ENVIRONMETRICS
Environmetrics, **10**, 395–411 (1999)

INTEGER VALUED AUTOREGRESSIVE MODELS FOR TIPPING BUCKET RAINFALL MEASUREMENTS

PETER THYREGOD,[1]* JACOB CARSTENSEN,[2] HENRIK MADSEN[1]
AND KARSTEN ARNBJERG-NIELSEN[2]

[1]*Department of Mathematical Modelling, Technical University of Denmark, 2800 Lyngby, Denmark*
[2]*Department of Environmental Science and Engineering, Technical University of Denmark, 2800 Lyngby, Denmark*

SUMMARY

A new method for modelling the dynamics of rain sampled by a tipping bucket rain gauge is proposed. The considered models belong to the class of integer valued autoregressive processes. The models take the autocorrelation and discrete nature of the data into account. A first order, a second order and a threshold model are presented together with methods to estimate the parameters of each model. The models are demonstrated to provide a good description of data from actual rain events requiring only two to four parameters. Copyright © 1999 John Wiley & Sons, Ltd.

KEY WORDS tipping bucket rain gauge; integer valued autoregressive model; INAR model

1. INTRODUCTION

Rainfall is an important process for the water cycle, and in urban areas the rainfall process interferes with many basic functions of today's society. From an environmental engineering point of view, a better understanding of the rainfall process is believed to lead to less flooding problems in the sewer system, reduced pollution discharge and improved efficiency of wastewater treatment plants.

The two most important aspects of modelling rain is forecasting and simulation. Forecasting can be used in real time control of urban hydrological systems with the objective of minimising peak loads. Such on-line control systems may require short-term rainfall forecasts as input. For most sewer systems the desired prediction horizon is between 30 and 60 min. Another application of rainfall models is design and analysis of urban drainage systems where there is a need for better understanding and description of the rainfall process. Long series of rainfall are taken as input to rainfall–runoff models to simulate long series of runoff, flow, water level and overflow to produce long-term extreme statistics. The simulated model output is analysed, and the system performance evaluated. However, there is a serious shortage of high-resolution measured rain series sufficiently long to produce long-term extreme statistics. Simulated rain series could be used with advantage. Before starting to simulate 30 years of rain data, a good understanding of single rain events is crucial.

* Correspondence to: P. Thyregod, Department of Mathematical Modelling, Technical University of Denmark, 2800 Lyngby, Denmark

CCC 1180–4009/99/040395–17$17·50

Received 5 March 1998
Accepted 12 December 1998

Most of the rain data collection is due to agricultural needs and for prediction of flooding in rivers. Hourly, or even daily, measurements are quite sufficient for these purposes. However, the temporal aggregation of rain data for use in urban hydrology should not exceed, say 5–10 min. Rain data is most often collected by means of a tipping bucket rain gauge. A tipping bucket rain gauge is a discrete sampler counting the number of times a bucket is filled in each sampling time interval. Knowing the volume of the bucket, the rain event can be reconstructed from these counts. Thus, when the rain is sampled in this way, the rain event is represented as a time series of counts.

In statistical time series literature very little attention has been given to the modelling of time series of counts. Most of the literature that does exist is concerning theoretical properties of such models, see e.g. MacDonald and Zucchini (1997). However, only very few applications of such models besides discrete time Markov chain models have been published. In the present paper an integer valued autoregressive model proposed by Al-Osh and Alzaid (1987) will be presented as a means of modelling tipping bucket rain data.

In Arnbjerg-Nielsen (1996) the modelling of single rain events was considered. The approach that was followed was to consider waiting times between consecutive tips of the bucket. These waiting times were modelled using traditional statistical time series models. One approach was using ARIMA models on the logarithm of the waiting times. Another was a full discrete time Markov chain with about 40 states representing the different waiting times.

2. RAIN DATA

In Denmark a large monitoring program was initiated in 1979 by the Danish Committee on Water Pollution Control (in Danish: Spildevandskomiteen). The objective was to obtain high resolution data on the rainfall process which could be used as input to hydrological and hydraulic models for simulating extreme events in the sewer system. For this reason a number of rain gauges of the tipping bucket type (see Figure 1) have been scattered over the country in order to obtain information on the regional variation as well. For a more detailed description see Harremoes and Mikkelsen (1995).

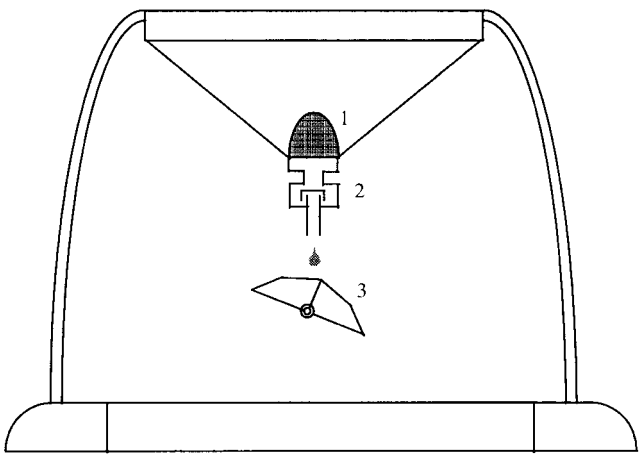

Figure 1. Tipping bucket rain gauge

A tipping bucket rain gauge operates by means of a pair of buckets. The basic principle of a tipping bucket rain gauge is as follows. The rain enters the gauge through the funnel, whereafter it goes through a filter (1) into a syphon (2). The purpose of the syphon is to ensure that the water always enters the bucket assembly (3) with the same momentum. When the syphon is full, all the water from the syphon enters the bucket assembly. The water enters one of the two buckets at a time, which overbalances directing the water into the second bucket. The flip-flop motion of the tipping buckets is transmitted to the recording device and provides a measure of rainfall intensity. The number of tips from the bucket assembly is registered with a 1 min sampling frequency.

