Practical Aspects of the Spectral Analysis of Irregularly Sampled Data With Time-Series Models

Piet M. T. Broersen

***Abstract*—Several algorithms for the spectral analysis of irregularly sampled random processes can estimate the spectral density for a low frequency range. A new time-series method extended that frequency range with a factor of thousand or more. The new algorithm has two requirements to give useful results. First, at least ten closest pairs of neighboring irregular observations should have a distance less than the minimum resampling distance for the chosen discrete-time frequency range. Second, a low-order time-series model should be appropriate to describe the global character of the data. The consequences and importance of this second demand are studied for irregular turbulence observations with narrow spectral details. Models of low orders are estimated from equidistant hot-wire observations and from irregularly sampled laser Doppler anemometer (LDA) data, which are obtained from the same turbulence process. The irregular data are resampled with the nearest neighbor method, both with and without slotting. Apart from the usual bias contributions of resampling irregular data, LDA data can give an additional spectral bias if the instantaneous sampling rate is correlated to the actual magnitude of the turbulent velocity. Making histograms of the amplitudes and the interarrival times provides useful information about irregularly sampled data.**

***Index Terms*—Autoregressive models, irregular sampling, order selection, slotted resampling, uneven sampling, velocity bias.**

I. Introduction

THE PROBLEM of estimating the power spectral density from irregularly sampled data arises in several measurement applications. Industrial, astronomical, meteorological, and medical data may consist of irregularly sampled observations. Spectra are determined from spectrographs with nonlinear wavelength scales in stellar physics, or time series of light curves of variable stars are observed in astronomical data [1]. In medical research, the heart rate variability is used to analyze the cardiovascular system [2]. Climate data are often incomplete and irregular, particularly if network data are considered [3]. Networks that are commonly available to connect measurement systems may give synchronization problems [4]. The observation times can be recorded with great precision, but the primitive synchronization makes the data irregularly sampled. Laser Doppler anemometry (LDA) is an optical nonintrusive flow-measuring technique that is widely used in turbulent flow research. Sampling times are given by the arrival of light-scattering seeding particles in the measurement volume and are irregular in turbulent flow [5].

It is well known that the analysis of irregularly spaced data sets is more complicated than that of equidistant data. A survey of methods shows the present limitations of all methods [5]. Direct Fourier transforms [5] and the method of Lomb–Scargle [6] converge to the usual periodogram if the observations would be equidistant. The methods suffer from a significant bias in the estimated spectra [5]. Equidistant resampling with the sample-and-hold (SH) algorithm can be used for frequencies well below the mean data rate f_0 [7]. However, a white noise source, called step noise, is added as bias in SH, and afterward, the true spectrum plus noise is filtered. This SH method is strongly biased for frequencies greater than $f_0/2\pi$ [7], where the spectral bias due to the filter error is already 50%. The SH filter error bias is only less than 10% for frequencies below $0.05f_0$. Refined methods can correct for this bias if the interarrival sampling distances have a Poisson distribution [5]. However, the useful spectral range remains limited until a maximum of about f_0. Higher frequencies may be important if periodicities are possible at very high frequencies, such as in irregular astronomical data.

The slotting principle has at first been used in the slotted autocorrelation estimation [8]. The product of two irregular observations contributes to a certain slot at lag $k\Delta$ of the autocorrelation function if their time distance falls within the slot between $(k-0.5)\Delta$ and $(k+0.5)\Delta$, where Δ is the slot width. The number of contributing pairs at $k\Delta$ will have variations for different k. Slotted autocorrelation estimates will always require very large data sets [8]. Unfortunately, slotted autocorrelations are not positive definite: they fail to produce spectra that are positive over the whole frequency range [9]. Until now, there is no satisfactory solution for that problem.

Nearest neighbor (NN) resampling without slotting substitutes the nearest irregular observation on each equidistant resampling grid point. It has an influence on the spectral estimation that is similar to SH. Slotted NN will only substitute an observation if that is less than half the slot width away from the grid point. Otherwise, the grid point is left empty. The resampled signal with gaps is a missing-data problem. A recent method is denoted ARMAsel-irreg as an acronym for *A*uto*R*egressive *M*oving *A*verage *sel*ection for *irreg*ular data. It uses multishift slotted NN resampling (MSSNNR) to transform an irregularly sampled signal into a couple of equidistantly resampled signals where data are missing [9]. ARMAsel-irreg is an algorithm that uses time-series models to estimate spectra

Manuscript received June 30, 2008; revised October 6, 2008. First published January 20, 2009; current version published April 7, 2009. The Associate Editor coordinating the review process for this paper was Dr. John Sheppard.

The author is with the Department of Multi-Scale Physics, Delft University of Technology, 2628 Delft, The Netherlands (e-mail: p.m.t.broersen@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIM.2008.2009201

and autocorrelation functions from that resampled irregular signal.

Bias due to slotting has separately been studied [10]. Four different bias contributions have been analyzed for the time-series models of irregular observations that are resampled to an equidistant grid: 1) aliasing; 2) shift of observation times to a grid; 3) truncation bias of incomplete models; and 4) a special missing-data bias. Solutions can be given to effectively reduce all those bias sources [10].

LDA data have an additional bias source, called velocity bias [11]. In a fixed time unit, more than average fluid (or gas) passes through the measurement volume if the fluid velocity is faster than average. This fluid contains more than the average number of scattering particles, which are observed by the LDA device [11]. Therefore, high velocities have a higher sampling rate. Likewise, low velocities have less than the average number of scattering particles per time unit. This creates a correlation between the amplitude of the fluid velocity and the data rate. Some corrections for the mean and variance have been suggested [12], [13].

This paper investigates the behavior of ARMAsel-irreg with MSSNNR [9] applied to practical turbulence data, measured with a constant temperature anemometer (CTA) or with LDA. Experimental data have been collected from the Internet [14], [15]. These data provide equidistant signals that are obtained with hot-wire or CTA measurements and irregular LDA data, for the same physical turbulent processes. It turned out that velocity bias destroyed the spectral accuracy at higher frequencies for those data [16]. A method to detect velocity bias is presented. It will be applied to practical LDA data [17]. Furthermore, a complete procedure is described for the analysis of new irregular data with unknown characteristics.

