Pruscha:

Some Forecast Methods in Regression Models for Categorical Time Series

Sonderforschungsbereich 386, Paper 3 (1995)

Online unter: http://epub.ub.uni-muenchen.de/

Some Forecast Methods in Regression Models for Categorical Time Series [1]

Helmut Pruscha
Mathematisches Institut der Universität München

Abstract

We are dealing with the prediction of forthcoming outcomes of a categorical time series. We will assume that the evolution of the time series is driven by a covariate process and by former outcomes and that the covariate process itself obeys an autoregressive law. Two forecasting methods are presented. The first is based on an integral formula for the probabilities of forthcoming events and by a Monte Carlo evaluation of this integral. The second method makes use of an approximation formula for conditional expectations. The procedures proposed are illustrated by an application to data on forest damages.

Keywords: Forecasting; Categorical Time Series; Regression Model; Cumulative Model; Monte Carlo Method; Forest Damage Data

1 Introduction

We are concerned with the problem of predicting forthcoming outcomes Y_{T+l} of a categorical time series Y_t, $t = 1, 2, \ldots$, if the history of the process up to time T was observed.

The evolution of the categorical response variable Y_t, where $Y_t \in \; J = \{1, 2, \ldots, m\}$, is assumed to be driven by

1. an $r-$dimensional covariate process Z_t, $t = 1, 2, \ldots$,
2. the last response Y_{t-1},
3. a vector summarizing the history before $t - 1$.

Thus, we are dealing with a transition type of regression model. The conditional probabilities

$$p_{t,j} = \mathbb{P}(Y_t = j \mid \mathcal{H}_t)$$

are modelled in the form $h_j(\eta_t)$, where h_j, $j \in J$, are response functions, η_t is a regression term depending on the regressors introduced in (1) to (3) above, and $\mathcal{H}_t$ comprises the variables

$$Z_1, Y_1, \ldots, Z_{t-1}, Y_{t-1}, Z_t.$$

Special attention is given to a cumulative regression model in the case where Y_t is measured on an ordinal scale.

[1] A version of this paper will appear in the Proceedings Volume of the 10th International Workshop on Statistical Modelling, Innsbruck,1995

To tackle the forecast problem we have to assume that the covariate process Z_t, $t = 1, 2, \ldots$, obeys an own autoregressive law, not influenced by the process Y_t, $t = 1, 2, \ldots$, of the response variables. We will be interested in the l-step predictor of the conditional probabilities $p_{t,j}$, i.e.

$$\hat{p}_{T,j}(l) = \mathbb{E}(p_{T+l,j} \mid \mathcal{F}_T), \quad \mathcal{F}_T = (\mathcal{H}_T, Y_T).$$

We will present two forecasting methods. The first is based on a multiple integral formula for $\hat{p}_{T,j}(l)$ and on its calculation by means of a recursive Monte Carlo algorithm. The second method is based on an approximation of the form

$$\mathbb{E}(h(\eta_{T+l}) \mid \mathcal{F}_T) \approx h(\mathbb{E}(\eta_{T+l} \mid \mathcal{F}_T))$$

and on an recursion formula for $\mathbb{E}(\eta_{T+l} \mid \mathcal{F}_T)$. We will close with an application of the forecasting methods to longitudinal data on forest damages. The responses Y_t are the levels of tree damages at time t, and the covariates Z_t refer to the trees, the site and the soil. From the observation in the period 1983 to 1992 and with an $AR(1)-$ law for the covariate process we try to determine the probability vector $\hat{p}_T(l)$ for the forthcoming damages ($T = 1992$, $l = 1, 2, \ldots$) as well as the mean damage values

$$\hat{\mu}_T(l) = \sum_{j=1}^{m} j \cdot \hat{p}_{T,j}(l).$$

2 Modelling

Let the components of

$$p_t = (p_{t,1}, \ldots, p_{t,m-1})^T$$

be positive, with a sum less than 1, let the vector response variable

$$W_t = (Y_{t,1}, \ldots, Y_{t,m-1})^T$$

be multinomially distributed with parameters 1 and p_t and put

$$Y_{t,m} = 1 - (Y_{t,1} + \ldots + Y_{t,m-1}), \quad p_{t,m} = 1 - (p_{t,1} + \ldots + p_{t,m-1}).$$