In Denmark two types of rainfall are predominant. Convective rain is high intensity rainfall usually of rather short duration, as seen for example during thunderstorms. An example of a typical convective rain event is shown in Figure 2. The other common type of rain is frontal rain. Frontal rain usually give rise to longer rain events with almost constant rain intensities, as seen for example during the passing of a front. An example of a typical frontal rain event is shown in Figure 3. Rain events are rarely pure in the sense that they can uniquely be said to belong to one of the two types mentioned previously. They often reflect behaviour of both types of rain.

The data used for testing the proposed class of models originates from tipping bucket rain gauge 20211 in Aalborg, Denmark. A total of 39 single rain events are considered, 16 of which are classified as being convective and the other 23 are classified as being frontal. The classification was carried out manually by a meteorological expert. Events reflecting both types of behaviour have mostly been classified as convective. A manual classification is both time-consuming and subjective, which may result in very different classifications depending on the expert and more important climatic differences between countries. The rain events cover a time span of 16 years; the first rain event is from 1979 and the last is from 1995. They have been selected using either extreme intensities (over 9 μm/s) or depths over 20 mm as criterion.

The 16 convective events selected have durations of between 75 and 564 min, and have depths in the interval [9·0; 47·4] mm. The 23 frontal events selected have durations between 308 and 1460 min, and have depths in the interval [20·2; 72·4] mm.

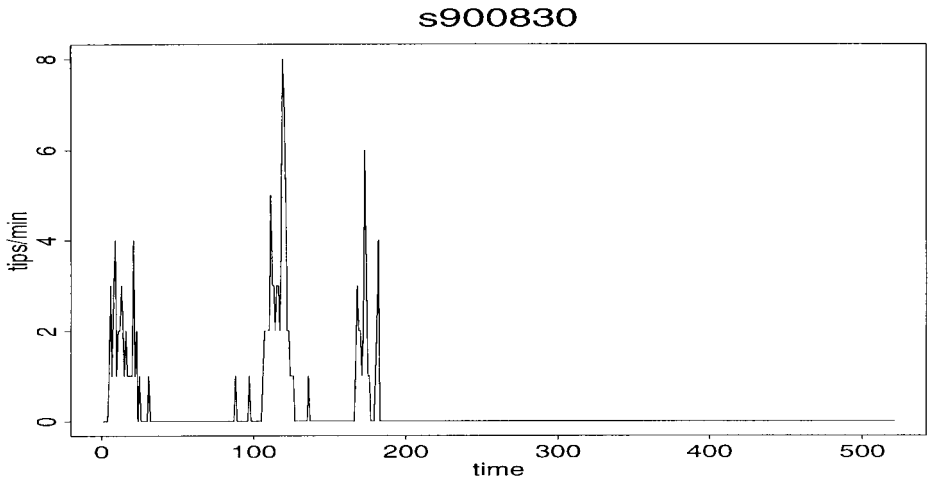

Figure 2. Convective rain event from August 30th 1990

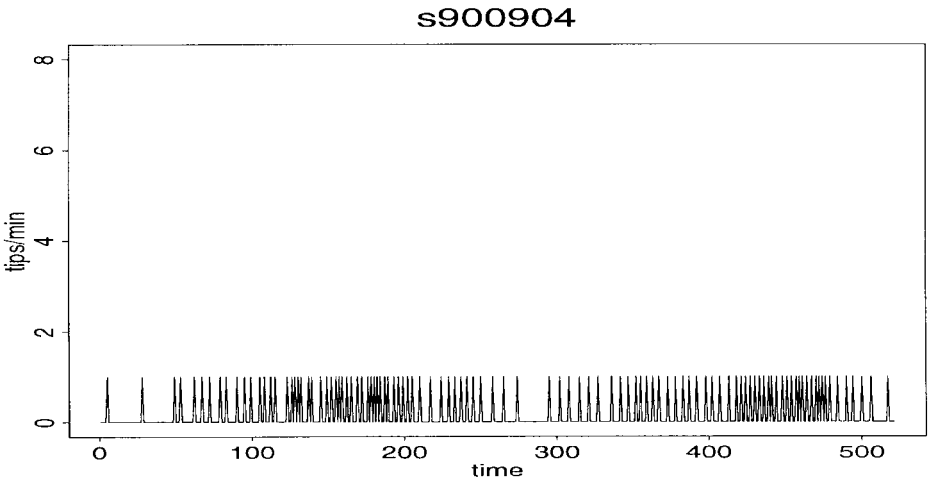

Figure 3. Frontal rain event from September 4th 1990

3. THE INAR CLASS OF MODELS

In time series analysis, modelling of processes with a continuous output space has gained the most attention. However, measurements are generally stored and processed in a discrete form. For the majority of applications the resolution of data is sufficiently good to justify the choice of a continuous probability distribution function. In the case of rainfall measurements from a tipping bucket rain gauge, data have a distinctive discrete output space requiring the use of discrete marginal distributions in the models. One model to consider for this type of data is the INteger valued AutoRegressive (INAR) process introduced by (Al-Osh and Alzaid 1987).

Although the properties of the INAR class of models have been studied extensively, it seems that only few applications have yet been published. One application of the INAR class of models is found in Franke and Seligmann (1993), where daily counts of epileptic seizures in one patient are considered.

In order to present the INAR class of models, first the definition of the $\circ$-operator will be presented. Let X be a non-negative integer valued random variable, then for any $\alpha \in [0, 1]$ the operator $\circ$ is defined by

$$\alpha \circ X = \sum_{i=1}^{X} Y_i$$

where Y_i is a sequence of iid random variables, independent of X, satisfying:

$$P\{Y_i = 1\} = 1 - P\{Y_i = 0\} = \alpha$$

In the following the INAR process of order 1 will be presented and further extended to include INAR-processes of order p.

