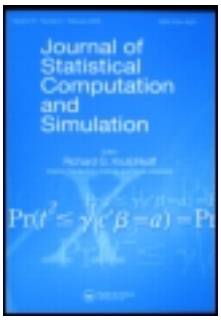

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

On count time series prediction

Vasiliki Christou [a] & Konstantinos Fokianos [a]

[a] Department of Mathematics & Statistics , University of Cyprus , Nicosia , Cyprus

Published online: 01 Aug 2013.

To cite this article: Vasiliki Christou & Konstantinos Fokianos , Journal of Statistical Computation and Simulation (2013): On count time series prediction, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2013.823612

To link to this article: http://dx.doi.org/10.1080/00949655.2013.823612

Journal of Statistical Computation and Simulation, 2013
http://dx.doi.org/10.1080/00949655.2013.823612

Taylor & Francis
Taylor & Francis Group

On count time series prediction

Vasiliki Christou and Konstantinos Fokianos*

Department of Mathematics & Statistics, University of Cyprus, Nicosia, Cyprus

(*Received 28 February 2013; final version received 7 July 2013*)

We consider the problem of assessing prediction for count time series based on either the Poisson distribution or the negative binomial distribution. By a suitable parametrization we employ both distributions with the same mean. We regress the mean on its past values and the values of the response and after obtaining consistent estimators of the regression parameters, regardless of the response distribution, we employ different criteria to study the prediction problem. We show by simulation and data examples that scoring rules and diagnostic graphs that have been proposed for independent but not identically distributed data can be adapted in the setting of count dependent data.

Keywords: calibration; prediction; probability integral transformation plot; quasi-likelihood; scoring rules; sharpness

1. Introduction

One of the most crucial aspects in time series analysis is the problem of prediction. This contribution studies the prediction problem in the context of count time series. More specifically, we follow the recent methodology of Czado et al.,[1] where various tools for predictive model assessment are developed for independent but not identically distributed data. We show that these methods can also be applied for count dependent data. We will take a similar point of view; that is predictors are probabilistic in the sense that a probability distribution can adequately describe their basic properties; see Dawid.[2]

To motivate our study consider the measles data shown in Figure 4. Note that this time series exhibits strong autocorrelation which decays slowly. This is an example of count time series. A sensible approach towards modelling such data is based on the Poisson assumption with a time varying mean which includes a feedback mechanism and depends on a vector of time invariant parameters, see Fokianos et al. [3] for instance. To predict future values of the response, the most natural way (in terms of mean square error) is to employ the conditional mean after estimating any unknown parameters by using maximum likelihood. However, the performance of such forecast is largely unknown in the literature; for an exception see Jung and Tremayne.[4] This contribution fills this gap by studying the behaviour of such predictor using a variety of scoring rules. In addition, we study the same problem for negative binomial time series. At the end, we can in fact assess which distribution is more suitable for the data at hand; for instance Figure 6 shows that the most appropriate model for the measles data is provided by the negative binomial distribution.

*Corresponding author. Email: fokianos@ucy.ac.cy

We focus on Poisson and negative binomial distributions since these are occurred in applications more frequently; however the methods can be applied to other discrete distributions provided that they are suitably parameterized. We also show that models which include a feedback mechanism considerably reduce the number of parameters for fitting and predicting models with strong autocorrelation, regardless of the chosen response distribution.

Section 2 discusses autoregressive models for count time series. In particular, we discuss models that include a feedback mechanism. The response distribution that is chosen for modelling the data is either the Poisson or the negative binomial distribution suitably parameterized to ensure that both mean processes are identical. In addition, we outline maximum (quasi) likelihood inference for the Poisson and negative binomial distribution, respectively. In Section 3 we assess the predictive performance of the proposed models by extending the tools developed by Czado et al. [1] for independent and not identically distributed data to count time series. More specifically, we address the problem of examining probabilistic calibration, marginal calibration and sharpness of the proposed models. Section 4 gives several simulation examples which illustrate the usefulness of the negative binomial distribution on modelling count time series. The article concludes with two real data examples.

2. Autoregressive models for count time series and inference

In what follows we will assume that $\{Y_t, t \geq 1\}$ denotes a count time series which will be the response. Furthermore, let $\{\lambda_t, t \geq 1\}$ be a sequence of mean processes. Denote by $\mathcal{F}_t^{Y,\lambda}$ the past of the process up to and including time t, that is $\mathcal{F}_t^{Y,\lambda} = \sigma(Y_s, s \leq t, \lambda_{1-q}, \ldots, \lambda_0)$, for some $q > 0$. We will study two important distributions that can be employed for conditional modelling of count time series data; the Poisson distribution given by

$$\mathrm{P}[Y_t = y \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \frac{\exp(-\lambda_t)\lambda_t^y}{y!}, \quad y = 0, 1, 2, \ldots \tag{1}$$

and the negative binomial distribution given by

$$\mathrm{P}[Y_t = y \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \frac{\Gamma(\nu + y)}{\Gamma(y+1)\Gamma(\nu)} \left(\frac{\nu}{\nu + \lambda_t}\right)^{\nu} \left(\frac{\lambda_t}{\nu + \lambda_t}\right)^{y}, \quad y = 0, 1, 2, \ldots, \tag{2}$$

where $\nu > 0$. Model (2) is motivated by the theory of mixed Poisson processes. To be more specific, suppose that $\tilde{N}$ is a standard homogeneous Poisson process (that is a Poisson process with rate equal to 1) and $\mu(t)$ is the mean value function of a Poisson process on $[0, \infty)$. Suppose further that Z is a positive random variable independent of $\tilde{N}$. Then, the process

$$N(t) = \tilde{N}(Z\mu(t)) = \tilde{N}(0, Z\mu(t)], \quad t \geq 0, \tag{3}$$

is called a mixed Poisson process (see [5], for instance). In particular, when Z is a Gamma random variable with mean 1 and variance $1/\nu$ and $\mu(t) = \lambda_t$ then we recover (2) as the probability mass function of the random variable which counts the number of events in the interval $(0, \lambda_t]$; see Christou and Fokianos [6] for more. It is clear that Equation (2) tends to Equation (1), when $\nu \to \infty$. It is also easy to see that the conditional mean of both distributions is equal to λ_t. However the conditional variance of the Poisson distribution is equal to λ_t, whereas the conditional variance of Equation (2) is equal to $\lambda_t + (\lambda_t)^2/\nu$. We conclude by this simple fact that the negative binomial distribution takes into account more properly overdispersion which is observed repeatedly in count dependent data. Although, simple Poisson models, as those studied by Fokianos et al.,[3] accommodate overdispersion, it is anticipated that negative binomial models will tend to fit data

better. The particular form of negative binomial distribution has been suggested by several other authors in the framework of generalized linear models; see Lawless,[7] Cameron and Trivedi,[8] among others. For the case of count time series, Zhu [9] suggested this particular parametrization without pointing out any possible implications for data analysis.