A regression model for categorical time series is defined by

$$p_t = h(\eta_t), \quad h : \mathbb{R}^{m-1} \to \mathbb{R}^{m-1} \tag{1}$$

where the regression term $\eta_t = (\eta_{t,1}, \ldots \eta_{t,m-1})^T$ is of the form

$$\eta_{t,j} = \alpha_j + \pi \cdot p_{t-1,j} + \lambda^T \cdot \Lambda(W_{t-1})_j - \beta^T \cdot Z_t. \tag{2}$$

It comprises as regressors the preceding probability vector p_{t-1}, a $q-$ dimensional function Λ of the preceding response W_{t-1} and the covariates Z_t. Unknown are the parameters

$$\alpha \in \mathbb{R}^{m-1}, \quad \pi \in \mathbb{R}, \quad \lambda \in \mathbb{R}^q, \quad \beta \in \mathbb{R}^r.$$

For such transition models see Fahrmeier and Kaufmann (1987), Zeger and Qaqish (1988) and recent surveys by Fahrmeir and Tutz (1994), Diggle et al (1994). In the case of an ordinal response it is useful to introduce cumulative probabilities

$$p_{t,(j)} = p_{t,1} + \ldots + p_{t,j}$$

as well as the cumulative quantities $\Lambda(\cdot)_{(j)}$ and $\eta_{t,(j)}$. Putting

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

model (1),(2) has the form of a cumulative regression model, see McCullagh (1980), namely

$$p_{t,(j)} = F(\eta_{t,(j)}), \quad \eta_{t,(j)} = \alpha_{(j)} + \pi \cdot p_{t-1,(j)} + \lambda^T \cdot \Lambda(W_{t-1})_{(j)} - \beta^T \cdot Z_t, \tag{3}$$

where the $\alpha_{(j)}$ stand in an increasing order, with $\alpha_{(m)} = \infty$, where F is a cumulative distribution function and where the parameters are restricted by

$$\pi \cdot p + \lambda^T \cdot \Lambda(w)_j > 0$$

for all $w, j, p = 0, 1$. The asymptotic theory of model (3) was given in some detail in Pruscha (1993). Two important special cases concerning Λ are

1. $q = m - 1, \quad \Lambda(W_t) = W_t,$ (lagged dummy variables)
2. $q = 1, \quad \Lambda(W_t) = Y_t = \sum_{j=1}^{m} j \cdot Y_{t,j}$ (lagged ordinal variable).

3 Forecasting. The General Set-Up

We adopt the following set-up from time series analysis, see Brockwell and Davies (1987, sec.5.1-5.5). If X_t, $t = 1, 2, \ldots$, is a time series and $\mathcal{F}_T$ comprises the information on $X_1, \ldots, X_T$, we define the $l-$step predictor, the prediction error and the prediction m.s.e, respectively, by

$$\begin{aligned}\hat{X}_T(l) &= \mathbb{E}(X_{T+l} \mid \mathcal{F}_T),\\ \Delta_T(l) &= X_{T+l} - \hat{X}_T(l),\\ V_T(l) &= \mathbb{E}(\Delta_T^2(l) \mid \mathcal{F}_T).\end{aligned}$$