3.1. The INAR process of order 1

Now the INAR(1) process can be defined for $t = 1, 2, \ldots$ as

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t \tag{1}$$

where $\alpha \in [0, 1]$, and $\{\varepsilon_t\}$ is a sequence of uncorrelated non-negative integer valued random variables with mean μ and variance σ^2. In this work the innovation, $\{\varepsilon_t\}$, is assumed to be a Poisson process, thus $\mu = \sigma^2 = \lambda$.

The form of the INAR(1) model is analogous to that of the standard AR(1) model with the scalar-multiplication replaced by the $\circ$ operator.

The INAR(1) model defined by equation (1) simply states that the count of the process at time t is the sum of the survivors of the process at time $t - 1$, each with a probability of survival α, and the new elements that entered the system in the interval $[t - 1, t]$.

The following results are of importance when calculating the variance and autocorrelation function of the process. The marginal distribution of the model (1) can be expressed in terms of the innovation sequence $\{\varepsilon_t\}$

$$X_t = \sum_{j=0}^{\infty} \alpha^j \circ \varepsilon_{t-j} \tag{2}$$

Another important property of the INAR(1) process is

$$\begin{aligned} X_t &= \alpha \circ (\alpha \circ (\cdots (\alpha \circ X_{t-k} + \varepsilon_{t-k+1}) \cdots)) \\ &= \alpha^k \circ X_{t-k} + \sum_{j=0}^{k-1} \alpha^j \circ \varepsilon_{t-j} \end{aligned} \tag{3}$$

These results are similar to those obtained for the ordinary AR(1) process. Another important property of the INAR(1) process is that X_t follows a Poisson distribution (Al-Osh and Alzaid (1987) if and only if ε_t follows a Poisson distribution. Consequently, the role which the distribution of ε_t plays in determining the distribution of X_t in the INAR(1) process is similar to the role played by the normal distributed errors in the AR(1) process.

The conditional mean and variance of the INAR(1) process, $\{X_t\}$, are directly found to be

$$E\{X_t | X_{t-1}\} = E\{\text{Bin}(X_{t-1}, \alpha) + \text{Po}(\lambda)\} = \sigma X_{t-1} + \lambda \tag{4}$$

$$\text{var}\{X_t | X_{t-1}\} = \alpha(1 - \alpha) X_{t-1} + \lambda \tag{5}$$

The unconditional mean of the INAR(1) process is then relatively easy found by recursive application of (4)

$$E\{X_t\} = \alpha E\{X_{t-1}\} + \lambda = \alpha^t E\{X_0\} + \lambda \sum_{j=0}^{t-1} \alpha^j$$

For large t this is equal to

$$E\{X_t\} = \alpha^t E\{X_0\} + \lambda \sum_{j=0}^{t-1} \alpha^j \simeq \lambda \frac{1}{1-\alpha}$$

Now the variance of the INAR(1) process can be found by applying the following text-book result

$$\text{var}\{X\} = E[\text{var}\{X|Y\}] + \text{var}[E\{X|Y\}]$$

This yields the following expression for the unconditional variance

$$\begin{aligned}\text{var}\{X_t\} &= \alpha^2 \text{var}\{X_{t-1}\} + \alpha(1-\alpha)E\{X_{t-1}\} + \lambda \\ &= \alpha^{2t}\text{var}\{X_0\} + (1-\alpha)\sum_{j=1}^{t} \alpha^{2j-1} E\{X_{t-j}\} + \lambda \sum_{j=1}^{t} \alpha^{2(j-1)}\end{aligned}$$

Even though the INAR(1) process possesses some of the same properties as the AR(1) process, the estimation of the parameters in an INAR(1) model is more complicated than is the case for an AR(1) model.

3.2. The INAR process of order *p*

A natural extension of the INAR(1) model is to extend the dependency to include lags of higher order than one.

The pth order integer autoregressive model (INAR(p)) is defined as

$$X_t = \sum_{i=1}^{p} \alpha_i \circ X_{t-i} + \varepsilon_t \tag{6}$$

where $\alpha_i \in [0, 1]$ and $\varepsilon_t \in \text{Po}(\lambda)$.

Though the form of the INAR(p) Model is very similar to that of an ordinary AR(p) model some important differences exist. The dependency across time of the $\circ$-operator is the reason for these differences.

The mutual dependence structure between the components of X_t, i.e.$\alpha \circ X_{t-i}$, $i = 1, 2, \ldots, p$ appearing at different times induces a moving-average structure into the process. In fact, it can be shown that the behaviour of the autocorrelation function of the INAR(p) process behaves like that of the ordinary ARMA($p, p-1$) process. See Alzaid and Al-Osh (1990) for details.

3.3. Self-exciting threshold INAR model

As mentioned previously, some of the rain events change during the event from being frontal to being convective. These events may lead to rather bad parameter estimates and possible misclassification if an INAR(1) model is used for classification. By combining two INAR(1) models in a threshold model, it is hoped that these critical rain events may be divided into more homogeneous parts. As an indicator of what regime to choose we observe the number of tips during the previous two 10-min samples.

Thus, the Self-exciting threshold INAR(1) (SETINAR) model can be formulated as

$$X_t = \begin{cases} \alpha_1 \circ X_{t-1} + \varepsilon_{1,t} \text{ for } \sum_{j=1}^{2} x_{t-j} \leqslant b \\ \alpha_2 \circ X_{t-1} + \varepsilon_{2,t} \text{ for } \sum_{j=1}^{2} x_{t-j} > b \end{cases}$$

where $\varepsilon_{1,t} \in \text{Po}(\lambda_1)$ and $\varepsilon_{2,t} \in \text{Po}(\lambda_2)$.

