Spectral Analysis of Irregularly Sampled Data with Time Series Models

Piet M.T. Broersen*

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

Abstract: Slotted resampling transforms an irregularly sampled process into an equidistant missing-data problem. Equidistant resampling inevitably causes bias, due to aliasing and the shift of the irregular observation times to an equidistant grid. Taking a slot width smaller than the resampling time can diminish the shift bias. A dedicated estimator for time series models of multiple slotted data sets with missing observations has been developed for the estimation of the power spectral density and of the autocorrelation function. The algorithm estimates time series models and selects the order and type from a number of candidates. It is tested with benchmark data. Spectra can be estimated until frequencies higher than 100 times the mean data rate.

1. INTRODUCTION

Continuous-time processes are sometimes observed at irregular observation times. Irregular intervals may be caused by wireless sensor networks of 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 [3]. LDA (laser Doppler anemometry) is an important application in measurement science, where the velocity can only be measured if a seeding particle passes through the measurement volume [4].

The continuous spectral density of irregular data can be computed at an arbitrary number of frequencies with the method of Lomb-Scargle [5, 6]. That method fits sine waves of selected frequencies to the data by minimizing the sum of squared errors. Applying this method to equidistant data would give the same result as the periodogram, if equidistant frequencies are chosen. However, various examples show a significant bias in the spectral estimates if the method is used for irregular data [4].

The true spectral density of continuous-time irregular data is infinitely wide in the frequency domain. A maximum likelihood (ML) approach has been described for the estimation of continuous-time AR models [7]. However, inspection of the surface of the likelihood, computed with that algorithm, as a function of the AR parameters showed that it was very rough [8], with many local maxima. Numerical problems prevented the convergence of the continuous-time ML estimates to a useful model for irregular data. No general applicable and reliable continuous-time spectral estimator is available yet for practical data.

Slotted autocorrelation estimation discretizes the distance between two observations to slots of width Δ. The product of two irregular observations contributes to the slotted autocorrelation at a certain lag $k\Delta$ if their distance is between $(k - 0.5)\Delta$ and $(k + 0.5)\Delta$ [9]. Unfortunately, slotted correlation estimates are not positive semi-definite and they do not fulfill the theoretical requirements for being a true autocorrelation function. Improvements have been introduced: local normalization [9] and fuzzy slotting. The spectral variance has been reduced further with variable windows [9]. However, no known variant of the estimated slotted autocorrelation functions is positive semi-definite. All methods fail to consistently produce a spectrum that is positive for all frequencies. The autocorrelation fit of slotting or its spectral quality is a matter of taste, not of any objective quality measure.

A large group of estimators uses resampling of the irregular data at equal time intervals. Equidistant resampling techniques replace an irregularly sampled continuous-time signal by an equidistant discrete-time signal, with only observations on a grid. After resampling, the discrete-time equidistant data can be analyzed with the conventional spectral analysis techniques [4] or with modern time series models [8]. Sample and Hold (S+H) reconstruction uses the *true* measured values of the irregular observations at shifted equidistant times. Intuitively, it seems to be preferable to interpolate the irregular observations and to substitute the value of the reconstructed signal on the grid times. This idea has been tested with simple linear interpolation and with more sophisticated methods like fractal reconstruction or the projection onto convex sets [10]. 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 S+H resampling [4, 10].

S+H reconstruction 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 error at that frequency is already 50 % [11]. These effects can in theory be eliminated by using a refined S+H estimator and subtracting the reconstruction noise from the spectral estimate [4]. This refinement explicitly uses a Poisson distribution for the observation instants to improve the estimate. It is not applicable to arbitrary distributions of arrival times. Refinement and noise suppression can take place in the time [4] or in the frequency domain [12]. If applicable, it can enlarge the useful fre-

*Address correspondence to this author at the Department of Multi Scale Physics, Delft University of Technology, The Netherlands; Tel: +31 15 2786419; E-mail: p.m.t.broersen@tudelft.nl

1876-8253/08

quency range of the spectral estimates somewhat, from $f_0/2\pi$ to a maximum of perhaps about f_0 [12].

Equidistant resampling will sometimes substitute the same irregular observation at more grid nodes. This occurs if the maximum distance between the irregular observations is greater than the resampling distance T_r. This multiple use of the same irregular observation creates a very large bias in correlation function and spectrum. That specific bias term can be avoided with the slotting principle. The slotting principle can be applied to the nearest neighbor (NN) resampling of irregular data on a regular time grid [8]. Slotting gives an equidistant signal, with data missing at those grid nodes that are further than half the slot width w away from an actual irregular observation. An observation is only accepted at a node if its distance is less than $w/2$. Bias is caused by the spectral aliasing of high frequencies and by the shifting the observation instants. The bias can be reduced by taking a higher resampling frequency $1/T_r$ or by making the slot width w smaller than T_r in multi-shift slotted NN resampling [8]. For $w = T_r/M$, M different discrete time series are obtained by shifting over the distance w. Those M signals can be used simultaneously to estimate time series models [8]. The highest spectral frequency $1/2T_r$ and the aliasing bias can remain the same while the smaller w reduces the shift bias.

