table.

Before dismissing the regressors $Y_{t-1,xy}$ and $Y_{t,(xy)^*}$ for the rest of the paper two remarks are in order.

1. The values of the (timely and spatially adjacent) neighbours are strongly correlated with the actual values $Y_{t,xy}$. Their inclusion in the regression equation can be recommended when assessing the individual covariates, as Göttlein & Pruscha (1996) did, since otherwise the significances of most explanatory variables are highly overestimated.

2. The approach (i) above includes the vector

$$\underline{Y}_{t,(xy)^*} = (Y_{t,(x-1,y)}, Y_{t,(x+1,y)}, Y_{t,(x,y-1)} Y_{t,(x,y+1)})$$

   of neighbouring values (assuming a regular grid for the moment). Thus it fixes a set of transition probabilities

$$(*) \qquad \mathbb{P}(Y_{t,xy} \leq j \,|\, \underline{Y}_{t,(xy)^*}), \quad xy \in \mathbb{D}.$$

   The question arises, wether the set $(*)$ of transition probabilities is at all consistent, and approach (i) with regressors $\underline{Y}_{t,(xy)^*}$ at all meaningful. In technical terms, we ask for the existence of a probability distribution $\mathbb{P}_t$ for $Y_{t,xy}$, $xy \in \mathbb{D}$, having $(*)$ as conditional probabilities. The existence is guaranteed for GLMs with *natural* link functions (Winkler, 1995, chap. 3). In our situation of a cumulative model with a nonnatural link function this theoretical problem remains open.

3 Global and Partial Residuals

Using the estimator $\hat{\theta}$ for the parameter vector θ, comprising all unknown model parameters, we calculate the predicted mean value for the response $Y_{t,xy}$ on the basis of model (1) by

$$\hat{m}_{t,xy} = \sum_{j=0}^{q} j \cdot p_{t,xy,j}(\hat{\theta}),$$

and the estimated standard deviation $s_{t,xy}$ by

$$s_{t,xy}^2 = \sum_{j=0}^{q} j^2 \cdot p_{t,xy,j}(\hat{\theta}) - (\sum_{j=0}^{q} j \cdot p_{t,xy,j}(\hat{\theta}))^2.$$

Global standardized residuals values are then defined by

$$\hat{e}_{t,xy} = \frac{1}{s_{t,xy}}(Y_{t,xy} - \hat{m}_{t,xy}), \quad t = 1, \ldots, T, \; xy \in \mathbb{D}. \tag{2}$$

The residuals $Y_{t,xy} - \hat{m}_{t,xy}$ in (2) support the *ordinal* nature of the response Y; they can be gained from general multivariate GLM residuals. To show this, we suppress the subscript xy and assume that the vector $W_t = (Y_{t,0}, \ldots, Y_{t,q-1})^\top$ is multinomially distributed with parameters 1 and $p_t = (p_{t,0}, \ldots, p_{t,q-1})^\top$. Defining for the following $\eta_t(\theta) = (\eta_{t(0)}(\theta), \ldots, \eta_{t(q-1)}(\theta))^\top$ and the *response vector function* $h^\top(\eta_t) = (h_0(\eta_t), \ldots, h_{q-1}(\eta_t))$, the *multivariate* GLM is

$$p_{t,j}(\theta) = h_j(\eta_t(\theta)), \quad t = 1, \ldots, T, \; j = 0, \ldots, q,$$

putting $p_{t,q} = 1 - (p_{t,0} + \ldots + p_{t,q-1})$ as probability for the category q. Multivariate GLM residuals $\underline{\hat{e}}_t = (\hat{e}_{t,0}, \ldots, \hat{e}_{t,q-1})^\top$ are of the form

$$\underline{\hat{e}}_t = D_t^{-\top}(\hat{\theta})\,(W_t - p_t(\hat{\theta})), \qquad D_t(\theta) = \left(\frac{d}{d\eta} h^\top(\eta_t)\right),$$

(Fahrmeir & Tutz, 1994, p. 98). In our case of a cumulative model (1) we have

$$h_j(\eta_t(\theta)) = F(\eta_{t(j)}(\theta)) - F(\eta_{t(j-1)}(\theta)),$$

and $D_t^{-\top}(\theta)$ is a lower triangular matrix, with

$$(1/F'(\eta_{t(j)}(\theta)), \ldots, 1/F'(\eta_{t(j)}(\theta)), 0, \ldots, 0)$$

as its j-th row. Hence we get

$$\hat{e}_{t,j} = \frac{1}{\hat{d}_{t,j}} \sum_{k=0}^{j} (Y_{t,k} - p_{t,k}(\hat{\theta})), \qquad \hat{d}_{t,j} = F'(\eta_{t(j)}(\hat{\theta})).$$