4. ESTIMATION

The estimation of INAR processes is more complicated than the straightforward estimation of AR-processes. This is due to that the conditional distribution of x_t given, $x_{t-1}, \ldots, x_{t-p}$ in the INAR(p) process is the convolution of the Poisson distribution of ε_t and p binomial distributions with scale parameters α_i and index parameter x_i. Thus, the conditional distribution of X_t in an INAR(p) process is the convolution of $p + 1$ distributions, which for large p requires a lot of computation. In the following, estimation procedures for the INAR(1), INAR(2) and SETINAR processes are given.

4.1. INAR(1)

In this section three methods for estimating the parameters in the simplest INAR model are presented. It is shown how the two parameters – α and λ – can be estimated from the moments of the process, least squares and likelihood function. For the two first methods, the similarity to the AR(1) is obvious. In the maximum likelihood method, the marginal distribution will be different.

4.1.1. Yule–Walker estimation

For any non-negative integer k the covariance at lag k, $\gamma(k)$, is

$$\begin{aligned} \gamma(k) = \text{cov}\left(X_t, X_{t-k}\right) &= \text{cov}(\alpha^k \circ X_{t-k} + \sum_{i=0}^{k-1} \alpha^i \circ \varepsilon_{t-i}, X_{t-k}\Big) \\ &= \text{cov}(\alpha^k \circ X_{t-k}, X_{t-k}) + \text{cov}\left(\sum_{i=0}^{k-1} \alpha^i \circ \varepsilon_{t-i}, X_{t-k}\right) \\ &= \alpha^k \text{var}\{X_{t-k}\} + \sum_{i=0}^{k-1} \text{cov}(X_{t-k}, \varepsilon_{t-j}) \\ &= \alpha^k \text{var}\{X_{t-k}\} = \alpha^k \gamma(0) \end{aligned} \tag{7}$$

the last equality is caused by the fact that $\text{cov}(X_{t-k}, \varepsilon_{t-j}) = 0$ for $j < k$.

Hence, the Yule–Walker estimator for α can be found from

$$\alpha = \frac{\gamma(1)}{\gamma(0)} \tag{8}$$

Thus, by replacing $\gamma(1)$ with the sample auto covariance function and $\gamma(0)$ with the sample variance, an estimator for α is

$$\hat{\alpha} = \frac{\sum_{t=0}^{n-1}(x_t - \bar{x})(x_{t+1} - \bar{x})}{\sum_{t=0}^{n}(x_t - \bar{x})^2} \tag{9}$$

An estimate for λ can be obtained by calculating $\hat{\varepsilon}_t = x_t - \hat{\alpha}x_{t-1}$. As ε_t is assumed to follow a Poisson distribution, a reasonable estimator for λ is

$$\hat{\lambda} = \frac{1}{n}\sum_{t=1}^{n}\hat{\varepsilon}_t \tag{10}$$

4.1.2. Conditional least squares

In (4) the conditional mean of X_t given X_{t-1} was found as

$$E\{X_t|X_{t-1}\} = \alpha X_{t-1} + \lambda$$

The conditional least squares (CLS) estimation is based on minimisation of the sum of squared derivations from the conditional expectation. Thus, the CLS estimates for α and λ are those values which minimise

$$S(\alpha, \lambda) = \sum_{t=1}^{n}[X_t - (\alpha X_{t-1} + \lambda)]^2 \tag{11}$$

Evaluation of $\partial S/\partial\alpha = 0$ and $\partial S/\partial\lambda = 0$ give the following estimators for α and λ

$$\hat{\alpha} = \frac{\Sigma x_t x_{t-1} - (\Sigma x_t \Sigma x_{t-1})/n}{\Sigma x_{t-1}^2 - (\Sigma x_{t-1})^2/n} \tag{12}$$

$$\hat{\lambda} = \frac{1}{n}(\Sigma x_t - \hat{\alpha}\Sigma x_{t-1}) \tag{13}$$

where all sums are over the interval $[0, n]$.

In Madsen (1995) it is shown that the conditional least squares estimate and the Yule–Walker estimate are asymptotically identical. All rain series used in the present work have first and last observation equal to zero. When this is the case, the two estimates will be exactly identical.

It should be stressed that Yule–Walker and Conditional Least Squares estimation can lead to negative values of α. This will violate the basis of the INAR model, the $\circ$ operator.

4.1.3. Maximum likelihood

The conditional density for an INAR(1) model is given by the convolution of a binomial and a Poisson distribution:

$$f_1(i) = \binom{x_{t-1}}{i}\alpha^i(1-\alpha)^{x_{t-1}-i} \tag{14}$$

$$f_2(i) = \frac{\lambda^i}{i!}\exp(-\lambda) \tag{15}$$

$$g(x_t|x_{t-1}) = f_1 * f_2 = \sum_{i=0}^{\infty} f_1(i)f_2(x_t - i) = \exp(-\lambda)\sum_{i=0}^{\min(x_t,x_{t-1})}\frac{\lambda^{x_t-i}}{(x_t-i)!}\binom{x_{t-1}}{i}\alpha^i(1-\alpha)^{x_{t-1}-i} \tag{16}$$

Thus, the likelihood function becomes

$$L(\alpha, \lambda; \boldsymbol{x}) = \prod_{t=1}^{n} g(x_t|x_{t-1}) \tag{17}$$

The estimates of α and λ are found as those values of α and λ that minimises the negative logarithm of the likelihood function.

$$\log L(\alpha, \lambda; \boldsymbol{x}) = \sum_{t=1}^{n} g(x_t|x_{t-1}) \tag{18}$$

4.2. INAR(2)

Below a method for maximum likelihood estimation in the INAR(2) case is presented. This is an extension to the work of Alzaid and Al-Osh (1990). The density of the INAR(2) process is the convolution of two binomial distributions and a Poisson distribution. The two binomial distributions have number of trials as X_{t-1} and X_{t-2} respectively.