2.1. *Modelling*

The proposed modelling approach that we take is along the lines of Fokianos et al.[3] We assume that the mean process depends upon its past values and the past values of the response. In other words, we assume that

$$\lambda_t = f(Y_{t-1}, \ldots, Y_{t-p}, \lambda_{t-1}, \ldots, \lambda_{t-q}), \tag{4}$$

where $f(\cdot,\cdot)$ is a parametric function defined on $\mathbb{N}_0^p \times \mathbb{R}_+^q$ and taking values on $(0,\infty)$, where $\mathbb{N}_0 = \{0,1,2,\ldots\}$. As we shall see below, there are several examples falling in this class of models; the linear model (see Example 1) being the most prominent. Furthermore, note that the mean process $\{\lambda_t\}$ is unobserved and we will assume that the starting values $\lambda_{1-q}, \ldots, \lambda_0$ are random. We consider models that include a feedback mechanism for the mean process $\{\lambda_t\}$ because this class of models successfully accommodates count time series whose autocorrelation function decays slowly. The inclusion of the feedback mechanism makes parsimonious modelling possible. In fact, we will see, by a simulated example, that models that include a feedback mechanism outperform models without the feedback mechanism in terms of prediction, even in cases where the true data generating process depends exclusively on a large number of lagged variables of the response. Here, it should be mentioned that both Equations (1) and (2) can be employed as conditional response distributions in connection with model (4). Some concrete examples follow.

Example 1 Consider the following linear model for Equation (4)

$$f(Y_{t-1}, \ldots, Y_{t-p}, \lambda_{t-1}, \ldots, \lambda_{t-q}) = d + \sum_{i=1}^{q} a_i \lambda_{t-i} + \sum_{j=1}^{p} b_j Y_{t-j}, \tag{5}$$

where $d, a_i, b_j > 0$ for all i, j so that the mean process is positive. With some slight abuse of terminology, we can refer to Equation (5) as an integer generalized autoregressive conditional heteroskedastic (INGARCH) model of order (p,q) at least for the case of the Poisson distribution where the conditional mean equals the conditional variance. This model has been studied by Ferland et al. [10] who showed that the process is second-order stationary, provided that $0 < \sum_i a_i + \sum_j b_j < 1$. In particular, when $p = q = 1$, then

$$\lambda_t = d + a_1 \lambda_{t-1} + b_1 Y_{t-1}. \tag{6}$$

This is an example of INGARCH model of order $(1,1)$, see Fokianos et al. [3] who proved that a perturbed version of this model is geometrically ergodic with moments of any order when $0 < a_1 + b_1 < 1$. The unperturbed version has been studied by Neumann [11] and Doukhan et al.[12] All the above contributions are under the assumption of Poisson distribution. Theorem 2.1 of Christou and Fokianos [6] shows that under the same conditions, model (6) is ergodic and stationary which posses moments of any order, but under the negative binomial assumption. The mean parametrization based on Equation (2) yields to a stationary region which is independent of the additional parameter ν. Therefore, we only need to be concerned with the appropriate choice of the regression parameters when fitting negative binomial models like model (6) to count time series. It is also an essential fact for developing optimization algorithms to fit such models. We

emphasize this point further by comparing the proposed parametrization (2) with that of Zhu [9] given by

$$\mathrm{P}[Y_t = y \mid \mathcal{F}_{t-1}^{Y,\lambda}] = \binom{y-r-1}{r-1} p_t^r (1-p_t)^y, \quad y = 0, 1, 2, \ldots \tag{7}$$

In the above expression r is a positive integer and p_t denotes the conditional probability of success. Zhu [9] models the odds ratio, denoted by $p_t/(1-p_t)$, as a function of past values of itself and past values of the response. Stationarity conditions for the model studied by Zhu [9] depend upon the parameter r of Equation (7). Therefore, it is challenging to optimize the log-likelihood function since such a maximization problem imposes severe restrictions on both regression parameters and r. Furthermore, it is well known that the estimation of odds ratio might be problematic when probabilities are either very small or very large. For related work see also Davis and Liu.[13]

Recall Equation (5). A particular case is given by setting $p = 5$ and $q = 0$; that is

$$\lambda_t = d + \sum_{i=1}^{5} b_i Y_{t-i}. \tag{8}$$

This is an example of an integer autoregressive model of order 5. It follows that the required conditions for model (8) to be stationary with moments of any order, is given by $0 < \sum_{i=1}^{5} b_i < 1$. We include this model for the sake of comparison with models that contain a feedback mechanism like Equation (6). We shall see that even though the data generating process follows model (8), it is still more efficient to work with model (6) in terms of prediction.

Example 2 An interesting example of a nonlinear model is given by (cf. [14])

$$\lambda_t = d\frac{1}{(1+\lambda_{t-1})^\gamma} + a_1\lambda_{t-1} + b_1 Y_{t-1}, \tag{9}$$

provided that all the parameters d, a_1, b_1, γ are positive. Inclusion of the parameter γ yields a nonlinear deviation, in the sense that small values of the parameter γ cause Equation (9) to approach model (6). Moderate values of γ introduce a stronger perturbation. Regardless of the chosen response distribution, it can be shown that when $\max\{a_1, d\gamma - a_1\} + b_1 < 1$, then model (9) is ergodic and stationary whose moments are finite, see Fokianos and Neumann.[15] We iterate again the point that this region is independent of the additional parameter ν of Equation (2).

2.2. *Inference*

Maximum likelihood inference for the Poisson model (1) and negative binomial model (2) has been developed by Fokianos et al. [3] and Christou and Fokianos,[6] respectively. Both the above studies develop estimation based on the Poisson likelihood function which for Poisson model (1) is obviously the true likelihood. However, for the negative binomial model (2) this method resembles similarities with quasi-likelihood inference developed for the estimation of generalized autoregressive conditional heteroskedasticity models. For instance Berkes et al.,[16] Francq and Zakoïan,[17] Mikosch and Straumann [18] and Bardet and Wintenberger [19] among others, study the Gaussian likelihood function irrespectively of the assumed distribution for the error sequence. This methodology avoids complicated likelihood function and it ensures that the regression parameters are estimated consistently since we only require the correct mean specification; Zeger and Liang,[20] Godambe and Heyde,[21] Heyde.[22]

Therefore, for defining properly the quasi-conditional maximum likelihood estimator (QMLE), consider the Poisson log-likelihood function, as in Fokianos et al.,[3] where $\boldsymbol{\theta}$ denotes the unknown parameter vector

$$l_n(\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}) = \sum_{t=1}^{n} (Y_t \log \lambda_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})). \tag{10}$$

The quasi-score function is defined by

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \frac{\partial l_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{t=1}^{n} \left(\frac{Y_t}{\lambda_t(\boldsymbol{\theta})} - 1 \right) \frac{\partial \lambda_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{11}$$

The solution of the system of nonlinear equations $\boldsymbol{S}_n(\boldsymbol{\theta}) = 0$, if it exists, yields the QMLE of $\boldsymbol{\theta}$ which we denote by $\hat{\boldsymbol{\theta}}$. Asymptotic properties of $\hat{\boldsymbol{\theta}}$ have been studied in detail by Fokianos et al. [3] for the case of the Poisson model (1) with the mean process given by Equation (6). For the case of the negative binomial distribution with model (4), see Christou and Fokianos.[6] In both cases, we obtain consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}$. Note that for the case of negative binomial response distribution, it is required to estimate the parameter ν. Towards this goal, we propose

$$\hat{\nu}_1 = \left(\frac{1}{n} \sum_{t=1}^{n} \frac{[(Y_t - \hat{\lambda}_t)^2 - \hat{\lambda}_t]}{\hat{\lambda}_t^2} \right)^{-1},$$

where $\hat{\lambda}_t = \lambda_t(\hat{\boldsymbol{\theta}})$. This is simply a moment-based estimator. Another estimator, say $\hat{\nu}_2$, is obtained by solving the equation

$$\sum_{t=1}^{n} \frac{(Y_t - \hat{\lambda}_t)^2}{\hat{\lambda}_t(1 + \hat{\lambda}_t/\nu)} = n - m,$$

where m denotes the dimension of $\boldsymbol{\theta}$. For more, see Lawless [7] and Christou and Fokianos.[6]

3. Assessment of the predictive performance

A major issue of time series analysis is to provide forecasts for future quantities. For count time series, the optimal predictor of Y_t given its past, is given by the conditional expectation λ_t, in terms of mean square error. Obviously, in applications we employ $\hat{\lambda}_t = \lambda_t(\hat{\boldsymbol{\theta}})$. We focus exclusively on this predictor and the associated probability models as introduced by Equations (1) and (2). Regardless the chosen response distribution, it is clear that $\hat{\lambda}_t$ is identical for both cases provided that we employ the QMLE obtained by Equation (11).