For $MA(\infty)-$ processes $\Delta_T(l)$ and $\mathcal{F}_T$ are independent, such that we have $V_T(l) = \mathbb{E}(\Delta_T^2(l))$ too. With the short-hand notation

$$\mathbb{E}_T(\cdot) = \mathbb{E}(\cdot \mid \mathcal{F}_T), \quad \mathrm{Var}_T(\cdot) = \mathbb{E}_T(\cdot - \mathbb{E}_T(\cdot))^2 \tag{4}$$

we can write

$$V_T(l) = \mathrm{Var}_T(X_{T+l}). \tag{5}$$

For regression models (1)-(3) we are firstly interested in forecasting $p_{t,j}$ and then in forecasting the derived quantity

$$\mu_t = \sum_{j=1}^{m} j \cdot p_{t,j} = \sum_{j=0}^{m-1} (1 - p_{t,(j)}).$$

For this reason we put

$$\mathcal{F}_T = (Y_1, \ldots, Y_T, Z_1, \ldots, Z_T)$$

and -with the definition of $\mathbb{E}_T$ and Var_T as in (4)- we introduce the $l-$step predictors

$$\hat{p}_{T,j}(l) = \mathbb{E}_T(p_{T+l,j}), \quad \hat{\mu}_T(l) = \mathbb{E}_T(\mu_{T+l}) = \sum_{j=1}^{m} j \cdot \hat{p}_{T,j}(l)$$

for $p_{T+l,j}$ and μ_{T+l}, respectively. From the equation

$$p_{T+l,j} = \mathbb{P}(Y_{T+l} = j \mid \mathcal{F}_{T+l-1}, Z_{T+l})$$

we immediately obtain

$$\hat{p}_{T,j}(l) = \mathbb{P}(Y_{T+l} = j \mid \mathcal{F}_T). \qquad (6)$$

Due to (5) the prediction m.s.e. $V_{T,j}(l)$ and $V_{T,\mu}(l)$ related to $\hat{p}_{T,j}(l)$ and $\hat{\mu}_T(l)$, respectively, are

$$V_{T,j}(l) = \mathrm{Var}_T(p_{T+l,j}), \quad V_{T,\mu}(l) = \mathrm{Var}_T(\mu_{T+l}).$$

For the rest of the paper we assume that the centered covariate process Z_t, $t = 1, 2, \ldots$, follows an $AR(p)$-equation of the familiar form

$$Z_t = R_1 \cdot Z_{t-1} + \ldots + R_p \cdot Z_{t-p} + e_t, \quad t = 1, 2, \ldots \qquad (7)$$

where the $r \times r$ - matrices R_i fulfil the causality criterion, see Brockwell and Davies (1987, sec.11.3), and the e_t are independently and $N(0, \Sigma_e)$-distributed.

4 Monte Carlo Simulation

In a first attempt to solve the forecast problem we will write down a precise expression for the $l-$ step predictor $\hat{p}_{T,j}(l)$ by using a multiple integral, and we will calculate the integral by means of Monte Carlo simulation. Note that this is not a forecast procedure in the classic sense. In the usual time series context, a unique path is generated, representing a best approximation to real forthcoming observations, while here many paths are generated and then averaged.

For the pair of covariates and response at time t let us write $x_t = (z_t, y_t)$, $X_t = (Z_t, Y_t)$, and let us denote by

$$f_T(x_{T+1}, \ldots, x_{T+l-1}, z_{T+l})$$

the conditional density of the regular conditional probability

$$\mathbb{P}(X_{T+1} \in B_1 \times \{y_1\}, \ldots, X_{T+l-1} \in B_{l-1} \times \{y_{l-1}\}, Z_{T+l} \in B_l \mid \mathcal{F}_T)$$

w.r.t. the measure $\nu = (\lambda \times \zeta)^{l-1} \times \lambda$, where λ is (only here) the Lebesgue-measure on $\mathbb{R}^r$ and ζ the counting measure on J. Then

$$\hat{p}_{T,j}(l) = \int \ldots \int f_T(x_{T+1}, \ldots, x_{T+l-1}, z_{T+l}) \cdot p_{T+l,j} \cdot \\ \cdot d\nu(x_{T+1}, \ldots, x_{T+l-1}, z_{T+l}), \qquad (8)$$

where the integration/summation is over $(\mathbb{R}^r \times J)^{l-1} \times \mathbb{R}^r$. The right hand side of (8) can now approximately be calculated by using Monte Carlo methods to generate repeatedly a sequence

$$(X_{T+1}, \ldots, X_{T+l-1}, Z_{T+l}).$$

To achieve this the following recursive algorithm is employed:

