SPECTRAL ANALYSIS OF IRREGULARLY SAMPLED DATA WITH AUTOREGRESSIVE MODELS

Piet M.T. Broersen, Robert Bos and Stijn de Waele

Signals and Systems Group, Department of Applied Physics
Delft University of Technology

Abstract: Irregular sampling of stochastic processes gives the theoretical possibility to estimate spectral densities up to very high frequencies. However, the methods developed tend to be heavily biased at higher frequencies or fail to produce a spectrum that is positive for all frequencies. A new estimator is introduced that applies autoregressive spectral estimation to unevenly spaced data. This estimator approximates the data by equidistant resampling with a special nearest neighbor algorithm, that only accepts data if the nearest irregular data point is within half the slot width of the resampling time grid. The algorithm searches for uninterrupted sequences of resampled data and analyzes those sequences using the Burg algorithm for segmented data. With sufficient data, results can be accurate at frequencies higher than the mean data rate. *Copyright © 2002 IFAC*

Keywords: autoregressive model, covariance, nearest neighbor resampling, slotting, spectrum estimation, time series analysis, turbulence data, uneven sampling,

1. INTRODUCTION

Astronomical data and turbulence data obtained by Laser-Doppler anemometry are often irregularly sampled, due to the nature of the observation system. This has the theoretical advantage that the highest frequency that can be estimated is higher than half the mean data rate, which is the upper limit for equidistant observations. Many estimation techniques for unevenly spaced data have been developed. Most can be considered as variations of slotting, of resampling methods or of the Lomb-Scargle estimator.

Benedict *et al*., (2000) give a recent survey of the developments in the different techniques. **Slotting** methods estimate an equidistant covariance function from the irregularly sampled data. Slotting algorithms have been refined with normalization and fuzzy slotting. *Local* normalization reduces the variance of the estimated covariance function. Fuzzy slotting produces a smoother covariance function by distributing products over multiple time slots. Covariance functions as estimated by present slotting techniques are not positive semi-definite. This results in spectra that can become negative at a large percentage of the frequencies where the power is weak. Using ad hoc windowing schemes such as variable windowing can sometimes reduce this effect. Variable windowing takes a wide spectral window at frequencies with low power and a smaller one for higher power. This can have a good performance in examples. Broersen *et al*., (2000) showed, however, that the effectiveness of such schemes depends strongly on the true characteristics of the data. Therefore, variable windowing requires experience and skill, as well as a good guess about the true spectrum that is sought. It cannot be used generally.

The second class of estimators resamples the data on an equidistant time grid. After **resampling**, the data can be analyzed using the periodogram or time series models for equidistant data. While resampling methods obtain positive spectra, estimates at higher frequencies will be severely biased. Adrian and Yao (1987) described Sample and Hold reconstruction as low-pass filtering followed by adding colored noise.

These effects can in theory be eliminated using the refined Sample and Hold estimator of Nobach *et al.*, (1996). This refined estimator tries to subtract the resampling noise from a modified periodogram and to undo the low-pass filtering. However, it does not guarantee a positive spectrum, especially not at frequencies where the resampling noise contribution is much greater than the spectral density of the true process. The refined resampling might give good results at frequencies where noise and process have the same power but not if the noise is much stronger than the process. Nearest Neighbor resampling has similar characteristics. In system identification, missing data are often not interpolated but reconstructed with an estimated data model; see Wallin *et al.*, (2000). Broersen *et al.*, (2000) and de Waele and Broersen (2000a) showed that the use of time series analysis instead of periodograms might produce improved spectral estimates. However, the noise and filtering effect of equidistant resampling set limits to the achievable accuracy of resampling methods. This precludes the accurate estimation of spectra at higher frequencies where the inevitable resampling noise exceeds all small details and hides spectral slopes.

A third category of estimators uses the method of **Lomb** (1976) and **Scargle** (1982), which directly estimates a spectrum from the irregularly sampled data. The Lomb-Scargle spectrum can be completely alias free and reduces to the periodogram if the data happens to be regularly sampled. It is a least squares fit of sines and cosines to the irregularly sampled data. The Lomb-Scargle method is able to accurately detect peaks up to very high frequencies in a very low noise environment, but fails to accurately describe any slopes and details in the spectrum.