II. Time-Series Models

A short introduction to the theory of time-series models is presented here. It clarifies how the estimated parameters can describe the spectral density of the data. A discrete-time autoregressive moving average ARMA(p,q) process can be written as [18], [19]

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

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 an equidistant stationary stochastic process. For resampled continuous-time data with the resampling distance T_r, the signal x_n is the observation at time nT_r. Other values for T_r would give ARMA(p',q') processes with different values for order, parameters, and σ_ε^2 in (1). The model (1) represents a parametric estimate of the autocorrelation function or of the power spectral density for measured data. For noisy data, the signal x_n denotes signal plus noise, and the estimated spectrum is always the spectrum of signal plus noise together.

The power spectral density $h(\omega)$ of the model and the frequency range depend on the resampling time T_r. The spectrum is fully determined by the parameters in (1), together with the variance σ_ε^2 and T_r, i.e.,

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

The autocorrelation function at lags that are multiples of T_r can be approximated by an inverse discrete Fourier transform of (2). However, exact formulas relating the autocorrelation to the parameters of (1) are available and more accurate [19]. Accuracy measures have been defined for time-series models of irregular data [9]. They are based on the squared error of one-step-ahead predictions.

The ARMASA toolbox contains a program ARMAsel that automatically selects the best model order and model type for measured data [20]. It computes a number of candidate AR, MA, and ARMA models and uses statistical criteria to select the best time-series model. It can be used only for complete contiguous equidistant observations.

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 nodes. This causes a large extra bias. The properties are similar to SH [7], and spectra are only more or less reliable until $f_0/2\pi$. Hence, NN resampled irregular data analyzed with ARMAsel are particularly suitable to estimate spectra of irregular data in a low frequency range.

The bias can be reduced with the slotting principle, where the slot width is taken equal to the resampling distance. Slotted NN accepts only 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. A dedicated ARMAsel-mis algorithm for missing-data signals has been developed [19]. This can deal with slotted resampling, where the slot width is equal to the resampling distance.

A further decrease 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 [9]. 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 nonempty resampling instants $nT_r + mw$, where an irregular observation falls within the slot width, are determined for the M signals by

$$nT_r + mw - 0.5w < t_i \le nT_r + mw + 0.5w, \quad m = 0, 1, \ldots, M-1 \tag{3}$$

where t_i denotes the time of an irregular observation. Now, all slots of width w are connected in time. M shifted starting

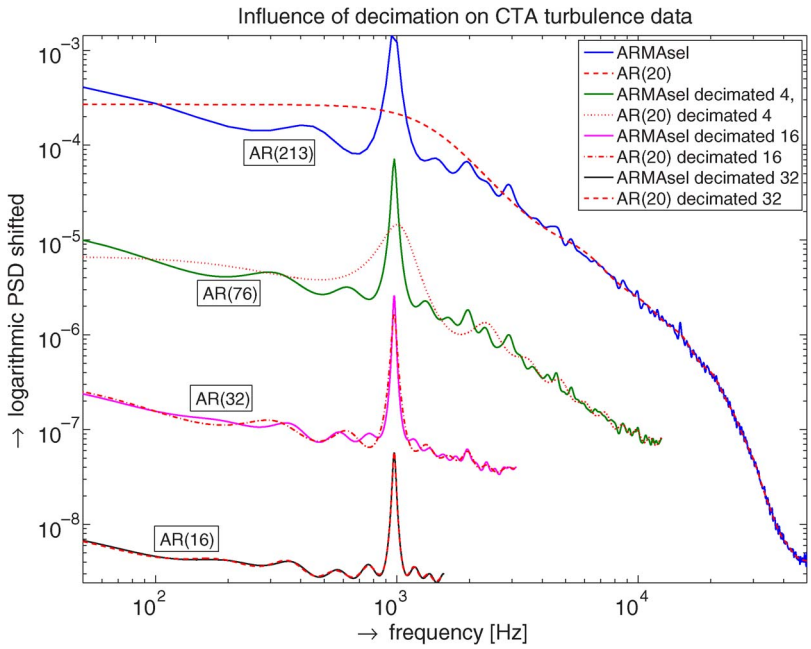

Fig. 1. Spectra of the selected time-series models and the AR(20) models of hot-wire cylinder wake data [14] for four decimated frequency ranges. The spectra are vertically shifted with a factor of 40 to improve the visibility of details. The selected AR orders are given for each frequency range.

points give M equidistant sequences, each with time step T_r. Data are missing in each signal. The likelihood function can individually be calculated for every signal with the ARMAsel-mis algorithm. However, it is possible to treat the likelihoods of the M missing-data signals together for the computation of a single time-series model for the irregular data [9]. The program ARMAsel-irreg will soon be available on the Internet. A general procedure to analyze irregular data will use NN and ARMAsel for low frequencies, and MSSNNR and ARMAsel-irreg for high frequencies.

III. Cylinder Wake Hot-Wire Data

Measurements have been made in the wake of the flow around a cylinder [14]. The Reynolds number was 12 000. The data have been collected in a measurement volume that was about four diameters downstream of the cylinder. At that location, the wake is expected to be fully turbulent. It has an additional oscillating component due to the periodic passage of vortices shed from the cylinder. The frequency of the vortices can be predicted with the Strouhal number. The dimensions of the turbulence experiment have been chosen to produce a peak at 1000 Hz. Hot-wire or CTA observations are continuous in time. They have been sampled with 100 kHz to obtain an equidistant reference signal. The mean data rate of the LDA measurements with irregular sampling was about 72.5 kHz. The two data types cannot simultaneously be obtained, but it has been expected that the spectral properties of the turbulent flow remain the same and are not disturbed by the method of observation.

First, it is investigated how the spectra of estimated fixed low-order AR(20) models of the hot-wire processes will approximate the spectral density for different frequency ranges. This order 20 is about the highest order that can be computed with ARMAsel-irreg. The hot-wire data turn out to have a much higher true AR order. As the true process was continuous in time, the sampling frequency can freely be chosen. No antialiasing filters have been used because those filters cannot be applied to irregular data either. The original data have a sampling frequency of 100 kHz. The data have been decimated to 25, 6.25, and 3.125 kHz with $K = 4$, 16, and 32, respectively. Decimation with a factor of K means taking only observations with $10^{-5}K$ seconds as intervals. The automatic ARMAsel program of the ARMASA toolbox [20] has been used for spectral estimation and order selection. Two spectra have been computed for each sampling frequency: the spectrum of the AR model with the automatically selected AR order and that of the fixed-order AR(20) model.