Spectral estimation is much simpler for equidistant signals with data missing for than irregular data. The discrete-time spectral density has a finite frequency interval, until half the resampling frequency. Equidistant *missing-data* problems have been investigated recently [13]. For missing data, Jones described an efficient method to calculate the true likelihood for autoregressive (AR) processes [14]. In practical computations, that has much more favorable properties than his continuous algorithm [7]. An automatic time series algorithm with AR, moving average (MA) and combined ARMA models outperformed all other methods that have been described for missing-data problems [13]. This best performing method for missing data, with AR, MA and ARMA models as candidates for selection, can be applied to irregular data, if they are resampled with the multi-shift slotting principle [8]. Due to the bias caused by aliasing and by shifting the irregular times to a grid, order selection has only biased models as candidates. The ARMAsel-irreg algorithm [8] has been developed with simulated data as an automatic spectral estimator for irregular data. The acronym *ARMAsel-irreg* denotes the automatic *selection* of the best fitting order and type from *AR*, *MA* or *ARMA* candidate models. The suffix *irreg* denotes irregularly spaced data. The algorithm is available at the internet [15]. Sometimes, a correction method can reduce the bias of the selected time series model.

A comparison of ARMAsel-irreg and the slotted correlation method has been reported [16]. This paper studies the quality of automatically selected time series models for the resampled irregular data. Benchmark data from a website with LDA examples [17] have been used as a test signal. This facilitates a future comparison with the accuracy of existing or new spectral estimators that can be applied to the same irregular benchmark data. The selected and corrected spectra are not based on any assumption about the sampling scheme or the distribution of the irregular intervals between data. Important questions are the minimum required sample size to obtain accurate spectral estimates and the highest frequency that can be estimated from given irregular data. It will be verified whether a bias correction that removes AR poles in the high frequency range can be used automatically, without adverse effects for true spectral details.

2. TIME SERIES MODELS

A discrete-time ARMA(p, q) model for equidistant discrete-time observations x_n can be written as [18, 19]

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

where ε_n is a purely random process of independent identically distributed stochastic variables with zero mean and the innovation variance σ_ε^2. It is assumed that the data x_n are a stationary stochastic signal. 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) for the same continuous-time signal. Theoretically, every stationary stochastic discrete-time process can be described with an AR, MA or ARMA model [19]. AR(p) models have $q = 0$ and $p = 0$ gives MA(q). The parameters themselves are not important, but they act as a parametric description of spectrum and autocorrelation function. The only requirement to apply this discrete-time model (1) to continuous-time irregular data is that the power spectral density in the discrete-time model is given by a stationary process.

The power spectral density $h(\omega)$ of the model (1) as well as its the frequency range depend on the resampling distance T_r. The spectrum of the ARMA(p, q) process is fully determined by the parameters in (1) together with the variance σ_ε^2 and T_r

$$h(\omega) = \frac{T_r \sigma_\varepsilon^2}{2\pi} \frac{\left| 1 + \sum_{i=1}^{q} b_i \, \mathrm{e}^{-j\omega i} \right|^2}{\left| 1 + \sum_{i=1}^{p} a_i \, \mathrm{e}^{-j\omega i} \right|^2} \ , \ -\frac{\pi}{T_r} < \omega \le \frac{\pi}{T_r} \cdot \qquad (2)$$

In addition, formulas have been given to compute the autocorrelation function at lags kT_r directly from the parameters of (1) [18]. For a given signal x_n, the parameters in (1) can in principle be estimated by minimizing the sum of squares of the residuals $\hat{\varepsilon}_n$, which replace the innovations ε_n in (1) if that equation is used for estimation. In practice, other estimation algorithms than least squares are often preferred [18]. The Matlab program ARMAsel has been developed for the automatic estimation of the AR and MA parameters and is available [15]. The automatic, approximate maximum likelihood program ARMAsel-mis estimates the AR and MA parameters of equidistant missing-data problems [15]. Some modifications of that program to ARMAsel-irreg for irregular data have been given [8].

The Generalized information criterion GIC(p) is a missing-data AR order selection criterion [8]:

$$\mathrm{GIC}(p) = LH + \alpha p \ . \qquad (3)$$

It uses twice the minimized negative log likelihood, denoted LH, with a penalty factor α that depends on the missing fraction: 3 for less than 25 % missing, 5 for more than 75 % missing and 4 in the range between. The missing fraction

is the percentage of empty grid points after slotted resampling. The fraction of non-empty grid points after resampling with the period T_r is approximately given by T_r / T_0 where T_0 is the average distance between samples. T_0 is equal to $1/f_0$ and f_0 is known as the mean data rate.

The same criterion (3) with the fixed penalty 3 can be used for MA(q) or ARMA(p, q) models

$$\mathrm{GIC}(p+q) = LH + 3(p+q). \quad (4)$$

For MA models, the number of estimated parameters is q, with $p = 0$.