Bos *et al.* (2001) introduced a new idea with time series analysis. Their estimator can be perceived as searching for sequences of data that are almost equidistant. The selected sequences of different lengths can be analyzed with an irregular version of the Burg (1967) algorithm for segments of de Waele and Broersen (2000b). This present paper introduces an improved estimator, combining the previous search for uninterrupted quasi-equidistant sequences with a refined slotted Nearest Neighbor resampling. On an equidistant grid, only observations are accepted if the original irregular sampling instant is not further than half a slot width from the resampling grid point. This yields several uninterrupted segments, which can be used for estimation of an autoregressive spectral model. The choices of the slot width as well as the automatic selection of the best order for the autoregressive spectral model are discussed.

2. SLOTTED NN RESAMPLING

The analysis of resampling methods shows that an important problem is the multiple use of a single irregular observation for more resampled data points.

This immediately creates a bias term in the estimated covariance function, because the covariance $R(0)$ leaks to estimated non-zero covariance lags. The analysis of Adrian and Yao (1987) shows that both the covariance and the spectrum suffer from bias in Sample and Hold resampling. Broersen *et al.*, (2000) evaluate Nearest Neighbor (NN) resampling. The total covariance can be calculated theoretically as the sum of two conditional expectations of lagged products in the covariance estimates, where contributions to the 'lagged' product use either the same irregular observation or two different irregular observations. Bias is caused by the shift of irregular time intervals to a fixed grid and by the multiple use of the same irregular observation. This second bias cause will be eliminated in *slotted* NN resampling.

The irregular signal x is measured at N irregular time instants $t_1...t_N$. The average distance between samples T_0 is given by $T_0 = (t_N-t_1)/(N-1) = 1/f_0$, with f_0 the mean data rate. The signal is resampled on a grid at N/k equidistant time instants at grid distance $T_g=kT_0$ (for simplicity in notation, the slot index k or 1/k is limited to integer numbers). The resampled signal exists only for $t=nT_g$ with n integer. The spectrum can be calculated up to frequency $f_0/2k$. The usual Nearest Neighbor resampling substitutes at all grid points nT_g the closest irregular observation $x(t_i)$, with

$$|t_{i-1}-nT_g| > |t_i-nT_g| \quad ; \quad |t_{i+1}-nT_g| > |t_i-nT_g|. \qquad (1)$$

Slotted Nearest Neighbor resampling accepts only a resampled observation at $t=nT_g$ if there is an irregular observation $x(t_i)$ with t_i within the time slot

$$(n-0.5)T_g < t_i \le (n+0.5)T_g . \qquad (2)$$

If there is more than one irregular observation within a slot, the closest one is selected for resampling; if there is no observation within the slot, the resampled signal is left empty. For small width of the slot, the number N_0 of non-empty resampled points nT_g becomes close to N because almost every irregular time point falls into another time slot. For larger slot width, more irregular observations may fall within one slot and only the one closest to the grid point survives in the slotted NN resampled signal.

3. AUTOREGRESSIVE MODELS

Autoregressive (AR) models are a class of models for all stationary stochastic processes; see Priestley (1981). The power spectrum and the covariance of the data can be expressed into the parameters of the AR model. The best model *order* (the number of parameters in the AR model) can be selected objectively and automatically, based on reliable statistical criteria; see Broersen (2000). This is an important advantage above periodogram estimates where no objective or statistical rule exists to select the best among different spectral estimates. An AR(p) process can be written as (Priestley 1981):

$$x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} = \varepsilon_n, \qquad (3)$$

where ε_n is a purely random process, a sequence of independent identically distributed stochastic variables with zero mean and variance σ_ε^2. Almost any *stationary* stochastic process can be written as an unique AR(∞) process, independent of the origin of the process. Hence, this model type can be applied to turbulence data and to similar physical phenomena. Broersen (2000) argues that AR parameters for different model orders of a stochastic process are preferably estimated with the method of Burg (1967). Selection criteria add a penalty for each estimated parameter to the logarithm of the residual variance of each estimated model and look for the minimum of the criterion; see Broersen (2000). Moreover, finite order models are quite well in describing true processes with infinite orders, because the true parameters are decreasing rapidly for most processes. The covariance function and the power spectral density of the data can be computed from the estimated model parameters. The power spectrum of the AR(p) model is given by (Priestley, 1981):

$$h(\omega) = \sigma_\varepsilon^2 / \left| A(e^{j\omega}) \right|^2. \qquad (4)$$

Here, $A(e^{j\omega})$ is defined as $1 + a_1 e^{-j\omega} + \ldots + a_p e^{-jp\omega}$ with the parameters a_i as defined in (3) .