Fig. 1 gives the results. The total number of hot-wire observations at 100 kHz was 131 072. The selected model orders for the different decimating rates were AR(213), AR(76), AR(32), and AR(16), respectively. It is obvious that larger frequency ranges require higher order models to represent the same details. In the lowest sampling frequency, namely, 3.125 kHz, the AR(20) model resembles the AR(213) model for the original data, sampled with 100 kHz. Roughly doubling the frequency range will often require twice the AR model order to give the same spectral details. This exactly applies in Fig. 1 for 6.25 and 3.125 kHz. However, if the data would have been generated with a low-order AR(1) process, an AR(1) model would be accurate for all frequency ranges.

The AR(20) model of the lowest sampling rate closely follows the true spectrum. The first low-order parameters are used to determine the general spectral shape for the whole frequency range. This smooth global shape is much more important than the spectral peak at 1 kHz if the frequency range goes beyond 10 kHz. The peak is only found in AR models of orders greater than 100. Fixed-order AR(10) models for the same data have

already been presented before [16]. The peaks at the sampling frequencies 25 and 6.25 kHz become higher and narrower with the AR(20) models than with AR(10). It is clear that the AR models of the fixed-order 20 can more easily detect peaks in a smaller frequency range.

The AR(20) model at the highest rate of 100 kHz closely follows the selected AR(213) spectrum for frequencies above about 2 kHz. Hence, the two spectra are close in 96% of the frequency domain and only significantly different in the first 4%. The peak is in that range and is invisible in the AR(20) spectrum. Some aliasing is visible at the end of the spectrum for decimation with 25 kHz. It is significant in the final 50% of the smaller frequency range. However, the peak around 1 kHz becomes clearly noticeable in the AR(20) model. The selected model shows the peak at 1 kHz and smaller peaks at 2 and 3 kHz. For 6.25-kHz sampling, the selected AR(32) time-series spectrum and the AR(20) spectrum stay close for all frequencies. The spectrum of the lowest resampling rate gives a dominant peak in the AR(20) spectrum, which is almost the same as the selected AR(16) model. Both the location and the shape of the peak are similar for the four selected AR models. Only the spectral peak without decimation is somewhat wider. It would require AR(300) or still higher order models to obtain the correct shape of the peak in spectra for the largest frequency range. However, those models give increased ripples for higher frequencies. The overall accuracy of the estimated AR model does not further increase if orders higher than 213 are selected, for the given sample size. Higher selected orders may be expected if many more data would be available. For an accurate shape, higher order models are required. These models will automatically be selected with ARMAsel if sufficient data are available.

Fig. 1 gives some interesting conclusions for the analysis of turbulence data with time-series models. Although they have been derived for equidistantly sampled data, they have implications for irregular data as well if they are resampled with NN. Low-order AR models cannot represent narrow spectral details in a wide frequency band. Peaks at low frequencies require AR models of orders higher than 100 to become visible. The same peak is found for lower order AR models in a smaller frequency band. As it is difficult and time consuming to estimate high AR orders in nonlinear optimization for irregular data, it becomes important to reduce the frequency range. Selected models for equidistant data will include all significant details if sufficient observations are available. Low-order AR models will describe the general shape of the spectrum over the whole frequency range, rather than the shape of spectral details. The model order to describe the shape of a spectral peak may depend on the (re)sampling rate. Accurate narrow shapes are more easily found in a smaller frequency range.

IV. Cylinder Wake LDA Data

Both hot-wire and LDA observations are available for the same process in the cylinder wake data [14]. This gives the opportunity to compare spectra obtained from irregular LDA observations with those from equidistant CTA data. Very high model orders can be selected from equidistant data sets. Irregular data can equidistantly be resampled with NN without slotting. Spectra will then be accurate until $0.05f_0$ and gradually lose their accuracy for higher frequencies. Therefore, it will always be advisable to use NN without slotting for resampling and the automatic ARMAsel time-series algorithm [20] for accurate spectral estimation in the lowest frequency range. Using a low resampling rate of about $0.1f_0$, the estimated spectra until $0.05f_0$ will be reliable and accurate. Spectra are biased by aliasing of the high frequency range, but they are not severely distorted by resampling. All details that are selected in this range are really present in the data, and all significant details will be visible in the spectral representation. Aliasing can move peaks to wrong frequencies if the true peak is above $0.05f_0$. However, this can be detected by evaluating the data in a wider frequency range, e.g., until $0.5f_0$. Although resampling with SH (or NN) suffers from step noise and a distorting filter, indications of spectral details can still be recognized in the selected time-series model for that wider frequency range. We will investigate what happens if the LDA data are resampled with NN with the mean data rate f_0.

First, the results of hot-wire CTA data are compared with NN resampled LDA data without slotting. Spectra are computed and selected with ARMAsel [20]. NN interpolation should be accurate for low frequencies, and it will be biased in a wider frequency range. Fig. 2 shows that the spectra of the three selected AR models of the hot-wire data and the NN resampled LDA data are close for most of their frequency ranges. This is unexpected for the LDA signal resampled with f_0, which should heavily be filtered above $f_0/2\pi$ [7]. It turns out to be a coincidence due to the interaction of step noise, NN filtering, and velocity bias. It has been verified in many examples with other data that the estimated NN spectrum is generally much weaker than the hot-wire spectrum for the high frequency range. However, the main purpose of Fig. 2 was to investigate whether the peak at 1 kHz is still present in NN data obtained with the higher resampling frequency f_0 and that is true. Even after NN resampling with a ten times higher resampling frequency, the peak could still be detected in the selected ARMAsel spectrum.

Fig. 3 compares NN resampling of the LDA data, with and without slotting, with the hot-wire spectrum as a reference. Slotting with the ARMAsel-irreg algorithm can only estimate AR models until order about 20 or 25 and requires much computing time. NN without slotting with the ARMAsel algorithm [20] can estimate AR models of orders higher than 10 000 if sufficient data are available and is fast. It selected order 297 for resampling with 100 kHz. ARMAsel-irreg selected order 8. The spectrum has some ripples at the high frequency end. The spectra of all estimated AR models with orders between 2 and 20 were very close for the 500 000 LDA observations. The ripples at the end of the frequency range become larger for higher order models, but the general level at the range is around $5 \cdot 10^{-7}$ for AR models from order 2 to 20. All those ARMAsel-irreg models are smooth for frequencies below 2 kHz, without any indication of the peak.