The accuracy of equidistant time series models can be evaluated with the prediction error or with the spectral distortion. That is the integral of the squared difference of the *logarithms* of the true and estimated spectra. Those and still several other measures are equivalent in practice [18]. The prediction error PE(p+q) of an ARMA(p, q) model is defined then as the squared error of the one step ahead prediction with the model in new fresh data. The normalized scaled version of the expectation $\mathrm{PE_s}(p+q)$ can be computed for benchmark signals without the generation of new data, by using only the known parameters and variance σ_ε^2 of the true generating process [18]:

$$\mathrm{PE_s}(p+q) = \frac{\mathrm{E}_y\left[\mathrm{PE}(p+q)\right]}{\sigma_\varepsilon^2}. \quad (5)$$

The expectation E_y in (5) denotes the expectation if an infinite length of fresh data would be used for the computation of $\mathrm{PE_s}(p+q)$. The minimum of the expectation of the normalized $\mathrm{PE_s}(p+q)$ with respect to the estimated parameters of an efficiently estimated unbiased ARMA(p, q) model, with $p + q$ parameters from equidistant data, is given by

$$\mathrm{E}\left[\mathrm{PE_s}(p+q)\right] = 1 + (p+q)/N. \quad (6)$$

Unbiased models have at least all truly non-zero parameters included, and furthermore they do not have any other bias source. The expectation for biased models will be greater than (6) and generally not depend on the sample size N. However, the $\mathrm{PE_s}(p+q)$ can be used as an objective accuracy measure for all time series models of irregular data, biased as well as non-biased [18]. For an efficient computation of (5), the true process parameters should be known for the frequency range of interest that is determined by the highest frequency $1/2T_r$. That information is available for the benchmark data that will be used in this paper [17].

3. ARMASEL-IRREG

In a first application of the ARMAsel-irreg algorithm to simulated irregular data, good results have been obtained for AR models where the order was known and less than five [8]. It has been demonstrated that using very high resampling rates is not a problem for the ARMAsel-irreg algorithm. The spectra of bubbly flow data [20] have been analyzed until frequencies that are more than 250 times higher than the mean data rate f_0. Good results could be found for low order AR orders for bubbly flow data. This showed that the algorithm can perform well until very high frequencies, even if the time instants of the observations are not Poisson distributed. The equidistant S+H resampling with refinement can only yield accurate spectra until about f_0 for Poisson distributed sampling [12]. The rough S+H spectra without refinement are only reliable until about $f_0/2\pi$ [11].

Slotted resampling replaces irregular sampling by equidistant sampling on a grid, with data missing. The missing fraction is determined by the number of empty grid points. It is approximately inversely proportional to the resampling distance T_r which determines the total number of grid nodes. Shifting the irregular observation times to a grid introduces shift bias and smaller shifts give less bias. The shift bias has been reduced by making w smaller than T_r [8] with multi-shift slotted NN resampling. The highest spectral frequency $1/2T_r$ with its aliasing bias can remain the same while the smaller w still reduces the shift bias. Making T_r a factor M times smaller gives a M times larger frequency range. Roughly, the AR model order should become about M times higher for the larger range to describe the same level of details in the smaller original frequency range belonging to T_r. However, the computational time and non-linear convergence problems of the likelihood calculation increase strongly with a greater missing fraction that belongs to a smaller T_r. A smaller slot width for a constant value of T_r reduces the shift bias without influence on the missing fraction. Taking $w = T_r / M$, with integer M, gives disjunct intervals where some irregular times t_i are not within any slot. Therefore, multi-shift slotted NN resampling (MSSNNR) has been developed, where M different equidistant missing data signals are extracted from one irregular data set. The equidistant sampling instants $nT_r + mw$ with non empty places for the M signals are given by

$$nT_r + mw - 0.5\,w < t_i \le nT_r + mw + 0.5\,w,\ m = 0, 1, \ldots, M-1, \quad (7)$$

where t_i denotes an irregular sampling instant. All slots of width w are connected in time. This MSSNNR signal is the input signal for ARMAsel-irreg, for $M \ge 1$. Those M signals from a single resampled irregular data set can be used simultaneously as if they are independent [8].

The first source of bias is due to aliasing. That bias is a contribution of frequencies above $1/2T_r$ to the discrete time frequency range below $1/2T_r$ [19]. The shift of irregular times to a grid introduces a second source of bias. A formula for the shift bias of MSSNNR has been given for Poisson distributed sampling instants [8, 21]. The bias in the frequency domain is similar to adding discrete-time white noise to the data. It has not much influence on the strong parts of the spectrum but it eliminates all weak parts below the noise level. It can be negligible for smooth spectra, but it will be influential for spectra with steep spectral slopes and deep valleys.

The smallest distances that are actually found between the irregular data give an upper limit to the frequency range and the smallest resampling distance that can be used. At least a few close observations with distance about T_r are required to estimate spectra up to the frequency $1/2T_r$. This is a hard limit for the highest resampling rate that can be used.