3.1 Burg for segments.

The original Burg (1967) algorithm for AR parameter estimation recursively estimates reflection coefficients k_p from equidistant data. The Burg algorithm for segments of de Waele and Broersen (2000b) is a modified version that allows the simultaneous AR estimation from S multiple segments, even if those segments are not of equal length. In the first step, the functions $f_0^s(n)$ and $b_o^s(n)$ are made equal to the S segmented data sequences $x^s(n)$, $s=1,\ldots,S$. The first reflection coefficient k_1 can now be estimated using the following equation, with $p=1$:

$$k_p = \frac{-2\sum f_{p-1}^s(n) b_{p-1}^s(n-p)}{\sum f_{p-1}^s(n)^2 + \sum b_{p-1}^s(n-p)^2}. \qquad (5)$$

New functions $f_p^s(n)$ and $b_p^s(n)$, called forward and backward residuals can be computed for each stage p.

$$f_p^s(n) = f_{p-1}^s(n) + k_p b_{p-1}^s(n-p)$$
$$b_p^s(n) = b_{p-1}^s(n) + k_p f_{p-1}^s(n+p). \qquad (6)$$

This means that only segments that contribute at stage p-1 appear in the residuals (6). These are used to estimate a new reflection coefficient k_{p+1} with (5). Only segments longer than p+1 points contribute in this stage. The total number of contributing residuals should at least be 25 for sufficient statistical

reliability. After the final parameter k_L has been computed, AR models of order 1 to L can be calculated by applying the recursive Levinson-Durbin formulas; see Stoica and Moses (1997). This modified Burg algorithm will be applied to uninterrupted segments from the slotted NN resampled data.

Order selection. After estimation of a number of AR(i) models, i=1..L, the problem is to select the best model order. Broersen and Wensink (1996) described a finite sample criterion FIC for equidistant data:

$$\text{FIC}(p,3) = \ln\left(\text{RES}(p)\right) + 3\sum_{i=0}^{p} v_p. \qquad (7)$$

RES(p) is the residual variance of the AR(p) model:

$$\text{RES}(p) = \sigma_x^2 \prod_{i=1}^{p} (1 - k_i^2), \qquad (8)$$

Further, v_p denotes the empirical formula for the variance of k_p. The empirical variance of k_p estimated with Burg's method from N observations can in equidistant time series analysis be approximated by:

$$v_p = 1/(N_p + 1). \qquad (9)$$

Here, N_p denotes the number of products N-p that is available to estimate the parameter k_p. When estimating parameters in irregularly sampled data, the variance v_p is still described accurately with the relation above. In irregularly sampled data, however, the actual number of available products N_p is much smaller. It tends to decrease exponentially with increasing orders p. The actual number N_p should be counted and used in the order selection criterion (7).

4. SLOTTED BURG IRREGULAR

It is possible to apply the Burg algorithm for segments directly to the slotted NN resampled signal. Segments are the uninterrupted sequences of resampled data. Each segment of length L can contribute to the computation of the reflection coefficients until k_{L-1} with (5). If the grid constant or slot was chosen as $T_g = kT_0$, the data are analyzed up to the frequency $f_0/2k$, where f_0 is the mean data rate. This can be repeated for different values of k. However, the bias due to shifting the irregular points to a grid increases for larger slot width kT_0. At the same time, the variance decreases, because more products become available. The increased bias causes the undesirable and unnecessary property that the accuracy of the spectral estimates may decrease if only a smaller low frequency area is investigated.