Following Gneiting et al.,[23] we take the point of view that predictions should be probabilistic in nature. In addition, they should strive to maximize the sharpness of the predictive distribution subject to calibration. Calibration, refers to the statistical consistency between the predictive distribution and the observations, and it is a joint property of the forecasts and the values that utilize. The notion of sharpness refers to the concentration of the predictive distribution and is a property of the forecasts only. It follows that if the predictive distribution is more concentrated, then the forecasts obtained are sharper. Therefore, this will yield better predictions subject to calibration. In this section, we provide diagnostic tools to evaluate the predictive performance.

3.1. *Assessment of probabilistic calibration*

The key tool in assessing probabilistic calibration is the nonrandomized probability integral transformation (PIT) histogram as discussed by Czado et al.[1] U-shaped histograms point at

underdispersed predictive distributions, hump or inverse U-shaped histograms indicate overdispersion and uniformity of histograms hints well-calibrated predictive distributions. An example of this diagnostic tool in the framework of modelling count time series is given in Christou and Fokianos,[6] where the authors present PIT histograms for a data collection of transactions of a certain stock. PIT plots are also shown in Figures 5 and 8. Both plots hint that the negative binomial distribution fits better the data than the Poisson distribution irrespectively of the chosen model; we will revisit this fact later on in this article. To avoid unnecessary repetition we refer to Christou and Fokianos [6] for details about the construction of the plot.

3.2. *Assessment of marginal calibration*

We now turn to the question of assessing marginal calibration. Suppose that the observed time series $\{Y_t, t \geq 1\}$ is stationary which follows marginally the cumulative distribution function (cdf) $G(\cdot)$. In addition, we assume that a candidate forecaster picks a probabilistic forecast in the form of a predictive cdf $P_t(x) = \mathrm{P}(Y_t \leq x \mid \mathcal{F}_{t-1}^{Y,\lambda})$. In the case that we study, we have $P_t(\cdot)$ to be either the cdf of a Poisson with mean $\hat{\lambda}_t$ or of a negative binomial distribution evaluated at $\hat{\lambda}_t$ and $\hat{\upsilon}$. We follow Gneiting et al. [23] to assess marginal calibration by comparing the average predictive cdf

$$\bar{P}(x) = \frac{1}{n}\sum_{t=1}^{n} P_t(x), \quad x \in \mathbb{R},$$

to the empirical cdf of the observations given by

$$\hat{G}(x) = \frac{1}{n}\sum_{t=1}^{n} 1(Y_t \leq x), \quad x \in \mathbb{R}.$$

To display the marginal calibration plot, we plot the difference of the two cdf

$$\bar{P}(x) - \hat{G}(x), \quad x \in \mathbb{R}, \tag{12}$$

for each of the various forecasters. If the marginal calibration hypothesis is true, then we expect minor fluctuations about 0. Major excursions from zero hint that the forecaster lacks marginal calibration. To see this, note that for x fixed and $P_t(x)$ as before, we obtain

$$\bar{P}(x) = \frac{1}{n}\sum_{t=1}^{n} P_t(x) \xrightarrow{p} \mathrm{E}(P(Y_t \leq x \mid \mathcal{F}_{t-1}^{Y,\lambda})) = P(Y_t \leq x),$$

as $n \to \infty$, because of ergodicity of $\{Y_t\}$. In addition,

$$\hat{G}(x) = \frac{1}{n}\sum_{t=1}^{n} 1(Y_t \leq x) \xrightarrow{p} \mathrm{E}(1(Y_t \leq x)) = P(Y_t \leq x),$$

as $n \to \infty$. Therefore, $\bar{P}(x)$ and $\hat{G}(x)$ converge in probability to the marginal cdf of $\{Y_t\}$. Hence, plot (12) would indicate agreement (or disagreement) between the predictive distribution and the marginal empirical distribution of the observed counts.

3.3. *Assessment of sharpness*

Addressing sharpness is accomplished via scoring rules. These rules provide numerical scores and form summary measures for the assessment of the predictive performance. In addition, scoring

rules help us to rank the competing forecast models. They are negatively oriented penalties that the forecaster wishes to minimize, see also Czado et al.[1]

Assume that for $t \geq 1$, $P_t(\cdot)$ denotes the predictive cdf as before. The score will be denoted, in general, by $s(P_t, Y_t)$. Put $p_y = \mathrm{P}(Y_t = y \mid \mathcal{F}_{t-1}^{Y,\lambda})$, for $y = 0, 1, 2, \ldots$; recall Equations (1) and (2). We consider seven different examples of scoring rules, as in Czado et al.,[1] for the case of count time series.

- Logarithmic score

$$\mathrm{logs}(P_t, Y_t) = -\log p_y. \tag{13}$$

- Quadratic or Brier score

$$\mathrm{qs}(P_t, Y_t) = -2p_y + \|p\|^2, \tag{14}$$

  where $\|p\|^2 = \sum_{y=0}^{\infty} p_y^2$.
- Spherical score

$$\mathrm{sphs}(P_t, Y_t) = -\frac{p_y}{\|p\|}. \tag{15}$$

- Ranked probability score

$$\mathrm{rps}(P_t, Y_t) = \sum_{x=0}^{\infty} (P_t(x) - 1(Y_t \leq x))^2. \tag{16}$$

- Dawid–Sebastiani score

$$\mathrm{dss}(P_t, Y_t) = \left(\frac{Y_t - \mu_{P_t}}{\sigma_{P_t}}\right)^2 + 2\log \sigma_{P_t}, \tag{17}$$

  where μ_{P_t} and σ_{P_t} are the mean and the standard deviation of the predictive distribution P_t, respectively.
- Normalized squared error score

$$\mathrm{nses}(P_t, Y_t) = \left(\frac{Y_t - \mu_{P_t}}{\sigma_{P_t}}\right)^2. \tag{18}$$

- Squared error score

$$\mathrm{ses}(P_t, Y_t) = (Y_t - \mu_{P_t})^2. \tag{19}$$

It is clear, from the above formulation, that the squared error score is identical for both the Poisson and the negative binomial distribution since the conditional means are equal. Note that the normalized square error score is formed by the so called Pearson residuals obtained after fitting a model. We will compare all those scores in the next sections.

4. Examples

In the following, we present a limited simulation study where we examine the predictive properties of all models considered in Examples 1 and 2. All simulations are based on 1000 runs and the data have been generated according to the negative binomial model (2) with mean specified by Equation (4) and $\nu = 2$. We consider sample sizes of $n = 200$ and $n = 1000$, throughout the simulations.

Table 1. Mean scores for models (6) and (8).

		Scoring rules						
	n	logs	qs	sphs	rps	dss	nses	ses
Model (6)	200	**1.871**	**−0.188**	**−0.434**	**1.056**	**2.521**	1.002	**4.620**
Model (8)		2.060	−0.165	−0.420	1.120	2.648	**0.998**	5.115
Model (6)	1000	**1.999**	**−0.178**	**−0.423**	**1.291**	**2.714**	1.004	**6.246**
Model (8)		2.352	−0.087	−0.365	1.412	2.979	**0.999**	7.492

Notes: The data have been generated by model (8). Models are compared by the mean logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, normalized squared error and squared error scores. Bold face numbers in each column indicate the minimum value obtained between the two models.