It is well known that NN resampling without slotting has an important filter bias plus a noise error for frequencies higher than about $f_0/2\pi$ Hz. The filter reduces the spectrum of SH with a factor of 2 at $f_0/2\pi$ or 11.5 kHz. Due to slotting, the

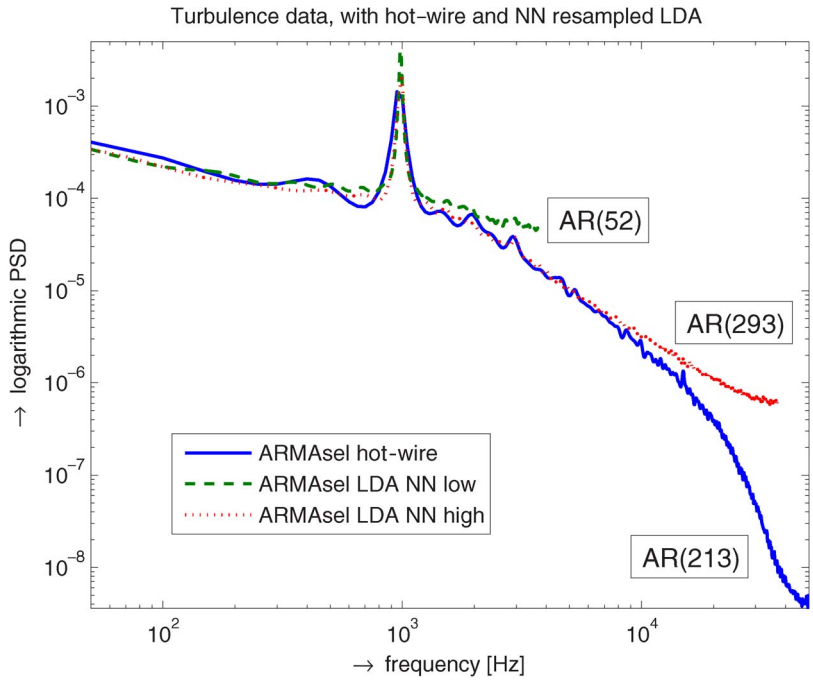

Fig. 2. Spectra of the selected time-series models of hot-wire CTA and irregular LDA cylinder wake data [14], with $f_0 = 72.5$ kHz. Five hundred thousand LDA observations are resampled with a low frequency of $0.1 f_0$ and a high frequency f_0. The selected AR orders are given.

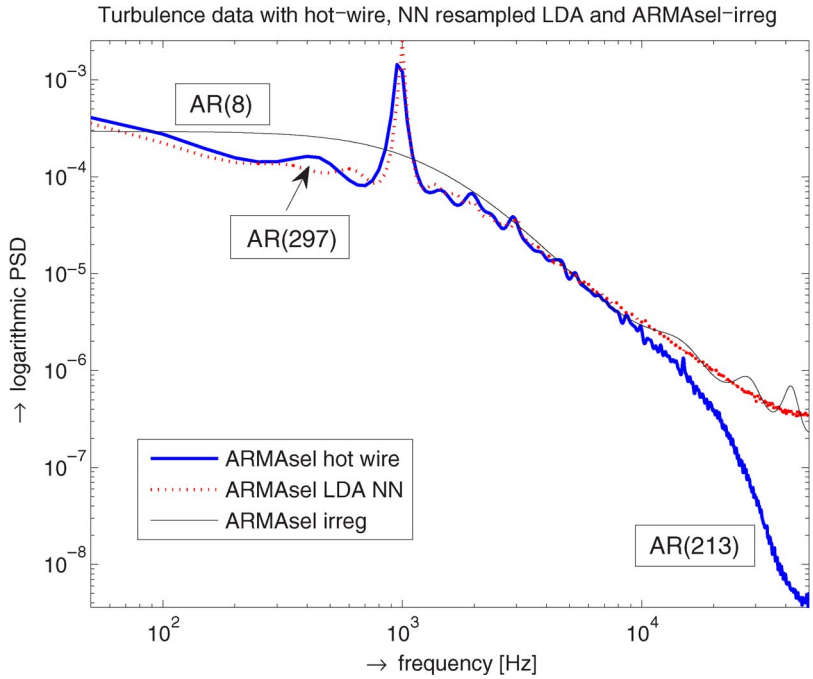

Fig. 3. Spectra of selected models of cylinder wake data obtained from 131 072 hot-wire and 500 000 LDA observations, with $f_0 = 72.5$ kHz. The LDA data are NN resampled with 100 kHz, with and without slotting. The slot width w was one fourth of the resampling distance.

bias becomes much smaller for ARMAsel-irreg because the maximum shift of the irregular time instants to a grid is reduced to half the slot width [9]. ARMAsel-irreg was expected to be close to the hot-wire spectrum. Unfortunately, this is not found in Fig. 3. In contrast, the spectrum of the AR(8) model of ARMAsel-irreg was close to the AR(8) model of NN. This cannot be in agreement with the assumption that the spectra of the hot-wire data and the LDA data are similar in the frequency range above 10 kHz. A possible explanation is that the velocity bias in LDA data causes a severe distortion of the LDA spectra in this example.

Fig. 4 gives the histograms of the hot-wire and LDA data that have been obtained from the same turbulent experiment [14]. It immediately explains that LDA velocities are quite different from hot-wire data, probably due to velocity bias. The mean values are 18.46 and 22.80, and the variances are 22.1 and 34.2.

