Meas. Sci. Technol. **19** (2008) 015103 (14pp) doi:10.1088/0957-0233/19/1/015103

Time series models for spectral analysis of irregular data far beyond the mean data rate

Piet M T Broersen

Department of Multi Scale Physics, Delft University of Technology, The Netherlands

E-mail: p.m.t.broersen@tudelft.nl

Received 11 July 2007, in final form 12 September 2007
Published 23 November 2007
Online at stacks.iop.org/MST/19/015103

Abstract
Slotted resampling transforms an irregularly sampled process into an equidistantly sampled signal where data are missing. Equidistant resampling always causes spectral bias, due to aliasing and to shifting of the observation times. The shift bias can be diminished by using a slot width that is smaller than the resampling time step. A special approximate maximum likelihood time series estimator has been developed to estimate the power spectral density and the autocorrelation function of multi-shift slotted nearest-neighbour resampled data sets with missing observations. The algorithm estimates several time series models and selects the best model order and model type from a number of candidates. It is tested with benchmark data. It can estimate spectra up to frequencies more than a thousand times higher than the mean data rate. It can be applied to various irregularly sampled data, including bubbly turbulent flow and very sparse climate or atmospheric data.

Keywords: autocorrelation estimation, autoregressive model, nearest-neighbour resampling, slotting, spectral estimation, uneven sampling

(Some figures in this article are in colour only in the electronic version)

1. Introduction

Continuous-time processes will be irregularly sampled if signals are observed at times given by irregular triggering events. Irregular intervals may be caused by wireless sensor networks in various applications, from astronomy to remote weather stations that are triggered by atmospheric events. Irregular sampling may arise naturally in geophysics, heart rate analysis [1], astronomy [2] and climate research. In LDA (laser Doppler anemometry), the velocity can only be measured if a seeding particle passes through the measurement volume [3]. Computing the spectral density of irregular data is simple with the direct method of Lomb–Scargle [4], which is a least squares fit of sine curves to the data. The method turned out to be severely biased for the turbulence spectra of irregular data [3]. A strong bias is also found with this method for equidistantly sampled data, if some observations are missing [5]. It is possible to detect peaks at high frequencies with the Lomb–Scargle method, but only if the signal has a periodic component with a very small amount of additive noise. Slopes in the spectrum could not be estimated and direct Fourier methods are not advised for the spectra of irregular LDA data [3].

Irregular sampling intervals imply a continuous-time process and models of this type are the first choice for spectral estimation. A continuous-time maximum likelihood (ML) approach has been developed for autoregressive (AR) models of irregularly sampled data [6]. However, inspection showed that the surface of the computed likelihood as a function of the continuous-time AR parameters was very rough [7], with many local maxima. No better continuous-time algorithm could be found and numerical problems prevented the convergence of this method. Similar problems have been reported before [6]. To the author's knowledge, no general applicable and reliable continuous-time ML estimation method for practical unevenly sampled data is yet available. In addition, continuous-time AR models have been estimated by approximating the derivative operator [8]. Promising results for low order models have been reported, but care had to be taken in the chosen approximation of the derivative [8]. The derivative is computed by using

a fixed number of neighbouring observations [8]. This will become less accurate in sparse data, if spectra have to be estimated for frequencies far above the mean data rate. To find a spectrum for those high frequencies, discrete-time solutions will be investigated.

The first discrete-time solution is the slotted autocorrelation estimation. The product of two observations contributes only to the slotted autocorrelation at a certain lag $k\Delta$ if their distance is between $(k - 0.5)\Delta$ and $(k + 0.5)\Delta$ [3, 9]. Unfortunately, slotted autocorrelation estimates are not positive semi-definite. Hence, slotted autocorrelation estimates do not fulfil the theoretical requirements for being an autocorrelation function. Some improvements have been introduced, local normalization [9] and fuzzy slotting [10], where every contribution is distributed over the two nearest lags. The spectral variance has been reduced further with variable windows [9]. Spectra obtained with variable windows often look well for strong peaks at low frequencies. However, no variant of the estimated slotted autocorrelation functions is positive semi-definite. They all fail to consistently produce a spectrum that is positive for all frequencies, especially for frequencies at weak spectral parts. The autocorrelation fit of slotting or its spectral quality is a matter of taste, not of any objective quality measure. No sensible or reliable quality measure could be defined for slotted autocorrelations and their spectra, not even for very large samples. Therefore, other estimation methods will be considered.

Resampling techniques can replace an irregularly sampled continuous-time signal by an equidistant discrete-time signal, with resampled observations at a grid of equal time intervals. After resampling, the discrete-time equidistant data can be analysed with the conventional spectral analysis techniques [3] or with modern time series models [7]. Sample and hold (SH) resampling is equivalent to low-pass filtering followed by adding white noise [11]. Spectral estimates are severely biased at frequencies higher than $f_0/2\pi$, where f_0 denotes the mean data rate. The filter effects can in theory be eliminated by using a refined SH estimator [3]. This refinement explicitly uses and is limited to a Poisson distribution for the observation instants. Refinement and noise suppression can take place in the time [3] or in the frequency domain [12]. If applicable, it can enlarge the useful frequency range somewhat, from $f_0/2\pi$ to about f_0 [12]. The variance of the estimated spectra, the deviations from Poisson-distributed arrival times and the limited accuracy of step noise removal limit this frequency range in practice until a maximum that is somewhat lower than f_0. Nearest-neighbour (NN) resampling has similar characteristics to SH and cannot estimate spectra at frequencies higher than f_0. Resampling irregular data on a fixed and dense grid will often have the problem that no irregular observation has been made close to a grid node. Observations further away have to be used in the resampled signal and the same irregular observation is substituted at several equidistant grid nodes. This multiple use of one observation is the main cause for the very large bias of SH and NN resampling. The advantage of simple resampling is that the signal processing is simple and easy with contiguous equidistant data. The disadvantage is that the bias is too large, unless the highest frequency of interest is much lower than the mean data rate.

Intuitively, it seems preferable to interpolate the irregular observations and to substitute the value of the reconstructed signal on the grid nodes. This idea has been tested with simple linear interpolation and with more sophisticated methods like fractal reconstruction or the projection onto convex sets [13]. The conclusion was that the visual appearance of the reconstructed signal looked promising, but the bias of spectral estimates could not be improved in comparison with SH resampling [3, 13].

Multiple uses of single observations in resampling are avoided with the slotting principle. The slotting principle can be applied to the NN resampling of irregular data on a regular time grid [7]. An observation is only accepted in resampling if a true observation is within half the slot width from an equidistant resampling grid point. Slotted resampling has bias properties that are similar to the bias of slotted estimates of the autocorrelation function. The bias can be reduced by taking a higher resampling frequency or by making the slot width smaller than the resampling distance [7]. Using the slotting principle gives an equidistant signal, with data missing at those grid nodes that are further than half the slot width away from an actual irregular observation.

Equidistant *missing-data* problems have been investigated as a separate problem [5]. Spectral estimation for equidistant observations with missing data is much simpler than the spectral analysis of continuous irregular data. For missing data, Jones described an efficient method to calculate the exact likelihood [14]. This equidistant likelihood algorithm has much more favourable numerical properties than his continuous algorithm [6]. An automatic time series programme uses this ML algorithm and outperformed other methods that have been described for missing-data problems [5]. This accurate method for missing data will be applied here to multi-shift slotted NN resampled (MSSNNR) irregular data [7].

This paper studies the bias properties of slotting, due to aliasing and to shifting of observation times to a grid. The ARMAsel-irreg algorithm [7] with the automatic selection of an AR, MA or ARMA model for irregular data has been developed to evaluate the MSSNNR data. However, order selection sometimes failed [7] and the results did not always become better if more data were available. These problems are investigated here. The quality of automatically selected time series models of an improved ARMAsel-irreg algorithm is established for the resampled irregular data of a turbulence-like benchmark example [15]. The behaviour of the algorithm for growing sample sizes is studied. The spectra of selected models are shown to be accurate at frequencies much higher than the mean data rate. Spectra can also be accurate if the irregular inter-arrival times do not obey a Poisson distribution.

2. Time series models

A discrete-time autoregressive, moving average ARMA(p, q) process can be written as [16, 17]

$$x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} = \varepsilon_n + b_1 \varepsilon_{n-1} + \cdots + b_q \varepsilon_{n-q}, \tag{1}$$

where ε_n is a purely random white noise process of independent identically distributed stochastic variables with zero mean and variance σ_ε^2. It is assumed that the data x_n represent a stationary stochastic process. For resampled continuous-time data with resampling distance T_r, the signal x_n is the observation at time nT_r. Other values for T_r would give different parameters and σ_ε^2 in (1).

Theoretically, every discrete-time stationary stochastic process can be represented by an ARMA(p, q) model, an AR(∞) or a MA(∞) model [16]. In practice, finite orders will be sufficient for estimated models because high order parameters are generally negligible and not statistically significant. It is important that it is in no way a restriction to suppose that data are represented by an ARMA(p, q) model, because all stationary random data have a model in that class. Data with additive noise are modelled together as a single ARMA(p, q) model. The model represents a unique parametric estimate of the autocorrelation function or of the power spectral density for the noisy data. An automatic ARMAsel algorithm has been developed to estimate the parameters and to select the best model type and model order for given equidistant data, without user interaction [17].

