Predictive Model Assessment for Count Data

Claudia Czado,[1,*] **Tilmann Gneiting**[2,†] **and Leonhard Held**[3,‡]

[1]Zentrum Mathematik, Technische Universität München, Boltzmannstr. 3, 85747 Garching, Germany

[2]Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98195-4322, U.S.A.

[3]Institut für Sozial- und Präventivmedizin, Abteilung Biostatistik, Universität Zürich, Hirschengraben 84, 8001 Zürich, Switzerland

September 5, 2007

SUMMARY. We discuss tools for the evaluation of probabilistic forecasts and the critique of statistical models for ordered discrete data. Our proposals include a non-randomized version of the probability integral transform, marginal calibration diagrams and proper scoring rules, such as the predictive deviance. In case studies, we critique count regression models for patent data, and assess the predictive performance of Bayesian age-period-cohort models for larynx cancer counts in Germany.

KEY WORDS: Calibration; Forecast verification; Model diagnostics; Predictive deviance; Probability integral transform; Proper scoring rule; Ranked probability score.

1. Introduction

One of the major purposes of statistical analysis is to make predictions, and to provide suitable measures of the uncertainty associated with them. Hence, forecasts ought to be

[*] *email:* cczado@mathematik.tu-muenchen.de
[†] *email:* tilmann@stat.washington.edu
[‡] *email:* leonhard.held@ifspm.unizh.ch

probabilistic in nature, taking the form of probability distributions over future quantities and events (Dawid, 1984).

Here, we consider the evaluation of probabilistic forecasts, or predictive distributions, for count data, as they occur in a wide range of epidemiological, ecological, environmental, climatological, demographic and economic applications (Christensen and Waagepetersen, 2002; Gotway and Wolfinger, 2003; McCabe and Martin, 2005; Elsner and Jagger, 2006; Frühwirth-Schnatter and Wagner, 2006; Nelson and Leroux, 2006). Our focus is on the low count situation in which continuum approximations fail; however, our results apply to high counts and rates as well, as they occur routinely in epidemiological projections (Knorr-Held and Rainer, 2001; Clements, Armstrong and Moolgavkar, 2005). To this date, statistical methods for the assessment of predictive performance have been studied primarily from biomedical, meteorological and economic perspectives (Pepe, 2003; Jolliffe and Stephenson, 2003; Clements, 2005), focusing on predictions of dichotomous events or real-valued continuous variables. Here, we consider the hybrid case of count data, in which methods developed for either type of situation continue to be relevant but require technical adaption.

Gneiting, Balabdaoui and Raftery (2007) contend that the goal of probabilistic forecasting is *to maximize the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between the probabilistic forecasts and the observations, and is a joint property of the predictive distributions and the events or values that materialize. Sharpness refers to the concentration of the predictive distributions, and is a property of the forecasts only.

In Section 2 we introduce tools for calibration and sharpness checks, among them a non-randomized version of the probability integral transform (PIT) that is tailored to count data, and the marginal calibration diagram. Section 3 discusses the use of scoring rules as omnibus performance measures. We stress the importance of propriety (Gneiting and Raftery, 2007), note examples, relate to classical measures of predictive performance, and

identify the predictive deviance as a variant of the proper logarithmic score. Section 4 turns to a cross-validation study, in which we apply these tools to critique count regression models for pharmaceutical and biomedical patents. The epidemiological case study in Section 5 evaluates the predictive performance of Bayesian age-period-cohort models for larynx cancer counts in Germany. We consider a recent suggestion by Baker and Bray (2005), according to which the inclusion of all age groups in the analysis, as opposed to older age groups only, leads to improved predictions. The paper closes with a discussion in Section 6.

For count data, a probabilistic forecast is a predictive probability distribution, P, on the set of the nonnegative integers. We denote its probability mass function by $(p_k)_{k=0}^{\infty}$ and the respective cumulative distribution function (CDF) by $(P_k)_{k=0}^{\infty}$. Generalizations of our proposed methodology to probabilistic forecasts for any type of ordered discrete data, as opposed to count data, are straightforward and given in an appendix. The tools are simple yet powerful, and they apply generally to problems of forecast evaluation, model criticism and model diagnosis.

2. Calibration and sharpness

As noted, probabilistic forecasts strive to maximize the sharpness of the predictive distributions subject to calibration. Calibration refers to the statistical consistency between the probabilistic forecasts and the observations, and its assessment requires frequentist thinking (Rubin, 1984). Gneiting et al. (2007) distinguish various modes of calibration and propose tools for the assessment of calibration and sharpness for probabilistic forecasts of continuous variables. Here, we adapt their proposals to the case of count data.

2.1 *Probability integral transform*

Dawid (1984) proposed the use of the *probability integral transform* (PIT) for calibration checks. This is simply the value that the predictive cumulative distribution function attains at the value that materializes. If the observation is drawn from the predictive distribution

— an ideal and desirable situation — and the predictive distribution is continuous, the PIT has a standard uniform distribution. Calibration then is checked empirically, by plotting the empirical CDF of a set of PIT values and comparing to the identity function, or by plotting the histogram of the PIT values and checking for uniformity (Diebold, Gunther and Tay, 1998; Gneiting et al., 2007). The PIT histogram is typically used informally as a diagnostic tool; formal tests can also be employed though they require care in their interpretation (Hamill, 2001; Jolliffe, 2007). Deviations from uniformity hint at reasons for forecast failures and model deficiencies. U-shaped histograms indicate underdispersed predictive distributions, hump or inverse-U shaped histograms point at overdispersion, and skewed histograms occur when central tendencies are biased.

In the case of count data, the predictive distribution is discrete. Here, the PIT is no longer uniform under the hypothesis of an ideal forecast, for which the observed count is a random draw from the predictive distribution. To remedy this, several authors have suggested a *randomized* PIT. Specifically, if P is the predictive distribution, $x \sim P$ is a random count and v is standard uniform and independent of x, then

$$u = P_{x-1} + v(P_x - P_{x-1}), \quad x \geq 1, \tag{1}$$

$$u = v P_0, \quad x = 0, \tag{2}$$

is standard uniform (Smith, 1985, pp. 286–287; Frühwirth-Schnatter, 1996, p. 297; Liesenfeld, Nolte and Pohlmeier, 2006, pp. 819–820). For time series data one typically considers one-step (or k-step) ahead predictions, based on a time series model fitted on past and current data, and checks for the independence of the randomized PIT, in addition to checks for uniformity.