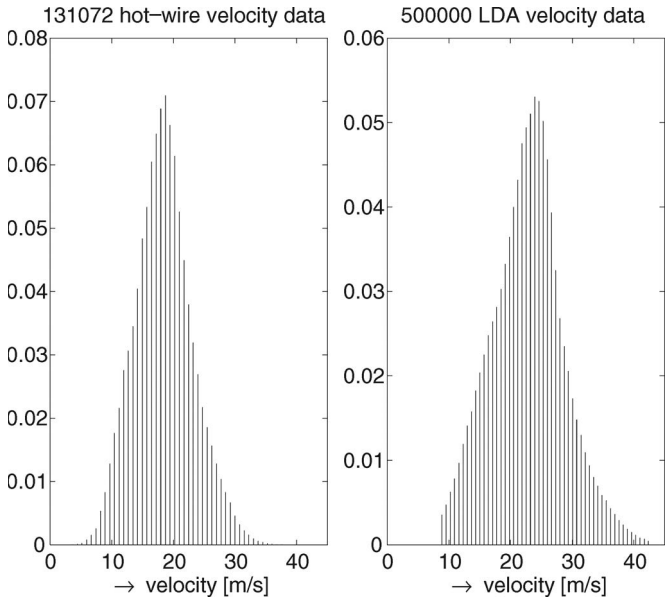

Fig. 4. Histograms of hot-wire and LDA data obtained from the same turbulence experiment. The measurement procedure has a strong influence on the observed velocities.

No further explanation can be given. It is evident that the two types of turbulent observations on the same physical experiment have different properties. Velocity bias can explain why higher velocities have a higher probability to be observed. However, it cannot explain why much higher velocities are found in LDA than in hot-wire experiments. Fig. 4 demonstrates that the analysis of irregular data requires a special treatment with much care for calibration and detection of anomalies. LDA data must be tested for the presence of velocity bias.

A new and independent set of hot-wire and LDA data has been tested before for further verification [16]. Axisymmetric free air jet data are available on the Internet [15]. This example yielded the similar result that LDA spectra did not resemble the hot-wire spectra [16].

ARMAsel-irreg has been applied to synthetic data with a known spectrum. Data with different spectral characteristics can be generated [21]. Many variations in the sampling scheme are possible. The benchmark program [21] has the possibility to add velocity bias in the irregularly generated data, as well as very large gaps or varying data rates. It has been shown in several simulation examples that ARMAsel-irreg could retrieve the known spectral characteristics of synthetic data [9], [16]. ARMAsel-irreg with slotted resampling is not sensitive to very large gaps or varying data rates. However, if velocity bias was added, the estimated spectrum of ARMAsel-irreg can sometimes be distorted. Therefore, ARMAsel-irreg may also give inaccurate spectra if velocity bias is present, such as all other existing methods.

V. Procedure for Analyzing New Data

If only irregular observations are available of some unknown process and if the spectral range of interest is not *a priori* known, it is advisable to use different resampling frequencies in ARMAsel and ARMAsel-irreg to analyze the irregular data. The mean data rate f_0 does not always indicate the important frequency range of the data. It is determined by the characteristics of the measurement equipment, not by the properties of the process data.

As a preliminary investigation, it is advisable to make a histogram of the observed data and the interarrival times. This simple analysis can indicate whether there are outliers in the data, or that velocity bias is probably present, or that there are anomalies in the sampling scheme [10]. This important information is anyhow required. Furthermore it gives an indication of the highest resampling rate that is possible. This is because at least ten pairs of irregular neighboring observations must have a distance less than or equal to the resampling distance. Otherwise, no spectral analysis is possible in that frequency range, and only a smaller range can be used for analysis of the spectral density. It should be realized, however, that in most experiments, the highest resampling rate of interest is much lower than what would be allowed by this limit.

The lowest frequency range can be investigated with NN resampling without slotting and the ARMAsel algorithm of the ARMASA toolbox [20] to determine the spectrum. Slotting has no advantages for very low resampling rates. It is a good idea to use the resampling frequencies $0.1f_0$ and f_0, such as in Fig. 2. This is an easy and fast method to detect spectral details in the frequency range until $f_0/2$ Hz. The spectrum is rather accurate until $f_0/2\pi$ Hz [7], but peaks and other details can still be detected beyond that range. The resampled signal is contiguous without data missing. Therefore, the computation is fast, and all available data can be used here.

MSSNNR [9] and ARMAsel-irreg can be used for higher frequencies. It is advisable to first do a rough search for the interesting frequency range. This can be made with low AR orders, a slot width w that is equal to the resampling distance T_r, and a range of resampling frequencies $f_0, 4f_0, 16f_0, 64f_0, \ldots$. This range will be continued as long as successive ranges produce the same spectral slope at the high frequency end, with the same transition points between the slopes. It is interrupted if the character of the spectra is no longer overlapping. This will require much computing time, and it is faster to use only about 5000 irregular observations in this exploratory stage. ARMAsel-irreg can automatically select the best model order from the candidates, which will be limited to AR(0), AR(1), AR(2), and AR(3) in this stage of the procedure. Of course, this preliminary procedure can entirely be skipped if similar data have been analyzed before. In that case, *a priori* knowledge is available from previous experiments.

Finally, after selection of the frequency range of interest $1/(2T_r)$, the irregular data will be resampled with the slot width $w = T_r/2$ or $w = T_r/4$, or even a smaller slot width if enough data are available. Data with a steep spectral slope at high frequencies may require a small slot width to reduce the bias [10]. For spectra with a slope less than about f^{-4} at high frequencies, $w = T_r/2$ will be small enough. AR orders higher than 3 are used as candidates in this final stage. A smaller slot gives a higher fraction of missing data. This will increase the estimation variance. This increase can be greater than the decrease of the bias, which is the intention of the smaller slot width [10].

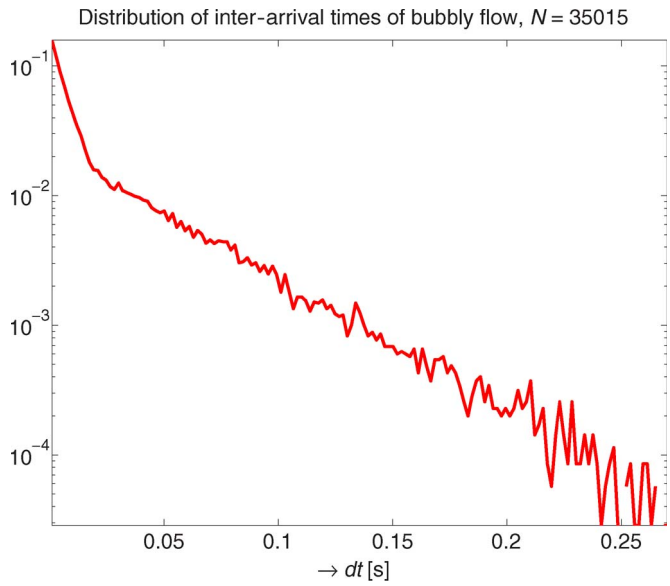

Fig. 5. Histogram of the interarrival times of 35 015 LDA data from the bubbly turbulence experiment. The mean data rate f_0 was 37.6 Hz, with $T_0 = 0.0266$ s.