The roots of the polynomial $A_p(z)$, built with the AR(p) parameters as coefficients,

$$A_p(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}, \tag{2}$$

are denoted as the poles of the AR(p) model. It is always assumed that data represent a stationary stochastic process, which is guaranteed if all poles of $A_p(z)$ are inside the unit circle. Likewise, the roots of the MA(q) part

$$B_q(z) = 1 + b_1 z^{-1} + \cdots + b_q z^{-q} \tag{3}$$

are called the zeros.

Spectral estimates describe the distribution of the power over the frequencies, but are also important for integral time scales, spectral peaks and slopes, dissipation rates and general scale information. The power spectral density $h(\omega)$ of the model and the frequency range depend on the resampling time T_r. The spectrum and the autocorrelation function are fully determined by the parameters in (1) together with the variance σ_ε^2 and T_r. The discrete-time spectrum is given by

$$h(\omega) = \frac{\sigma_\varepsilon^2 T_\mathrm{r}}{2\pi} \frac{|B_q(\mathrm{e}^{\mathrm{j}\omega i})|^2}{|A_p(\mathrm{e}^{\mathrm{j}\omega i})|^2} = \frac{\sigma_\varepsilon^2 T_\mathrm{r}}{2\pi} \frac{\left|1 + \sum_{i=1}^{q} b_i \,\mathrm{e}^{-\mathrm{j}\omega i}\right|^2}{\left|1 + \sum_{i=1}^{p} a_i \,\mathrm{e}^{-\mathrm{j}\omega i}\right|^2}, \qquad -\frac{\pi}{T_\mathrm{r}} < \omega \leqslant \frac{\pi}{T_\mathrm{r}}. \tag{4}$$

The true continuous-time spectrum has an infinitely wide frequency range, which cannot be represented exactly with (4). Two approximations are available for the model of the true process spectrum in the limited discrete-time frequency range in (4). They give different sets of true parameters and σ_ε^2. The first option is to determine parameters that describe the true continuous infinitely wide spectrum without aliasing for only the given frequency range until $\omega = \pi/T_\mathrm{r}$ and to consider the true spectrum outside that frequency range as zero. The autocorrelation function at lags kT_r can be approximated in this option by an inverse discrete Fourier transform of (4). The spectrum is maintained in a part of the frequency range at the cost of a distorted autocorrelation function. The second possibility gives the aliased spectrum for the frequency range that is determined by the resampling frequency T_r. Aliasing is the effect that the continuous-time contributions to the spectrum, for frequencies outside the discrete-time frequency range, are folded back into that limited range [16]. The spectrum is distorted in this option. However, the autocorrelation is given now by the true autocorrelation function sampled at the lags kT_r. The formulae that relate the autocorrelation to the parameters of (1) are the same for both options, but different parameter sets give different results for the autocorrelation. Autocorrelations of a MA(q) process are zero for lags greater than q; other processes have in principle an infinitely long, decaying, autocorrelation function.

The goal would be to estimate the true power spectral density without aliasing in the discrete-time frequency range. This is not possible with equidistantly resampled observations. Therefore, the practical purpose will be to choose the resampling frequency $1/T_\mathrm{r}$ high enough to have only a small influence of aliasing. This happens if the power of the true spectrum is as small as possible for the frequency range $f > 1/(2T_\mathrm{r})$, which agrees with the radial frequency range $\omega > \pi/T_\mathrm{r}$. However, a higher resampling frequency gives more equidistant resampling grid points for the same number of irregular observations. Therefore, a compromise between a small aliasing effect and a low resampling frequency is necessary.

Time series models can also be used for missing-data problems [5, 18] and for irregularly sampled data [7]. The automatic, approximate maximum likelihood, ARMAsel-mis program has been developed for the estimation of ARMA(p, q) models for equidistant missing-data problems [5]. Modifications of that missing-data algorithm to the time series program ARMAsel-irreg for irregular data have been described before [7]. Sometimes, selection of the model order had unexpected results when applied to irregular data.

An AR order selection criterion GIC has been developed for missing-data problems [5]:

$$\mathrm{GIC}(p) = LH + \alpha p. \tag{5}$$

It uses twice the minimized negative log-likelihood, denoted by LH, with a penalty factor α that depends on the missing fraction. Extensive simulations for missing-data problems have been used to determine which values of α give the best results in order selection. Penalties between 2 and 20 have been considered [18]. It has been advised to use $\alpha = 3$ if less than 25% is missing, $\alpha = 5$ for more than 75% missing and $\alpha = 4$ in the range between. The same penalty will be used for equidistantly resampled irregular observations.

The number of grid points after resampling with the slot width w is approximately given by NT_0/w, where $T_0 = 1/f_0$ is the average distance between the irregular samples. The fraction of non-empty grid points is approximately given by w/T_0. The application of this order selection criterion to irregular data is investigated in this paper by applying it to benchmark data where the true process, both with and without aliasing, is known.

NN resampling replaces an unevenly sampled signal by an equidistant signal with the resampling distance T_r. At

every resampling node, the closest irregular observation is substituted. If necessary, the same observation will be used for more resampled data. This causes a large extra bias. The properties are similar to SH [11], and spectra are only more or less reliable until $f_0/2\pi$. The bias can be reduced with the slotting principle, where the slot width is taken equal to the resampling distance. Slotted NN only accepts an observation if it is within half the slot width from the resampling time. If no irregular observation falls within the slot width, the grid node is left empty as a missing observation. If more observations are present within the slot, only the one closest to the grid point is used. It excludes the possibility that a single irregular observation is used at different resampling points and strongly reduces the bias.

A further improvement of the bias due to the shifting of observations to a grid is found with MSSNNR, where the slot width w is made smaller than the resampling distance [7]. Taking $w = T_r/M$, where M is an integer number, gives disjoint intervals for the slots and several irregular observation times t_i are not within the small slot around $t = nT_r$. MSSNNR extracts M different equidistant missing-data signals from one irregular data set by using M shifted starting points at mw with distance w, $m = 0, 1, \ldots, M-1$. The non-empty resampling instants $nT_r + mw$, where an irregular observation falls within the slot width, are determined for the M signals by

$$nT_r + mw - 0.5w < t_i \leqslant nT_r + mw + 0.5w, \quad m = 0, 1, \ldots, M-1, \tag{6}$$

where t_i denotes the time of an irregular observation. Now, all slots of width w are connected in time. M shifted starting points give M equidistant sequences, each with a time step T_r. Data are missing in each signal. The likelihood function can be calculated for every signal individually with the ARMAselmis algorithm.

If a continuous-time representation of a signal were available for all t, it is possible to make a discrete-time representation by sampling it at the resampling instants nT_r. M different discrete-time signals could be extracted by using the resampling instants $nT_r + mw$, $m = 0, 1, \ldots, M-1$, with $w = T_r/M$. These M signals would be very similar for a small value of T_r, and the M likelihoods that can be computed for the M signals would be strongly dependent. One single signal for one value of m would contain almost all valuable information and there is hardly anything gained by evaluating all M signals. Nothing is lost either, because all signals would give approximately the same time series model. This changes if the M signals are obtained with MSSNNR of (6) for irregular data. Now, each signal is a missing-data record, with many missing data if a small value for T_r is used. Generally, it can be expected that the places where some close irregular observations at distances of a multiple of T_r are present are on different and independent locations for the M missing-data signals. This is certainly the case for Poisson-distributed irregular sampling instants, and also for many other irregular sampling schemes. Nearby observations can best be predicted and give the largest influence on the negative log-likelihood that is minimized in parameter estimation. Hence, the influential parts may be at different places for the M MSSNNR signals. These M likelihoods are added to a single likelihood value in the ARMAsel-irreg algorithm, as if they were independent. It does no harm if they are dependent, but it can give a much better accuracy than a single signal if they are independent. Therefore, it will reduce the estimation variance of the parameters. The AR parameters are estimated by maximizing the sum of the M likelihoods together as a function of the parameters [7]. MA and ARMA models are computed from those AR parameters. Afterward, (4) is used to compute the spectrum.

3. Accuracy measures

The sum or the integral of the squared differences between spectra is equal to the sum of the squared differences of their autocorrelation functions, according to Parseval's law for Fourier transforms [16, 17]. Large relative errors in small spectral values, say less than 0.001 of the peak value of the spectrum, have almost no influence on the sum of squared differences of autocorrelations or spectra. The sum of squares can be small for very large relative spectral errors. Therefore, the sum of squares is not an acceptable measure for the spectral or autocorrelation accuracy of arbitrary data.

The spectral distortion (SD) is a relative integral spectral error measure that has been defined as [17]

$$\mathrm{SD} = \frac{0.5T_r}{2\pi}\int_{-\pi/T_r}^{\pi/T_r} [\ln\{h(\omega)\} - \ln\{\hat{h}(\omega)\}]^2\,\mathrm{d}\omega = \frac{0.5T_r}{2\pi}\int_{-\pi/T_r}^{\pi/T_r}\left\{\ln\frac{h(\omega)}{\hat{h}(\omega)}\right\}^2\mathrm{d}\omega. \tag{7}$$

where h is true and $\hat{h}$ denotes the estimated spectral density.