$$\begin{aligned} f_1(i) &= \binom{x_{t-1}}{i}\alpha_1^i(1-\alpha_1)^{x_{t-1}-i} \\ f_2(i) &= \binom{x_{t-2}}{i}\alpha_2^i(1-\alpha_2)^{x_{t-2}-i} \\ f_3(i) &= \frac{\lambda^i}{i!}\exp(-\lambda) \end{aligned} \tag{19}$$

The density for $U = \text{Bin}(x_{t-1}, \alpha_1) + \text{Bin}(x_{t-2}, \alpha_2)$ is:

$$\begin{aligned} h(u) = f_1 * f_2 &= \sum_{i=0}^{\infty} f_1(i)f_2(u-i) \\ &= \sum_{i=\max(0,u-x_{t-2})}^{\min(u,x_{t-1})}\binom{x_{t-1}}{i}\binom{x_{t-2}}{u-i}\alpha_1^i\alpha_2^{u-i} \times (1-\alpha_1)^{x_{t-1}-i}(1-\alpha_2)^{x_{t-2}-(u-i)} \end{aligned} \tag{20}$$

The limits of the summations are found as those values of i for which f_1 and f_2 are defined

$$\left.\begin{array}{l} \binom{x_{t-1}}{i} \Rightarrow x_{t-1} \geqslant i \\ \binom{x_{t-2}}{u-i} \Rightarrow x_{t-2} \geqslant u-i \Leftrightarrow i \geqslant u - x_{t-2} \\ \binom{x_{t-2}}{u-i} \Rightarrow u - i \geqslant \ \Leftrightarrow i \leqslant u \end{array}\right\} \Rightarrow \tag{21}$$

$$i \in [\max(0, u - x_{t-2}); \min(u, x_{t-1})] = I_1$$

The resulting density is that of $X = U + \mathrm{Po}(\lambda)$ which becomes

$$\begin{aligned} g(x|x_{t-1}, x_{t-2}) = (f_1 * f_2) * f_3 &= \sum_{j=0}^{\infty} h(j) f_3(x-j) \\ &= \exp(-\lambda) \sum_{j=0}^{x} \frac{\lambda^{x-j}}{(x-j)!} \sum_{i \in I_1} \binom{x_{t-1}}{i} \binom{x_{t-2}}{j-i} \\ &\quad \times \alpha_1^i \alpha_2^{j-i} (1-\alpha_1)^{x_{t-1}-i} (1-\alpha_2)^{x_{t-2}-(j-i)} \end{aligned} \tag{22}$$

Thus, giving the following likelihood function

$$L(\alpha_1, \alpha_2, \lambda; \boldsymbol{x}) = \prod_{t=1}^{n} g(x_t|x_t|x_{t-1}, x_{t-2}) \tag{23}$$

For $\alpha_2 = 0$ the terms of the inner sum of the density $g(\,)$ are seen to be different from zero only for $i = j$, in which case the value is

$$\binom{x_{t-1}}{j} \binom{x_{t-2}}{0} \alpha_1^j \alpha_2^0 (1-\alpha_1)^{x_{t-1}-j} 1^{x_{t-1}-(x_t-j)} = \binom{x_{t-1}}{j} \alpha_1^j (1-\alpha_1)^{x_{t-1}-j} \tag{24}$$

The right-hand side of (24) are exactly the components of the sum for the INAR(1) model, thus making the INAR(1) model a true sub-model of the INAR(2) model.

4.3. SETINAR

The number of tips in a 20-min interval is used to determine the appropriate regime. Thus, the following two index sets are defined

$$T_1 = \left\{ t | \sum_{i=1}^{2} x_{t-i} \leqslant b \right\}$$

and

$$T_2 = \left\{ t | \sum_{i=1}^{2} x_{t-i} > b \right\}$$

Let $g(x_t|x_{t-1}, \alpha_1, \lambda_1)$ denote the one-step conditional density of the INAR(1) process. The likelihood function for each of the two regimes is

$$L(\alpha_1, \lambda_1; \boldsymbol{x}) = \prod_{t \in T_1} g(x_t|x_{t-1}, \alpha_1, \lambda_1)$$

and

$$L(\alpha_2, \lambda_2; \boldsymbol{x}) = \prod_{t \in T_2} g(x_t|x_{t-1}, \alpha_2, \lambda_2)$$

Hence, the joint likelihood function of all observations may be expressed as

$$L(\alpha_1, \alpha_2, \lambda_1, \lambda_2; \boldsymbol{x}) = L(\alpha_1, \lambda_1; \boldsymbol{x}) \times L(\alpha_2, \lambda_2; \boldsymbol{x}) \tag{25}$$

5. RESULTS

The parameters of the INAR(1) model have been estimated for each of the 39 rain events by optimising the maximum likelihood function derived in the previous chapter. Only the maximum likelihood is used as this makes it possible to compare the results of the three models. The results are presented in plots showing connected values of the two parameters. Those rain events classified as being frontal have been marked with a 'F' and those classified as being convective are marked with a 'C'.

5.1. INAR(1)

When estimating the parameters of the INAR(1) model using series of one-minute sampled values it was found that these series do not contain enough information for modelling in the time domain. Many events do not have any minute-to-minute correlation that is worth considering. Hence, all the series have been aggregated into 10 min sampling intervals. Aggregating data corresponds to a low pass filtering making the fast variations less dominating, thus bringing out the slower dynamics of the process.

The parameter space is restricted to $\alpha \in [0; 1]$. The expression for the logarithm of the likelihood found in (18) is also seen only to be defined for values of α in this interval. The numerical optimisation routine used for minimising the negative log-likelihood function is called NPSOL. NPSOL is a procedure for constrained optimisation and uses a quadratic programming algorithm with a BFGS quasi-Newton update of the Hessian.