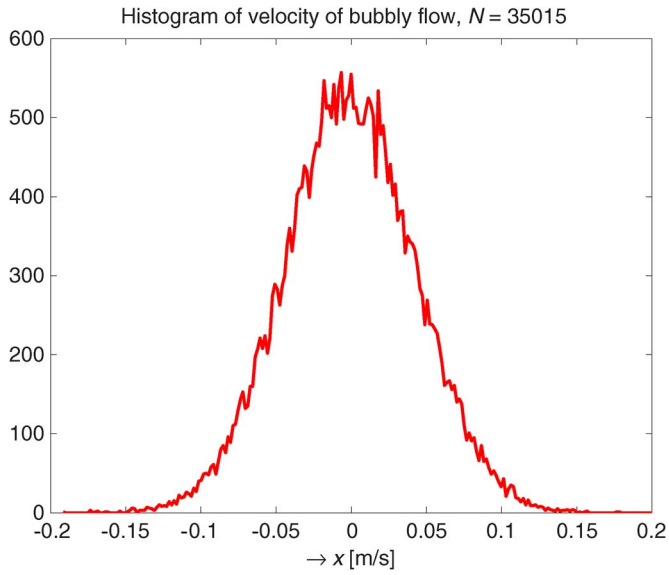

Fig. 6. Histogram of the velocity of LDA data from the bubbly turbulence experiment. The mean velocity of 0.2 m/s has been subtracted.

VI. Bubbly Flow Data

Irregularly sampled data with gaps in the interarrival times are the result of the LDA measurements of turbulent bubbly flow [17]. The gaps disturb the Poisson distribution of the arrival times and preclude the use of refined analysis techniques [5]. These data are used to demonstrate the analysis procedure for new data.

Fig. 5 gives the probability density function of the interarrival times. A Poisson distribution would give a straight line in the logarithmic presentation. The actual histogram gives approximately two straight lines, with different slopes. Very short intervals are present, and the signal can be analyzed until high frequencies.

Fig. 6 gives the histogram of the observed velocities, after subtraction of the mean value. There are no outliers present. The symmetrical global shape of the histogram is a strong indication that velocity bias is not very important for those LDA data. This is generally the case if the mean velocity is much greater than the variations around the mean. Although the mean value for the data is only about four times the standard deviation here, no strong asymmetric distribution of the velocities is found.

The analysis of the spectra for NN resampling with ARMAsel gives a remarkable result in Fig. 7. For resampling with $f_0/10$ Hz, the selected model order was 0. No details are found for very low frequencies. With f_0 (37.6 Hz), the AR(2) model was selected. This follows the filtering that is expected for NN resampling above $f_0/2\pi$ Hz [7]. The different levels are due to aliasing, because both spectra in Fig. 7 have the same integral. The same type of analysis in Fig. 2 showed the presence of a sharp peak at a low frequency.

MSSNNR with slotting has been applied in six frequency ranges, starting with f_0 and each time a factor of 4 higher. The results of the selected models from 5000 irregular observations are given in Fig. 8. The selected AR orders were 1, 2, 2, 3, 3, and

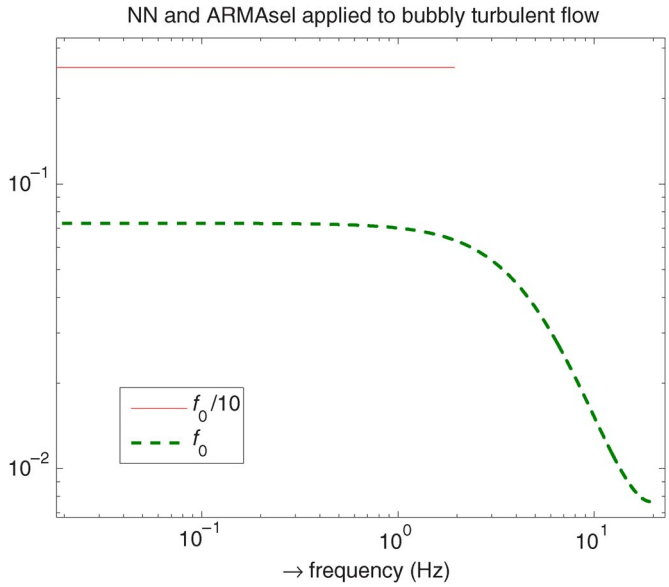

Fig. 7. Spectra obtained after NN resampling without slotting and spectral analysis with ARMAsel.

3, respectively. The spectrum for the resampling frequency f_0 is rather flat over the whole frequency range. It explains why the selected model for NN resampling with $f_0/10$ Hz was AR(0), because the spectrum was flat for frequencies lower than 2 Hz.

The transition from the horizontal slope to a decreasing spectrum is at about the same frequency for all six spectra in Fig. 8. On the other hand, the details at higher frequencies are at different frequencies in the three lower spectra. Hence, those details are not reliable for the LDA measurements of turbulent bubbly flow. For those three resampling frequencies, the AR(3) models were selected. The AR(3) models for the three lowest resampling frequencies were almost equal to the selected AR(1) and AR(2) models. As the lower order models give the same spectra with less parameters, those models are selected with an order selection criterion [19].