The accuracy of equidistant time series models can be evaluated with the prediction error [17]. The prediction error PE(p', q') of an ARMA(p', q') model with arbitrary orders p' and q' is defined as the squared error of the one-step-ahead prediction with that model in new fresh data. The FPE of Akaike [19] can be computed as an estimate for PE from equidistant data. The PE for irregular data, however, can only be computed if the parameters of the true ARMA(p, q) process and of the estimated model with arbitrary orders p' and q' are known. This occurs in simulations and in benchmark experiments [17]. It does not require the availability of new data. It has been shown that the prediction error is strongly related to the SD. Its value is given for an ARMA(p, q) process with the true parameter polynomials $A_p(z)$ and $B_q(z)$, estimated polynomials $\hat{A}_{p'}(z)$ and $\hat{B}_{q'}(z)$ and known variance σ_ε^2 by [17]

$$\mathrm{PE}(p', q') = \frac{\sigma_\varepsilon^2 T_r}{2\pi}\int_{-\pi/T_r}^{\pi/T_r}\frac{h(\omega)}{\hat{h}(\omega)}\,\mathrm{d}\omega = \frac{\sigma_\varepsilon^2 T_r}{2\pi}\int_{-\pi/T_r}^{\pi/T_r}\left|\frac{\hat{A}_{p'}(\mathrm{e}^{\mathrm{j}\omega})B_q(\mathrm{e}^{\mathrm{j}\omega})}{A_p(\mathrm{e}^{\mathrm{j}\omega})\hat{B}_{q'}(\mathrm{e}^{\mathrm{j}\omega})}\right|^2\mathrm{d}\omega, \tag{8}$$

where (4) has been used for the computation of the true and for the estimated spectra. If the quotient of h and $\hat{h}$ is close to 1, the logarithm of $1 + \delta$ in (7) can be approximated by δ and the differences between (7) and (8) are mainly in scaling and an additional constant.

Dividing PE(p', q') by σ_ε^2 makes the new scaled error measure PE$_s$(p', q') independent of the variance level of the data. The same result is obtained by normalizing σ_ε^2 to the fixed value 1 in simulation experiments. PE$_s$(p', q') can be used for pure AR models where $q' = 0$, and for MA models with $p' = 0$. The expectation of the PE$_s$(p', q') for efficiently estimated *unbiased* ARMA(p', q') models from equidistant data is given by the number of estimated parameters and the sample size N [17]:

$$\mathrm{E}[\mathrm{PE}_s(p', q')] = \mathrm{E}\left[\frac{\mathrm{PE}(p', q')}{\sigma_\varepsilon^2}\right]$$
$$\approx 1 + \frac{p' + q'}{N}, \qquad p' \geqslant p, \quad q' \geqslant q. \tag{9}$$

The approximation (9) can be used for models of the true order or higher. Every additional parameter gives a contribution $1/N$ due to its estimation variance. Models of lower orders are truncated and biased because they do not have sufficient parameters to describe the exact true spectrum (4). They have a higher PE$_s$(p', q') than given by (9), and the truncation bias contribution does not depend on N. Bias is an important reason why spectral estimates will not become better if more data are available.

4. Bias of multi-shift slotted NN resampling

After equidistant resampling, the frequency range of the discrete-time spectrum is determined by the resampling frequency. The discrete-time autocorrelation function is the sampled version of the continuous autocorrelation $R(\tau)$. The aliased true spectrum can be computed with [16]

$$h(\omega) = \int_{-\infty}^{\infty} R(kT_r)\, \mathrm{e}^{-\mathrm{j}\omega k}\, \mathrm{d}k, \qquad -\frac{\pi}{T_r} < \omega \leqslant \frac{\pi}{T_r}. \tag{10}$$

It can also be found by superimposing folded ranges of the spectrum [16]. The aliased spectrum has the same integrated power over its limited frequency range as the continuous spectrum over the infinite range. This follows because the integral over the frequency range concerned is given by the same value $R(0)$. Aliasing has a strong influence on the spectrum if that is rather flat. The influence can be small if the spectrum has a steep slope at higher frequencies. The smallest possible influence for continuous spectra is a difference of a factor 2 at $\omega = \pi/T_r$. It is not possible to reduce the influence of aliasing on irregular data by using anti-aliasing filters.

In addition, shifting the irregular observation times to an equidistant grid has influence on the autocorrelation function and the spectrum. For a smaller slot width or increasing M, the bias of MSSNNR in (6) is reduced in comparison to slotted NN with $M = 1$ and still more in comparison to ordinary NN, where the multiple use of observations occurs. For Poisson sampling instants, the bias of multi-shift-slotted NN has been described with the probability density function $f(\tau)$ of the continuous-time lags τ of the continuous autocorrelation function $R(\tau)$ that contribute to the resampled autocorrelation $R_{res}(kT_r)$ [7]. The expectation of the resampled autocorrelation becomes

$$R_{res}(nT_r) = \int_{-w}^{w} R(nT_r + \tau) f(\tau)\, \mathrm{d}\tau, \qquad n \neq 0$$
$$R_{res}(0) = R(0). \tag{11}$$

The resampled spectrum follows by substituting $R_{res}(kT_r)$ in (10) for $R(kT_r)$. It has been demonstrated how the bias in spectra is diminished by using a smaller value for T_r if $M = 1$ and with greater values for M if T_r is kept fixed [7].

For very dense irregular sampling, if the slot width $w = T_r/M$ is much greater than the average distance T_0 between the irregular observations, several observations will mostly fall within the slot width and only the one closest to the centre of the slot nT_r is accepted. This means that the probability density $f(\tau)$ is much narrower than the slot width w. If w is much smaller than T_0, however, the influence of the probability density function of the observation instants on $f(\tau)$ is no longer important. The width of $f(\tau)$ is determined by $2w$ and the shape of $f(\tau)$ turns out to become almost triangular. A triangle is a fair approximation for the shape of $f(\tau)$ for $w > T_0$ and a very good approximation for still smaller w. The explanation is easy. If there is only one observation within the slot width, it can be anywhere. This means that the density function of the observation time within the slot is uniform over w. Combining two independent observations in different slots to compute the autocorrelation function gives the convolution of two uniform rectangular densities for the combined density $f(\tau)$. The convolution of two rectangles gives a triangle. This was also the density to which Poisson distributions converged for very small w.

The triangular shape of $f(\tau)$ is largely independent of the arrival times if the slot width of MSSNNR is smaller than the average time between the irregular observations. Therefore, the bias of multi-shift slotting can be approximated with

$$f(\tau) = \frac{w - |\tau|}{w^2}, \qquad 0 < |\tau| < w$$
$$f(\tau) = 0, \qquad |\tau| > w. \tag{12}$$

This density can be substituted in (11) for arbitrary densities of the observation times. Formula (12) loses its accuracy only if irregular observations are found more often within one slot width.

In principle, this bias of shifting time instants to a grid can be made as small as desired by using denser and denser grids, with a smaller slot width w. The limiting value is the aliasing bias that follows with (10). The disadvantage of smaller slots is that the number of grid points increases for the same number of irregular observations, and more grid points are left empty. This gives a greater missing fraction, and the estimation of time series parameters becomes less accurate and takes more computing time [7]. It turns out in many examples that the bias, due to shifting irregular times to a regular grid, is always rather small if the dynamic range in the spectrum is less than about 100. Only spectral details that are less than 0.01 times the peak values are lost by the bias.

At least four sources of bias have been defined for time series models of irregular data. The first is aliasing given in (10). The second is the shifting bias that follows from (11). The third is called the truncation bias. This is caused by estimating underfitting models, with a lower order than the true process. Not all details can be represented by such models. The magnitude of the truncation bias is independent of the sample size N, like the aliasing bias and the bias due

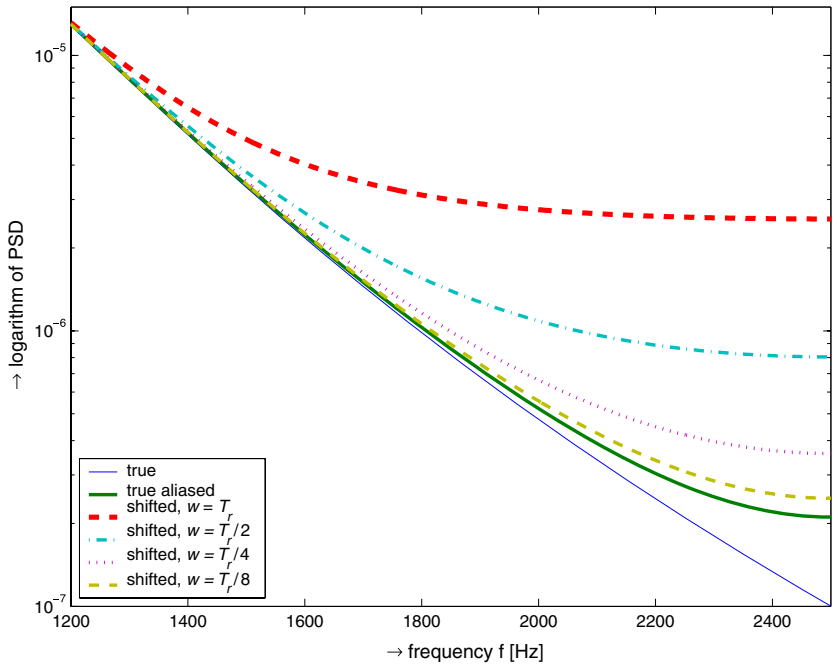

Figure 1. The different true spectra that represent an irregular benchmark process of type 3 of Nobach [15]. The true spectrum is given, the spectrum resampled with $T_r = 1/5000$ s with only aliasing bias and four different spectra with the additional shifting bias, for $w = T_r/M$, with $M = 1, 2, 4$ and 8.