A comparison of the estimated spectra of time series models and of slotted autocorrelations has been made [16]. The slotted correlation never produced a spectrum that was positive over the whole frequency range. No positive semi-definite estimator with slotted correlations is available in the literature. Next to the negative spectral parts, always local

spectral peaks were found. That hampers the interpretation of the slotted autocorrelation spectra. However, in some examples and in a small part of the frequency range where the spectrum is strong, the spectra may look well for large sample sizes [9]. This is a subjective quality measure. Objective accuracy measures like PE_s of (5) are based on or equivalent to the logarithm of the spectrum and they cannot be applied to spectra with negative parts. This paper concentrates on the objective quality in the whole frequency range, with the PE_s of (5).

Theoretically, the ARMAsel-irreg time series algorithm can be used for all irregularly sampled stationary stochastic continuous-time signals, independent of the sampling distribution. The sampling instants must preferably be independent of the values of the measured signal itself. Irregularities in the arrival times will not distort the spectral estimates of ARMAsel-irreg. However, velocity bias is an example in LDA analysis where the probability of an observation depends on the amplitude of the signal. It has been observed that this type of bias may sometimes distort the estimated ARMAsel-irreg spectra [21]. Also examples without any distortion have been seen and no general rule can be given yet for amplitude-dependent sampling.

The bias due to aliasing and to the shifting of the irregular times to a grid can have a peculiar effect in order selection. Order selection has been developed to reduce the *truncation* bias [21] that is caused by incomplete lower order models without statistically significant parameters. Roughly speaking, the selected model order includes all significant parameters that give a reduction of the likelihood that is greater than the statistical inaccuracy of estimating those parameters. Sometimes, models are selected that look apparently worse than other candidate models that are not selected. It has been verified that no errors have been made in the computation of the likelihood function that is minimized in the estimation of the parameters of AR models [8]. An explanation can be given. For spectra with strong peaks in the lower frequency range, low order AR models describe the strongest spectral details that are found at low frequencies. The estimated spectra at higher frequencies are just an extrapolation of the strong low order model parameters estimated for primarily the low frequency range. The extrapolation has steep slopes at high frequencies. Higher order estimated models will correct that by introducing additional peaks in the high frequency range. Therefore, higher order models can give a closer approximation to the spectrum with shift and aliasing bias included. Order selection has only biased models as candidates, with the observations shifted to a grid node. It has been found already that the removal of peaks of the selected model in the highest part of the discrete-time frequency range could be beneficial to reduce the shift bias. This removal will be tested here as an additional feature of the automatic ARMAsel-irreg algorithm. One would like to select the estimate that is closest to the true spectrum, without the resampling bias of aliasing and time shifting, but that is not possible. Order selection has only candidates with bias available.

4. BENCHMARK DATA, TYPE 5

Several test signals can be generated with a benchmark generator; the generating program and its description are available online [17]. Spectral type 5 of the benchmark gives an excellent example to demonstrate the possibilities of ARMAsel-irreg, because low order AR models give the best fit to those simulation data, until high frequencies. A strong peak dominates the spectrum. No significant details are present in the high frequencies of the true spectrum, just a steep constant slope in the log-log representation of the spectrum. Therefore, the effect of aliasing is very small here and all details that are found in estimated spectra at higher frequencies are wrong. They are probably caused by the bias due to the shifting of observation times.

Fig. (**1**) shows the true and two estimated spectra of 100 irregular observations. The not corrected spectrum cannot belong to a sampled continuous-time spectrum, unless it is very strongly aliased. Continuous spectra are infinitely wide and should decrease at higher frequencies. Otherwise, the integrated power over the whole frequency range cannot remain finite. Therefore, the selected AR(3) spectrum has a spurious peak at the end of the discrete-time frequency range, which is caused by the shifting bias [21].

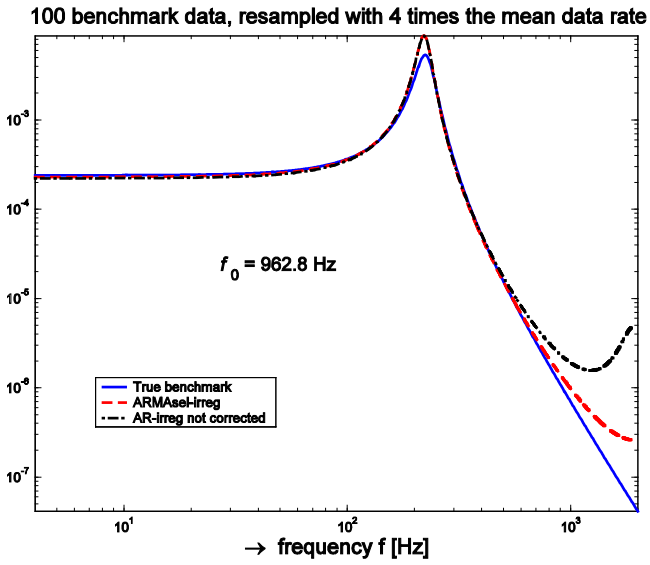

Fig. (1). True benchmark spectrum of type 5 until 1925.6 Hz, the automatically selected AR(3) estimate for $T_r = T_0/4$ and $w = T_r/2$, with $PE_s(3) = 1.834$ and the correction to the ARMAsel-irreg AR(2) estimate with $PE_s(2) = 1.072$. The estimated AR(2) model had $PE_s(2) = 1.094$.