Here we propose a *non-randomized* yet uniform version of the PIT histogram. To this end, we replace the randomized PIT value in (1) and (2) by its conditional cumulative

distribution function given the observed count x, that is, by

$$F(u) = \begin{cases} 0, & u \leq P_{x-1}, \\ (u - P_{x-1})/(P_x - P_{x-1}), & P_{x-1} \leq u \leq P_x, \\ 1, & u \geq P_x, \end{cases} \tag{3}$$

if $x \geq 1$, and

$$F(u) = \begin{cases} u/P_0, & u \leq P_0, \\ 1, & u \geq P_0, \end{cases} \tag{4}$$

if $x = 0$, similarly to the *discrete grade transformation* in relative distribution methodologies for the social sciences (Handcock and Morris, 1999, p. 180). Calibration can then be assessed by aggregating over the predictions and comparing the mean PIT,

$$\bar{F}(u) = \frac{1}{n} \sum_{i=1}^{n} F^{(i)}(u), \quad 0 \leq u \leq 1, \tag{5}$$

where $F^{(i)}$ is based on the predictive distribution $P^{(i)}$ and the observed count $x^{(i)}$, to the distribution function of the standard uniform law, that is, the identity function.

We prefer to perform this comparison by plotting a non-randomized PIT histogram, which can be interpreted diagnostically in the ways described above. Specifically, we pick the number of bins, J, compute $f_j = \bar{F}(\frac{j}{J}) - \bar{F}(\frac{j-1}{J})$ for equally spaced bins $j = 1, \ldots, J$, plot a histogram with height f_j for bin j, and check for uniformity. Under the hypothesis of calibration, that is, if $x^{(i)} \sim P^{(i)}$ for all forecast cases $i = 1, \ldots, n$, it is straightforward to verify that $\bar{F}(u)$ has expectation u, so that we expect uniformity. Typically, $J = 10$ or $J = 20$ are good choices for the number of bins in the PIT histogram.

It is important to note that uniformity of the PIT histogram is a strong requirement, being equivalent to all prediction intervals showing nominal coverage. In particular, the sometimes practice of tabulating empirical coverage for selected prediction intervals can be

interpreted as a special case of the use of the PIT histogram. Consider the prediction interval with lower and upper probability limits α and β. The nominal coverage is $\beta - \alpha$. Using randomization, we find the empirical coverage as the frequency of randomized PIT values that fall into the interval $[\alpha, \beta]$. We prefer to use the non-randomized approach, in which the empirical coverage is computed as the difference $\bar{F}(\beta) - \bar{F}(\alpha)$ of the mean non-randomized PIT (5) at the probability limits α and β. Of course, this difference is the expected value of the empirical coverage computed on the basis of the randomized PIT, when the expectation is taken with respect to the randomization.

2.2 *Marginal calibration diagram*

We now consider what Gneiting et al. (2007) refer to as *marginal calibration*. The idea is straightforward: If each observed count is a random draw from the respective probabilistic forecast, and if we aggregate over the individual predictive distributions, we expect the resulting mixture distribution and the histogram of the observed counts to be statistically compatible. A *marginal calibration diagram* illustrates the predicted probability mass for specific x values or intervals $(x_a, x_b]$, when averaged over the predictive distributions, along with the respective empirical frequency of count observations. Major discrepancies hint at reasons for forecast failures and model deficiencies. An example of this type of diagnostic tool is shown in Figure 4 below.

2.3 *Sharpness*

Sharpness refers to the concentration of the predictive distributions. In the context of prediction intervals, this can be rephrased simply: The shorter the intervals, the sharper, and the sharper the better, subject to calibration. Prediction intervals for continuous predictive distributions are uniquely defined, and Gneiting et al. (2007) suggest to tabulate their average width, or to plot *sharpness diagrams* as a diagnostic tool. Sharpness continues to be critical for count data; however, we have found these tools to be less useful for discrete

predictive distributions, for the ambiguities in specifying prediction intervals. Our preferred way of addressing sharpness is indirectly, via proper scoring rules; see below.

2.4 *Simulation study*

We consider the negative binomial distribution $\mathrm{NB}(\lambda, a)$ with mean $\lambda \geq 0$ and dispersion parameter $a \geq 0$, hence variance $\lambda(1 + a\lambda)$. If $a = 0$, this is simply the Poisson distribution $\mathrm{P}(\lambda)$. We sample 200 counts from an $\mathrm{NB}(5, \frac{1}{2})$ distribution, and consider probabilistic forecasters whose predictive distribution is $\mathrm{NB}(5, 0) = \mathrm{P}(5)$, $\mathrm{NB}(5, \frac{1}{2})$ and $\mathrm{NB}(5, 1)$. Figure 1 shows non-randomized PIT histograms with $J = 10$ equally spaced bins for these three cases. The PIT histograms are U-shaped, uniform and inversely U-shaped, indicating underdispersed, well-calibrated and overdispersed predictive distributions, respectively.

3. Scoring rules

Scoring rules provide summary measures in the evaluation of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes. We take scoring rules to be negatively oriented penalties that a forecaster wishes to minimize. Specifically, if the forecaster quotes the predictive distribution P and the count x materializes, the penalty is $s(P, x)$. We write $s(P, Q)$ for the expected value of $s(P, \cdot\,)$ under Q. In practice, scores are reported as averages over suitable sets of probabilistic forecasts, and we use upper case to denote a mean score; say

$$S = \frac{1}{n} \sum_{i=1}^{n} s(P^{(i)}, x^{(i)}),$$

where $P^{(i)}$ and $x^{(i)}$ refer to the ith predictive distribution and the ith observed count, respectively. In particular, Table 1 below shows mean scores.

3.1 *Propriety*

Suppose, then, that the forecaster's best judgement is the predictive distribution Q. The forecaster has no incentive to predict any $P \neq Q$, and is encouraged to quote her true belief, $P = Q$, if

$$s(Q, Q) \leq s(P, Q) \tag{6}$$

with equality if and only if $P = Q$. A scoring rule with this property is said to be *strictly proper*. If $s(Q, Q) \leq s(P, Q)$ for all P and Q, the scoring rule is said to be *proper*. Propriety is an essential property of a scoring rule that encourages honest and coherent predictions (Bröcker and Smith, 2007; Gneiting and Raftery, 2007). Strict propriety ensures that both calibration and sharpness are being addressed.

A scoring rule s for count data is *regular* if $s(P, x)$ is finite, except possibly that $s(P, x) = \infty$ if $p_x = 0$. Let $\mathcal{P}$ denotes the class of probability measures on the set of the nonnegative integers. The Savage representation theorem (Savage, 1971; Gneiting and Raftery, 2007) states that a regular scoring rule S for count data is proper if and only if

$$s(P, x) = h(P) - \sum_{k=0}^{\infty} h'_k(P) p_k + h'_x(P)$$

where $h : \mathcal{P} \to \mathbb{R}$ is a concave function and $h'(P)$ is a subgradient of h at the point P, for all $P \in \mathcal{P}$. The statement holds with proper replaced by strictly proper, and concave replaced by strictly concave.

Phrased slightly differently, a regular scoring rule s is proper if and only if the expected score function $h(P) = s(P, P)$ is concave on $\mathcal{P}$, and the sequence $(s(P, k))_{k=0}^{\infty}$ is a subgradient of h at the point P, for all $P \in \mathcal{P}$. The expected score function allows for an interpretation as a generalized entropy function (Gneiting and Raftery, 2007).