to shifting irregular times to a grid. It can be measured with the PE_s of (9) and will be a constant. The fourth bias type is a special bias that will only occur when data are missing [5]. It has a simple explanation. For contiguous equidistant observations, only observations at distance T_r contribute to the estimation of the AR(1) parameter. If the true correlation at distance T_r were zero, the expected AR(1) estimate could also be zero and it is independent of the autocorrelation at greater lags. This does not apply to missing data because observations at multiples of T_r will also contribute to the likelihood estimate of that AR(1) parameter if data are missing. This causes an additional bias in estimated missing-data models as long as the order is lower than the true process order. The same reasoning has been applied to an AR(2) missing-data signal with two non-zero parameters [5]. The explanation is only qualitative. No quantitative theoretical derivation for the bias is known. It will depend on the probability density function of the observation times. This bias type is no longer present if models of the true order or higher orders are estimated.

Three different representations can be given for a true spectral density in benchmark tests:

- **T**: the part of the infinitely wide true continuous-time spectrum that falls in the frequency range after resampling with T_r which is given by $-\pi/T_r < \omega \leqslant \pi/T_r$,
- **A**: the aliased true spectrum in that frequency range which is obtained with (10),
- **S**: the aliased spectrum with the additional shifting bias obtained by shifting the observation times to a regular grid with width $w = T_r/M$, which can be computed by substitution of the correlation function $R_{res}(kT_r)$ of (11) in (10), where $f(\tau)$ is given by (12).

All three representations can be used as the true parameter polynomials $A_p(z)$ and $B_q(z)$ in (8), to compute the prediction error. To indicate the mutual difference, the prediction error is denoted $PE_s^{\mathbf{T}}$, $PE_s^{\mathbf{A}}$, $PE_s^{\mathbf{S}}$. By computing the prediction error for an estimated model with the three measures, it can be seen whether estimated spectra are close to the true, the aliased or the shifted expectations of the true spectrum.

The different spectra have been computed for an example that will be studied in this paper. It is spectral type 3 of the benchmark generator [15], which gives a Heisenberg spectrum with two constant slopes in the double logarithmic spectrum, proportional to $f^{-5/3}$ from $f = 100$ Hz and to f^{-7} from $f = 1000$ Hz. Figure 1 gives the spectra for $T_r = 1/5000$; only the part of the frequency range where they are different is shown. In this example, the influence of aliasing is small. Only a very small part of the total power is above the highest frequency 2500 Hz for this spectrum that is proportional to f^{-7}. The aliased spectrum is a factor 2 greater than the true continuous spectrum at the end of the range, and aliasing has almost no effect for frequencies below 1800 Hz. In other examples or in the same example if it were resampled with a frequency lower than 1000 Hz, the influence of aliasing could be much greater. The shifting bias depends on the slot width. The spectrum converges to the aliased spectrum if the slot width is small enough. The $PE_s^{\mathbf{T}}$ of aliasing is 1.010 here and the $PE_s^{\mathbf{T}}$ of the shifting bias for $w = T_r/M$ is 1.408, 1.134, 1.040 and 1.016 for $M = 1, 2, 4$ and 8, respectively. This implies that the best accuracy $PE_s^{\mathbf{T}}$ of spectra estimated with $w = T_r$ can only be 1.408. It is certainly worthwhile to try a smaller slot width. $PE_s^{\mathbf{A}}$ of the spectra with a shifting bias is 1.327, 1.085, 1.013 and 1.001 for $M = 1$, 2, 4 and 8, respectively. This shows that all spectra with shifting bias are closer to the aliased true

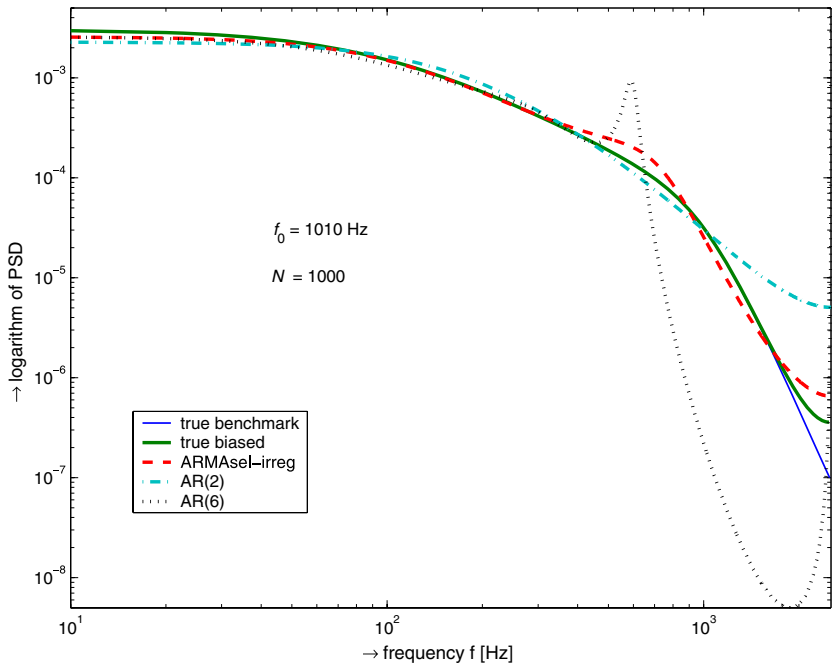

Figure 2. Estimated and true spectra of 1000 irregular benchmark observations of a type 3 process of Nobach [15], with a mean data rate of 1010 Hz. Slotted NN used $T_r = T_0/5$ and $w = T_r/4$. ARMAsel-irreg selected the AR(3) model that gives the best fit of all models estimated for those data.

spectrum than to the true continuous spectrum. They converge to the aliased spectrum for very small w.

5. Benchmark data with Poisson sampling

In the first application of the ARMAsel-irreg algorithm to irregular simulated data, satisfactory results have been obtained with estimated AR models with a known and low order [7]. It has been demonstrated that using very high resampling rates is not a computational problem for the ARMAsel-irreg algorithm. The spectra of practical bubbly flow data [20] have been analysed up to frequencies which are 250 times higher than the mean data rate f_0 [7]. However, order selection was unreliable in that example and the accuracy of the estimates could not be evaluated because the true spectrum was not known for those real-life data. Therefore, irregular benchmark signals with a known true spectrum will be used here.

Several test signals with Poisson-distributed arrival times can be generated with a benchmark generator; program and description are available on the Internet [15]. Spectral type 3 of the benchmark [15] gives a Heisenberg spectrum with two constant slopes in the double logarithmic spectrum, proportional to $f^{-5/3}$ from $f = 100$ Hz and to f^{-7} from $f = 1000$ Hz. This process has already been used in figure 1 to evaluate the biases. The true process is theoretically AR(∞). However, all parameters of orders higher than 6 are less than 0.1 and above order 80 they are all less than 0.001. The parameters of the AR(500) model have been used as the true process parameters to compute the prediction error with (8).

Many simulation runs have been examined, for different N, T_r and w. No numerical problems have been encountered. The surface of the computed likelihood function has been investigated for the MSSNNR data, because the numerical result for continuous likelihood was very rough, with many local minima [7]. The surface of the AR likelihood of the MSSNNR data was always smooth, and it did not have local minima close to the global minimum. It has been verified that the minimization procedures did converge to the same global optimum from different starting values for the parameters. If all zeros are used as starting values, it converged to the same minimum. But sequential estimation with increasing AR model orders is preferable. Using the AR model of the previous order and an additional zero as starting values for the new order in ARMAsel-irreg generally gives a faster overall convergence, as well as the spectral estimates for all lower order AR models. Furthermore, this gives the likelihoods of all model orders that are required in (5) for order selection.