In Figure 4 the maximum likelihood estimates are shown for the aggregated series. It is noticed that there is still a noticeable separation of the events. When the fast variations are no longer dominating, the rain events classified as being frontal seem to be associated with a higher degree of memory than the rain classified as being convective. The opposite is the case for the innovation

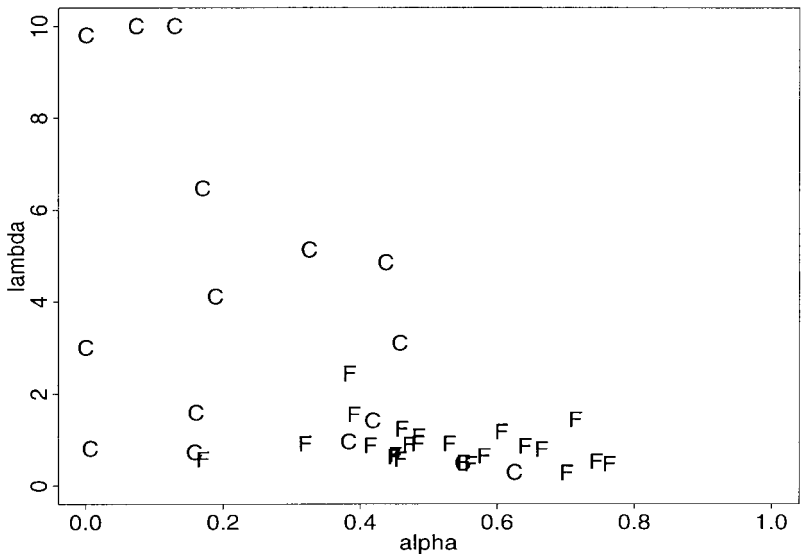

Figure 4. ML estimates of aggregated series

parameter, λ, where the convective events tend to give rise to higher estimates than frontal events. As the innovation parameter, λ, describes what can not be described by the carry-over effect, it can be claimed that the frontal events are better described by the INAR(1) model than the convective events.

These findings are in very good agreement with the common meteorological understanding that convective rain events are innovation processes, whereas frontal rain events have more memory. The parameter estimates from a rain event provide a good means for characterising the event. The above findings indicate that classification of rain can be carried out by evaluating the parameters in an INAR(1) model.

As the results from using 10-min aggregated data seems reasonable, in the rest of this chapter, where extensions of the INAR(1) model will be considered, only 10 min aggregated values will be used.

5.2. INAR(2)

When estimating the parameters of the INAR(2) model, it was found that only half the events gave rise to estimates of α_2 that were significantly different from zero.

In Figure 5 the likelihood ratios against INAR(1) are shown. Naturally the events having α_2s not significantly different from zero will have likelihood ratio test statistics close to zero, which can also be seen from the figure. Only very few of the events are found to give a significantly better description of the data at a 95% level.

5.3. SETINAR

For the SETINAR model, first the threshold has to be estimated. Utilising the discrete nature of the problem and the fact that the sum of some maxima is again a maximum, this is solved using a

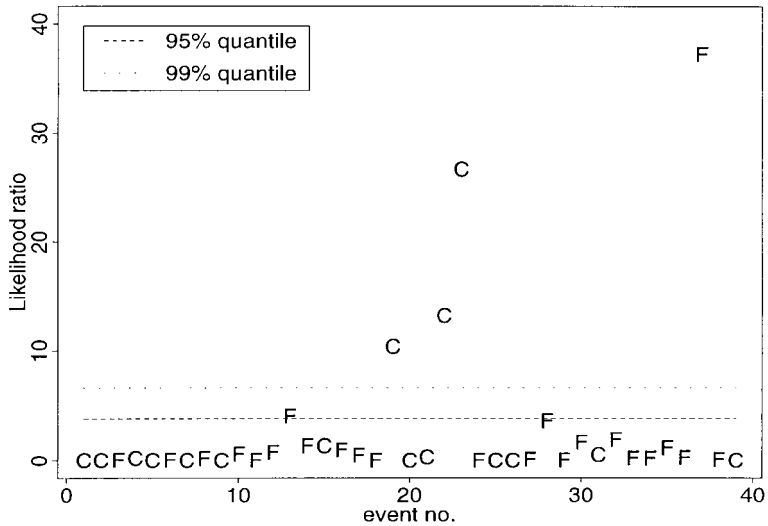

Figure 5. Likelihood ratios for INAR(2)/INAR(1)

Table I. Maximised likelihood functions for seven different threshold values

Threshold	$-\Sigma \log L$
$b = 6$	3218·454
$b = 7$	3214·453
$b = 8$	3205·839
$b = 9$*	3203·911
$b = 10$	3212·592
$b = 11$	3214·339
$b = 12$	3215·469

maximum likelihood approach. For a given value of the threshold, the sum of the logarithm of the maximised likelihood functions for all events is calculated. This is done for seven different values of the threshold. This resulted in the values in Table I, hence a threshold of nine tips in the past two 10-min samplings was chosen.

The resulting parameter estimates are shown in Figures 6 and 7. Now there are four parameters for each event. In Figure 6 the parameters have been marked with the type of rain and in Figure 7 they have been marked with '1' and '2' according to the character of the regime.

The parameters marked with a '2' belong to the regime determined by $\Sigma_{i=1}^{2} x_{t-i} > 9$. The parameters of that group are characterised by a higher innovation, just as expected for the convective parts of the rain events. The tendency, however, is not as clear as anticipated. One reason for this might be that some of the events do not change regime at all, thus giving rise to unreliable parameter estimates. Another reason might be that many of the convective events when aggregated have become rather short. Roughly 10 observations are rather on the small side for estimating four parameters.