3.2 *Examples of proper scoring rules*

The *logarithmic score* is defined as

$$\text{logs}(P, x) = -\log p_x. \tag{7}$$

This is the only proper scoring rule that depends on the predictive distribution P only through the probability mass p_x at the observed count (Good, 1952). The associated expected score or generalized entropy function is the classical Shannon entropy.

There is a close relationship between the logarithmic score and the *predictive deviance*, defined as

$$\text{dev}(P, x) = -2\log p_x + 2\log f_x,$$

where f_x is "some fully specified standardizing term that is a function of the data alone" (Spiegelhalter, Best, Carlin and van der Linde, 2002, p. 587). If the predictive distribution is a member of a one-parameter exponential family, such as the binomial or Poisson, the standardizing term is routinely taken to be the saturated deviance (McCullagh and Nelder, 1989, pp. 33-34; Knorr-Held and Rainer, 2001, p. 114; Spiegelhalter et al., 2002, p. 606; Clements et al., 2005, p. 581). However, when the predictive distributions come from possibly distinct parametric or non-parametric families, it is vital that the standardizing terms in the deviance are common (Spiegelhalter et al., 2002, p. 634). We contend that the choice is rather arbitrary and propose for simplicity that the standardizing term is taken to be zero (Gschlößl and Czado, 200x, sect. 6.3), which corresponds to the use of the logarithmic score.

Let $\|p\|^2 = \sum_{k=0}^{\infty} p_k^2$, which can frequently be computed analytically, as shown in Appendix A for the Poisson and negative binomial distributions. The *quadratic score* or *Brier score* and the *spherical score* are then defined as

$$\text{qs}(P, x) = -2p_x + \|p\|^2 \tag{8}$$

and

$$\mathrm{sphs}(P, x) = -\frac{p_x}{\|p\|}, \tag{9}$$

respectively. Wecker (1989) proposed the use of the quadratic score in the assessment of time series predictions of counts.

The *ranked probability score* (Epstein, 1969) was originally introduced for ranked categorical data. It is easily adapted to count data, by defining

$$\mathrm{rps}(P, x) = \sum_{k=0}^{\infty} \left\{P_k - \mathbf{1}(x \le k)\right\}^2. \tag{10}$$

Equation (14) in Gneiting and Raftery (2007) implies an alternative representation expressed in terms of expectations, which we now assume to be finite, namely

$$\mathrm{rps}(P, x) = E_P|X - x| - \frac{1}{2} E_P|X - X'|,$$

where X and X' are independent copies of a random variable with distribution P. The ranked probability score generalizes the absolute error, to which it reduces if P is a point forecast. Hence, it provides a direct way of comparing point forecasts and predictive distributions. The scores introduced in this section are strictly proper, except that the ranked probability score requires Q to have finite first moment for strict inequality in (6) to hold.

There is no automatic choice of a proper scoring rule to be used in any given situation, unless there is a unique and clearly defined underlying decision problem. However, in many types of situations probabilistic forecasts have multiple simultaneous uses, and it may be appropriate to use a variety of diagnostic tools and scores, to take advantage of their differing emphasis and strengths. For instance, there is a distinct difference between the ranked probability score and the other scores discussed in this section, in that the former blows up score differentials between competing forecasters in case predicted and/or observed counts

are unusually high. Hence, a few, or even a single high count case can dominate and obscure differences in the mean score. We will see an example of this in Section 5 below. This type of behavior might be desirable, if the high count cases are the crucial ones, or might be undesirable, depending on the application at hand.

3.3 *Classical measures of predictive performance*

We now discuss traditional summary measures of predictive performance. For simplicity, we assume hereinafter that all moments considered are finite. Suppose first that $\mu \in \mathbb{R}$ is point forecast and the count x materializes. Typically, one uses the absolute error, $\mathrm{ae}(\mu, x) = |x - \mu|$, or the squared error, $\mathrm{se}(\mu, x) = (x - \mu)^2$, as a measure of predictive performance, averaging, again, over suitable sets of forecasts, to obtain the mean absolute error and mean squared error, respectively. Of course, these measures apply to probabilistic forecasts as well. For example, we can define the *squared error score*,

$$\mathrm{ses}(P, x) = (x - \mu_P)^2 , \tag{11}$$

where μ_P is the mean of the predictive distribution P. Viewed as a scoring rule for probabilistic forecasts, this score is proper, but not strictly proper.

We now turn to studentized errors. It has frequently been argued that the *squared Pearson residual* or *normalized squared error score*,

$$\mathrm{nses}(P, x) = \left(\frac{x - \mu_P}{\sigma_P} \right)^2 , \tag{12}$$

where μ_P and σ_P^2 denote the mean and the variance of P, ought be approximately one when averaged over the predictions (Carroll and Cressie, 1997, p. 52; Liesenfeld et al., 2006, pp. 811, 818). Gotway and Wolfinger (2003, p. 1423) call the mean normalized squared error score the average *empirical-to-model variability ratio*, arguing also that it should be close to

one. One way of justifying this is by noting that the function

$$f(\mu_P, \sigma_P^2) = (\text{nses}(P, Q) - 1)^2$$

has a minimum at $\mu_P = \mu_Q$ and $\sigma_P^2 = \sigma_Q^2$. The normalized squared error score and its expectation $\text{nses}(P, Q)$ when P is the predictive distribution and $x \sim Q$ realizes, depend on P only through the first two moments, so the function f is well-defined. Still, we follow Frühwirth-Schnatter (1996, p. 297) in arguing that the PIT histogram is a more informative and more robust tool for unmasking dispersion errors.

The scores in this section depend on the predictive distribution P only through the first two moments. Dawid and Sebastiani (1999) provide a comprehensive study of proper scoring rules for which this property holds. A particularly appealing example is the scoring rule

$$\text{dss}(P, x) = \left(\frac{x - \mu_P}{\sigma_P}\right)^2 + 2 \log \sigma_P, \tag{13}$$

to which we refer as the *Dawid-Sebastiani score*. It was proposed by Gneiting and Raftery (2007) as a proper alternative to the improper *predictive model choice criterion* of Gelfand and Ghosh (1998).

3.4 *Simulation study*

We now return to the simulation study in Section 2.4. We sample 200 counts from an $\text{NB}(5, \frac{1}{2})$ distribution, and suppose that the predictive distribution is $\text{NB}(5, 0) = \text{P}(5)$, $\text{NB}(5, \frac{1}{2})$ and $\text{NB}(5, 1)$, respectively. For each probabilistic forecast and each of six scoring rules (logarithmic, quadratic, spherical, ranked probability, Dawid-Sebastiani and normalized squared error scores), Figure 2 summarizes the scores for the 200 individual forecasts. The first five scoring rules, which are proper, show the lowest scores for the $\text{NB}(5, \frac{1}{2})$ forecast, which is correctly identified as superior. A similar statement holds for the normalized squared

error score, which is closest to its target value one for this forecast. The respective mean normalized squared error scores are 3.84, 1.10 and 0.64, thereby supporting the dispersion assessments that the PIT histograms in Figure 1 make more powerfully. Of course, the predictive mean, and therefore the mean squared error, is the same for all three forecasts.