As an example, the results of a single run are presented. Figure 2 gives the true spectrum, the expectation of the spectrum with shift bias for $N = 1000$, $T_r = T_0/5$ and $w = T_r/4$ and three estimated AR spectra, for the orders 2, 3 and 6. The frequency axis has been made logarithmic. The AR(3) model was selected in this run. The spectra of the AR(2) and the AR(4) models have a regular appearance. Models of order 5 and higher have spurious spectral peaks and they have never been selected in other runs with the same example. AR models can also be estimated for a smaller frequency range than up to 2500 Hz. However, these models would miss the characteristic spectral slopes of the process. The best estimated spectrum that can be expected with the MSSNNR resampling for given values of T_r and w is a spectrum close to the true spectrum with the shift bias included, because that is the actual spectrum of the resampled data after the shifts of the irregular observation times to a resampling grid. Therefore, the proper choice of

Table 1. Several accuracy measures for estimated AR models computed for the 1000 irregular data used in figure 2 with $T_r = T_0/5$ and $w = T_r/4$, as a function of the model order. The final columns give the PE_s^T results for the true truncated continuous process in the limited frequency range of the resampled signal and of the true aliased spectrum with the shift bias added.

AR order	PE_s^T	GIC	LH	PE_s^A	PE_s^S	PE_s^T(true)	PE_s^T(shift)
0	18.76	1570.0	1570.0	17.50	15.75	18.76	18.76
1	2.61	1408.0	1403.0	2.44	2.20	2.52	2.53
2	1.81	1405.9	1395.9	1.70	1.55	1.52	1.52
3	1.14	1405.1	1390.1	1.09	1.05	1.13	1.13
4	1.29	1410.0	1390.0	1.23	1.15	1.05	1.06
5	3.18	1413.6	1388.6	3.02	2.81	1.02	1.04
6	5.91	1413.6	1383.6	5.64	5.36	1.01	1.04

T_r and w is important. If reliable *a priori* knowledge about the true spectral shape is available, this can probably be used in the construction of the likelihood function to diminish the influence of the shift bias.

The missing fraction is determined by the choices of T_r and w. Taking $T_r = T_0/K$ and $w = T_r/M$ makes the number of grid points KM times greater than the number of observations that have an average distance of T_0. Therefore, the remaining fraction γ is approximately $1/KM$ and the missing fraction is $1 - 1/KM$. The effective number of observations is for Poisson-distributed inter-arrival times approximately given by γN. The effective number of observations is the expectation of the number of observation pairs in the MSSNNR data with a distance of T_r, $2T_r$, $3T_r$ and so on. Asymptotically, these expected numbers are all equal for Poisson distributions. A smaller slot width gives a better approximation of the aliased spectrum in figure 1 at the cost of a larger missing fraction and a smaller effective number of observations. To obtain reliable estimates with ARMAsel-irreg, the effective number of observations γN must at least be as large as would be required from contiguous equidistant data of the same process. That number depends on the true process characteristics, but it will generally be much greater than 10. This limits the choices of T_r and w for a measured irregular data set, for which N and T_0 are given constants.

Table 1 gives information about the accuracy of estimated AR models and of the exact truncated lower order models derived from the true process parameters. AR models are truncated to lower order models by keeping only the values of the lower order reflection coefficients. Reflection coefficients are defined as the negatives of partial correlation coefficients for increasing orders [17]. Parameters for increasing orders can be computed from the reflection coefficients with the Levinson–Durbin recursion [17]. Estimated AR models are stationary if all reflection coefficients are less than 1 in absolute value. This property has been used in the likelihood function that is computed in the ARMAsel-irreg algorithm to guarantee the stationarity of the estimated AR model [7]. AR models of orders higher than 6, MA and ARMA models were never selected and they are not interesting for this type of data. A missing-data analysis showed that the AR model type is mostly selected if more than 50% of the data is missing [18]. Therefore, MA and ARMA are not discussed here, although they have been estimated with ARMAsel-irreg.

The columns in table 1 with PE values use the definition of (8) and require knowledge about the true AR(500) process. Therefore, they can only be given in simulation experiments where the true process is exactly known. The column with LH gives the minimized LH for practical data and GIC is the order selection criterion that is derived from LH with (5). GIC is the practical measure, and the order with the smallest GIC is selected. The three columns with PE_s^T, PE_s^A and PE_s^S give the accuracy of the estimated model in comparison with the true continuous spectrum, the true aliased spectrum and the expectation of the spectrum with the shift bias included, respectively. The final two columns give the accuracy of truncated true AR(m) processes in comparison with the true continuous AR(500) spectrum, as a function of the model order m. PE_s^A is somewhat smaller than PE_s^T in table 1 for all model orders because the estimated AR model fits better to the aliased spectrum of the resampled data. For all model orders, PE_s^S is the smallest, which demonstrates that the estimated models have shifting bias.

Order 3 is selected because GIC has its minimum there. The expected decrease of the likelihood for high order AR models above the true order is always 1 for each additional parameter, independent of the sample size or of the true process parameters or the mean data rate. The decrease of 167 for order 1 is very significant, but the decrease of LH for higher orders is rather small. The PE_s^T value for the AR(0) process was so high in repeated simulations that AR(1) was always better. The likelihood always gives a significant decrease to select at least the order 1 with (5). On the other hand, the decrease of the likelihood was never significant from order 3 to 4, which is remarkable because order selection criteria generally allow a small probability of selecting too high orders. Therefore, order selection will give varying results in different simulation runs, but order 0 will never be selected.

The error PE_s^T of the selected AR(3) model is 1.14, very close to the expectation 1.13 that belongs to the truncated true or shifted AR(3) process in the last columns of table 1. The minimum expectation of the prediction error in (9) is still closer to 1, but that is only applicable to unbiased models. No AR(3) model will have a smaller PE_s than the expectation 1.13, independent of the number of observations, because the true AR process has an order higher than 3. The increase of PE_s^T for higher order estimated AR models is due to the estimation variance. Due to this variance, the large variations of the parameters cause high values of PE_s^T, but only small

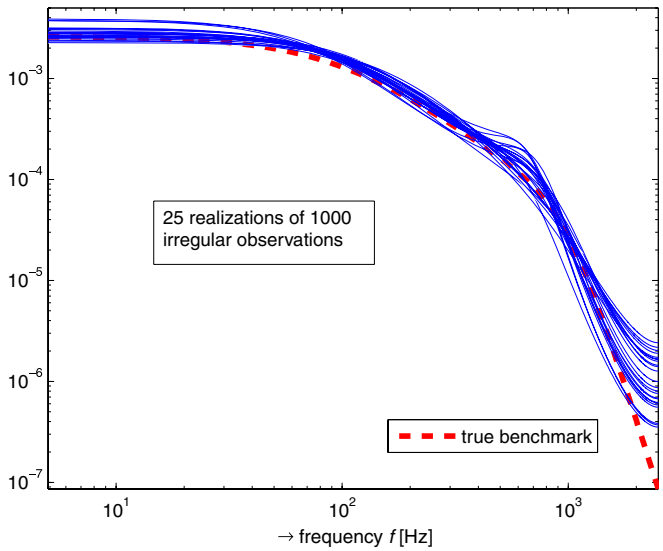

Figure 3. True spectrum of 1000 irregular benchmark observations and 25 estimated AR(3) spectra for $N = 1000$ with a mean data rate of 1000 Hz. MSSNNR used $T_r = T_0/5$ and $w = T_r/2$.

variations of the likelihood for higher AR orders in table 1. PE_s^T due to the truncation bias of the true AR(4) model is 1.05 for the given values of T_r and w. The truncation bias is negligible for truncated true AR models of an order higher than 6 and PE_s^T(true) converges to 1. PE_s^T(shift) converges to the final value 1.04 for high order models with aliasing and shift bias included for the given values of T_r and w. For $N = 1000$, the estimation variance strongly disturbs PE_s^T of estimated AR models of order 4 and higher. The truncation bias for a given AR order will not decrease for more data; only the variance may become smaller.

The missing-data bias has a remarkable effect in table 1. Estimated AR(1) models have a PE_s^T value that is always somewhat greater than the theoretical value 2.52 of the truncated true AR(500) process. The bias is much stronger for AR(2) models, with PE_s^T almost 0.3 greater than the value for the truncated true process. This bias will not diminish for greater values of N. Bias is the reason that AR(2) models of this process will always have poor quality and they will not become better if more observations are available. The truncation bias gives $PE_s^T = 1.13$ for the truncated true AR(3) process and the value of the estimated AR(3) model is very close here. This indicates that the effect of the missing-data bias is very strong for the AR(2) model and less for higher order models.

Repeated simulations with $N = 1000$ give diverse results for $T_r = T_0/5$ and $w = T_r/4$, both in selected model orders and in model quality. Order 3 is selected in about 50% of the runs, with PE_s^T close to 1.14. Order 2 was selected in about 40% of the runs, with PE_s^T about 1.7 and in the remaining runs AR(1) was selected with PE_s^T about 2.6. AR(0) has never been selected. The effective number of observations was around 50, which is not sufficient to give the same results in all simulation runs for this benchmark example. Sometimes, AR(3) may have a lower PE_s^T than AR(2), but AR(2) is selected if the order selection criterion GIC(3) of (5) has a higher

Table 2. PE_s^T for the estimated AR models in a single simulation run as a function of the model order and of the sample size. The data are N irregular benchmark observations of a type 3 process of Nobach [15], sampled with a mean data rate of about 1000 Hz. Slotted NN used $T_r = T_0/5$ and $w = T_r/4$. The selected order is printed in bold face.