4.1. *Models with and without feedback mechanism*

To show that in fact models with a feedback mechanism are more parsimonious than models without a feedback mechanism, in terms of prediction, consider models (6) and (8), for instance. We will use the scoring rules, as outlined in Section 3.3, to measure their predictive power.

Suppose that we sample data from the negative binomial distribution (2) with $\nu = 2$ and the mean process λ_t is modelled by means of Equation (8) with $d = 1$, $b_1 = 0.01$, $b_2 = 0.05$, $b_3 = 0.01$, $b_4 = 0.2$, $b_5 = 0.3$. For this example, we fit models (6) and (8) following the methodology outlined in Section 2.2. After obtaining the QMLE, we calculate all scoring rules based on negative binomial probabilistic prediction. Table 1 shows the results of the scoring rules for the two models. All rules point out to the adequacy of parsimonious model (6); the difference between the normalized squared error scores of models (6) and (8) is not significant. This simple exercise shows that the inclusion of the feedback mechanism in Equation (6) takes into account dependence more properly, even though the data have been generated by Equation (8). Hence, model (6) seems more suitable to fit data that show slowly decaying sample autocorrelation function and additionally offers a parsimonious parametrization.

4.2. *Linear and nonlinear models*

We generate now data according to the linear model (6) and the nonlinear model (9). For the linear model (6), data are generated with true values $(d, a, b) = (0.5, 0.4, 0.5)$ from Equation (2) with $\nu = 2$. For each simulation we divide the data into two sets; a training dataset and a test dataset. The training dataset is used to fit the model and to estimate the unknown parameters, while the test dataset is employed for prediction. The training dataset consists of the first 65%, 75% or 85% of the observations of the full data collection and the remaining data points form the test dataset for each split. We also use the whole observed time series to study the in-sample performance of the predictor. Calculation of the maximum likelihood estimators, regardless whether we split the data or not, is carried out by applying the quasi-likelihood methodology as outlined in Section 2.2. After obtaining the QMLE $\hat{\boldsymbol{\theta}} = (\hat{d}, \hat{a}, \hat{b})$, and using the fact that $\mathrm{E}(Y_t \mid \mathcal{F}_{t-1}^{Y,\lambda}) = \lambda_t$, we predict Y_t from $\hat{\lambda}_t = \lambda_t(\hat{\boldsymbol{\theta}})$. We consider two different probabilistic forecasters whose predictive distribution is either the Poisson or the negative binomial. For each split of the data, both forecasters predict the response of the corresponding test dataset. Thereafter, we calculate the mean score of six scoring rules given by Equations (13)–(18). Then, we compute the pairwise difference between mean scores for negative binomial and Poisson for each of the numerical measures; in other words we compute the mean score obtained from the Poisson minus the mean score obtained from the negative binomial.

Table 2. Proportions of the positive differences between the scoring rules obtained from the Poisson distribution and the scoring rules obtained from the negative binomial distribution.

	Split percentage					Split percentage			
Score	65%	75%	85%	Full data	Score	65%	75%	85%	Full data
$n = 200$									
logs	0.954	0.907	0.844	1.000	logs	0.892	0.841	0.773	0.999
qs	0.687	0.689	0.704	0.686	qs	0.637	0.656	0.647	0.611
sphs	0.671	0.660	0.700	0.680	sphs	0.618	0.629	0.603	0.607
rps	0.762	0.750	0.719	0.803	rps	0.680	0.645	0.650	0.656
dss	0.828	0.759	0.686	1.000	dss	0.734	0.699	0.609	0.998
nses	0.999	1.000	1.000	1.000	nses	0.997	0.998	0.999	1.000
$n = 1000$									
logs	1.000	1.000	1.000	1.000	logs	1.000	0.999	1.000	1.000
qs	0.813	0.792	0.779	0.836	qs	0.732	0.729	0.746	0.766
sphs	0.815	0.786	0.771	0.847	sphs	0.732	0.733	0.749	0.748
rps	0.938	0.936	0.928	0.961	rps	0.894	0.857	0.860	0.936
dss	1.000	0.999	0.998	1.000	dss	0.999	0.994	0.980	1.000
nses	1.000	1.000	1.000	1.000	nses	1.000	1.000	1.000	1.000

Notes: For the left columns, data are generated from the linear model (6) for $(d, a, b) = (0.5, 0.4, 0.5)$, $\nu = 2$ and for sample sizes $n = 200$ and $n = 1000$. For the right columns, data are generated from the nonlinear model (9) for the same set of parameters and setting $\gamma = 0.5$. Results are based on 1000 simulations.

Since the true data generating process follows the negative binomial distribution, we expect that the forecaster whose predictive distribution is assumed to be the negative binomial, will obtain smaller values for each mean score than those obtained from the forecast whose predictive distribution is Poisson. To verify this assertion, using the results from all simulations, we produce a table of proportions of positive differences for each split of the data. For purposes of comparison, we construct all the above results for the full data collection as well. The left columns of Table 2 show the results.

Most of the proportions of these tables are approaching unity. This fact implies that for most of the simulation runs, the negative binomial forecast outperforms the corresponding Poisson prediction.

Note that both the quadratic and spherical scores (see Equations (14) and (15), respectively) yield, in most of the cases, the smaller values for the obtained proportions. In addition, the logarithmic score (13), the Dawid–Sebastiani score (17) and the normalized squared error score (18) yield proportions which approximate unity, regardless of the splitting of the data. We conclude that these scores seem to be more appropriate for correctly identifying the true data generating process. Figure 1 shows the boxplots of all scores (13)–(18) for the two forecasters and for each split of the data for the case of linear model (6). White boxplots correspond to the Poisson prediction; grey boxplots correspond to the negative binomial forecast. It is obvious that the negative binomial forecast is superior to the Poisson forecast for each case. In addition, Figure 1 indicates that the performance of all scoring rules does not depend upon the particular partition of the data set.

Table 3 reports the scoring rules calculated from the full data collection in this simulation experiment, where the two forecasters are compared by the mean scores given by Equations (13)–(19). Scores from this table show clear preference towards the negative binomial over the Poisson prediction. Note that the predictive mean is equal – by (3) – for both distributions and therefore the mean squared error score is the same for the two forecasts. Table 4 reports results when the true data generating process follows the Poisson distribution. As we note, all scoring rules are similar and therefore we suggest that it is preferable to employ the Poisson forecast when such phenomenon is observed.