In Figure 8 the likelihood ratios are shown. The values should be compared with quantiles of a $\chi^2(2)$ distribution. The 95% and 99% quantiles are 5·991 and 9·210, respectively. As the

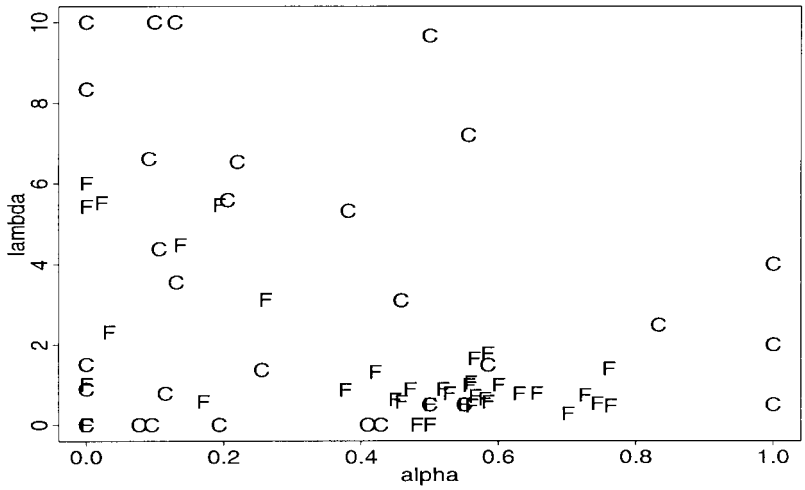

Figure 6. Parameter estimates for 39 rain events for SETINAR

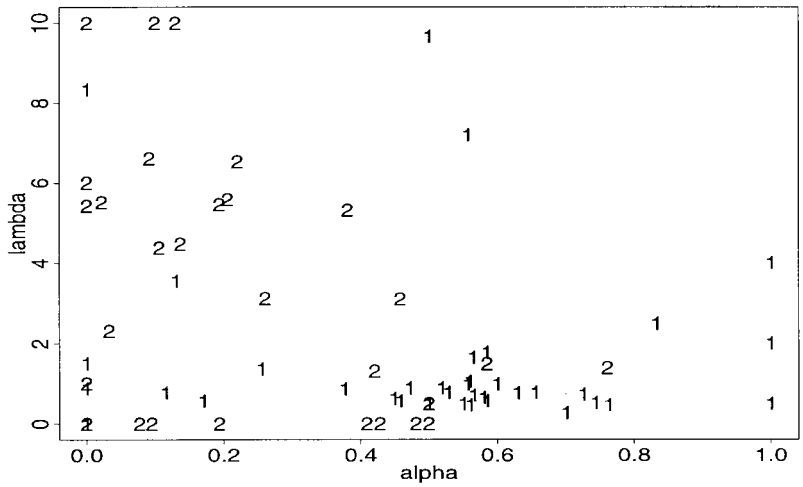

Figure 7. Parameter estimates for 39 rain events for SETINAR

likelihood ratios for a majority of the convective events are larger than 10, the SETINAR model is seen to give a significantly better description of a large fraction of the convective rain events. On the other hand, the rain events classified as being frontal do not benefit from the extension in quite the same way, indicating that they are not that prone to shift between regimes.

6. APPLICTION OF INAR-MODELS

The INAR-model as well as the SETINAR-model can be used to generate synthetic rain, for instance used for assessing long term extreme statistics. The simple structure of the model makes the simulation procedure straightforward. In each step the value of X_t is found as the sum of a

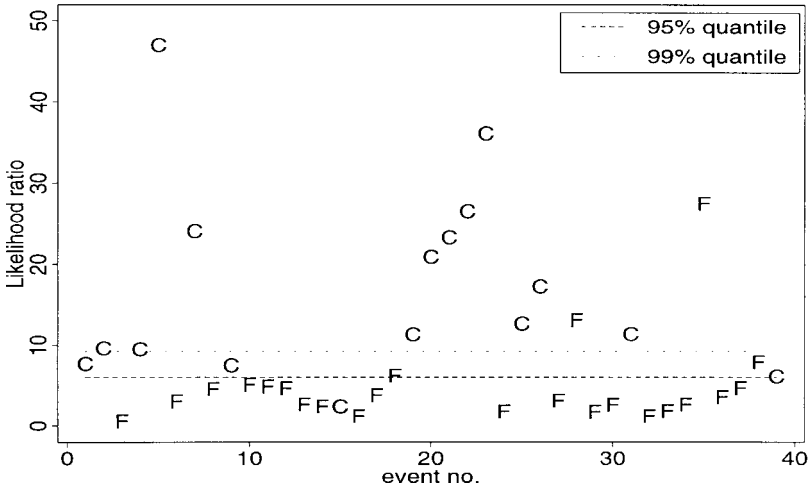

Figure 8. Likelihood ratios for SETINAR/INAR(1)