N	AR(1)	AR(2)	AR(3)	AR(4)	AR(5)	AR(6)
100	**2.84**	1.70	1.19	2.90	4.95	6.68
300	2.59	2.02	1.31	**2.98**	5.43	6.11
1 000	2.61	1.81	**1.14**	1.29	3.18	5.91
3 000	2.57	1.86	**1.17**	1.15	1.63	2.95
10 000	2.57	1.85	**1.15**	1.23	1.15	1.24
30 000	2.58	1.84	**1.19**	1.12	1.02	1.17
100 000	2.58	1.83	1.17	1.26	**1.07**	1.15
200 000	2.58	1.83	1.17	1.16	**1.02**	1.25
300 000	2.58	1.81	1.17	1.18	**1.04**	1.16

value than GIC(2). Variation of the penalty factor α in GIC between 2 and 5, trying greater values for the slot width w and using different resampling frequencies can solve ambiguities in marginal situations. For those simulation runs where AR(1) was selected with $w = T_r/4$, it turned out that the selected order became AR(3) for $w = T_r/2$, with PE_s^T between 1.16 and 1.22. A larger slot width gives a smaller missing fraction and more reliability in order selection. Generally, the selected AR order and the shape of the spectrum are almost the same for several choices of w. It is a subject of future research to select the best value for w automatically.

Figure 3 gives an idea about the variability of the ARMAsel-irreg algorithm [7] if 1000 irregular observations are used in each simulation run. The AR(3) spectrum has been presented for each run, with 1.28 as an average value for PE_s^T. AR(3) has been selected with GIC in 14 runs; in the other 11 runs AR(2) has been selected. The average quality PE_s^T for selected orders was 1.42. The plot for selected orders would look similar to figure 3, with a couple of additional lines that are somewhat higher at the end of the interval because they belong to AR(2) models. Figure 3 gives the results for $w = T_r/2$, with an effective number of observations γN approximately 100. The results for $w = T_r$ give a higher bias and a smaller variance and for $w = T_r/4$ the bias would be smaller at the cost of a higher variance. However, the average model quality PE_s^T is about the same for all values of w, also for other series of 25 runs. It has been verified that more observations give less variability in the estimated spectra, which converge to biased spectra, with shift bias and probably missing-data bias included. Repeated simulations with $N = 30\,000$ and $w = T_r/4$ had as average PE_s^T in 25 runs the value 1.186 and all realizations had a PE_s^T value between 1.15 and 1.22. This value is greater than 1.13 of table 1 because of missing-data bias. The individual estimated AR(3) spectra were almost identical for $N = 30\,000$. The biased average was similar to the AR(3) estimate in figure 2. The individual variations had about the range of the plotted line width of the true benchmark spectrum in figure 3. The selected order was AR(3) in all runs.

In practice, it is always advisable to try various values for w. In examples where the shifting bias is smaller than the aliasing effect, $w = T_r$ would be the best choice. In contrast, if

the shift bias is greater as in figure 1 and if many observations are available, a smaller slot width would be preferable. The best choice for w is a compromise. For $N = 1000$ and variation of w, estimated AR(3) models gave the PE_s^T values 1.30, 1.16, 1.14, 1.18, 1.14, 1.51, 1.62 and 2.66 for $T_r w = 1$, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64 and 1/128, respectively, for the data that have been used in figure 2. The selected orders of ARMAsel-irreg, however, would be 3, 3, 3, 2, 1, 1, 2 and 1. For a smaller slot width, the effective number of observations decreases too much, which gives less accurate parameter estimates. It is not yet possible to use automatic methods to select the best slot width w. However, the selected order was 3 for the three largest slot widths and the AR(3) spectra were close for some slots that were still smaller.

Table 2 gives the accuracy of estimated AR models for increasing sample sizes. The effective number of observations is about $N/20$ for the given resampling time and slot width. This means that the missing fraction in the multi-shift-slotted NN resampled signals is about 95%, independent of N. With much less than 100 observations, the effective number is too small to obtain reliable estimates. It is remarkable that the accuracy of AR models of orders up to 4 hardly improves if more data are available. The truncated true values for PE_s^T are 2.52, 1.52 and 1.13 for the orders 1, 2 and 3, respectively. It is evident that missing-data bias is present for all sample sizes. The realized values for PE_s^T are already close to some biased expectation with aliasing bias, truncation bias, shift bias and missing-data bias included, for $N = 100$ for the orders 1, 2 and 3. It cannot become much smaller for greater sample sizes. The truncation bias is the largest bias contribution for AR models up to order 4. For higher orders, the main bias contribution would be the bias in (11) of shifting the observation times to a grid. However, for small sample sizes, the variance contributions are the dominant contributions PE_s^T in models of order 4 and higher. All model qualities are given for a single run. The variation in the columns demonstrates that the AR models of orders 4, 5 and 6 will still need more than 300 000 observations before they converge in all simulation runs. However, it should be realized that all estimated spectra with PE_s^T less than 1.1 are rather accurate.

In repeated simulations with $N = 100$, $T_r = T_0/5$ and $w = T_r$, the order 0 with $PE_s^T = 18.76$ was never selected, mostly the order 1 or 2, and sometimes the order 3, with $PE_s^T = 1.19$ for this value of w. The effective number of observations is about 20 then. Therefore, even for this very small sample size, some relevant information about the spectral shape is obtained. The estimated AR(1) spectrum is quite close to the true spectrum for the frequency range below 1000 Hz, in all simulation runs. Only for N less than about 15 will order 0 be selected for $T_r = T_0/5$ and $w = T_r$.

Eventually, for much greater values of N, the value of PE_s^T will converge to 1.04 for all model orders greater than 5 for the slot width $w = T_r/4$. The bias will not decrease if more observations are available; only the estimation variance will become smaller. Due to this estimation variance, the accuracy of the estimated AR(6) model is still significantly worse than its biased expectation for N equal to 300 000. The accuracy of the estimated AR models of orders higher than 5 will become better for larger sample sizes, but never better than 1.04 if $w = T_r/4$. Asymptotically, only a smaller slot width can give an improvement.

For higher model orders, the estimated spectra converge to the true biased spectrum, as given by the Fourier transform of the biased autocorrelation in (11). The estimated AR(5) spectra with PE_s^T equal to 1.02 or 1.04 in table 2 are almost identical with their biased expectation in figure 1, within the line width for $N > 100\,000$. The accuracy of estimated autocorrelation functions is also given by PE_s^T. For N greater than 100 000, order 5 was selected in most simulation runs. Often, but not always, the AR model which has its estimated spectrum closest to some biased expectation is selected with the criterion GIC(p) of (5), not the model that is closest to the true continuous spectrum.

6. Benchmark data with disturbed sampling

The ARMAsel-irreg algorithm estimated spectra up to very high frequencies, even in bubbly flow where the time instants of the observations are not Poisson distributed. Theoretically, the only demand for the irregular time series algorithm seems to be that the true continuous signal is stationary and stochastic. The smallest inter-arrival distance between the irregular observations limits the resampling rate and the possible frequency range belonging to it. There is no reason to suspect that other distributions than Poisson in the arrival times will strongly influence the spectral estimates of ARMAsel-irreg.

The benchmark data [15] have the possibility to add some practical deviations from a regular Poisson distribution for the sampling times for the irregular data. The options chosen were to include [15]

- dropouts, where some time intervals have fewer observations, like in practice in bubbly flows;
- varying data rate;
- processor delay to simulate the finite measurement volume in LDA data;
- lower mean data rate, about 100 Hz;
- 10 000 observations instead of the default value of 100 000.

Results of a single run with all deviations mentioned are given in figure 4. They are representative of many other runs with the same characteristics of data and slot width. Due to the possible variation of the irregular sampling times, unexpected results can sometimes be found in such data sets. If there are no pairs with the distance T_r in a simulation run, the result differs from what happens if a couple of such closest pairs exist. The expectation of the number of pairs would be 100 for 10 000 Poisson-distributed sampling instants and the given data, resampling rate and slot width. The varying data rate gives more close pairs here; 132 for 10 000 observations for $w = T_r/2$. Obviously, if the irregular observation times happened to be equidistant in one run, the spectrum cannot be established further than the half the mean data rate. However, this theoretical possibility will not likely occur. There can be many reasons why an estimated AR model does not converge

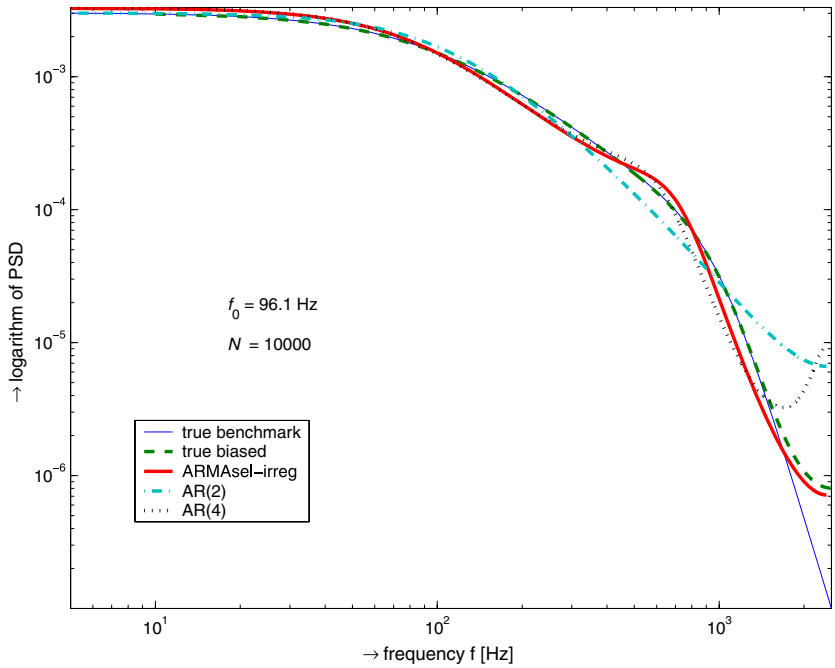

Figure 4. Estimated and true spectra of 10 000 irregular benchmark observations of a type 3 process of Nobach [15] with a disturbed sampling scheme, with a mean data rate of 96.1 Hz. MSSNNR used $T_r = T_0/50$ and $w = T_r/2$. ARMAsel-irreg selected the AR(3) model with $PE_s^T = 1.14$, while AR(2) and AR(4) give 2.02 and 1.78, respectively.