The correction to lower order models is based on elimination of specified poles of the selected AR polynomial $A_p(z)$. That is defined as

$$A_p(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}. \qquad (8)$$

Poles of $A_p(z) = 0$ with a positive real part belong to peaks in the first half of the frequency domain from 0 to $1/(4T_r)$ Hz and poles with negative real parts belong to frequencies from $1/(4T_r)$ until $1/(2T_r)$. The selected AR(3) model has a pole on the negative real axis. That pole causes the peak in Fig. (**1**) at the end of the frequency domain. The correction consists of giving all poles with a negative real part the value zero. Afterward, the AR parameters are computed from the remaining poles, which are all in the first half

of the discrete-time frequency range. In this way, the order of the selected model is reduced, but only if there are estimated poles with negative real parts. In all simulation runs of the benchmark spectrum in Fig. (**1**) where AR(3) has been selected, the PE_s value improved very much by removal of the negative pole and became better than the estimated AR(2) model. The explanation is simple. In equidistant AR estimation, reflection coefficients for increasing model orders are estimated successively, where all previous reflection coefficients remain the same [18]. In missing-data or irregular sampling problems, all reflection coefficients are estimated simultaneously and vary for increasing orders [13]. The AR(2) model has been estimated for the whole frequency range. The AR(3) model, selected in Fig. (**1**), has an additional pole at the highest frequency. That pole has mainly influence on the last part of the frequency range. That gives the opportunity for the first two parameters to estimate a better description for the first part of the frequency range. That follows from the fact that the corrected lower order AR(2) model almost always gives a smaller $PE_s(2)$ than the AR(2) model that has been estimated directly from the data. This has been observed in simulations with many different examples: the best lower order models are found from corrected higher order models [16].

The estimated PE_s value of the corrected model, with reduced shift bias, is very close to the $PE_s(2) = 1.065$ that can be computed from the parameters of the truncated true aliased AR(2) process. That means that the best possible AR(2) model could already be estimated from only 100 irregular observations. It would not become better if more data would be available, because of the bias. The true order of the generating process is AR(∞), with small and very small parameters for all orders greater than 5. The estimated AR(3) model that was selected here was much less accurate for those 100 observations due to the variance of estimation from very small samples, with $PE_s(3) = 1.834$. For very large data sets, with $N > 100000$, AR(3) would become selected more often, with as lower limit $PE_s(3) = 1.004$ for the truncated true AR(3) process parameters. The influence of aliasing is very small in this example. The PE_s between the continuous spectrum and the aliased spectrum is only 1.015 for the frequency range of Fig. (**1**).

The simulation of Fig. (**1**) presents only one possible realization from many runs. For other realizations of 100 irregular observations with the given mean data rate, the AR(2) model was selected in about 90% of all runs, sometimes ARMA(2,1) was selected and AR(3) was selected in about 10% of the runs. There was always a spectral peak at the end of the frequency range if AR(3) was selected. The value for $PE_s(2)$ was always lower than 2 for the selected AR(2) model; it was often about 1.20 and only occasionally a value greater than 1.5 has been found. However, it was never less than 1.065, because the value for the AR(2) model is always greater than the value 1.065 that belongs to the truncated true aliased AR(2) process. Only higher order models can have a smaller value for PE_s.

$PE_s(0)$ is 92.8 for this example and $PE_s(1)$ is about 10.9. Therefore, those low model orders with very poor quality are never selected for 100 observations or more. The estimation and order selection are easy in this example because the likelihood *LH* and the prediction error are so much reduced by the estimated parameters of orders 1 and 2. For $N > 8$, the AR(2) model was selected for a particular run, with $PE_s(2)$ less than 2 for all sample sizes. For $N \leq 8$, order 0 was always selected. In other simulation runs with the same process, the AR(0) was sometimes selected for *N* about 20. For $N > 50$, always models with a spectral peak near 220 Hz have been selected. It should be noticed that 50 irregular observations is a very small sample size for irregular data, particularly if resampling at a frequency higher than the mean data rate is used. The variability of spectral estimates for small irregular samples is much greater than for the same sample size in equidistant observations. That is mainly caused by the actual shift of the most influential irregular observations, which are those at inter-arrival distances of only a few times T_r. The actual distribution of the irregular sampling times has a strong impact in very small samples.

It seems to be a limitation that poles with a negative real part are always eliminated by the correction of ARMAsel-irreg. However, the highest spectral frequency becomes twice larger by taking a double resampling rate. Details in the second half of the frequency domain are moving to the first half for the higher resampling rate. In practice, they can always be included by using a higher resampling rate. In simulations, like in this paper, it is known where spectral details can be expected. If the range of interesting frequencies is unknown, as happens in practical data, the frequency range can be extended by repeatedly doubling the resampling rate. Unfortunately, higher order models are required for the more densely resampled data. The best compromise is still an open question.