4. Case study: Model critique for count regression

Count data often show substantial extra variation or overdispersion relative to a Poisson regression model (Dean and Lawless, 1989; Winkelmann, 2005). Various alternatives have been suggested to accommodate this, such as negative binomial and mixed Poisson models (Lawless, 1987). In this section, we investigate whether the non-randomized PIT histogram, the marginal calibration diagram and proper scoring rules are effective tools for model criticism (O'Hagan, 2003) in this context. We adopt a leave-one-out cross-validation approach, in which the prediction for each observation is based on a count regression model fitted on the remaining data only.

We study the relationship of the number of patent applications to research and development (R&D) spending and sales using data from 1976 for 70 pharmaceutical and biomedical companies (Hall, Cummins, Laderman and Mundy, 1988). This data set was also studied by Wang, Cockburn and Puterman (1998), who used a mixed Poisson regression approach to address the overdispersion that is commonly observed in patent counts (Hausman, Hall and Griliches, 1984; Czado, Erhardt and Min, 2006). Here we take a simpler approach and compare Poisson regression to negative binomial regression, using the specification

$$\log \lambda = \beta_0 + \beta_1 \frac{\text{R\&D}}{\text{sales}} + \beta_2 \left(\text{R\&D}\right)^{1/5}$$

for the predictive mean λ. Figure 3 shows non-randomized PIT histograms based on the leave-one-out predictive distributions, using Poisson and negative binomial count regression

models fitted with the R functions GLM() and GLM.NB() (Venables and Ripley, 1997, Section 7.4). The PIT histogram for the Poisson case indicates under-dispersion of the Poisson regression model. The histogram for the negative binomial case does not show any lack of model fit. Figure 4 shows a marginal calibration diagram, as introduced in Section 2.2. We see that the predicted frequencies for the negative binomial regression are closer to the observed ones, when compared to the Poisson regression.

Figure 5 uses boxplots to display scores for the two competing methods. The five proper scores all prefer the negative binomial over the Poisson regression model. Superficially, the same appears to be true for the normalized squared error score. However, the mean normalized squared error can blow up under outliers, which is the case here. Outliers beyond the range of the boxplot lead to mean normalized squared error scores of 12.5 and 91.6 for the Poisson and negative binomial case, respectively. This is in stark contrast to the PIT histograms in Figure 3 and the boxplots in Figure 5, and provides an example of potentially misleading inferences that non-robust performance measures may suggest.

In conclusion, our diagnostic tools all point at the superiority of the negative binomial regression model. The non-randomized PIT histogram and the marginal calibration diagram furthermore allow us to diagnose the reason for this critique, in that the Poisson model is strongly underdispersed.

5. Case study: Predicting cancer incidence

Bayesian age-period-cohort models are used increasingly to project cancer incidence and mortality rates. Data from younger age groups (typically age < 30 years) for which rates are low are often excluded from the analysis. However, a recent empirical comparison (Baker and Bray, 2005) based on data from Hungary suggests that age-specific predictions based on full data are more accurate. A natural question arises here in how to quantify the quality of the predictive distributions.

Baker and Bray (2005, p. 799) predict mortality rates, using what they call the *sum of squared standardized residuals* to assess the quality of the forecasts. From personal communication with the authors, the standardization is not based on the predictive variance, so the aforementioned residuals are not the squared Pearson residuals in (12). Instead, Baker and Bray (2005) use the traditional standard error of a rate estimate for standardization and argue, in discussing their Table 1, that smaller values of this quantity correspond to more accurate predictions. Clements et al. (2006) question the assessment in Baker and Bray (2005) and suggest the use of the predictive deviance, originally proposed by Knorr-Held and Rainer (2001). They argue that "the Bayesian age-period-cohort model suffers from very wide credible intervals", but do not relate the width of the intervals to properties of calibration, and do not specifically recommend the use of proper scoring rules.

In this section, we use scoring rules to investigate whether the conclusion drawn by Baker and Bray (2005) applies to larynx cancer data from Germany, 1952–2002. Our assessment is based on counts rather than rates. We fit four different predictive models depending on whether or not data from age groups < 30 years have been included in the analysis, and whether or not the model allows for overdispersion, as shown in Table 1.

Let n_{ij} be the number of persons at risk in age group i and year j. We assume that the respective number of deaths X_{ij} is binomially distributed with parameters n_{ij} and π_{ij}. A Poisson model would be a nearly identical choice. Following Besag, Green, Higdon and Mengersen (1995) and Knorr-Held and Rainer (2001), we decompose the logarithmic odds $\eta_{ij} = \log\{\pi_{ij}/(1-\pi_{ij})\}$ additively, into an overall level μ, age effects θ_i, period effects φ_j and cohort effects ψ_k, namely

$$\eta_{ij} = \mu + \theta_i + \varphi_j + \psi_k.$$

Note that there is a problem in defining cohorts because age groups (in 5-year steps) and periods (in 1-year steps) are not on the same grid. We follow Fienberg and Mason (1979)

and Knorr-Held and Rainer (2001) and use the cohort index $k = 5 \cdot (I - i) + j$, where I is the number of age groups.

Typically, parametric time trends for age, period and cohort effects are too restrictive. On the other hand, period and cohort effects as factors cause instability of the maximum likelihood estimates, possibly resulting in a saw-tooth pattern, as noted first by Holford (1983). Also, it is not obvious how to use maximum likelihood estimates for prediction. Osmond (1985) suggested to compute unknown period and cohort effects for future periods by linear regression applied to a subjectively chosen number of the most recent estimates on each scale. One criticism of this method is that it is arbitrary in the choice of the number of past values to use. In a recent comparative study, Bray (2002, pp. 161–162) concludes that "empirical projections based on the method of Osmond (1985) are poor."

Here we use non-parametric smoothing priors within a hierarchical Bayesian framework, for which model-based extrapolation of period and cohort effects for future periods is straightforward (Besag et al., 1995). This choice has the additional advantage that adjustments for overdispersion are easy to make. Inference and prediction based on Markov chain Monte Carlo techniques is done as described in Knorr-Held and Rainer (2001).

Observed and fitted/predicted numbers of deaths from larynx cancer per 100,000 males are displayed in Figure 6. We show posterior means and 90% pointwise prediction intervals based on Model 4, which does not adjust for overdispersion and includes data only from age group 30–34 onward. For better comparison, incidence rates per 100,000 men are shown. The subsequent analysis is based on counts.

To assess the predictive performance of the different models, we predict mortality counts for the five years 1998–2002. For all different models, we consider predictions in the 12 age groups with age > 30 years. Table 1 shows mean scores, averaged over all $12 \cdot 5 = 60$ projections. Interestingly, the scores do not agree. One set of scores (logarithmic, quadratic, spherical, Dawid-Sebastiani and normalized squared error scores) points to Model 1 as the

best, which includes data from the very young age groups and adjusts for overdispersion. The other set of scores (ranked probability and squared error scores) prefers Model 4.