Table 3. PE_s^T for the estimated AR models in a single simulation run as a function of the slot width w for $T_r = T_0/50$. The data are the 10 000 irregular benchmark observations of a type 3 process of Nobach [15], sampled with a mean data rate of about 96.1 Hz, that have been used for figure 4. The selected order is printed in bold face. The final column gives the number of pairs at distance T_r.

w/T_r	AR(1)	AR(2)	AR(3)	AR(4)	Pairs T_r
1	2.59	2.01	**1.29**	1.83	258
1/2	2.58	2.02	**1.14**	1.78	132
1/4	2.58	2.02	**1.14**	1.12	59
1/8	2.61	1.85	**1.37**	2.48	37
1/16	2.59	1.61	**1.68**	2.04	15
1/32	**2.73**	1.64	1.60	2.18	7

to a good spectrum with 10 000 irregular observations. The average quality of AR(3) models is close to 1.14 for repeated runs. However, the average quality of automatically selected models will be higher, because different model orders are selected in some runs.

It is remarkable that dropouts, varying or low data rates and processor delay have hardly any influence on the accuracy of the spectral estimates, as long as the true continuous process is a stationary stochastic process. To the author's knowledge, no other algorithms have this property. Furthermore, other algorithms do not estimate spectra at frequencies much higher than the mean data rate and require many more data to estimate useful spectra.

Table 3 gives the results of the data of figure 3 if different slot widths are used. No accurate models can be estimated any longer if the effective number of observations, quantified by the number of consecutive pairs in the MSSNNR signal, is too small. The quality of the spectra and of order selection is not very sensitive to the slot width here. Generally, order selection is most reliable for the largest slot width, if the influence of bias on the spectra is not dominant. Always, lower orders will be selected if the slots become so narrow that the number of pairs decreases too much.

With small effective numbers of observations, order selection is not very accurate or reliable. Too many observations are so far from their predecessor that they cannot be predicted at all. Hence, they do not contribute to a reduction of the likelihood of the data. The likelihood in (5) is only reduced if observations can be predicted. Therefore, and because of the high penalty in (5) for large missing fractions, too low orders are often selected.

7. Sparse data with Poisson sampling

Irregular data are called sparse if the desired resampling rate is much higher than the mean data rate. This means that the missing-data fraction after MSSNNR is close to 1 for sparse data. The resampling rate was 50 times higher than the mean data rate in the example of figure 4 with disturbed sampling. It is unknown what might be the highest possible resampling rate for given irregular data. This will depend on the number of pairs at the resampling distance T_r, which depends on the probability density of the sampling instants. A sparse example with a Poisson distribution will be investigated here, with one very significant spectral peak that is generated by a continuous-time AR(2) process, without other spectral details [15]. The best discrete-time approximation for this process is an ARMA(2,1) process [16], which in turn is equivalent to AR(∞). Higher order true AR parameters will be small. Therefore, it can be expected that the discrete-time AR(2) model is a reasonable approximation with very

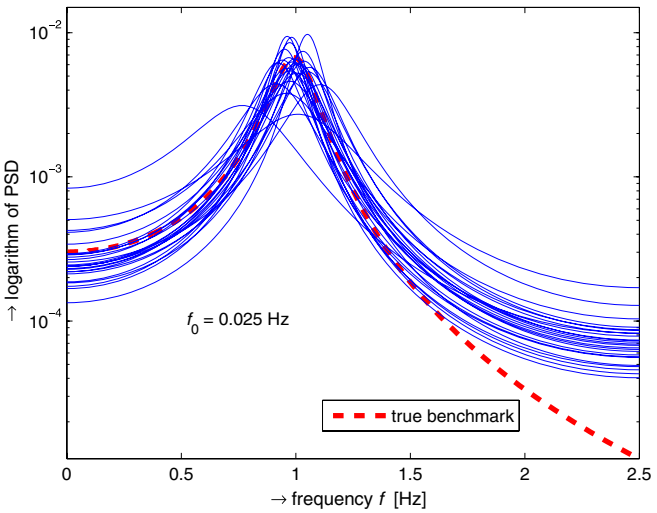

Figure 5. The true spectral density and 25 estimates of the AR(2) spectrum of MSSNNR, estimated with ARMAsel-irreg from $N = 2000$ sparse observations. The resampling frequency is 200 times the mean data rate f_0, with $T_r = 0.005T_0$, $w = T_r$. The average effective number of observations γN is 10 here, which is enough to compute some very significant AR parameters.

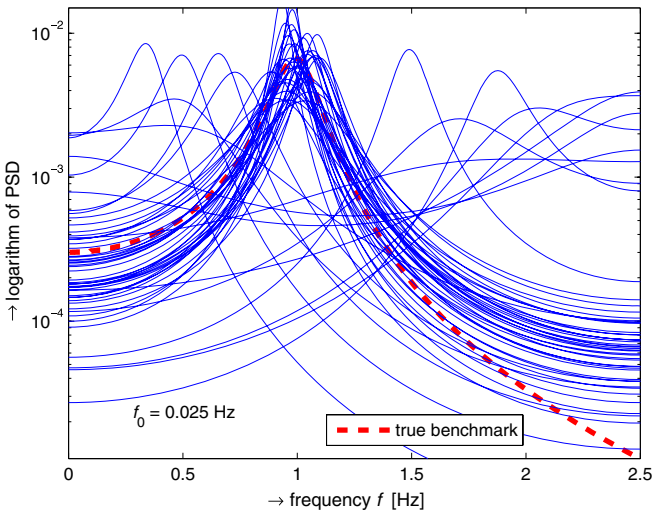

Figure 6. The true spectral density and 50 estimates of the AR(2) spectrum of MSSNNR, estimated with ARMAsel-irreg from $N = 1000$ sparse observations that are obtained by splitting each of the 25 data sets of figure 5 into two halves. The resampling frequency is 200 times the mean data rate f_0, with $T_r = 0.005T_0$, $w = T_r$. The average effective number of observations γN is 5 here, which is not always enough to compute an accurate spectrum.

significant values for the two parameters. However, the actual truncation and aliasing bias contributions will depend on the resampling frequency. The slot width is taken equal to the resampling distance to obtain as many pairs as possible at the resampling distance T_r. This specific example has been chosen because a few effective observations will be enough to obtain good estimates for the AR(2) model.

Figure 5 gives the true continuous spectrum and 25 AR(2) spectra estimated from 2000 irregular observations each. The resampling rate was 5 Hz which is 200 f_0, the maximum frequency of the estimated spectrum is 100 f_0 and the data have a spectral peak at about 1 Hz which is 40 f_0. The average effective number of observations NT_r/T_0 is 10 for each run if the slot width w is equal to T_r. Figure 5 shows the remarkable capacity of ARMAsel-irreg to estimate spectra for frequencies far above the mean data rate. In each individual estimate, a spectral peak near 1 Hz is found. The AR(1) model cannot describe a spectral peak, but the peak is present in the estimated higher order models. The spectrum becomes very inaccurate for higher AR orders, but those orders would never be selected. Even much higher resampling frequencies $1/T_r$ can be used for the same mean data rate if more observations were available, as long as the effective number of observations NT_r/T_0 is greater than about 5 or 10. In MSSNNR for a Poisson distribution, this effective number of observations can be seen as the number of observations with a distance of T_r. It is obvious that the occurrence of some pairs of observations with that distance is the minimum requirement to estimate the spectrum up to the frequency $f = 1/2T_r$. For the process of figure 5, ten effective observations are sufficient. However, other processes may require a higher effective number of observations, e.g. because they need higher AR orders for a reliable spectral estimate.

ARMAsel-irreg can also deal with disturbed sampling rates as long as at least a couple of pairs are found at the distance of about T_r.