1. From $\mathcal{F}_T$ calculate Z_{T+1} according to (7) by drawing an $N(0, \hat{\Sigma}_e)$ random vector, $\hat{\Sigma}_e$ being an estimator of Σ_e
2. From $(\mathcal{F}_T, Z_{T+1})$ calculate

   $$\eta_{T+1} = \alpha + \pi \cdot p_T + \lambda^T \cdot \Lambda(Y_T) - \beta^T \cdot Z_{T+1}$$

   and then $p_{T+1} = h(\eta_{T+1})$
3. Draw Y_{T+1} according to the probability vector p_{T+1}.

Continue with 1.-3. after increasing T to $T+1$. After l steps we arrive at the vectors $\eta^{(1)}_{T+l}$ and $p^{(1)}_{T+l}$. K repetitions of this algorithm lead to vectors

$$\eta^{(k)}_{T+l} \quad \text{and} \quad p^{(k)}_{T+l}, \quad k = 1, \ldots, K,$$

then to the averaged vectors $\bar{\eta}_{T+l}$ and $\bar{p}_{T+l}$, where we have set $\bar{a} = \sum_{k=1}^{K} a^{(k)}/K$, and to $\bar{\mu}_{T+l} = \sum_{j=1}^{m} j \cdot \bar{p}_{T+l,j}$. Now $\bar{\eta}_{T+l}$, $\bar{p}_{T+l}$ and $\bar{\mu}_{T+l}$ are the Monte Carlo solutions for

$$\hat{\eta}_T(l) = \mathbb{E}_T(\eta_{T+l}), \quad \hat{p}_T(l) \quad \text{and} \quad \hat{\mu}_T(l), \quad \text{resp.}$$

In the application below we will further make use of the prediction m.s.e of $\hat{\mu}_T(l)$ estimated by

$$\hat{V}_{T,\mu}(l) = \sum_{k=1}^{K} (\mu^{(k)}_{T+l} - \bar{\mu}_{T+l})^2/(K-1). \tag{9}$$

5 Approximation Procedure

Our second approach comes closer to the spirit of the classic time series forecasting methods. We want to gain a predictor for p_{T+l} by interchanging conditional expectation and response function h. Here we have to make use of the predictors of the covariate process Z_t. For the $AR(p)$-process Z_t an $l-$ step predictor of Z_{T+l} will be denoted by $\hat{Z}_T(l)$, see Brockwell and Davies (1987, sec.11.4). We will now calculate the $l-$ step predictor

$$\hat{p}_T(l) = \mathbb{E}_T(h(\eta_{T+l}))$$

by the approximation formula $\hat{p}_T(l) \approx \check{p}_T(l)$, where

$$\check{p}_T(l) = h(\hat{\eta}_T(l)), \quad \hat{\eta}_T(l) = \mathbb{E}_T(\eta_{T+l}). \tag{10}$$

One successively derives, by using (6) and $\hat{Z}_T(l) = \mathbb{E}_T(Z_{T+l})$,

$$\hat{\eta}_T(1) = \alpha + \pi \cdot p_T + \lambda^T \cdot \Lambda(Y_T) - \beta^T \cdot \hat{Z}_T(1)$$

$$\cdots$$

$$\hat{\eta}_T(l) = \alpha + \pi \cdot \hat{p}_T(l-1) + \sum_j \hat{p}_{T,j}(l-1)\lambda^T\Lambda(j) - \beta^T \cdot \hat{Z}_T(l), \tag{11}$$

where $\hat{p}_T(l-1)$ is approximated by $\check{p}_T(l-1) = h(\hat{\eta}_T(l-1))$ at each step. Thus equation (11) allows a recursive calculation of $\hat{\eta}_T(l)$ and -via (10)- of $\hat{p}_T(l) \approx \check{p}_T(l)$.