The disagreement can be explained as follows. The scores in the first set are roughly independent of the size of the counts in the different age groups. This is most obvious for the normalized squared error score, but approximately also true for the other scores. In contrast, the ranked probability and squared error scores are highly dependent on the size of the counts. Hence the results in the mid age groups, where the counts are highest and Model 4 is more competitive, dominate the mean score.

These results might support Baker and Bray's (2005) contention that age-specific predictions based on full data yield sharper predictive distributions, and more accurate point forecasts for the younger age groups that benefit from the strong cohort effect that is present here, particularly for the younger birth cohorts. Of course, these findings are tentative, being based on 60 dependent predictions only, and further experiments are called for.

6. Discussion

We have introduced a toolbox for the assessment of the predictive performance of probabilistic forecasts for count data, which includes a non-randomized probability integral transform (PIT) histogram, the marginal calibration diagram and proper scoring rules. Simplicity, generality and interpretability are attractive features of these tools; they apply both in parametric and non-parametric settings and do not require models to be nested, nor be related in any way. Typically, they are used diagnostically, to identify model deficiencies and facilitate model comparison and model selection. Formal inference is often feasible (Clements, 2005; Jolliffe, 2007), but may not be the goal.

The toolbox applies to two apparently distinct, yet closely related tasks. One is the evaluation of probabilistic forecasts that take the form of predictive distributions for future counts. Here, the PIT histogram and the marginal calibration diagram are employed di-

agnostically, and proper scoring rules allow us to rank competing forecasters. The other task is the critique of statistical models (O'Hagan, 2003); frequently, models can be fitted in cross-validation mode, and can be assessed based on the quality of the ensuing probabilistic forecasts. We have demonstrated the use of these tools in case studies in both types of situations. It is our belief that they can provide similar guidance in a very wide range of applied statistical problems for ordered discrete data.

Acknowledgements

Claudia Czado was supported by the German Research Foundation. Tilmann Gneiting acknowledges support by the National Science Foundation under Award DMS-0706745 and by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

References

Abramowitz, M. and Stegun, I. A. (1970). *Handbook of Mathematical Functions*. Dover, New York.

Baker, A. and Bray, I. (2005). Bayesian projections: What are the effects of excluding data from younger age groups? *American Journal of Epidemiology* **162**, 798–805.

Besag, J. E., Green, P. J., Higdon, D. M. and Mengersen, K. L. (1995). Bayesian computation and stochastic systems (with discussion). *Statistical Science* **10**, 3–66.

Bray, I. (2002). Application of Markov chain Monte Carlo methods to projecting cancer incidence and mortality. *Applied Statistics* **51**, 151–164.

Bröcker, J. and Smith, L. A. (2007). Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting* **22**, 382–388.

Carroll, S. S. and Cressie, N. (1997). Spatial modeling of snow water equivalent using covariances estimated from spatial and geomorphic attributes. *Journal of Hydrology* **190**, 42–59.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics* **58**, 280–286.

Clements, M. P. (2005). *Evaluating Econometric Forecasts of Economic and Financial Variables*. Palgrave Macmillan, Basingstroke, Hampshire, UK.

Clements, M. S., Armstrong, B. K. and Moolgavkar, S. H. (2005). Lung cancer rate predictions using generalized additive models. *Biostatistics* **6**, 576–589.

Clements, M. S., Hakulinen, T. and Moolgavkar, S. H. (2006). Re: "Bayesian projections: What are the effects of excluding data from younger age groups?". *American Journal of Epidemiology* **164**, 292–293.

Czado, C., Erhardt, V. and Min, A. (2006). Zero-inflated generalized Poisson models with regression effects on the mean, dispersion and zero-inflation level applied to patent outsourcing rates. Technical Report no. 482, SFB 386 (`http://www.stat.uni-muenchen.de/sfb386/`). To appear in *Statistical Modelling*.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A, General* **147**, 278–292.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics* **27**, 65–81.

Dean, C. and Lawless, J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association* **84**, 467–472.

Diebold, F. X., Gunther, T. A. and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* **39**, 863–883.

Elsner, J. B. and Jagger, T. H. (2006). Prediction models for annual U.S. hurricane counts. *Journal of Climate* **19**, 2935–2952.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**, 985–987.

Fienberg, S. E. and Mason, W. M. (1979). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology* **10**, 1–67.

Frühwirth-Schnatter, S. (1996). Recursive residuals and model diagnostics for normal and non-normal state space models. *Environmental and Ecological Statistics* **3**, 291–309.

Frühwirth-Schnatter, S. and Wagner, H. (2006). Auxiliary mixture sampling for parameter-driven models of time series of counts with applications to state space modelling. *Biometrika* **93**, 827–841.

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85**, 1–11.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **69**, 243–268.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society, Series B, Methodological* **14**, 107–114.

Gotway, C. A. and Wolfinger, R. D. (2003). Spatial prediction of counts and rates. *Statistics in Medicine* **22**, 1649–1660.

Grammig, J. and Kehrle, K. (2007). A new marked point process model for the federal funds rate target: Methodology and forecast evaluation. Technical report, SSRN (http://ssrn.com/abstract=888543).

Gschlößl, S. and Czado, C. (200x). Modelling count data with overdispersion and spatial effects. *Statistical Papers,* DOI 10.1007/s00362-006-0031-6.

Hall, B. H., Cummins, C., Laderman, E. and Mundy, J. (1988). The R&D master file docu-

mentation. Technical Report no. 72, National Bureau of Economic Research, Cambridge.

Hamill, T. M. (2001). Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129**, 550–560.

Handcock, M. S. and Morris, M. (1999). *Relative Distribution Methods in the Social Sciences*. Springer, New York.

Hausman, J. A., Hall, B. H. and Griliches, Z. (1984). Econometric models for count data with an application to the patents–R&D relationship. *Econometrica* **52**, 909–938.

Holford, T. R. (1983). The estimation of age, period and cohort effects for vital rates. *Biometrics* **39**, 311–324.

Jolliffe, I. T. (2007). Uncertainty and inference for verification measures. *Weather and Forecasting* **22**, 637–650.

Jolliffe, I. T. and Stephenson, D. B. (2003). *Forecast Verification: A Practicioner's Guide in Atmospheric Science*. John Wiley and Sons, Chichester, UK.

Knorr-Held, L. and Rainer, E. (2001). Projections of lung cancer mortality in West Germany: A case study in Bayesian prediction. *Biostatistics* **2**, 109–129.

Krzysztofowicz, R. and Sigrest, A. A. (1999). Calibration of probabilistic quantitative precipitation forecasts. *Weather and Forecasting* **14**, 427–442.

Lawless, J. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics* **15**, 209–225.

Liesenfeld, R., Nolte, I. and Pohlmeier, W. (2006). Modeling financial transaction price movements: A dynamic integer count data model. *Empirical Economics* **30**, 795–825.

McCabe, B. P. M. and Martin, G. M. (2005). Bayesian predictions of low count time series. *International Journal of Forecasting* **21**, 315–330.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd edition.