The average PE_s^T for the 25 simulation runs in figure 5 was 1.23. The individual accuracies varied between 1.12 and 1.57 for the spectra in figure 5. The expectations of PE_s^T for the truncated true AR models are 3.75, 3.26, 1.117, 1.026 and 1.009 for the orders 0, 1, 2, 3 and 4, respectively, for the resampling frequency that is used here. The accuracy of the AR(2) model for a simulation run with 50 000 irregular observations with the same resampling and mean data rates was $PE_s^T = 1.12$, close to the true truncated value. The estimated spectra of repeated runs with many more than 2000 observations converged to a narrower spectral range, with an average accuracy about 1.12 due to the truncation bias.

Taking fewer observations or a still higher resampling rate gives as the result of ARMAsel-irreg that the white noise model or AR(0) will be selected in the example of figure 5. This result of order selection can always be expected, for all sorts of irregular data, if the effective number of observations is only 3 or less. If no pairs of observations are found within the length of the autocorrelation function, the irregular observations look like uncorrelated white noise. This is the limiting case for extremely sparse data of all possible processes.

Each of the 25 simulated signals that are used in figure 5 has been divided into two signals of length 1000. From those 50 shorter signals, AR(2) spectra have been estimated with ARMAsel-irreg. The spectra are shown in figure 6. The variability is much greater than in figure 5. The average PE_s^T for the 50 simulation runs in figure 6 is 5.3. The individual accuracies varied between 1.12 and 13.4 for the spectra in figure 6. The average value 5.3 was computed from 35 realizations with $PE_s^T < 1.5$ and 15 realizations with $PE_s^T > 1.5$. These 15 realizations can all be recognized individually in figure 6. Taking fewer observations will give still more variability between the simulation runs, until at last no realization will find a spectral peak around 1 Hz.

However, even if the expected effective number of observations is zero, one of many realizations can have a couple of nearby observations and still deliver a reasonable spectral estimate. It has been verified that this may occur, even for 100 observations of the Poisson-distributed data in this section, where sometimes a spectral peak near 1 Hz has been found, with $PE_s^T \approx 1.2$. If observation times are not Poisson distributed but have clusters at some places, no general rules can be given.

Some simulations have been carried out with $N = 30\,000$ observations, a mean data rate $f_0 = 0.001\,25$ Hz and $T_r = 0.2$ s to compute the spectrum until 2.5 Hz. The expected effective number NT_0/T_r is 7.5 here. The results are comparable with single estimated spectra in figure 6. Sometimes an accurate spectrum was found and in other runs there was no peak or a peak at the wrong frequency. The highest frequency in the discrete-time spectrum was 2.5 Hz, which is 2000 times the mean data rate. Projected on a different frequency scale, this means that one can study daily variations from Poisson-distributed observations with a mean data rate of less than one observation per year.

The highest discrete-time frequency that can be studied for Poisson-distributed observations with a given mean data rate is a final question. Although the effective number γN is very important, it is not the only factor. Also, the total number of observations has a strong influence. The likelihood LH is in practice diminished only by nearby observations. Very roughly speaking, there are about γN influential contributions and $(1-\gamma)N$ remaining observations. These have no perceptible contribution to the likelihood because no other observation was close enough to give a significant reduction of the likelihood by any AR model. However, these $(1-\gamma)N$ observations together will always give some variations that mask the γN effective contributions. Therefore, repeated simulations with a mean data rate less than 0.0025 Hz and more observations, such that $\gamma N \approx 10$, are less reliable than the spectra in figure 5. On the other hand, no matter how low the mean data rate of Poisson-distributed observations, there will always be a sample size N for which γN for a chosen resampling rate $1/T_r$ is great enough to obtain accurate spectra in theory. The computation time might become a limiting factor then.

The Lomb–Scargle method [4] can also detect high frequencies, but only if a truly periodic signal with a single frequency is sampled with a very small amount of white noise. For strong spectral peaks generated with AR processes, the bias of the Lomb–Scargle method is too large [5]. It has been verified that the Lomb–Scargle method did not detect the peak in the spectrum of figure 5 for the given data, not even if the data rate were 40 times higher. Both the slotted resampled data and the original irregular data have been tried in the Lomb–Scargle algorithm. No matter how many observations are available, the Lomb–Scargle method did not detect a peak for irregular observations of this process if the mean data rate was less than the peak frequency 1 Hz. Only if the mean data rate is higher than the frequency of the peak has a slight indication of a periodicity been found in the estimated Lomb–Scargle spectrum for $N = 2000$. No other method is known in the literature that can estimate spectra at those very high frequencies from relatively short data sets. Slotted correlation methods [3, 9] would obtain an expected number of contributions for each correlation lag that is given by the effective number γN for a given resampling rate and mean data rate. About ten contributions for each lag are certainly not sufficient to obtain a useful spectral estimate. ARMAsel-irreg is able to detect periodicities in very sparse and irregular physical, astrophysical, geophysical, medical or meteorological data and is ready for interesting applications.

8. Conclusions

Irregular data can be transformed into an equidistant missing-data problem by MSSNNR (multi-shift slotted nearest neighbour resampling). The ARMAsel-irreg estimator fits AR, MA and ARMA models and automatically selects the best model order and model type. That model is used to compute the autocorrelation function and the spectral density.

The bias caused by shifting the irregular observation times to a regular resampling grid can be diminished by using a slot width that is smaller than the resampling distance. The spectra of the selected models are mostly close to a biased true spectrum, including the bias effects of aliasing, shift and missing-data bias. For models of too low AR orders, the truncation bias can be significant.

The algorithm always computed a model for the data without numerical problems. It performed well on benchmark data. Low order AR models can give an accurate description for various spectral shapes, with a steep slope or with a strong peak. Spectra can be computed up to frequencies higher than a thousand times the mean data rate. In simulations with few irregular data or with strong deviations from the Poisson distribution for the sampling instants, the results of ARMAsel-irreg are much better than what can be obtained with any other known spectral estimation technique.

Areas for future investigations include the automatic choice of the slot width, the selection of the model order if high order AR candidates are allowed, the missing-data bias, the computing time and the application to real practical data.

References

[1] Mateo J and Laguna P 2000 Improved heart rate variability signal analysis from the beat occurrence times according to the IPFM model *IEEE Trans. Biomed. Eng.* **47** 985–96

[2] Thiebaut C and Roques S 2005 Time-scale and time-frequency analyses of irregularly sampled astronomical time series *Eurasip J. Appl. Signal Process.* **15** 2486–99

[3] Benedict L H, Nobach H and Tropea C 2000 Estimation of turbulent velocity spectra from laser Doppler data *Meas. Sci. Technol.* **11** 1089–104

[4] Scargle J D 1982 Studies in astronomical time series analysis: II. Statistical aspects of spectral analysis of unevenly spaced data *Astrophys. J.* **263** 835–53

[5] Broersen P M T, de Waele S and Bos R 2004 Autoregressive spectral analysis when observations are missing *Automatica* **40** 1495–504

[6] Jones R H 1981 Fitting a continuous time autoregression to discrete data *Applied Time Series Analysis II* ed D F Findley pp 651–82

[7] Broersen P M T and Bos R 2006 Estimating time-series models from irregularly spaced data *IEEE Trans. Instrum. Meas.* **55** 1124–31
[8] Larsson K L and Söderström T 2002 Identification of continuous-time AR processes from unevenly sampled data *Automatica* **38** 709–18
[9] Tummers M J and Passchier D M 1996 Spectral estimation using a variable window and the slotting technique with local normalization *Meas. Sci. Technol.* **7** 1541–6
[10] van Maanen H R E, Nobach H and Benedict L H 1999 Improved estimator for the slotted autocorrelation function of randomly sampled LDA data *Meas. Sci. Technol.* **10** L4–7
[11] Adrian R J and Yao C S 1987 Power spectra of fluid velocities measured by laser Doppler velocimetry *Exp. Fluids* **5** 17–28
[12] Simon L and Fitzpatrick J 2004 An improved sample-and-hold reconstruction procedure for estimation of power spectra from LDA data *Exp. Fluids* **37** 272–80
[13] Müller E, Nobach H and Tropea C 1994 LDA signal reconstruction: application to moment and spectral estimation *Proc. 7th Int. Symp. on Applications of Laser Technology to Fluid Mechanics (Lisbon)* paper 23.2 1–8
[14] Jones R H 1980 Maximum likelihood fitting of ARMA models to time series with missing observations *Technometrics* **22** 389–95
[15] Nobach H 2001 LDA Benchmark Generator III, http://www.nambis.de
[16] Priestley M B 1981 *Spectral Analysis and Time Series* (London: Academic)
[17] Broersen P M T 2006 *Automatic Autocorrelation and Spectral Analysis* (London: Springer)
[18] Broersen P M T 2006 Automatic spectral analysis with missing data *Digit. Signal Process.* **16** 754–66
[19] Akaike H 1970 Statistical predictor identification *Ann. Inst. Stat. Math.* **22** 203–17
[20] Harteveld W K, Mudde R F and van den Akker H E A 2005 Estimation of turbulence power spectra for bubbly flows from laser Doppler anemometry signals *Chem. Eng. Sci.* **60** 6160–8