Therefore, a modeling time scale T_m is introduced, with $T_m = mT_0$. The model scale m is restricted in the slotted Burg irregular algorithm to an integer multiple of the grid scale k, giving m=k, m=2k, m=3k,…Mk, with as highest spectral frequency $f_0/2k$,

$f_0/4k, f_0/6k, \ldots f_0/2Mk$ respectively. In the current practice, k is often given the values 1/32, 1/16, 1/8, ¼, ½ and 1 and m is limited to k, 2k, 4k, 8k, 16k and 32k. The maximum value for m equals 1, where the frequency range goes up to $f_0/2$. That is the upper limit for equidistant data. A smaller slot width k than 1/32 would give an under-modeling bias and a high variance because too few contributions N_p remain. Slots greater than 1 give too much bias due to shifts of irregular time instants to the fixed grid. The upper limit m=1 to the model time step yields the spectrum up to $f_0/2$. This limit is not necessary but it reflects the practical fact that the spectrum until $0.15f_0$ or $0.25f_0$ can be found more reliably by ordinary Sample and Hold or Nearest Neighbor resampling, followed by equidistant time series analysis; see the theoretical analysis of Adrian and Yao (1987), the survey of Benedict et al. (2000) and the simulation results of de Waele and Broersen (1999).

For m=k, uninterrupted segments can be determined directly from the NN resampled signal at nT_g. For m=ik, i=2, 3, …, there are i different, slightly shifted, starting points for the i equidistant subsequences: T_g+niT_g, $2T_g+niT_g$, $3T_g+niT_g$, …, iT_g+niT_g, respectively. For integer values of n, this yields all possible equidistant segments at the model scale $mT_0 = ikT_0 = iT_g$. In this way, all slotted NN resampled observations are used in the modeling at time scale T_m.

The number of products that is available in the segments depends very much on the model order and on the slot width $T_g=kT_0$ and is largely independent of the modeling time scale T_m. Fig.1 shows that the lines are almost straight in this log-linear representation, indicating an exponential relation. For Poisson distributed distances they are approximated with

$$N_p = N_0(1-e^{-k})^p \approx N_0 k^p. \qquad (10)$$

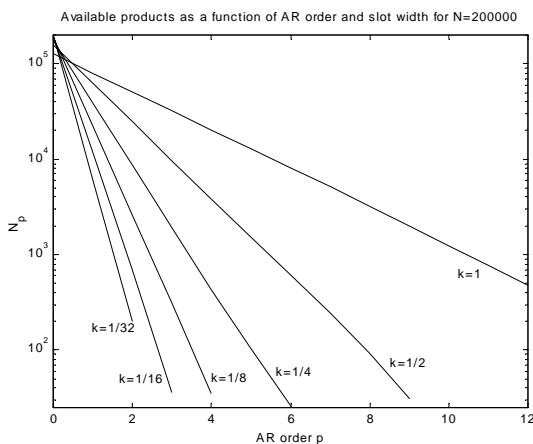If N_p would become less than 25, the AR parameter

Fig.1. Number of products N_p that is available for estimation of the reflection coefficient of order p with the slotted Burg irregular algorithm for different slot widths.

of order p will no longer be computed to maintain statistical reliability. It is clear that a small value of k, giving a small bias due to shifting the time instants to a regular grid, gives only a small number of available products and a low maximum AR order. At the same time, a small grid value k gives extra information because it enables the evaluation on many modeling time steps, whereas k=1 can only give a model at m=1 within the limits set for slotted Burg irregular.

5. PERFORMANCE OF THE NEW ESTIMATOR

The performance of the slotted Burg irregular algorithm has been tested in simulations. It has also been compared to the similar algorithm of Bos et al. (2001). That algorithm, without slotted resampling, has more contributions N_p for the first few values of p and less for greater values of p. It will generally be slightly more accurate if the best AR order for p is 1 or 2, because more products are available for estimation then. However, the problem in practice is mostly that not enough products are available for higher AR orders p; in those cases the slotted Burg irregular algorithm is to be preferred. Simulations with a known (aliased) spectrum are a first step in testing new algorithms.