Nelson, K. P. and Leroux, B. G. (2006). Spatial models for autocorrelated count data.

Statistics in Medicine **25**, 1413–1430.

O'Hagan, A. (2003). HSSS model criticism. In Green, P. J., Hjort, N. L. and Richardson, S., editors, *Highly Structured Stochastic Systems*, pages 423–44. Oxford University Press.

Osmond, C. (1985). Using age, period and cohort models to estimate future mortality rates. *International Journal of Epidemiology* **14**, 124–129.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association* **66**, 783–801.

Sloughter, J. M., Raftery, A. E., Gneiting, T. and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review* **135**, 3209–3220.

Smith, J. Q. (1985). Diagnostic checks of non-standard time series models. *Journal of Forecasting* **4**, 283–291.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **64**, 583–639.

Venables, W. N. and Ripley, B. D. (1997). *Modern Applied Statistics with S-PLUS*. Springer, New York, 2nd edition.

Wang, P., Cockburn, I. M. and Puterman, M. L. (1998). Analysis of patent data — a mixed-Poisson-regression-model approach. *Journal of Business and Economic Statistics* **16**, 27–41.

Wecker, W. B. (1989). Assessing the accuracy of time series model forecasts of count observations. *Journal of Business and Economic Statistics* **7**, 418–419.

Winkelmann, R. (2005). *Econometric Analysis of Count Data*. Springer Verlag, Berlin, 4th edition.

Appendix A

Computation of $\|p\|^2$

Let $\|p\|^2 = \sum_{k=0}^{\infty} p_k^2$. For the Poisson distribution with parameter λ, we find that $\|p\|^2 = e^{-2\lambda} I_0(\lambda)$ where I_0 is a modified Bessel function (Abramowitz and Stegun, 1970, p. 374). For the negative binomial distribution with mean λ and dispersion parameter $a \geq 0$, we have

$$\|p\|^2 = (1 + 2a\lambda)^{-1/a} \, L_{-1/a}\left(1 + \frac{2a^2\lambda^2}{1 + 2a\lambda}\right),$$

where L is a Legendre function of the first kind (Abramowitz and Stegun, 1970, p. 332).

Appendix B

Probabilistic forecasts of ordered discrete data

The tools proposed in this paper generalize easily to probabilistic forecasts for arbitrary ordered discrete data, which are not necessarily counts. Without loss of generality, we consider the prediction of a quantity x that can attain a countable number of real numbers $(x_k)_{k=-\infty}^{\infty}$, where $x_{k-1} < x_k < x_{k+1}$ for all k. Let P be a probabilistic forecast for this quantity. We denote the probability mass function and cumulative distribution function for the predictive distribution P by $(p_k)_{k=\infty}^{\infty}$ and $(P_k)_{k=-\infty}^{\infty}$, respectively. Note that hereinafter the index $k \geq 0$ corresponds to x_k, which in general is not an integer. If the quantity can only attain a finite number of values, we have $p_k = 0$ for all but finitely many indices k. In the case of count data, we have $x_k = k$ for $k \geq 0$ and $p_k = 0$ for $k < 0$.

The probability integral transform (PIT) generalizes easily to this situation. We first

consider its randomized version. If $x = x_k$ obtains, we put

$$u = P_{k-1} + v(P_k - P_{k-1}) \tag{B.1}$$

where v is standard uniform and independent of x, which reduces to (1) and (2) in the case of count data. Grammig and Kehrle (2007) apply this device to assess probabilistic forecasts of a discrete economic variable. Similar tools have been used to assess the calibration of probabilistic quantitative precipitation forecasts, which typically take the form of a mixture of a point mass at zero and a predictive density on the positive half-axis (Krzysztofowicz and Sigrest, 1999; Sloughter, Raftery, Gneiting and Fraley, 2007).

To generalize the non-randomized PIT, we proceed as follows. If $x = x_k$ realizes, we put

$$F(u) = \begin{cases} 0, & u \leq P_{k-1}, \\ (u - P_{k-1})/(P_k - P_{k-1}), & P_{k-1} \leq u \leq P_k, \\ 1, & u \geq P_k, \end{cases} \tag{B.2}$$

which reduces to (3) and (4) in the case of count data. Again, we aggregate as in (5) and check the PIT histogram for uniformity.

The marginal calibration diagram does not require any adjustments, nor do the scoring rules, with the obvious exceptions that $\|p\|^2 = \sum_{k=-\infty}^{\infty} p_k^2$ and that the ranked probability score (10) is now computed as

$$\mathrm{rps}(P, x) = \sum_{k=-\infty}^{\infty} \left\{P_k - \mathbf{1}(x \leq x_k)\right\}^2 . \tag{B.3}$$

Table 1

Four predictive models for larynx cancer counts in Germany, 1998–2002, and the respective mean scores. The best value in each column is shown in bold face.

Model	age	disp	LogS	QS	SphS	RPS	DSS	SES	NSES
1	+	+	**4.27**	**−0.041**	**−0.153**	14.0	**6.74**	852.9	**1.66**
2	+	–	4.35	−0.040	−0.152	12.9	6.89	684.4	2.05
3	–	+	4.29	−0.040	−0.152	14.2	6.78	870.0	1.69
4	–	–	4.35	−0.039	−0.151	**12.2**	6.90	**564.8**	2.12

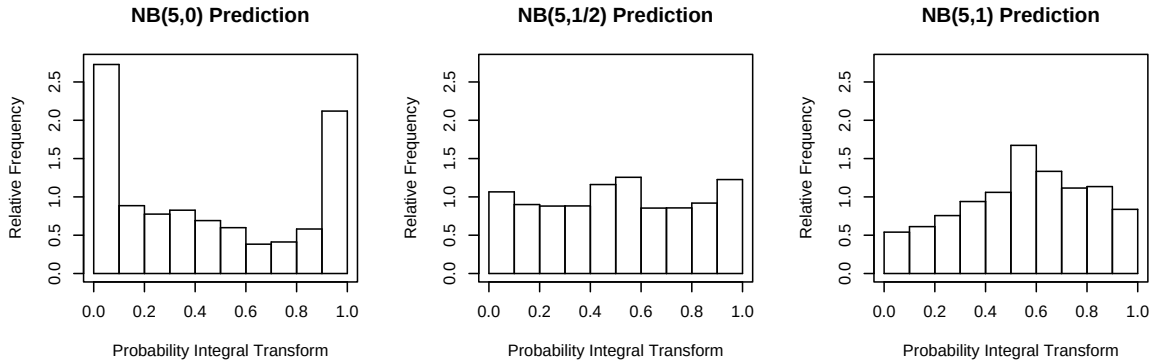

Figure 1. Non-randomized PIT histograms for probabilistic forecasts for a sample of 200 counts from a negative binomial distribution $\mathrm{NB}(\lambda, a)$ with mean $\lambda = 5$, dispersion parameter $a = \frac{1}{2}$ and variance $\lambda(1 + a\lambda)$. The predictive distribution is negative binomial with mean $\lambda = 5$ and dispersion parameter $a = 0$, $a = \frac{1}{2}$ and $a = 1$ (from left to right). The PIT histograms are U-shaped, uniform and inversely U-shaped, indicating underdispersed, well-calibrated and overdispersed predictive distributions, respectively.