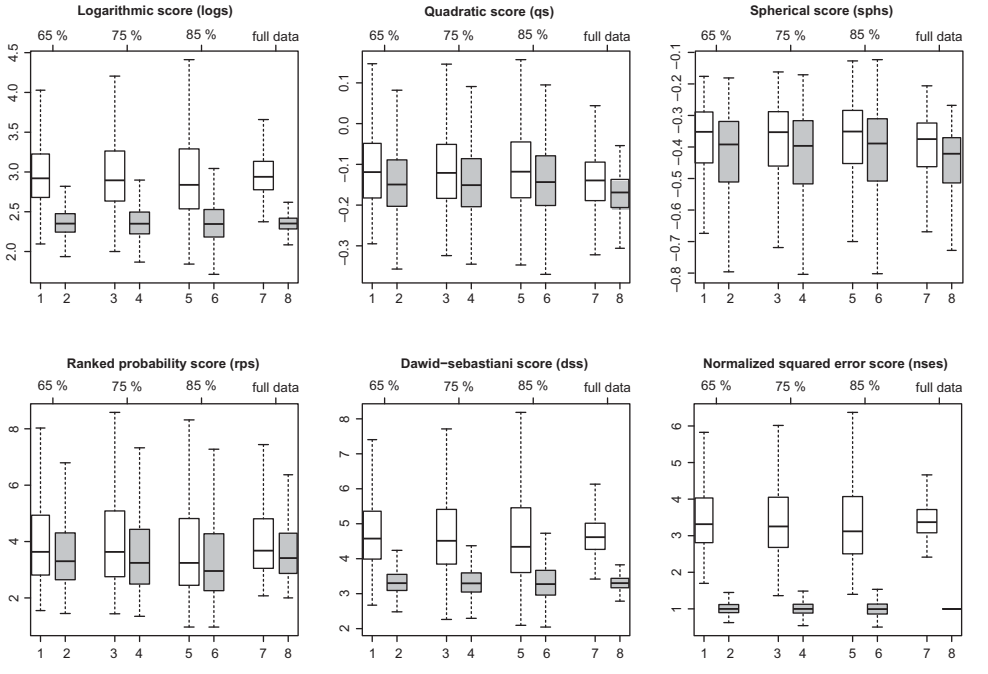

Figure 1. Boxplots for the mean scores given by Equations (13)–(18). White plots are for the Poisson forecast, while grey plots are for the negative binomial prediction. Data are generated from the linear model (6) when the true values are $(d, a, b) = (0.5, 0.4, 0.5)$ and $\nu = 2$. The results are based on 1000 data points and 1000 simulations.

Table 3. Scoring rules calculated for the linear model (6) and the nonlinear model (9) when $d = 0.5, a = 0.4, b = 0.5, \gamma = 0.5, \nu = 2$ and $n = 200$ or $n = 1000$.

			Scoring rules						
	Forecaster	n	logs	qs	sphs	rps	dss	nses	ses
Linear model (6)	Poisson	200	1.714	−0.267	−0.566	1.276	2.077	1.786	3.919
	Negative binomial		**1.599**	**−0.283**	**−0.571**	**1.183**	**1.819**	**0.985**	**3.919**
Nonlinear model (9)	Poisson		1.425	−0.334	−0.624	0.938	1.488	1.554	2.303
	Negative binomial		**1.359**	**−0.345**	**−0.629**	**0.885**	**1.338**	**0.984**	**2.303**
Linear model (6)	Poisson	1000	2.989	−0.139	−0.423	4.625	4.722	3.480	44.951
	Negative binomial		**2.354**	**−0.170**	**−0.469**	**3.958**	**3.307**	**0.997**	**44.951**
Nonlinear model (9)	Poisson		2.234	−0.226	−0.529	2.676	3.167	2.466	19.181
	Negative binomial		**1.921**	**−0.250**	**−0.564**	**2.360**	**2.466**	**0.997**	**19.181**

Notes: Results are based on 1000 simulations. For both models, the two forecasters are compared by the mean logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, normalized squared error and squared error scores. Bold face numbers in each column indicate the minimum value obtained between the two forecasters.

Furthermore, Figure 2 depicts the marginal calibration plot for both forecast distributions; recall Equation (12). It is clear that the negative binomial is superior to the Poisson distribution. The solid line of the graph (which corresponds to the negative binomial forecast) illustrates small deviations from zero, as it should be expected. Similar conclusions hold for the nonlinear model (9). Using the same set of parameters as for the linear model and setting $\gamma = 0.5$, we verified empirically from the right columns of Table 2, Table 3 and Figure 3 the superiority of the negative binomial prediction.

Table 4. Scoring rules calculated for the linear model (6) and the nonlinear model (9) when data are generated from the Poisson distribution with true values $d = 0.5, a = 0.4, b = 0.5, \gamma = 0.5$ and $n = 200$ or $n = 1000$.

			Scoring rules						
	Forecaster	n	logs	qs	sphs	rps	dss	nses	ses
Linear model (6)	Poisson	200	2.123	−0.152	−0.414	2.302	2.463	0.991	4.954
	Negative binomial		2.122	−0.152	−0.415	2.290	2.462	0.947	4.954
Nonlinear model (9)	Poisson		1.775	−0.224	−0.503	1.494	1.824	0.989	2.713
	Negative binomial		1.774	−0.224	−0.504	1.485	1.823	0.945	2.713
Linear model (6)	Poisson	1000	2.128	−0.152	−0.420	2.441	2.476	0.998	4.991
	Negative binomial		2.128	−0.152	−0.420	2.436	2.475	0.980	4.991
Nonlinear model (9)	Poisson		1.780	−0.228	−0.507	1.580	1.838	0.999	2.750
	Negative binomial		1.780	−0.228	−0.507	1.575	1.837	0.980	2.750

Notes: Results are based on 1000 simulations. For both models, the two forecasters are compared by the mean logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, normalized squared error and squared error scores.

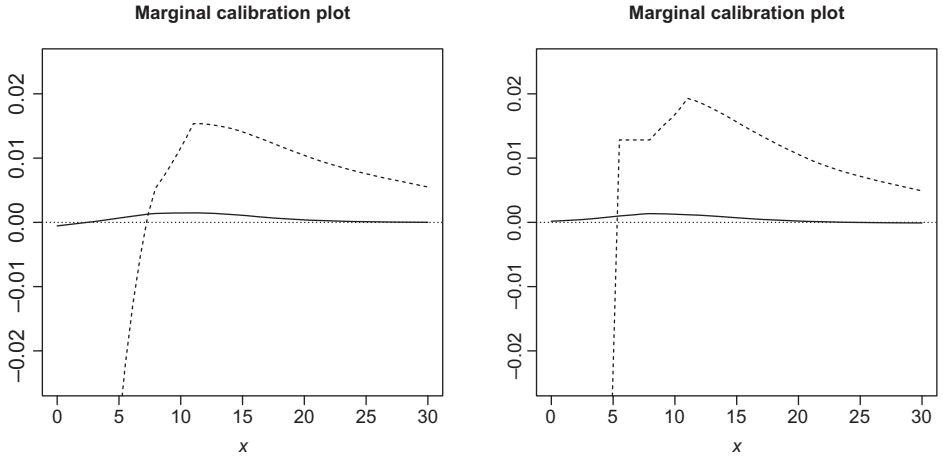

Figure 2. Left plot: marginal calibration plot for the linear model (6). Data are generated with true values $(d, a, b) = (0.5, 0.4, 0.5)$ and $\nu = 2$. Right plot: marginal calibration plot for the nonlinear model (9). Data are generated with true values $(d, a, \gamma, b) = (0.5, 0.4, 0.5, 0.5)$ and $\nu = 2$. The results are based on 1000 data points and 1000 simulations. Solid line corresponds to the negative binomial prediction, while dashed line is for the Poisson forecast.

5. Case studies

In order to examine whether or not negative binomial is a better forecaster than the Poisson predictor for some real data applications, we use the machinery developed in Section 3, namely the nonrandomized PIT histogram, the marginal calibration plot and the various scoring rules. We fit the linear model (6) and the nonlinear model (9) to the following two different data collections introduced below.