5. BENCHMARK DATA, TYPE 2

Spectral type 2 of the benchmark is a more challenging example. It gives a spectrum with a constant negative slopes in the logarithm of the spectrum, that is given by $\sim 2^{-f/300}$ [17]. Extra difficulties in the sampling scheme have been applied in the generated data to verify that ARMAsel-irreg is not sensitive to the (not Poisson) sampling distribution. The chosen options are [17]

- drop outs, where some periods in the time have much less observations, which occurs in LDA practice in bubbly flow
- varying data rate, which is common in practice
- processor delay, to simulate that only one particle at a time can be observed in the finite measurement volume of LDA
- low data rate
- much less than 100 000 data, which is the default value

It is always possible to estimate 10 AR parameters, often 20 can be estimated and occasionally 50 from irregular data. However, the computing time increases dramatically and a more efficient computer program would be required to estimate so many parameters. However, the non-linear search of 50 parameters simultaneously will always remain time consuming. Models for irregular observations cannot be estimated recursively, in contrast with equidistant observations with efficient recursive algorithms [18]. The estimated parameters of the lower order models have an extra missing-

data bias as long as the current model order is lower than the true order and can only serve as non-linear starting values for higher order models [13].

It has been verified that drop outs, processing delay and varying low data rates have hardly any influence on the accuracy of the spectral estimates of ARMAsel-irreg because the true process is stationary stochastic. No other positive semi-definite algorithms have this property to the author's knowledge. Furthermore, other algorithms cannot reliably estimate spectra at frequencies much higher than the mean data rate and they require many more data [4].

Fig. (**2**) gives the true continuous spectrum in the chosen frequency range, the aliased true spectrum that belongs to that range and the selected AR(1) spectrum. The AR(1) pole is positive and no correction takes place. The spectrum has no strong details and is rather flat. The values for the truncated true aliased spectrum are $PE_s(0)$ = 1.1441, $PE_s(1)$ = 1.0007 and $PE_s(2)$ = 1.0006. It would require much more than 10000 equidistant observations to let the AR(2) model be statistically significant in (6). PE_s values are computed with the aliased true process as reference. That is done because the aliasing effect cannot be reduced by filtering in irregularly sampled data. It will always be present in estimated spectra if equidistant resampling is used. The PE_s of the true continuous spectrum from 0-1000 Hz is 1.0183 in Fig. (**2**). In almost all simulation runs, the estimated spectrum of the selected model is much closer to the aliased spectrum than to the true continuous-time spectrum.

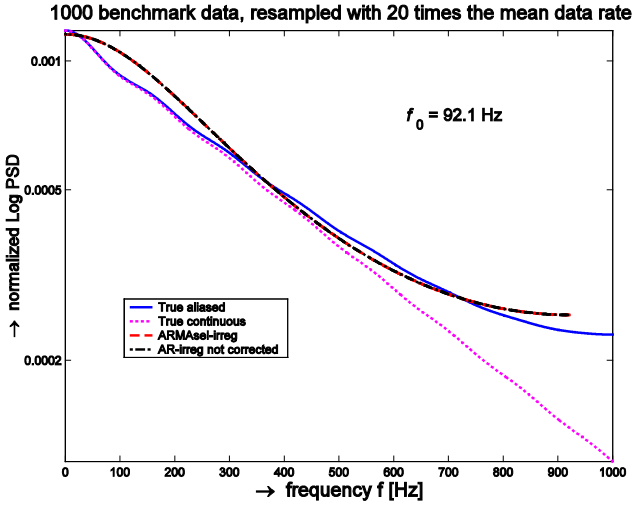

Fig. (2). True benchmark spectrum of type 2 until 1000 Hz, the automatically selected AR(1) estimate for $T_r = T_0/20$ and $w = T_r$, with $PE_s(1)$ = 1.0007 and the correction that is identical with AR(1).

From 1000 irregular observations, AR(1) is selected in the majority of the simulation runs, with $PE_s(1)$ values between 1.0007 and 1.03. In about 10 % an MA(1) model was selected, with $PE_s(1)$ about 1.05. Sometimes AR(0) was selected, although the accuracy of the AR(1) estimate was better. In those cases, the likelihood had a peculiar behavior due to the deviation from the Poisson scheme, the drop outs and other irregularities. However, taking N = 900 or N = 1100 observations from the same simulation runs would generally select the AR(1) model with (3). This demonstrates the sensitivity of the likelihood function to the precise irregular observation times in those examples. If N is less than 500, AR(0) is more often selected and for still smaller sample sizes, AR(0) is almost always selected for this benchmark example. N > 1000 will generally select the AR(1) model. The accuracy of the AR(1) model is very good and the shape of the AR(1) spectrum is similar to the shape of the aliased spectrum. Therefore, higher order models are no close competitors in order selection in this example.

The missing fraction is 95 % for $w = T_r$, with resampling at 20 times the mean data rate. Taking $w = T_r/2$ or smaller for the same 1000 data would give the white noise AR(0) model as the result of order selection. A smaller w gives a larger missing fraction and that is not necessary in this example because the shifting bias is very small here. Shift bias is only important in spectra with a greater dynamic range, with large differences between strong and weak parts of the spectrum, e.g. greater than a factor 1000. In most simulation runs with N = 1000, the spectra of AR models of orders 1 and 2 were smooth, AR(3) had sometimes a spurious peak at the end of the frequency range and higher order models were full of spurious details. However, the high order models were almost never selected. Fig. (**3**) gives an example of the estimated spectra until the AR order 7. The spectra of orders one and two are very accurate. AR(1) was selected. Spectra of AR orders three and higher are irregular, with peaks in the second half of the frequency range. The fit of estimated models of orders five and higher is worse than the fit of the white noise AR(0) model.