We want to give now an estimate $B_T^{(2)}(l)$ of the bias

$$B_T(l) = \hat{p}_T(l) - \check{p}_T(l) = \mathbb{E}_T h(\eta_{T+l}) - h(\hat{\eta}_T(l))$$

produced by approximation formula (10). To this end we assume that h is twice continuously differentiable, and we start with the second-order expansions

$$h_j(\eta_{T+l}) = h_j(\eta_T) + h_j'(\eta_T) \cdot x_{T,l} + (x_{T,l})^T \cdot h_j''(\eta_T) \cdot x_{T,l}/2 + R_T(l),$$

where $x_{T,l} = \eta_{T+l} - \eta_T$, and with the same expansion for $h_j(\hat{\eta}_T(l))$, i.e.

$$h_j(\hat{\eta}_T(l)) = h_j(\eta_T) + h_j'(\eta_T) \cdot \hat{x}_{T,l} + (\hat{x}_{T,l})^T \cdot h_j''(\eta_T) \cdot \hat{x}_{T,l}/2 + \hat{R}_T(l),$$

where $\hat{x}_{T,l} = \hat{\eta}_T(l) - \eta_T$ and $R_T(l)$, $\hat{R}_T(l)$ are remainder terms. This leads to

$$B_{T,j}^{(2)}(l) = \frac{1}{2}\,\mathbb{E}_T[(\eta_{T+l})^T \cdot h_j''(\eta_T) \cdot \eta_{T+l}] - \frac{1}{2}\,\hat{\eta}_T(l)^T \cdot h_j''(\eta_T) \cdot \hat{\eta}_T(l). \tag{12}$$

In the special case of the cumulative regression model (3) we can simplify formula (12). In fact, the second-order approximation for the bias $B_{T,(j)}(l) = \hat{p}_{T,(j)}(l) - \check{p}_{T,(j)}(l)$ amounts to

$$B_{T,(j)}^{(2)}(l) = \frac{1}{2}\,F''(\eta_{T,(j)}) \cdot \mathrm{Var}_T(\eta_{T+l,(j)}). \tag{13}$$

Taking as an example the logistic distribution function $F(s) = 1/(1+e^{-s})$, then $B_{T,(j)}^{(2)}(l)$ turns out to be positive/negative, if $\eta_{T,(j)}$ is negative/positive.

Formulas (12) and (13) can be applied to correct the bias of the approximation (10), if estimators for

$$\mathbb{E}_T[(\eta_{T+l})^T \cdot h_j''(\eta_T) \cdot \eta_{T+l}] \quad \text{and} \quad \mathrm{Var}_T(\eta_{T+l,(j)}), \quad \text{resp.,}$$

are available. To establish explicit expressions for them seems difficult. Numerically, they can be gained as by-products of the Monte-Carlo method of sec.4. The approximation method of this section, however, was introduced to get forecasts without the computer intensive method of Monte-Carlo simulation.

6 Application

6.1 Forest Damage Data

The cumulative logistic regression model (3) is now applied to three longitudinal data sets on damages in beech, oak and pine trees, respectively. These data were gathered by Dr.A.Göttlein, University of Bayreuth, during the last 12 years in a forest district of Spessart (Bavaria). The damage Y_t in the year t was measured on an ordinal scale consisting of $m = 8$ categories of needles/leaves lost. The longitudinal structure of the data is determined by the observation period of 12 years (1983 - 1994) and by N sites (N = 80 beech sites, N = 25 oak sites, N = 14 pine sites). For each site and each year a vector Z_t of $r = 20$ covariates were recorded concerning the trees (age, canopy, stand), the site (gradient, height, exposition), the climate and the soil (type, moisture, pH-values), see Göttlein and Pruscha (1992) and (1995) for more details. The parameter of the model were estimated from the longitudinal data by the m.l. method for each species separately. Concerning the function Λ we made the special choice $\Lambda(W_t) = Y_t$, see special case 2 in sec.2. We further put $\pi = 0$. The covariate process Z_t is assumed to be driven by an AR(1)-equation.