5.1. *Measles data*

The left plot of Figure 4 shows the monthly number of measles at Sheffield for the period between 8 September 1978 and 17 April 1987. The total number of observations is 450. The right plot of Figure 4 shows the autocorrelation function of those data. Apparently, the autocorrelation plot shows the high degree of correlation among observations and hence we anticipate that either linear

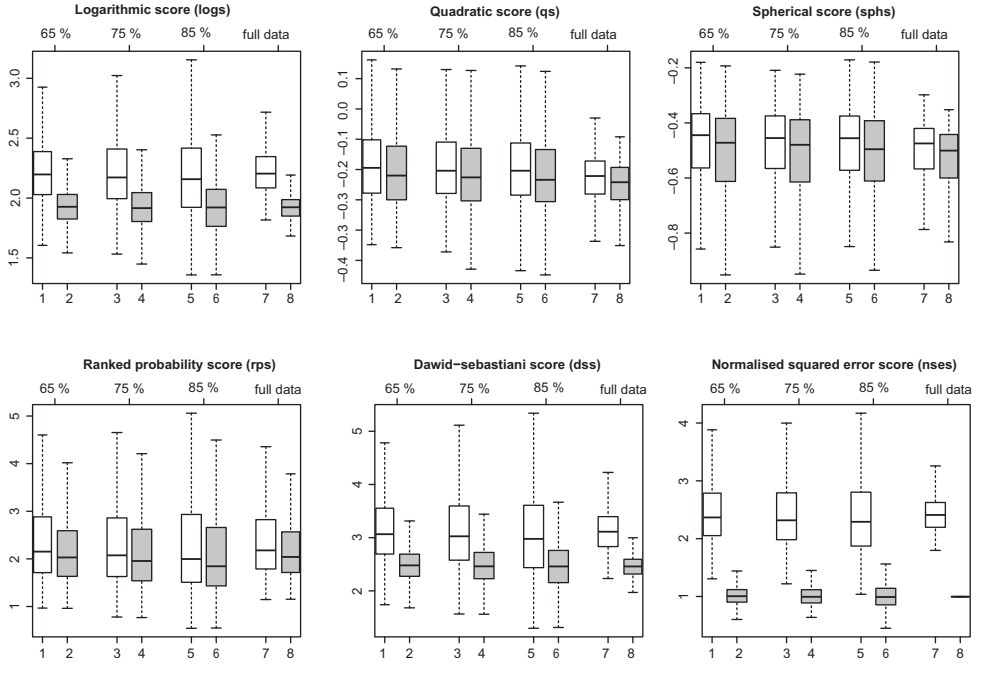

Figure 3. Boxplots for the mean scores given by Equations (13)–(18). White plots correspond to the Poisson forecasts, while grey plots correspond to the negative binomial prediction. Data are generated from the nonlinear model (9) when the true values are $(d, a, \gamma, b) = (0.5, 0.4, 0.5, 0.5)$ and $\nu = 2$. The results are based on 1000 data points and 1000 simulations.

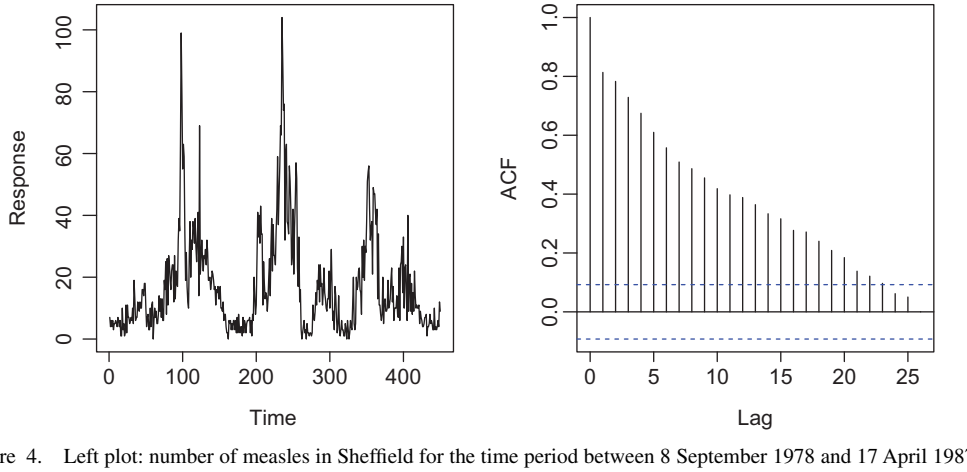

Figure 4. Left plot: number of measles in Sheffield for the time period between 8 September 1978 and 17 April 1987. Right plot: the autocorrelation function for the measles data.

or a nonlinear model would accommodate suitably those data. We fit the linear model (6) and the nonlinear model (9) to the data by using the quasi-likelihood methodology outlined in Section 2.2 and we obtain quasi-maximum likelihood estimators for the regression parameters. The results are summarized in Table 5. Together with the estimators of d, a and b, we give standard errors where the first row of standard errors are under the negative binomial distribution and the second line corresponds to the case of Poisson distribution. Standard errors for $\hat{\nu}_i$, $i = 1, 2$ have been computed by employing parametric bootstrap.

Table 5. Maximum likelihood estimators and their standards errors (in parentheses) for the linear model (6) and the nonlinear model (9) with $\gamma = 0.5$, for the monthly number of measles at Sheffield for the period between 8 September 1978 and 17 April 1987.

	Maximum likelihood estimators			Estimators of ν	
	$\hat{d}$	$\hat{a}$	$\hat{b}$	$\hat{\nu}_1$	$\hat{\nu}_2$
Linear model (6)	0.720	0.490	0.469	4.853	5.309
	(0.235)	(0.057)	(0.055)	(0.617)	(0.650)
	(0.122)	(0.024)	(0.023)		
Nonlinear model (9)	1.549	0.506	0.469	4.816	5.167
	(0.524)	(0.055)	(0.056)	(0.631)	(0.649)
	(0.292)	(0.023)	(0.022)		

Note: The total number of observations is 450.

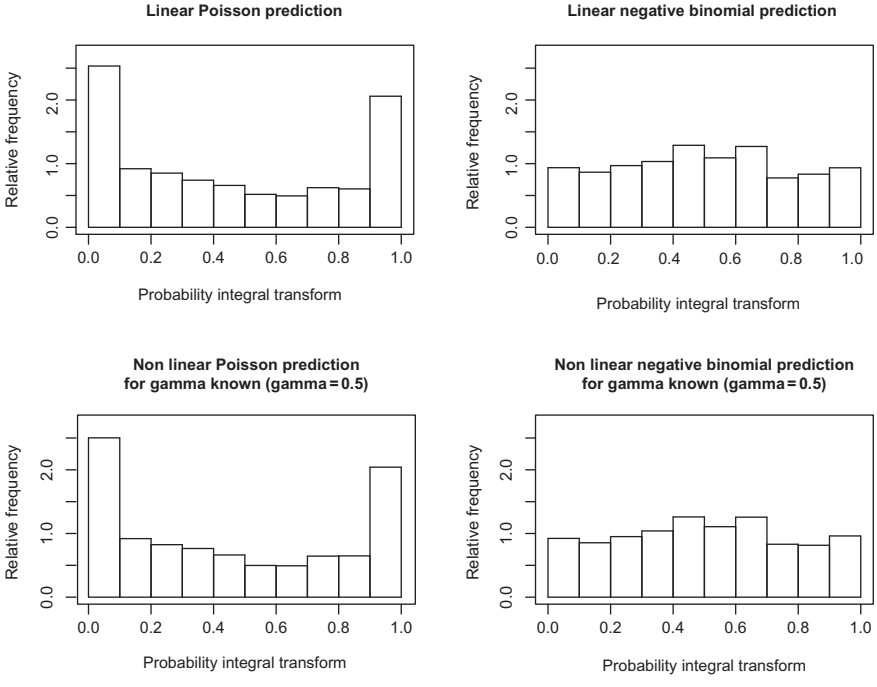

Figure 5. PIT histograms applied to the total number of measles in Sheffield for the time period between 8 September 1978 and 17 April 1987. From top to bottom: PIT histograms for model (6) and model (9) for $\gamma = 0.5$. Left plots: the conditional distribution is Poisson. Right plots: the conditional distribution is negative binomial.