The simulated data had a background spectrum consisting of two declining slopes. The first slope descends at a rate of $\sim f^{-5/3}$ from $0.01f_0$, and the second at a rate of $\sim f^{-7}$ for frequencies f above $0.1f_0$. This type of spectrum is representative of turbulence data. To test the ability of the new estimator to detect spectral details peaks have been added to the background spectrum. Test data was generated using the following procedure. First 128N equidistant data points were generated using a high order AR process. Then, randomly 127N data points were discarded. Each data point had a probability of 127/128 to be discarded. The resulting irregular data was non-equidistant and time intervals between arrivals were roughly Poisson distributed. In simulations the true properties of the data are known; hence, the quality of estimated results can be established. A quality measure for the fit is the aliased model error, ME_T that has been defined by de Waele and Broersen (2000a) as:

$$ME_T(p) = N\left[\frac{PE_T(p)}{\sigma_{\varepsilon,T}^2} - 1\right] \qquad (11)$$

It uses the true correlation at time scale $T_m = mT_0$ to compute how well the process x(t) can be predicted by using $x(t-T_m)$, $x(t-2T_m)$, ⋯.

Fig.2 gives a comparison of the ordinary Nearest Neighbor resampling, followed by equidistant time series analysis with the ARMAsel program of Broersen (2001) and the new slotted Burg irregular algorithm. The 200000 irregular observations have

Comparison of Nearest Neigbor resampling and slotted Burg irregular
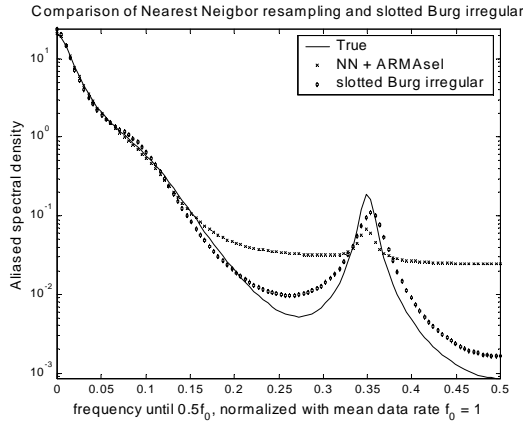
Fig.2. Comparison of Nearest Neighbor resampling with ARMAsel time series analysis and the slotted Burg irregular where only uninterrupted segments are used for AR estimation. Omission of many data improves the accuracy.

been generated with the turbulence background and an additional peak at $0.35f_0$, which has 1% of the total background power. The theoretical Nearest Neighbor resampled spectrum is closely followed by theselected ARMA(8,7) model over the whole frequen-cy range. The filtering and the added noise effects of resampling are not noticeable for frequencies below $0.15f_0$, where the ARMAsel (see Broersen, 2001) spectrum is a close approximation of the true (aliased) spectrum. Above $0.15f_0$, the resampled spectrum is dominated by distortion and the slotted Burg irregular is a much better approximation there. It shows that taking only uninterrupted segments discards a lot of data, but yields improved accuracy. Refined reconstruction of Nobach *et al.* (1996), using Sample and Hold for resampling and the time series spectral model for undoing the filtering and for subtracting the resampling noise has also been investigated. However, none of the variants with a guaranteed positive spectral density at all frequencies gave the same accuracy as slotted Burg irregular.

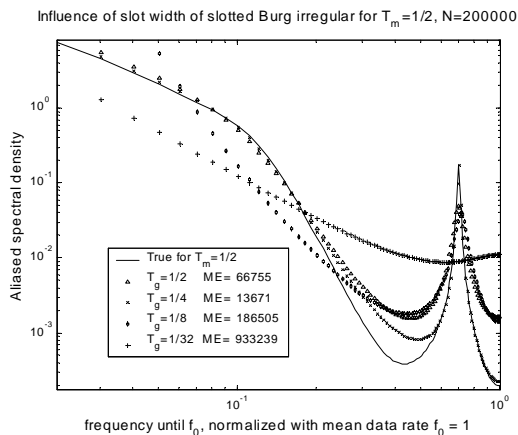Influence of slot width of slotted Burg irregular for $T_m=1/2$, N=200000

Fig.3. Influence of slot width on the accuracy of the slotted Burg irregular spectrum for a given model scale $T_m=1/2$. Wide slots give too much bias and narrow slots have not enough products available for AR models of sufficiently high order.