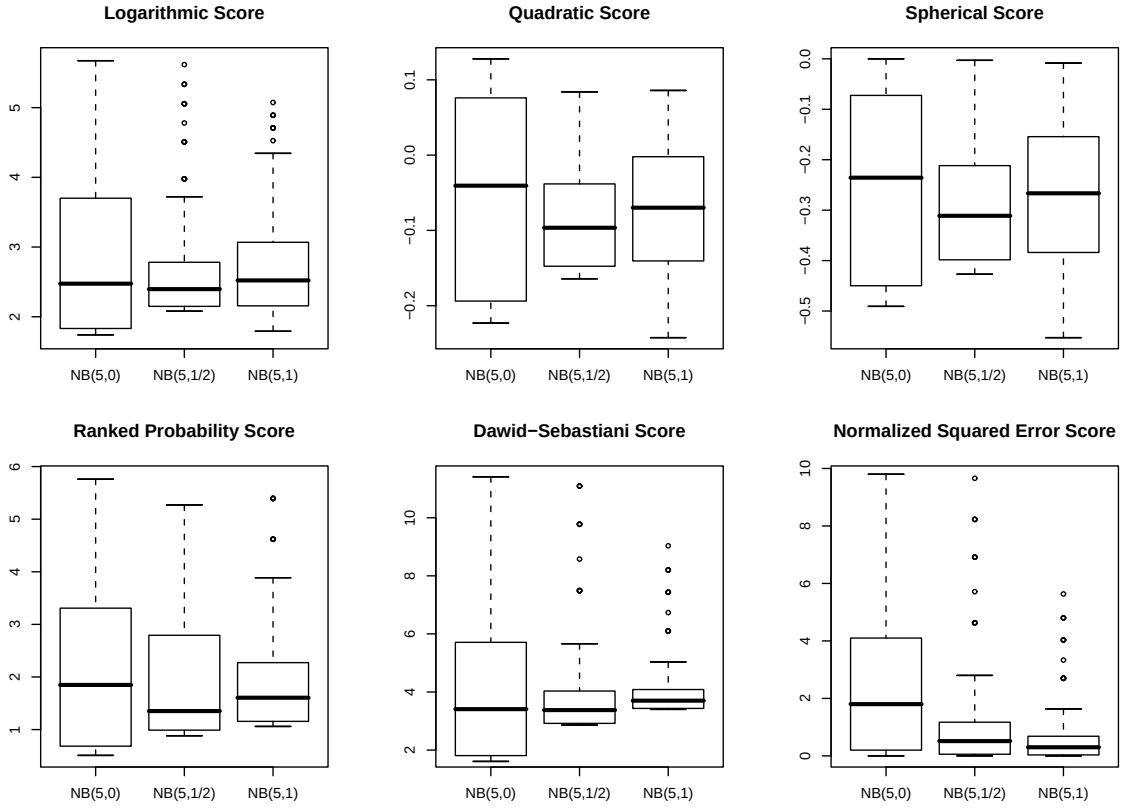

Figure 2. Boxplots for various scores in the situation of Figure 1.

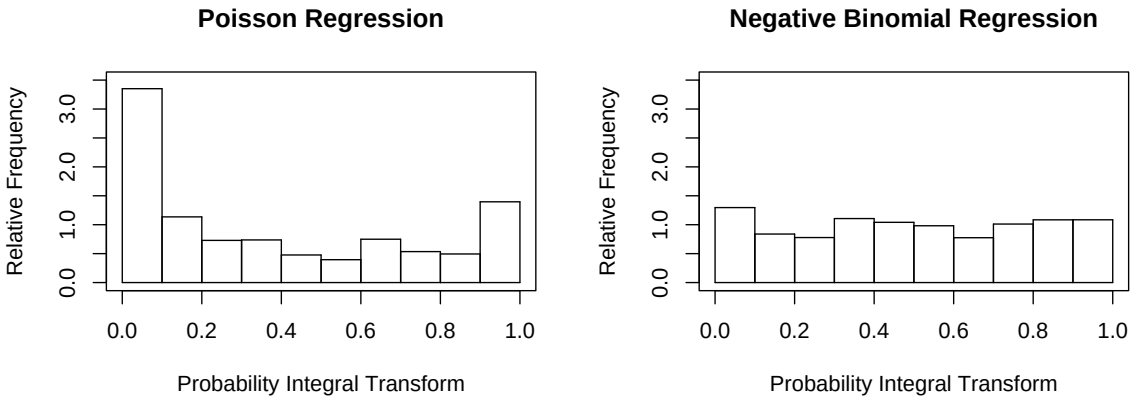

Figure 3. Non-randomized PIT histograms for patent data count regressions.

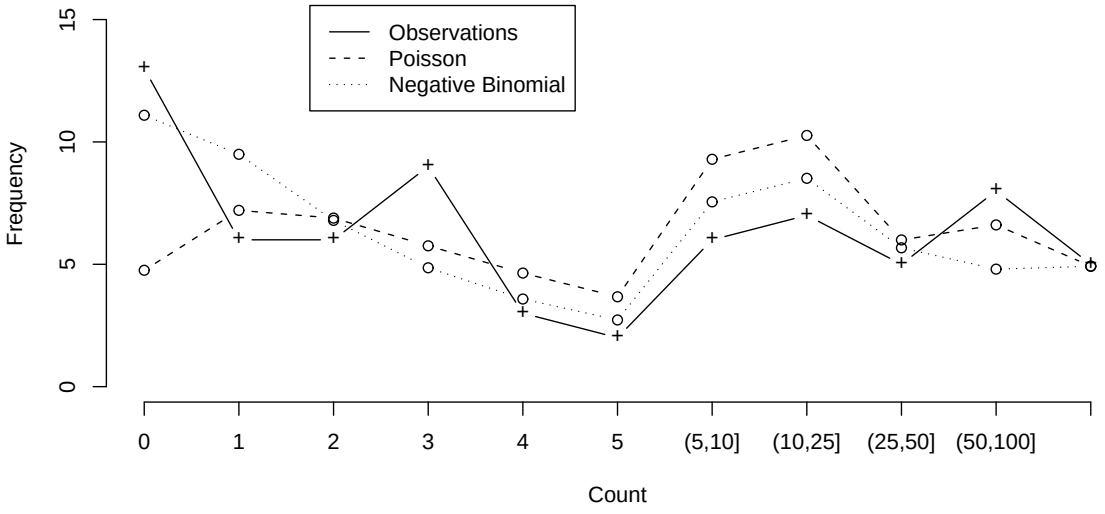

Figure 4. Marginal calibration diagram for patent data count regressions.