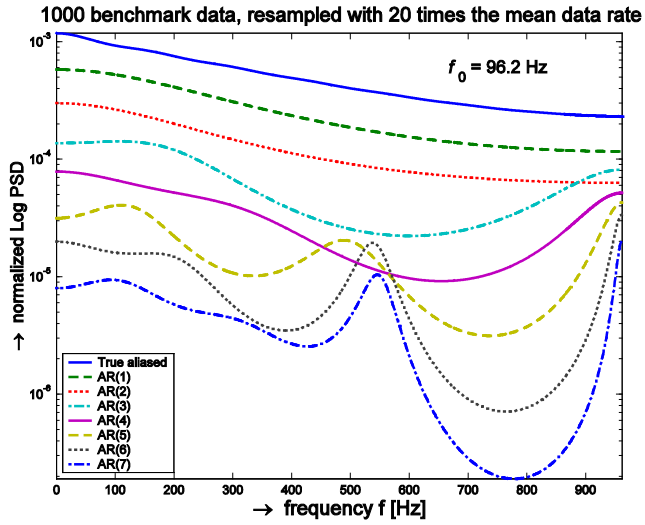

Fig. (3). True benchmark spectrum of type 2 until 962 Hz and the estimated AR spectra for the orders 1 to 7, $T_r = T_0/20$ and $w = T_r$, The spectra are shifted with a factor 2 to enhance visibility of details. AR(1) was selected. The PE_s was 1.0015, 1.0020, 1.1046, 1.1115, 1.1559, 1.3920 and 1.7510 for orders 1 to 7, respectively. $PE_s(0)$ = 1.144.

The behavior in Fig. (**3**) of estimated models as a function of the model order is also more or less representative for other examples. The first parameters are quite accurate if strong details are present in the true spectrum, like in Fig. (**1**). The spectra of Fig. (**2**) require more observations to give accurate parameter estimates. If all significant details are

included, the parameter estimates for higher orders are quite unpredictable. This occurs for orders greater than 3 in Fig. (**1**) and for orders greater than 2 in Fig. (**2**).

Only for a few simulation runs, MA or ARMA models have been selected in the two examples. However, if data are generated with true MA or ARMA processes, those model types are better than AR and they are often selected. The MA and ARMA models are estimated from the parameters of AR models as an intermediate model in ARMAsel-irreg [8], [18]. They are computed very fast in comparison with the AR parameters that are found by computing the likelihood for the given data. Therefore, ARMAsel-irreg will automatically compute as many MA and ARMA models as can be done with the estimated AR parameters [18].

The two examples show that ARMAsel-irreg *automatically* selects the correct spectral shape for very small data sets at high resampling rates. The estimated spectra always look like the biased true aliased spectra. The influence of the aliasing bias is generally much greater than that of the estimation variance. The variance diminishes if more observations are available, but the bias remains the same. Therefore, the quality of estimated spectra of ARMAsel-irreg does not improve much if at least a minimum number of observations is available.

A known lower limitation for the sample size N for a certain resampling rate is that at least about 5 or 10 pairs of observations with distance T_r or smaller should be available. Otherwise, N is too small, or the resampling time has been chosen too small or the processor delay is too large. For the resampling distance β times the average distance T_0 between irregular observations, the effective number of observations is about βN if the observation times have a Poisson distribution. The effective number of observations is a measure for the number of pairs at distances T_r, $2T_r$, … and so on.

A simulation sample of $N = 5000$ observations have been generated from the benchmark true spectrum of Fig. (**2**), with the mean data rate $f_0 = 10$ Hz. The irregular sampling had a Poisson distribution. Those data have been resampled with $200f_0$ to have the same frequency range until 1000 Hz as in Fig. (**2**). The missing fraction was now 99.5% and the effective number of observations at distance T_r was 28. ARMAsel-irreg selected the AR(1) model with $PE_s(1) = 1.0078$.

A still less densely sampled example with $N = 50000$ observations with mean data rate $f_0 = 1$ Hz, resampled with $2000f_0$ have the missing fraction 99.95% and the effective number of observations at distance T_r was 25. ARMAsel-irreg selected the AR(1) model with $PE_s(1) = 1.012$ for those sparse data. In this example, AR(1) is almost always selected as long as the effective number of observations is greater than 50. Even for smaller effective numbers, AR(1) has often been selected. For a resampling rate Lf_0, the number of Poisson distributed observations should be greater than $50L$ for a guaranteed successful evaluation with ARMAsel-irreg for this example. The benchmark with the peak in section IV requires still less data.