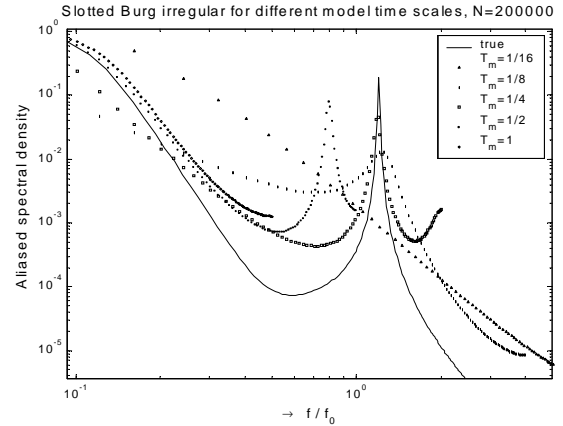Slotted Burg irregular for different model time scales, N=200000

Fig.4. Evaluation of AR models estimated with the slotted Burg irregular algorithm for some model scales. For the large scale $T_m=1$, the peak is drowned in the aliased spectrum. For small model scales, more data are required to estimate enough AR parameters to describe the spectral peak.

Fig.3 gives a simulation result with different slot widths of the slotted Burg irregular algorithm for the turbulence background with an additional peak at $0.7f_0$ with as power 1% of the total power. For every slot width, the AR order is selected with FIC(p,3) of (7). Also the FIC values for different slot values have been compared. They are –3.82, -4.48, -3.85 and –3.18 for the grids ½, ¼, 1/8 and 1/32, respectively. A comparison with the ME values in Fig.3 reveals that the lowest value of FIC is found for the slot width ¼ with the smallest ME value. Comparing FIC values for different slot widths or grids for a fixed model time scale gives surprisingly a very good choice. FIC has not been developed for this selection but it gives satisfactory results in most situations. For every time scale, the best order and the best slot width can be selected automatically with FIC(p,3). It turns out that the slot width, mostly selected in this way, is the smallest that gives enough contributions N_p for the AR order p that is minimally required to describe the significant details. Smaller slots would give bias due to under-modeling and wider slots increase bias without sufficient reduction of variance.

It remains to find a good modeling time scale. No automatic solution to this problem has been found. No general rule can be given for all circumstances. Sometimes, it is clear that the frequency of a peak shifts for different model time scales. In those cases, a comparison will show where the true peak frequency belongs without aliasing. In the simulation of Fig.4, a peak at $1.2f_0$ with 1% of the total power had been added to the background spectrum. Two estimated spectra, at scale 1 and at scale 1/16, show no peak. Scale 1/4 and 1/8 show a peak at $1.2f_0$ and scale ½ gives a peak at $0.8f_0$. This is exactly an alias of the peak at the two smaller scales. In practice, some skill will be required to interpret and to combine the information. However, looking at different time scales produces valuable information.
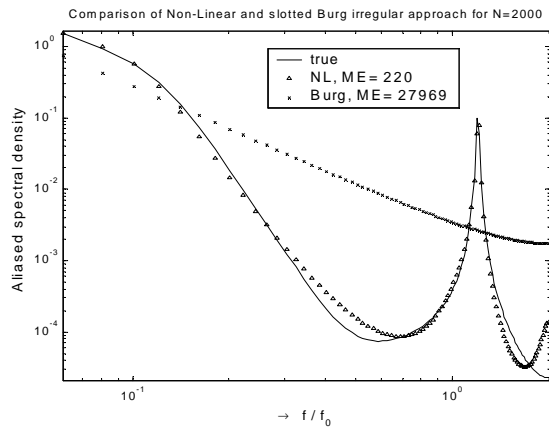
Fig.5. A non-linear approximate maximum likelihood model can, if it converges, give good estimates with much less data than slotted Burg irregular.

It requires many irregular observations to estimate AR models of moderate orders. Most information in the signal cannot be used because only uninterrupted sequences are considered in the slotted Burg irregular algorithm. For this type of algorithm, this requirement is necessary to maintain the positive semi-definite property that makes spectra non-negative at all frequencies. It is tempting to use also the information in the many interrupted sequences. The obvious solution, filling the gaps with Nearest Neighbor interpolation, is shown in Fig.2 to give a poor spectrum. An approximate maximum likelihood algorithm is being developed with non-linear optimization. It uses the same slotted data as the slotted Burg irregular algorithm. With a recursive conditional approximation of the likelihood function, also interrupted data can be used. So far it needs very much computing time and its convergence is not guaranteed. But it has in certain examples given a considerable improvement, in the whole frequency range. Making this algorithm more robust is a future development. However, the algorithm requires the results of AR order, model scale and parameters of the slotted Burg irregular algorithm as initial information and it is not yet computationally reliable.