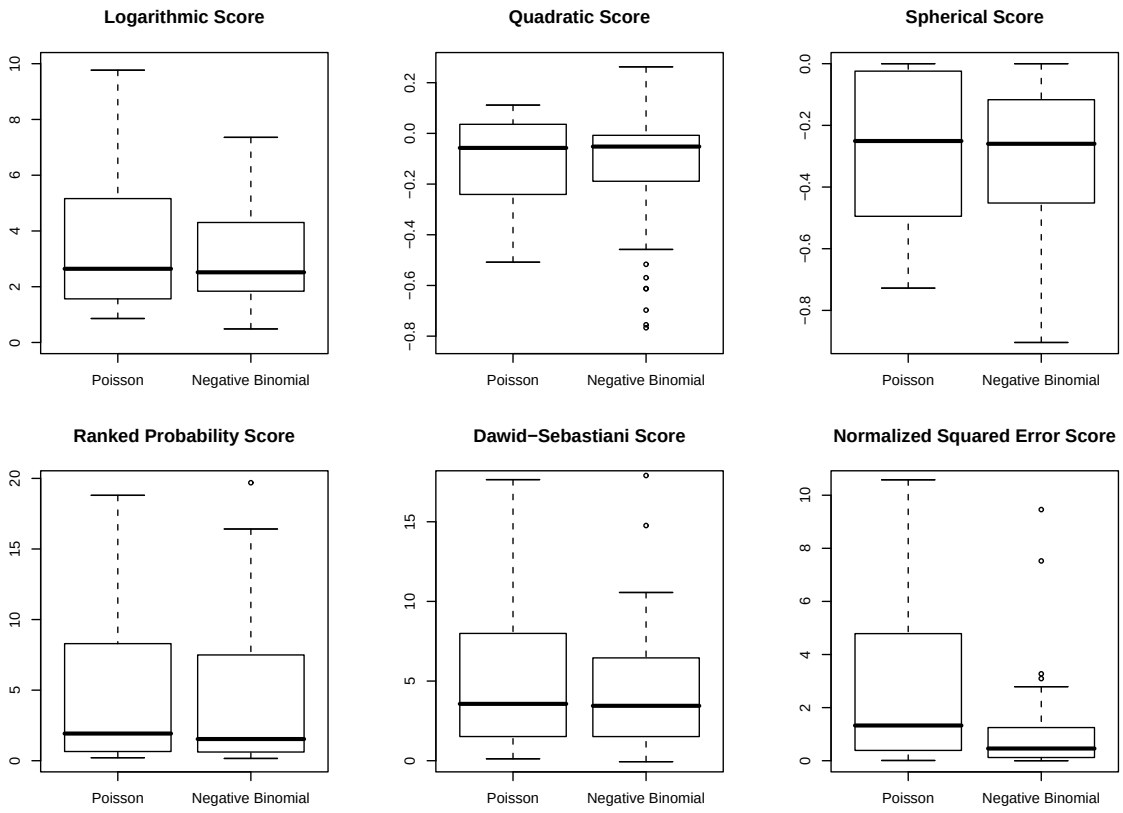

Figure 5. Boxplots for various scores for patent data count regressions.

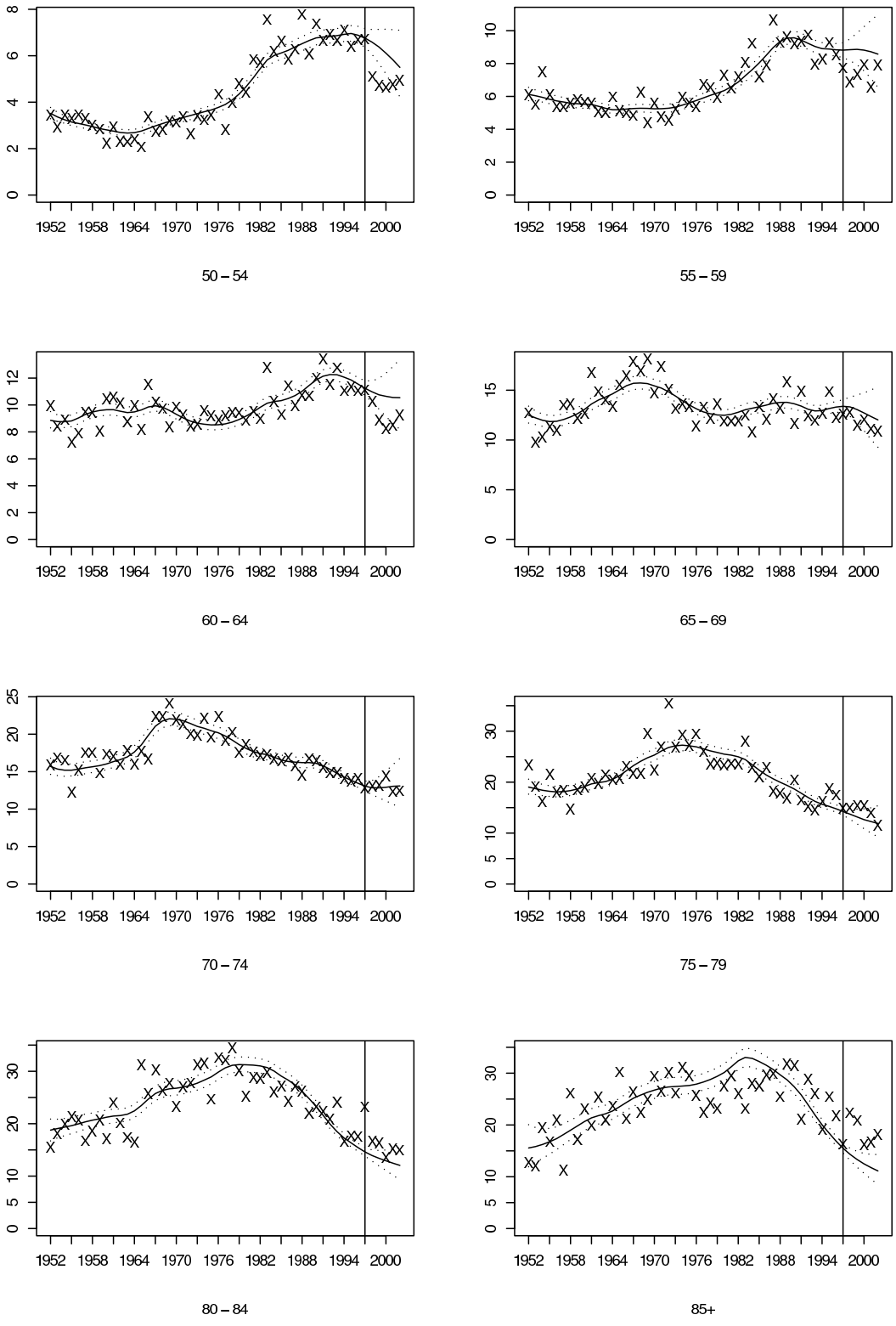

Figure 6. Observed (×) and fitted/predicted number of deaths from larynx cancer per 100,000 males in Germany in age groups 50–54, 55–59, . . . , 85–, based on Model 4.