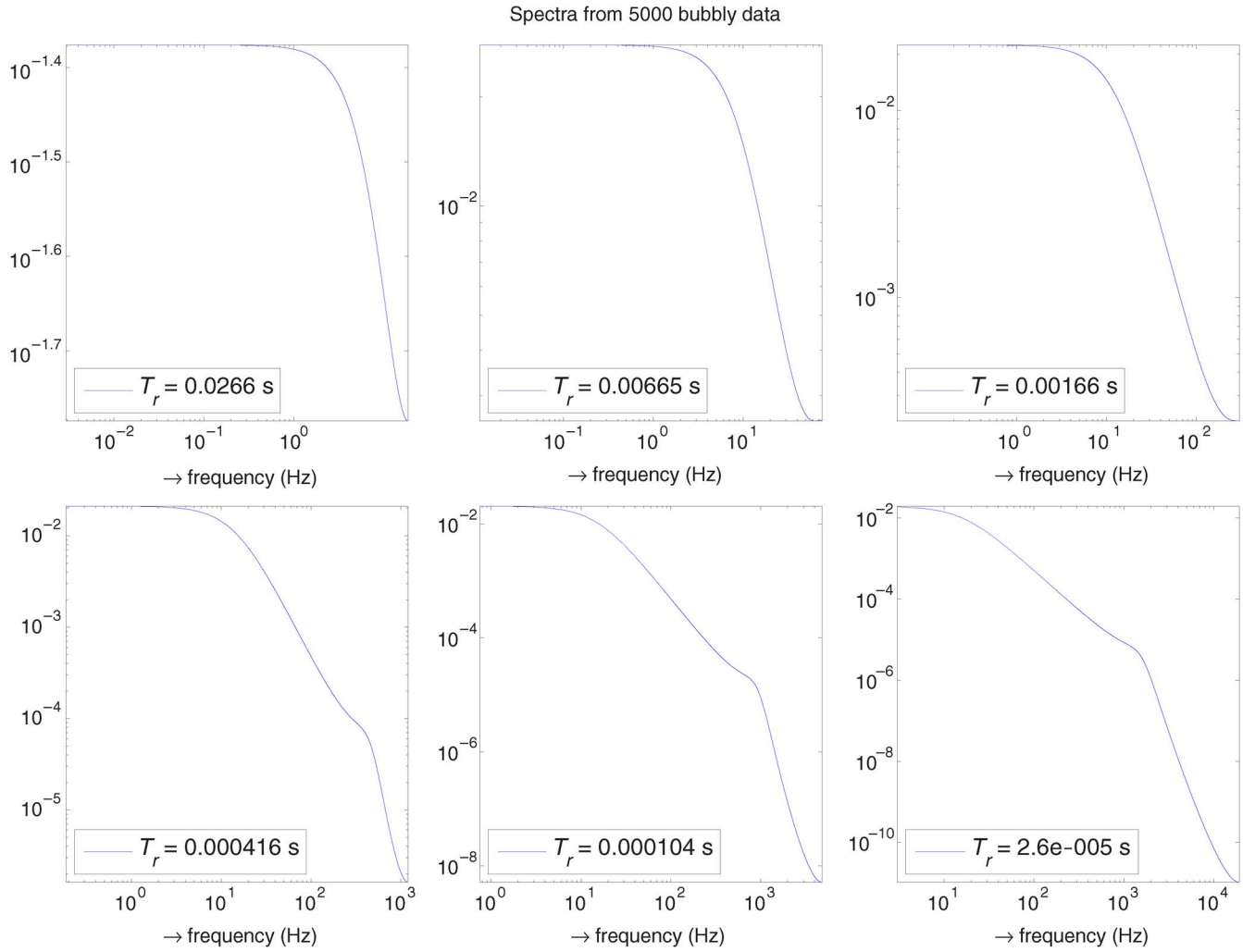

Fig. 8. Spectra from 5000 irregular observations, estimated for increasing sampling frequencies with orders selected by ARMAsel-irreg from candidates AR(0), AR(1), AR(2), and AR(3).

The agreement for low frequencies is easily seen in Fig. 9, where the same spectra of Fig. 8 have been presented together. All estimated and selected spectra coincide until 100 Hz and start to diverge at higher frequencies.

The missing fraction in the MSSNNR signal is very small for resampling with T_0 and almost 99.9% for resampling with $T_0/1024$. Only 11 pairs at that small distance were available, and this explains why spectra can become less reliable for high resampling rates. Higher order AR models would give many spurious spectral details at higher frequencies above 200 Hz.

Fig. 10 shows the detailed analysis for the resampling frequency $16f_0$, with the slot width $T_0/2$. A rather large slot is chosen here because the spectra in Fig. 9 do not have steep slopes. In this example, higher order models are not selected. The preliminary analysis already produced a very accurate model that was not improved by the analysis with a smaller slot and more data. However, the high-order models can develop spurious spectral details here, with sharp peaks. These peaks were not present in high-order models from the LDA data in Fig. 3. They are specific for the bubbly flow data. The behavior of estimated spectra is not reliable for frequencies above 200 Hz

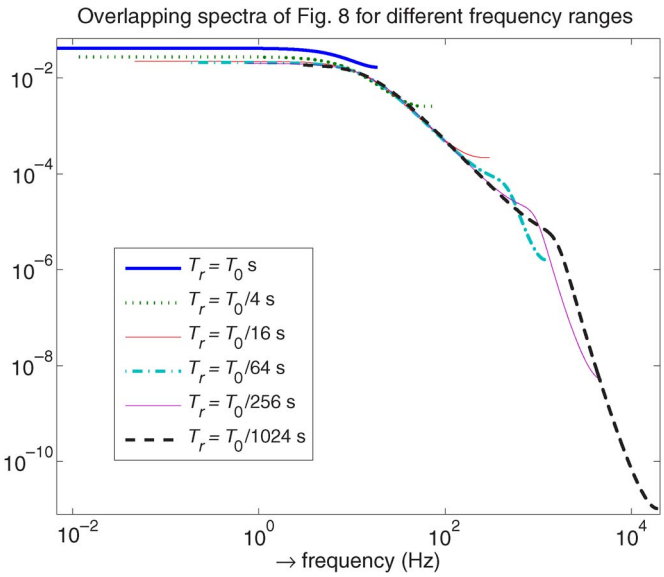

Fig. 9. Spectra from 5000 irregular observations, estimated for increasing sampling frequencies. The spectra for the three smallest ranges coincide, except for aliasing. Transitions to different slopes at higher frequencies occur at different locations and are not reliable.