Putting $\tilde{e}_t = \sum_{j=0}^{q} \hat{d}_{t,j}\, \hat{e}_{t,j} \,/\, \hat{d}_t$, $\hat{d}_t = \sum_{j=0}^{q} \hat{d}_{t,j}$, we finally arrive at

$$\tilde{e}_t = -\frac{1}{\hat{d}_t}(Y_t - \hat{m}_t), \tag{3}$$

with $Y_t = \sum_{j=0}^{q} j \cdot Y_{t,j}$, $Y_{t,q} = 1 - \sum_{j=0}^{q-1} Y_{t,j}$. Equation (3) amounts to (2) after exchanging the denominator and the sign, which is a matter of convention only. For details of the above derivation see Pruscha, (1994). The denominator in (3) is more appropriate for partial residual analysis, which follows now: If $\eta_{t(j)} = \kappa_{(j)} + \eta_t^{(1)} + \eta_t^{(2)}$ is a decomposition of the regression term, *partial* residuals *from* regression on part $\eta_t^{(1)}$ are defined by $\tilde{e}_t + \eta_t^{(2)}(\hat{\theta})$. We prefer the opposite sign, arriving at partial residuals of the form

$$\hat{e}_{t,xy}^{(1)} = \frac{1}{\hat{d}_{t,xy}} (Y_{t,xy} - \hat{m}_{t,xy}) - \eta_{t,xy}^{(2)}(\hat{\theta}), \tag{4}$$

the indices referring again to the (x, y) information. For the special case of binary logistic model see Landwehr et al. (1984).

4 Residuals in the Space-Time Analysis

We resume our program to analyse the prediction of the outcome Y, and to separate the contributions of the covariates Z from the contribution of the space and time functions $a(x, y)$ and $s(t)$. To this end, we build ordinal residual values from the regression on the covariates Z and check then the role of the smooth functions a and s. This is done in a twofold manner, namely by applying the *semiparametric* model, including the terms Z, a and s, and form partial residuals from Z (in 4.1 below) and by applying the *parametric* model, including Z only, and build global residuals (in 4.2 below).

For the interpretation of (standardized) residuals $\hat{e}$ one should note, that a high positiv residual value (say $\hat{e} \geq 1$) indicates a surplus of damage which cannot be traced back to the explanatory variables Z. A negative value is interpreted analogously, while a residual value around zero tells us that the damage level can quite good be "explained" by the values of Z.

4.1 Partial residuals from regression on the covariates Z

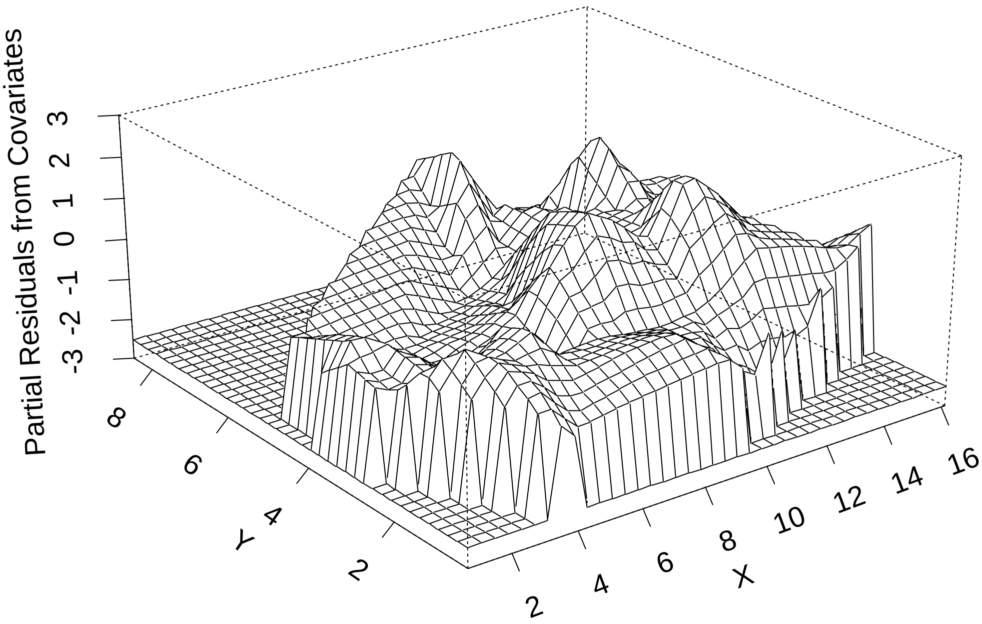
We apply model (1) with the *semiparametric* regression term

$$\eta_{t,xy(j)} = \kappa_{(j)} + \beta^\top \cdot Z_{t,xy} + a(x, y) + s(t)$$

to the data and form *partial* residuals $\hat{e}_{t,xy}^{(1)}$ from regression on Z according to equation (4). We plot these residual values over the (x, y)-coordinates and over the time t resulting in Fig. 3a and Fig. 3b.