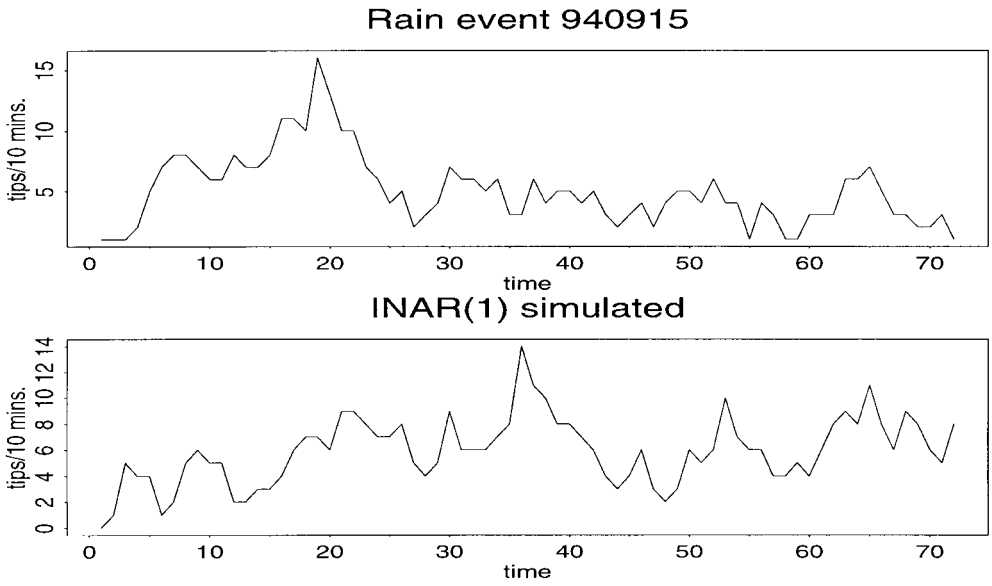

Figure 9. Simulation of rain with the INAR(1) process compared to an observed rain event

randomly generated binomial variable with number of trials X_{t-1} and probability α, and a randomly generated Poisson variable with intensity parameter λ.

In Figure 9 a rain event has been simulated using the estimated parameters of one of the 39 events. The simulated rain series is compared to the original series. Although, the simulated and original rain series do not reflect exactly the same dynamics, it appears that the behaviour of the two rain series is similar. It should be stressed that simulation of the rain process with an

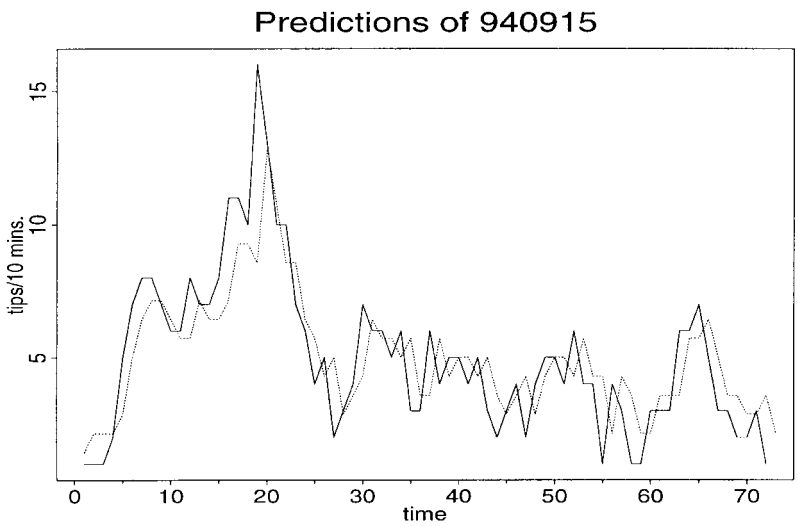

Figure 10. One step predictions using an INAR(1)-model (predictions: dotted line)

INAR-model should be restrained to a fixed length of simulation, because the INAR-process will generate innovations of rain for as long as the simulation runs.

Short term predictions of rain is useful for instance for on-line control of sewer systems and treatment plants. Figure 10 shows the one-step predictions of the INAR(1)-model for the same rain event as in Figure 9. The resemblance to the AR(1) process is clearly visible.

7. CONCLUSION

In this paper a class of models for integer valued processes has been suggested to describe the rainfall measurement process, and the models have been estimated and tested on observed rainfall data. The suggested INAR-models are new within the field of rainfall modelling. Only models with a few parameters are investigated. Models with many parameters have the disadvantage that they require a large amount of data for estimation with a reasonable precision. Furthermore, models with many parameters tend to have a high degree of substitutability, i.e. the parameter estimates tend to be highly correlated, and therefore the identification of the individual and independent parameters may be doubtful. Models with few parameters, on the other hand, provide better possibility of interpreting and comparing the estimates of the parameters. Finally, models with less parameters are often more robust.

Models were estimated for each of the available 39 rain events. It is shown that when estimating a set of parameters based on one rain event only, the parameters reflect properties of this particular rain event. Specifically, the estimated parameters can be used as an objective method for classifying the rain event as being either frontal or convective. Such classification is today done manually and thus subjectively, but the estimated parameters of the models presented can be used in this classification.

One extension of the INAR model considered in this paper was to incorporate two first order INAR models into a threshold structure. In this way the model has two carry-over parameters, and more importantly two innovation parameters. The threshold INAR model showed an

evident improvement for a large fraction of the convective rain events, whereas no improvements were seen for the frontal rain events.

Finally, the INAR model has been used for simulation and prediction. The simulated series show the same characteristics as the actual series of rainfall measurements. The INAR model produces short-term predictions which can be used as input to rainfall–runoff models in real time control of sewer systems.

ACKNOWLEDGEMENT

Financial support for this work was partially provided by the centre CINTEM under the Danish Ministry for Business and Industry.

REFERENCES

Al-Osh, M. and Alzaid, A. (1987). 'First-order integer-valued autoregressive process'. *Journal of Time Series Analysis* **8**.

Alzaid, A. and Al-Osh, M. (1990). 'An integer valued pth-order autoregressive structure process'. *Journal of Applied Probability* **27**.

Arnbjerg-Nielsen, K. (1996). Statistical analysis of urban hydrology with special emphasis on rainfall modelling. PhD thesis, DTU.

Franke, J. and Seligmann, T. (1993). 'Conditional maximum-likelihood estimates for inar(1) processes and their application to modelling epileptic seizure counts'. *Developments in Time Series Analysis*, ed. T. S. Rao. London: Chapman & Hall.

Harremoes, P. and Mikkelsen, P. (1995). 'Properties of extreme point rainfall I: Results from a rain gauge system in Denmark'. *Atmospheric Research* **37**.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time Series*. London: Chapman & Hall.

Madsen, H. (1995). *Tidsraekkeanalyse*. DTU: Institute of Mathematical Modelling (in Danish).