6.2 Forecasting μ_t

Fixing the outcomes of the years 1983 - 1992 as known, we try to forecast the values of the mean damage category

$$\mu_t = \sum_{j=1}^{m} j p_{t,j}$$

for the years 1993 - 1998, letting the years 1993 and 1994 -for which we have observations- as control. That is, we put $T = 10$ and we are interested in the $l-$ step predictors $\hat{\mu}_T(l), l = 1, 2, \ldots, 6$. The calculations of the forecasts, leading to the Fig.1, are performed separately for each of the three species, beech, oak and pine trees, and for each site $i = 1, 2, \ldots, N$, followed by averaging over the N sites of the species.

First, the Monte Carlo method (MOCA) of sec.4 is applied, with $K = 200$ repetitions to calculate $\bar{\mu}_{T+l}$ as Monte Carlo solution for $\hat{\mu}_T(l)$ and the corresponding m.s.e. $\hat{V}_{T,\mu}(l)$ according to (9). A 95 per cent confidence interval for the averaged μ_{T+l} is established by the confidence limits

$$\hat{\mu}_T(l) \pm \sqrt{\hat{V}_{T,\mu}(l)} \cdot 1.960/\sqrt{N}$$

holding approximately for the individual years $l = 1, 2, \ldots, 6$.

Secondly, the approximation method (APPR) of sec.5 is employed. For all three species the forecasts $\hat{\mu}_T(l)$ produced by the MOCA and by the APPR method run very similar over the 6 years 1993 to 1998, with the APPR curve below the MOCA curve. The upward trend of the pine curve at the end of the observation period is continued in a strongly damped form.

To compare the forecast solutions with the observation data of the years 1986 to 1994, we include plots for $\bar{Y}_t$ and $\bar{\mu}_t$, where Y_t is the observed category, $\mu_t = \sum j p_{t,j}$ is the predicted mean value at year t (as predicted on the basis of the estimated cumulative regression model) and the bar means averaging over the N sites of the tree species. Note the lag-effect which is produced by the term Y_{t-1} in the regression model, especially in the oak data: a zig zag run of the $\bar{Y}_t$ values becomes apparant in the run of the $\bar{\mu}_t$ values with a lag of one year.