Substituting these estimators to the expression of λ_t, we estimate the mean process. After obtaining $\hat{\lambda}_t$, we construct the PIT histograms, the marginal calibration plots and mean scores for the two models. The left plots of Figure 5 show the PIT histograms when the fit is based on Poisson distribution, for both linear and nonlinear models. Apparently, the plots show deviations from the Poisson distribution indicating underdispersed predictive distribution. The right plots indicate no apparent deviations from the uniformity; these plots are based on the negative binomial distribution.

Furthermore, to assess marginal calibration and sharpness of the prediction, we compute the scoring rules (13)–(19) and we construct the marginal calibration plot for these data. Table 6 shows the mean scores for the linear and nonlinear model. In addition, Figure 6 depicts the marginal

Table 6. Scoring rules calculated for the measles data after fitting the linear model (6) and the nonlinear model (9) for $\gamma = 0.5$.

		Scoring rules						
	Forecaster	logs	qs	sphs	rps	dss	nses	ses
Linear model (6)	Poisson	3.920	−0.015	−0.189	11.435	6.242	3.677	78.844
	Negative binomial	**3.254**	**−0.033**	**−0.197**	**11.160**	**4.853**	**0.993**	**78.844**
Nonlinear model (9)	Poisson	3.925	−0.014	−0.189	11.482	6.257	3.697	79.349
	Negative binomial	**3.255**	**−0.032**	**−0.196**	**11.213**	**4.854**	**0.993**	**79.349**

Notes: The two forecasters are compared by the mean logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, normalized squared error and squared error scores. Bold face numbers in each column indicate the minimum value obtained between the two forecasters.

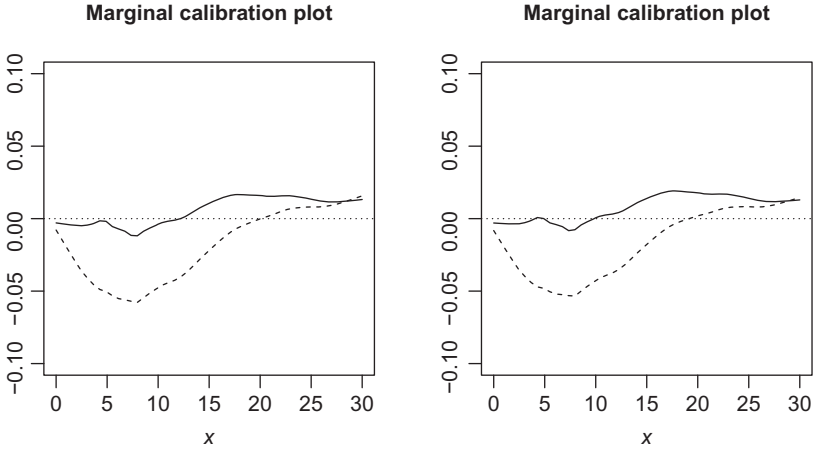

Figure 6. Left plot: marginal calibration plot for the measles data if we fit the linear model (6). Right plot: marginal calibration plot for the measles data if we fit the nonlinear model (9) for $\gamma = 0.5$. Solid line corresponds to the negative binomial prediction, while dashed line is for the Poisson forecast.

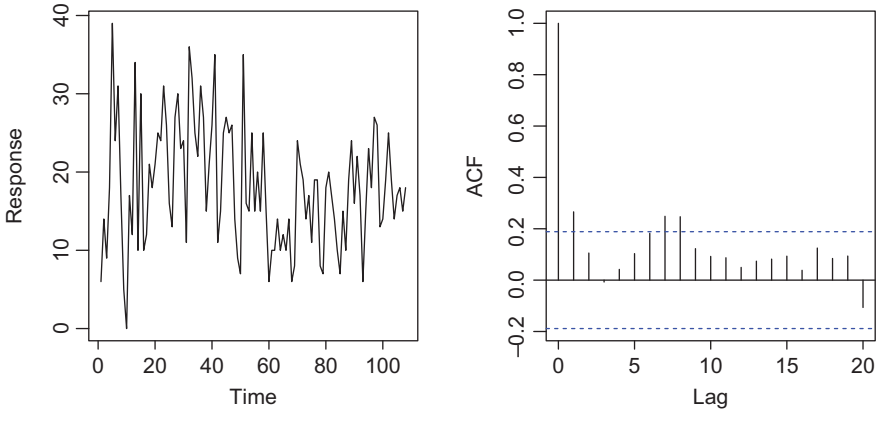

Figure 7. Left plot: number of monthly breech births in Edendale hospital of Pietermaritzburg in South Africa from February 1977 to January 1986. Right plot: the autocorrelation function for the births data.

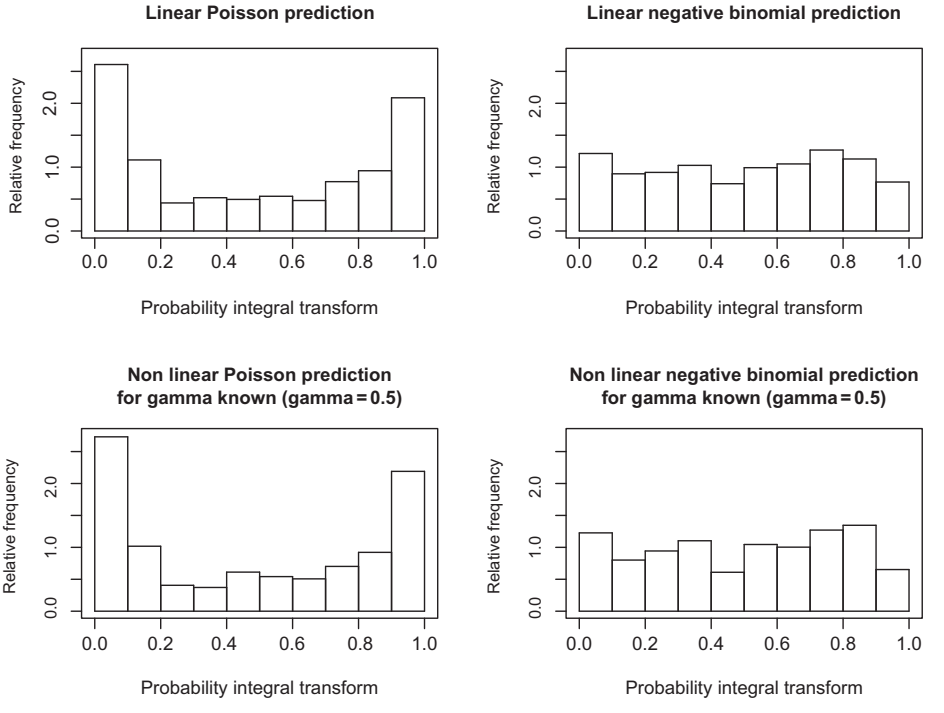

Figure 8. PIT histograms applied to the number of breech births in Edendale hospital from February 1977 to January 1986. From top to bottom: PIT histograms for model (6) and model (9) for $\gamma = 0.5$. Left plots: the conditional distribution is Poisson. Right plots: the conditional distribution is negative binomial.