First we discuss the spatial distribution of the partial residuals (Fig. 3a): The covariates mostly fail to explain the high/medium damages around the village Rothenbuch (see the four peaks around $(x, y) = (10, 5)$) The low damages in the west part of the area and in the north of Rothenbuch at $(10.2, 6.0) - (10.2, 8)$ are more or less correctly predicted by the covariates.

Now to the the development in time (Fig. 3b): The peak in the year 1988 is badly explained, the residual values at 1988 are large, too. This is different with the large



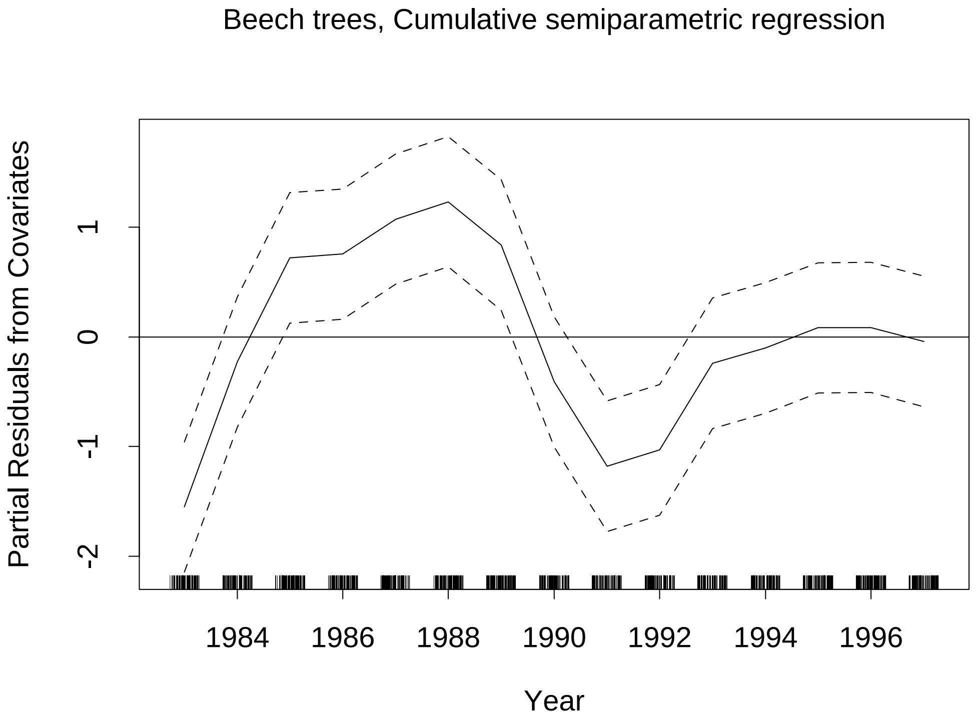


Figure 3: **a)** above: Spatial distribution of *partial* residual values over the area of 80 sites (averaged over the 15 years). The residuals are gained by using a *semiparametric* logistic model and by building partial residuals from regression on the set of covariates. The plot was produced by a very close local polynomial fitting (`S+, loess(x,y,span=0.1)`). **b)** below: Timely development of residual values over the 15 years, together with a 2× standard-error band (same residual method as above).