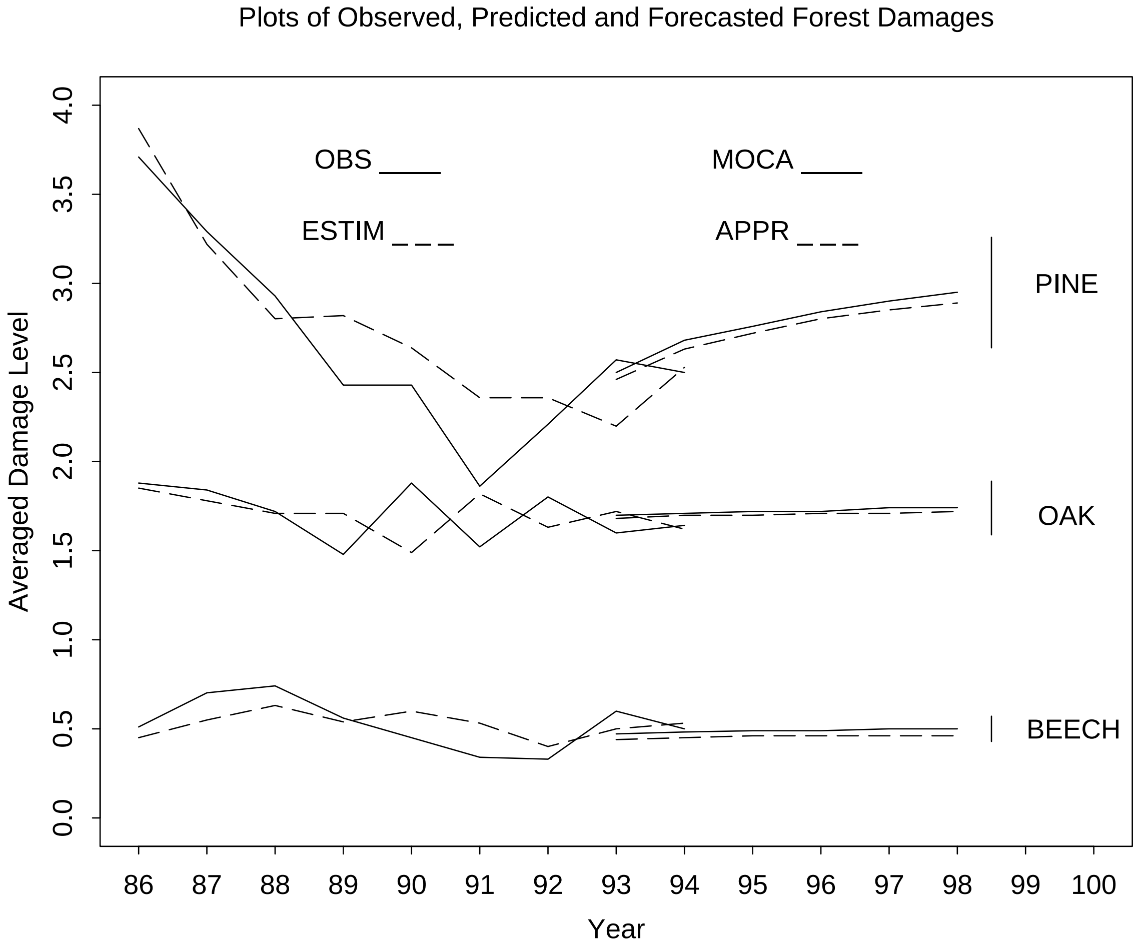

Figure 1: Observed, predicted and forecasted forest damages

6.3 Some Remarks to Fig.1

Observed, predicted and forecasted damages are shown in Fig.1 separately for three tree species, beech, oak and pine trees, respectively.

Over the years 1986 to 1994 are plotted
- the OBServed damage category Y_t and the predicted mean category μ_t, as predicted from the ESTIMated cumulative regression model.

Over the years 1993-1998 are plotted
- the forecasted mean category $\hat{\mu}_T(l)$,$l = 1, \ldots, 6$, as produced by the MOCA and the APPR methods of sec.4 and sec.5, resp.

All values are averaged values over the N sites of the tree species. Further, at the end of the forecast curve, a 95 per cent confidence interval for μ_{T+l} is indicated by an vertical bar, holding approximately for the last forecast step, i.e for $l = 6$.

Acknowledgments: This work is partially supported by the DFG (SFB 386).

References

Brockwell,P.J. and Davis,R.A. (1987). *Time Series: Theory and Methods.* Springer, N.Y.

Diggle,P.J.,Liang,K.-Y. and Zeger,S.L. (1994). *Analysis of Longitudinal Data.* Claredon Press, Oxford.

Fahrmeir,L. and Kaufmann,H. (1987). Regression models for non stationary categorical time series. *Journal of Time Series Analysis*, **8**, 147-160.

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

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

Göttlein,A. and Pruscha,H. (1995). Der Einfluss von Topographie, Standort, Klima und Bestand auf die Entwicklung des Waldzustandes im Bereich Rothenbuch (submitted).

McCullagh,P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society*, Series B, **42**, 109-142.

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

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

Address of Author
Helmut Pruscha
Mathematisches Institut der Universität München
Theresienstr. 39
D 80333 München
e-mail: pruscha@rz.mathematik.uni-muenchen.de