6. CONCLUSIONS

Irregular data are transformed into an equidistant missing data problem by multi-shift slotted nearest neighbor resampling. The ARMAsel-irreg estimator fits AR, MA and ARMA models and automatically selects the best model order and model type. Especially in simulations with few irregular data, the results of ARMAsel-irreg are better than what can be obtained with other known spectral estimation techniques. In addition, good results are obtained for data with strong deviations from the Poisson distribution for the sampling instants. ARMAsel-irreg has only a problem if very high order AR models, with orders higher than about 20, are required to give a satisfactory spectral model.

Spectra can be estimated until frequencies higher than 100 times the mean data rate, as long as they can be represented well by a low order AR model in the discrete-time frequency range. The highest possible frequency $1/(2T_r)$ in the discrete-time spectrum is determined by the smallest T_r with at least 5 or 10 irregular data pairs at about that distance.

Elimination of spurious poles from impossible spectral estimates can be included in an automatic algorithm. By eliminating all poles in the second half of the resampled frequency range, order selection will select one of the best AR models from the computed candidates.

REFERENCES

[1] J. Mateo and P. Laguna, "Improved heart rate variability signal analysis from the beat occurrence times according to the IPFM model", *IEEE Trans. Biomed. Eng*., vol. 47, no. 8, pp. 985-996, Aug. 2000.

[2] C. Thiebaut and S. Roques, "Time-scale and time-frequency analyses of irregularly sampled astronomical time series", *Eurasip J. Appl. Sig. Proc*., vol. 15, pp. 2486-2499, Aug. 2005.

[3] S.M. Robeson, "Influence of sampling and interpolation on estimates of air temperature change", *Clim. Res.*, vol. 4, pp. 119-126, Aug. 1994.

[4] L.H. Benedict, H. Nobach and C. Tropea, "Estimation of turbulent velocity spectra from Laser Doppler data", *Meas. Sci. Technol.*, vol. 11, pp. 1089-1104, 2000.

[5] N.R. Lomb, "Least squares frequency analysis unequally spaced data", *Astrophys. Space Sci.*, no. 39, pp. 447-462, 1976.

[6] J. D. Scargle, "Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly spaced data", *Astrophys. J.*, no. 263, pp. 835-853, 1982.

[7] R.H. Jones, "Fitting a continuous time autoregression to discrete data", *Appl. Time Series Analysis II*, Ed. D.F. Findley, 1981. pp. 651-682.

[8] P.M.T. Broersen and R. Bos, "Estimating time-series models from irregularly spaced data", *IEEE Trans. Instrum. Meas.*, vol. 55, no. 4, pp. 1124-1131, Aug. 2006.

[9] M.J. Tummers and D.M. Passchier, "Spectral estimation using a variable window and the slotting technique with local normalization", *Meas. Sci. Technol*., vol. 7, pp. 1541-1546, 1996.

[10] E. Müller, H. Nobach and C. Tropea, "LDA signal reconstruction: application to moment and spectral estimation", Proceedings 7th International Symposium on Applications of Laser technology to Fluid Mechanics, Lisbon, paper 23.2, pp. 1-8, 1994.

[11] R.J. Adrian and C. S. Yao, "Power spectra of fluid velocities measured by Laser Doppler velocimetry", *Exper. Fluid*, vol. 5, pp. 17-28, 1987.

[12] L. Simon and J. Fitzpatrick, "An improved sample-and-hold reconstruction procedure for estimation of power spectra from LDA data", *Exper. Fluids*, vol. 37, pp. 272-280, 2004.

[13] P.M.T. Broersen, S. de Waele and R. Bos, "Autoregressive spectral analysis when observations are missing", *Automatica*, vol. 40, pp. 1495-1504, 2004.

[14] R.H. Jones, "Maximum likelihood fitting of ARMA models to time series with missing observations", *Technometrics*, vol. 22, pp. 389-395, 1980.

[15] P.M.T. Broersen, ARMASA, and ARMAsel for Irregular or Missing Data: Matlab Toolboxes [Online]. Available: http://www.

mathworks.com/matlabcentral/fileexchange, then select as search term armasa, 2008.

[16] P.M.T. Broersen, "Spectral estimation from irregularly sampled data for frequencies far above the mean data rate", Proceedings IEEE/IMTC Conference, Warsaw, Poland, paper 7108, pp. 1-6, May 2007.

[17] H. Nobach, "LDA benchmark generator III" [Online]: Available: http://www. nambis.de/benchmark, 2001.

[18] P.M.T. Broersen, Ed., *Automatic Autocorrelation and Spectral Analysis*. Springer: London, 2006.

[19] M.B. Priestley, Ed., *Spectral Analysis and Time Series*. Academic Press London: U.K, 1981.

[20] W.K. Harteveld, R.F. Mudde and H.E.A. van den Akker, "Estimation of turbulence power spectra for bubbly flows from laser doppler anemometry signals", *Chem. Eng. Sci.*, vol. 60, pp. 6160-6168, 2005.

[21] P.M.T. Broersen, "Five separate bias contributions in time series models for equidistantly resampled irregular data", *IEEE Trans. Instrum. Meas.*, accepted 2009.

Received: November 11, 2008 Revised: November 24, 2008 Accepted: November 25, 2008