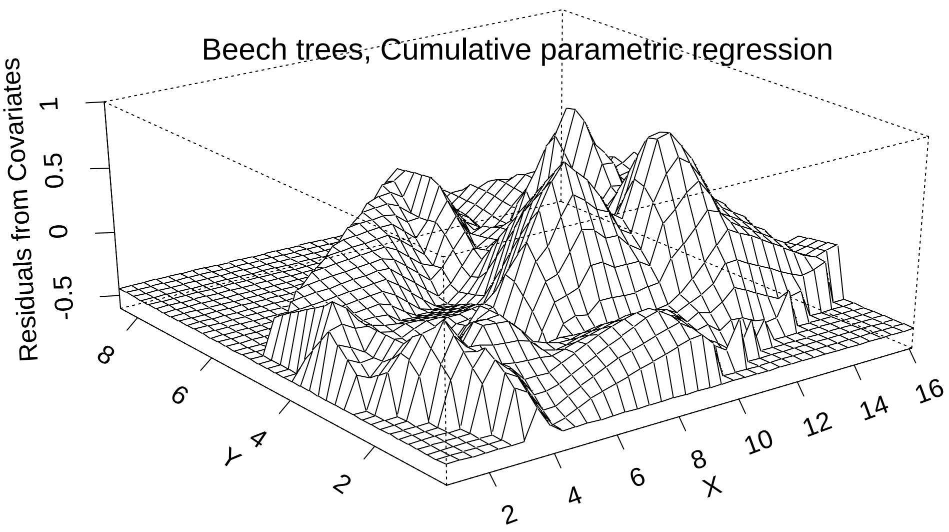


Figure 4: Spatial distribution of *global* residual values over the area of 80 sites (averaged over the 15 years). The residuals are gained by using a *parametric* logistic model and by building global residuals from regression on the set of covariates.

damage values in the period 1993 - 1997, where the residuals are around zero: The damages in this period can be traced back to the development of covariates.

4.2 Global residuals from regression on the covariates Z

We apply model (1) with the *parametric* regression term

$$\eta_{t,xy(j)} = \kappa_{(j)} + \beta^\top \cdot Z_{t,xy}$$

to the data and form *global* residuals $\hat{e}_{t,xy}$ from regression on Z according to equation (2). The plot of these residual values over the (x, y) area, that is Fig. 4, can be compared to Fig. 3a). It shows the peaks of "unexplained" damages around the village Rothenbuch even clearer. The plot of these residual values over time is nearly identical to Fig. 3b) and omitted. Fig. 5 combines the space and time information: the mean residual values $\hat{e}$ of the three consecutive 5-years periods are coded into a 3 digit number, as we already did with the damage values Y in Fig. 2.

Now we can perform a fine analysis of individual sites which seems quite useful from the forestry point of view. This will be done by comparing –for a specific site– the profiles of damages and of residuals, given in Fig. 2 and Fig. 5 (showing the three 5-years-averages of damage values Y and of residual values $\hat{e}$, resp.). Let us consider 3 different sites, with (x, y) coordinates $(1.5, 5.0), (4.2, 1.4)$ and $(8.1, 4.0)$. They all have the same damage profile 111 (Fig. 2), but different residual profiles, namely 333, 111 and 553, resp. (Fig. 5). The following table presents the numerical values of damages and residuals instead of the coded values, as Fig. 2 and 5 do.