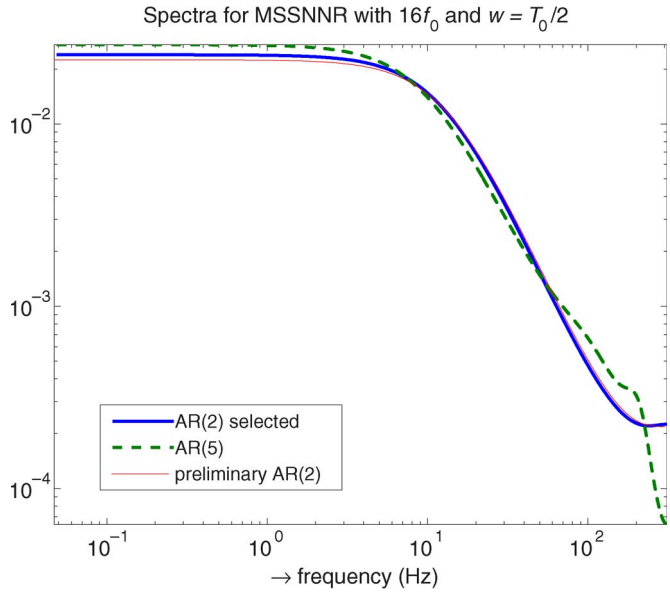

Fig. 10. Selected AR(2) and the AR(5) spectra from 35 015 irregular observations. As a comparison, the preliminary estimate from Fig. 8 for the same resampling frequency has been given.

in this example. This can already be concluded from the preliminary spectra in Fig. 8.

VII. Conclusion

ARMAsel-irreg has no problems in estimating spectra from practical data. If sufficient data are available and there is no velocity bias, it computes a spectrum that is close to the true spectrum if that can be approximated by a model with less than about 15 parameters. Large gaps or varying data rates do not disturb the quality of the estimated spectrum.

The differences found between experimental hot-wire and LDA data show that the two measured signals from the same turbulence experiment that have been evaluated in this paper have different spectra. The velocity bias that is present in the LDA data has a very strong influence. The hot-wire data without velocity bias give a much better representation for a turbulent process. For those equidistant hot-wire data and for equidistantly NN resampled irregular data without slotting, ARMAsel can select the best spectral model.

A procedure has been described that includes all stages required in the analysis of new irregularly sampled data with unknown spectral contents. It can be used for the choice of a resampling rate. It is important to include the significant spectral information without dubious peaks at high frequencies.

References

[1] C. Thiebaut and S. Roques, "Time-scale and time–frequency analyses of irregularly sampled astronomical time series," *EURASIP J. Appl. Signal Proc.*, vol. 2005, no. 15, pp. 2486–2499, 2005.

[2] 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.

[3] S. C. Sherwood, "Simultaneous detection of climate change and observing biases in a network with incomplete sampling," *J. Clim.*, vol. 20, no. 15, pp. 4047–4062, Aug. 2007.

[4] L. Barford, "Filtering of randomly sampled time-stamped measurements," *IEEE Trans. Instrum. Meas.*, vol. 57, no. 2, pp. 222–227, Feb. 2008.

[5] L. H. Benedict, H. Nobach, and C. Tropea, "Estimation of turbulent velocity spectra from laser Doppler data," *Meas. Sci. Technol.*, vol. 11, no. 8, pp. 1089–1104, Aug. 2000.

[6] J. D. Scargle, "Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly spaced data," *Astrophys. J.*, vol. 263, no. 2, pp. 835–853, 1982.

[7] R. J. Adrian and C. S. Yao, "Power spectra of fluid velocities measured by laser Doppler velocimetry," *Exp. Fluids*, vol. 5, no. 1, pp. 17–28, Jan. 1987.

[8] 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, no. 11, pp. 1541–1546, Nov. 1996.

[9] P. M. T. Broersen, "Time series models for spectral analysis of irregular data far beyond the mean data rate," *Meas. Sci. Technol.*, vol. 19, no. 1, Article no. 015 103, 2008.

[10] P. M. T. Broersen, "Bias contributions in time series models for resampled irregular data," in *Proc. IMTC*, Victoria, BC, Canada, 2008, pp. 882–889.

[11] M. J. Tummers and D. M. Passchier, "Spectral analysis of biased LDA data," *Meas. Sci. Technol.*, vol. 12, no. 10, pp. 1641–1650, Oct. 2001.

[12] D. K. McLaughlin and W. G. Tiedeman, "Biasing correction for individual realization of laser anemometer measurements in turbulent flows," *Phys. Fluids*, vol. 16, no. 12, pp. 2082–2088, 1973.

[13] T. J. McDougall, "Bias correction for individual realisation LDA measurements," *J. Phys. E, Sci. Instrum.*, vol. 13, no. 1, pp. 53–60, Jan. 1980.

[14] P. Gjelstrup, H. Nobach, F. E. Jørgensen, and K. E. Meyer, "Experimental verification of novel spectral analysis algorithms for laser Doppler anemometry data," presented at the 10th Int. Symp. Applications Laser Techniques Fluid Mechanics, Lisbon, Portugal, 2000, Paper 3.2.

[15] H. Nobach, E. Müller, and C. Tropea, "Refined reconstruction techniques for LDA data analysis," presented at the 8th Int. Symp. Applications Laser Techniques Fluid Mechanics, Lisbon, Portugal, 1996, Paper 36.2.

[16] P. M. T. Broersen, "Application of time series models to the spectral analysis of irregular turbulence data," in *Proc. IMTC*, Victoria, BC, Canada, 2008, pp. 33–38.

[17] 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, no. 22, pp. 6160–6168, 2005.

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

[19] P. M. T. Broersen, *Automatic Autocorrelation and Spectral Analysis*. London, U.K.: Springer-Verlag, 2006.

[20] P. M. T. Broersen, *ARMASA Matlab Toolbox*, 2002. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange

[21] H. Nobach, *LDA Benchmark Generator III*, 2001. [Online]. Available: http://www.nambis.de

Piet M. T. Broersen was born in Zijdewind, The Netherlands, in 1944. He received the M.Sc. degree in applied physics and the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 1968 and 1976, respectively.

He is currently with the Department of Multi-Scale Physics, Delft University of Technology. He developed statistical measures to let the measured data speak for themselves in estimated time-series models. This provides a practical and accurate solution for the spectral and autocorrelation analysis of stochastic data. Software for the automatic estimation of spectra and autocorrelation functions of equidistant random data, for missing-data problems, and for irregularly sampled observations is available online. His main research interest is in the automatic and unambiguous identification of the character of stationary random measurement data.