6. CONCLUSIONS

A new estimator is introduced that fits an AR model to slotted resampled segments of irregularly sampled data. The new slotted Burg irregular algorithm combines a spectrum that is guaranteed to be positive with accurate results at higher frequencies. In simulations, the results are better than those that can be obtained with known existing techniques.

Order selection for the slotted Burg irregular algorithm requires a new selection criterion for the AR model order. It is based on the decreasing number of products that is available for each order. The same criterion can be used to determine the best slot width for a given model time scale. Finally, the best model time scale is still determined by visual inspection of spectra for several time scales.

REFERENCES

Adrian R. J. and C. S. Yao (1987). Power spectra of fluid velocities measured by laser Doppler velocimetry. *Exper. in Fluids*, **vol. 5**, pp. 17-28.

Benedict, L. H., H. Nobach and C. Tropea (2000). Estimation of turbulent velocity spectra from Laser Doppler data. *Measurement Sci. Technol.*, **vol. 11**, pp. 1089-1104.

Bos, R., S. de Waele and P. M. T. Broersen (2001). AR spectral estimation by application of the Burg algorithm to irregularly sampled data. *Proc. IEEE/IMTC Conference*, Budapest, Hungary, pp. 1208-1213.

Broersen, P. M. T. (2001). *ARMASA Toolbox*, from http://www.tn.tudelft.nl/mmr/downloads.

Broersen, P. M. T. (2000). Finite sample criteria for autoregressive order selection. *IEEE Trans. on Signal Process.*, **vol. 48**, no. 12, pp. 3550-3556.

Broersen, P. M. T., S. de Waele and R. Bos (2000). The accuracy of time series analysis for Laser Doppler velocimetry. *Proc. 10th Int. Symp. On Laser Techn. to Fluid Mechanics*, Lisbon, 12 pp.

Broersen, P. M. T. and H. E. Wensink (1996). On the penalty factor for autoregressive order selection in finite samples. *IEEE Trans. on Signal Processing*, **vol. 44**, no. 3, pp. 748-752.

Burg, J. P., (1967). Maximum entropy spectral analysis. *Proc 37th Meeting Soc. Of Exploration Geophysicists,* 6 pp.

Lomb, N. R. (1976). Least squares frequency analysis unequally spaced data. *Astrophysics and Space Science*, no. 39, pp. 447-462.

Nobach, H., E. Müller and C. Tropea (1996). Refined reconstruction techniques for LDA analysis", *Proc. 7th Int. Symp. on Appl. of Laser Techn. To Fluid Mechanics*, Lisbon, 8 pp.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London, U.K.: Academic Press.

Scargle, J. D. (1982). Studies in astronomical time series analysis II. statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysics Journal*, no. 263, pp. 835-853.

Stoica P. and R. L. Moses (1997). *Introduction to Spectral Analysis*. Upper Saddle River, NJ: Prentice Hall.

Waele, S. de and P. M. T. Broersen (2000a). Error measures for Resampled Irregular Data. *IEEE Trans. on Instrumentation and Measurement*, **vol. 49**, no. 2, pp. 216-222.

Waele, S. de and P. M. T. Broersen (2000b). The Burg algorithm for segments. *IEEE Trans. on Signal Process.*, **vol. 48,** no. 10, pp. 2876-2880.

Waele, S. de and P. M. T. Broersen (1999). Reliable LDA-spectra by resampling and ARMA-modeling. *IEEE Trans. on Instrumentation and Measurement*, **vol. 48**, no. 6, pp. 1117-1121.

Wallin, R., A.J. Isaksson and L. Ljung (2000). An iterative method for identification of ARX models from incomplete data. *Proc. CDC/IEEE Conference*, Sydney, Australia, pp. 203-208.