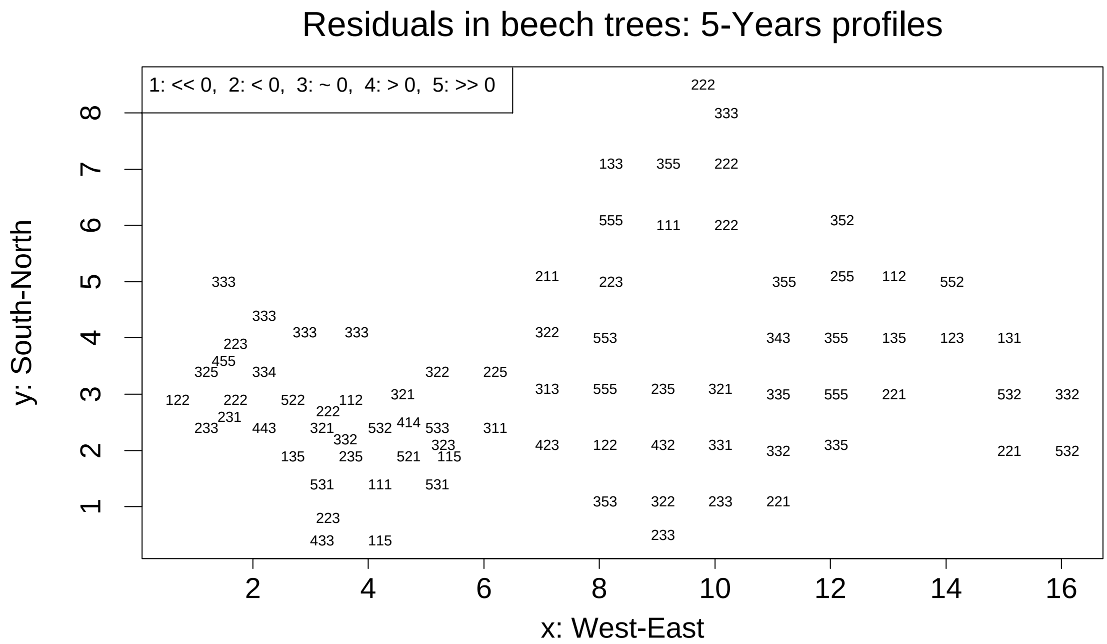


Figure 5: The residual profile for each of the 80 sites, coded by a 3 digit number, giving average values for the three consecutive 5-years periods. Code number 1: residual value $\hat{e} \leq -s$, number 2: $-s < \hat{e} \leq -s/2$, number 3: $-s/2 < \hat{e} \leq s/2$, number 4: $s/2 < \hat{e} \leq s$, number 5: $\hat{e} > s$ (s the standard deviation of the residual values).

Site	(x, y)	Damage values	Residual values	age	canopy %	fertil	NSpec
(1)	1.5 5.0	0.0 0.0 0.0	-0.08 -0.09 -0.10	43-57	100	1	3
(30)	4.2 1.4	0.0 0.0 0.0	-0.82 -0.83 -0.81	64-78	90	0	1
(47)	8.1 4.0	0.4 0.4 0.2	1.45 1.30 0.20	26-40	100-90	0	3
(48)	8.1 3.1	1.2 1.4 0.8	1.78 2.84 1.36	158-172	90-80	1	3

At site (1) the residual values are near zero: Damage values approximately match the values predicted by the covariates.
At site (30) the residual values are strongly negative: the beeches are more healthy than predicted ((i) no fertilizing and (ii) the beech as sole species are normally two indicators for higher damages).
At site (47) the residuals for the first 10 years are strongly positive: the beeches are less healthy than young trees within a mixing forest stand should be. In the last 5-years period the decrease of canopy and the increase in age seems (among others) to be responsible for the closeness of predicted and observed damage value.
As a last example: The medium damages at site (48) –with profile 332 in Fig 2– are too high taking into account the covariates (age, fertilizing, mixing forest stand) leading to high residual values and thus to a residual profile of 555 in Fig. 5.

Acknowledgments

This work is partly supported by the Deutsche Forschungsgemeinschaft (SFB 386).

References

Fahrmeir, L. & Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, N.Y.

Göttlein,A. & Pruscha,H. (1992). Ordinal time series models with application to forest damage data. In: *Lecture Notes in Statistics*, **78**. Springer, New York, 113-118.

Göttlein,A. & Pruscha,H. (1996). The influence of stand characteristics, topography, site and weather on tree health in the forest district Rothenbuch. *Forstw. Cbl.*, **115**, 146-162.

Kaufmann,H. (1987). Regression models for nonstationary categorical time series: asymptotic estimation theory. *Ann. Statist.*, **15**, 79-98.

Kublin,E. (1987). Statistische Auswertmodelle für Waldschadensinventuren - Methodische Überlegungen. *Forstw. Cbl.*, **106**, 57-68.

McCullagh,P. (1980). Regression models for ordinal data (with discussion). *J. R. Statist. Soc. B*, **42**, 109-142.

Landwehr,J.M., Pregibon,D. & Shoemaker,A.C. (1984). Graphical methods for assessing logistic regression models. *J. Am. Statist. Ass.*, **79**, 61-71.

Li,W.K. (1994). Time series models based on generalized linear models: some further results. *Biometrics*, **50**, 506-511.

Pritscher,E., Bäumler,A. & Fahrmeir,L. (1995). Marginale Regressionsmodelle für ordinale Waldschadensdaten mit räumlicher Korrelation. *Forstw. Cbl.*, **113**, 367-378.

Pruscha,H. (1993). Categorical time series with a recursive scheme and with covariates. *statistics*, **24**, 43-57

Pruscha,H. (1994). Partial residuals in cumulative regression models for ordinal data. *Statist. Papers*, **35**, 273-284.

Venables,W.N. & Ripley,B.D. (1997). *Modern Applied Statistics with Splus*. 2nd ed. Springer-Verlag, N.Y.

Winkler,G. (1995). *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Springer-Verlag, Berlin.

Zeger, S. L. & Qaqish, B. (1988). Markov regression models for time series: a quasi likelihood approach. *Biometrics*, **44**, 1019-1031.

Authors' address:
H.Pruscha, Mathematisches Institut der Universität München, Theresienstr.39, D 80333 München.
e-mail: pruscha@rz.mathematik.uni-muenchen.de
A.Göttlein, Lehrbereich Waldernährung und Wasserhaushalt der Technischen Universität München, Am Hochanger 13, D 85354 Freising. e-mail: goettlein@boku.forst.uni-muenchen.de