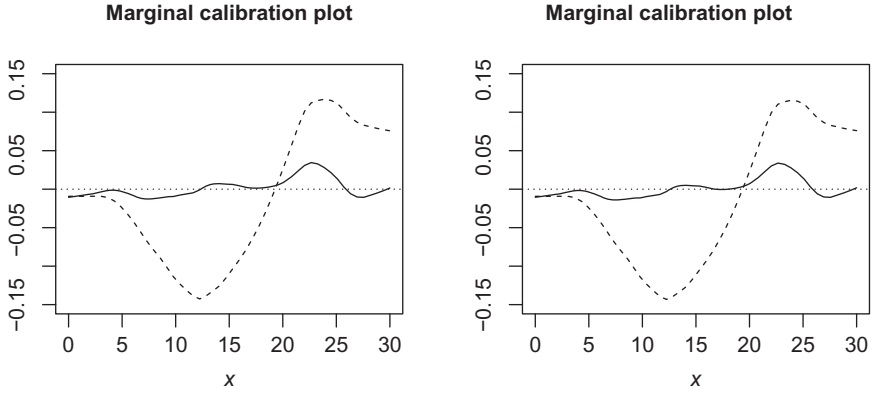

Figure 9. Left plot: marginal calibration plot for the breech births data if we fit the linear model (6). Right plot: marginal calibration plot for the births data if we fit the nonlinear model (9) for $\gamma = 0.5$. Solid line corresponds to the negative binomial prediction, while dashed line is for the Poisson forecast.

calibration plots for the measles data. The results in all cases indicate that the negative binomial distribution fits the data better than the Poisson distribution.

5.2. *Breech births data*

The second example to be considered is a count time series reported by Zucchini and MacDonald.[24] These data correspond to the number of monthly breech births in Edendale

Table 7. Maximum likelihood estimators and their standards errors (in parentheses) for the linear model (6) and the nonlinear model (9) if $\gamma = 0.5$, for the number of monthly breech births in Edendale hospital of Pietermaritzburg in South Africa from February 1977 to January 1986.

	Maximum likelihood estimators			Estimators of ν	
	$\hat{d}$	$\hat{a}$	$\hat{b}$	$\hat{\nu}_1$	$\hat{\nu}_2$
Linear model (6)	11.753	0.099	0.261	7.885	7.686
	(4.209)	(0.254)	(0.103)	(1.963)	(1.869)
	(2.406)	(0.144)	(0.055)		
Nonlinear model (9)	12.703	0.662	0.181	8.031	7.376
	(5.222)	(0.121)	(0.086)	(2.027)	(1.658)
	(3.138)	(0.069)	(0.046)		

Note: The total number of observations is 108.

Table 8. Scoring rules calculated for the breech births data after fitting the linear model (6) and the nonlinear model (9) for $\gamma = 0.5$.

		Scoring rules						
	Forecaster	Logs	qs	sphs	rps	dss	nses	ses
Linear model (6)	Poisson	3.970	−0.001	−0.151	4.754	6.109	3.208	57.232
	Negative binomial	**3.447**	**−0.030**	**−0.178**	**4.621**	**5.088**	**0.972**	**57.232**
Nonlinear model (9)	Poisson	3.984	−0.001	−0.151	4.674	6.102	3.199	57.647
	Negative binomial	**3.452**	**−0.030**	**−0.177**	**4.579**	**5.080**	**0.972**	**57.647**

Notes: The two forecasters are compared by the mean logarithmic, quadratic, spherical, ranked probability, Dawid–Sebastiani, normalized squared error and squared error scores. Bold face numbers in each column indicate the minimum value obtained between the two forecasters.

hospital of Pietermaritzburg in South Africa from February 1977 to January 1986. The size of this particular time series is $n = 108$. Figure 7 shows the data and the corresponding autocorrelation function. Here, we note that there is a reduced degree of autocorrelation among successive observations. Nevertheless, we can operate as in the previous example and following the same methodology, we obtain Figures 8 and 9 and Tables 7 and 8. The data analysis in this case shows again that the negative binomial prediction is superior to the Poisson forecaster.

Acknowledgements

We would like thank a reviewer for several constructive comments on a previous version. Work supported by Cyprus Research Promotion Foundation TEXNOLOGIA/THEPIS/0609(BE)/02.

References

[1] Czado C, Gneiting T, Held L. Predictive model assessment for count data. Biometrics. 2009;65:1254–1261.
[2] Dawid AP. Statistical theory: the prequential approach. J R Stat Soc Ser A. 1984;147:278–292.
[3] Fokianos K, Rahbek A, Tjøstheim D. Poisson autoregression. J Am Stat Assoc. 2009;104:1430–1439.
[4] Jung R, Tremayne A. Useful models for time series of counts or simply wrong ones? AStA Adv Stat Anal. 2011;95: 59–91.
[5] Mikosch T. Non-life insurance mathematics. 2nd ed. An introduction with the Poisson process. Berlin: Springer-Verlag; 2009
[6] Christou V, Fokianos K. Quasi-likelihood inference for negative binomial time series models; under review.
[7] Lawless JF. Negative binomial and mixed Poisson regression. Can J Stat. 1987;15:209–225.
[8] Cameron AC, Trivedi PK. Regression analysis of count data. 1st ed. Cambridge: Cambridge University Press; 1998.
[9] Zhu F. A negative binomial integer-valued GARCH model. J Time Series Anal. 2011;32:54–67.
[10] Ferland R, Latour A, Oraichi D. Integer-valued GARCH processes. J Time Series Anal. 2006;27:923–942.
[11] Neumann M. Absolute regularity and ergodicity of Poisson count processes. Bernoulli. 2011;17:1268–1284.

[12] Doukhan P, Fokianos K, Tjøstheim D. On weak dependence conditions for Poisson autoregressions. Stat Probab Lett. 2012;82:942–948.
[13] Davis RA, Liu H. Theory and inference for a class of observation-driven models with application to time series of counts; 2012. Available from: ArXiv e-prints, http://adsabs.harvard.edu/abs/2012arXiv1204.3915D
[14] Gao J, King M, Lu Z, Tjøstheim D. Specification testing in nonlinear and nonstationary time series regression. Ann Stat. 2009;37:3893–3928.
[15] Fokianos K, Neumann MH. A goodness-of-fit test for Poisson count processes. Electron J Stat. 2013;7:793–819.
[16] Berkes I, Horváth L, Kokoszka P. GARCH processes: structure and estimation. Bernoulli. 2003;9:201–227.
[17] Francq C, Zakoïan J-M. Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. Bernoulli. 2004;10:605–637.
[18] Mikosch T, Straumann D. Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: a stochastic recurrence equations approach. Ann Stat. 2006;34:2449–2495.
[19] Bardet J-M, Wintenberger O. Asymptotic normality of the quasi-maximum likelihood estimator for multidimensional causal processes. Ann Stat. 2010;37:2730–2759.
[20] Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986;42: 121–130.
[21] Godambe VP, Heyde CC. Quasi-likelihood and optimal estimation. Int Stat Rev. 1987;55:231–244.
[22] Heyde CC. Quasi-likelihood and its applications: a general approach to optimal parameter estimation. New York: Springer; 1997.
[23] Gneiting T, Balabdaoui F, Raftery AE. Probabilistic forecasts, calibration and sharpness. J R Stat Soc Ser B (Stat Methodol). 2007;69:243–268.
[24] Zucchini W, MacDonald LI. Hidden Markov models for time series. Boca Raton, FL: CRC Press; 2009.