Generating Data With Prescribed Power Spectral Density

Piet M. T. Broersen and Stijn de Waele

***Abstract*—Data generation is straightforward if the parameters of a time series model define the prescribed spectral density or covariance function. Otherwise, a time series model has to be determined. An arbitrary prescribed spectral density will be approximated by a finite number of equidistant samples in the frequency domain. This approximation becomes accurate by taking more and more samples. Those samples can be inversely Fourier transformed into a covariance function of finite length. The covariance in turn is used to compute a long autoregressive (AR) model with the Yule–Walker relations. Data can be generated with this long AR model. The long AR model can also be used to estimate time series models of different types to search for a parsimonious model that attains the required accuracy with less parameters. It is possible to derive objective rules to choose a preferred type with a minimal order for the generating time series model. That order will generally depend on the number of observations to be generated. The quality criterion for the generating time series model is that the spectrum estimated from the generated number of observations cannot be distinguished from the prescribed spectrum.**

***Index Terms*—$1/f$ noise, ARMA process, linear filtering, order selection, spectral analysis, time series models.**

I. Introduction

THE ACCURACY of many experiments is limited by the presence of measurement noise in the observations. It will certainly be a problem in sensitive satellite measurement data [1]. Often, a good characterization of the noise spectrum is known from previous experiments or from a detailed physical description of the sensor and its environment. In such circumstances, it is possible to test the signal processing sequence in advance by using a noise realization of a stochastic process that has the known spectral density. The purpose may be to verify that the proposed signal processing allows for the desired accuracy. One specific application is the new gravity field and ocean circulation explorer mission, to be launched in 2006 [1]. This mission aims at the development of an improved model of the earth's gravity field. The presence of colored observation noise in a huge number of observations leads to a difficult numerical regression problem, demanding a weighted least squares solution. The weighting matrix is the inverse of the extremely large covariance matrix of the noise. The power spectral density of the colored noise is concentrated in the low frequency part. A time series description of the known noise spectral density gives the possibility to generate noise realizations. The same time series model can be used to design an inverse filter for the colored noise; for real-time implementation, it is important that the filter has a low order. The filtered noise becomes uncorrelated, which is an important advantage in the remaining regression problem [1]. Other examples for the generation of data emerge in turbulence, where the spectrum is proportional to $f^{-5/3}$ and in other physical problems, e.g., with $1/f$ noise.

Time series modeling is a parametric description of spectral densities. The three model types that can be used for time series models are autoregressive (AR), moving average (MA), and combined ARMA models. All stationary stochastic processes can be characterized by AR or by MA models of infinite order [2]. AR models are more suitable for spectral peaks; MA models are better for valleys. In practice, most processes can be described adequately by AR(p), MA(q), or ARMA(p, q) processes of finite orders p and/or q [3].

Generating stationary data for a given ARMA process requires some care. Using zeros or any arbitrary values as initial conditions, generated signals become stationary after the duration of the impulse response. Unfortunately, the impulse response is only finite for a MA process. It is infinitely long for AR and ARMA processes. Therefore, this primitive method of data generation without care for initial conditions can only be exact for MA processes. It is at best an approximation for AR or ARMA processes. A better method is found by separating the generation into an AR and a MA part. Consider the joint probability density function of a finite number of AR observations with a prescribed correlation function. Data can be generated that obey that prescribed AR correlation. A realization of the ARMA process is obtained by filtering AR data with the MA polynomial.

So far, it has been assumed that the spectral requirements are already formulated as a time series model. However, prescriptions for data can also be given in terms of correlation functions or power spectral densities. In principle, N observations can be considered as a single period of a multisine signal, that can be given the prescribed spectral density at maximally $N/2+1$ discrete frequencies $\omega = 2\pi k/N$, $k = 0, 1, \ldots, N/2$. Multisine signals are particularly appropriate as input signals for system identification at a limited number of frequencies, but not for applications like $1/f$ noise. The discrete spectrum would require a new signal specification for each value of N, at different frequencies. It would only have the desired spectral properties for sample sizes that are multiples of N, if care is taken with the transients of the initial conditions. Therefore, the treatment of these types of prescriptions for data also preferably uses time series models, which produce stationary stochastic data with a continuous spectral density that can be generated for different values of N with the same model. A solution has been given to

Manuscript received June 15, 2002; revised April 16, 2003.

The authors are with the Department of Applied Physics, Delft University of Technology, Delft, The Netherlands (e-mail: broersen@tn.tudelft.nl).

Digital Object Identifier 10.1109/TIM.2003.814824

fit an AR, MA, or an ARMA model to a finite number of sampled spectral values [4]. It uses the inverse Fourier transform of the spectrum as the desired covariance function. An AR model of high order is calculated then from those covariances with the Yule–Walker relations [2]. The AR model is also a basis to compute other time series models. Models can often be simplified by using the long AR model as input for a reduced statistics analysis [4].

This paper describes the joint probability density function (pdf) of an arbitrary number of observations of an AR(p) process. That is the basis for the generation of data. Prescribed spectral densities can be based on estimated or on exact knowledge. It is shown that order selection criteria can be adapted to the reliability or the accuracy of the spectral prescriptions. Furthermore, it is shown that an adequate generating process for the prescribed spectral density may depend on the number of observations to be generated.

II. ARMA Models

An ARMA(p, q) process is defined as [2]

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

where ε_n represents a series of independent, identically distributed, zero mean white noise observations. The process is purely AR for $q = 0$ and MA for $p = 0$. The ARMA process can also be written with polynomials of AR and of MA parameters as

$$A(z)x_n = B(z)\varepsilon_n, \qquad z^{-1}x_n = x_{n-1} \quad (2)$$

with $A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}$ and $B(z) = 1 + b_1 z^{-1} + \cdots + b_q z^{-q}$. Models may have estimated polynomials of arbitrary orders, not necessarily equal to the true p and q. Processes and models are stationary if the poles, the estimated roots of $A(z)$, are inside the unit circle; they are invertible if the zeros, the roots of $B(z)$, are inside the unit circle. The spectral density is given by [2]

$$h(\omega) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{\left|B(e^{j\omega})\right|^2}{\left|A(e^{j\omega})\right|^2}. \quad (3)$$

This shows that the parameters of a time series model, together with the variance of the exciting white noise, determine the spectral density. The covariance function is the inverse *integral* Fourier transform of (3). It can be approximated by an inverse discrete Fourier transform of a sampled $h(\omega)$. However, it is also possible to obtain an exact covariance function for a given ARMA(p, q) process by direct computations in the time domain [2]. Therefore, the parameters of a time series model are a good representation for the characteristics of a process, exact in both the time and in the frequency domain.

The quality of estimated ARMA(p', q') models with polynomials indicated by $\wedge$ is measured with the model error ME [5]. This measure can be used in simulations where an omniscient experimenter knows the true process ARMA(p, q) parameters that generated the observations. Likewise, it can be used to evaluate the difference between a prescribed spectral density, expressed as an ARMA process and the approximating ARMA spectrum. The ME is a scaled transformation of the expectation of the squared error of prediction PE

$$\mathrm{ME}\left(\frac{\hat{B}_{q'}(z)}{\hat{A}_{p'}(z)}, \frac{B(z)}{A(z)}\right) = N\left(PE/\sigma_\varepsilon^2 - 1\right). \quad (4)$$

The model error is asymptotically equal to N times the spectral distortion SD, defined as

$$\mathrm{SD} = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(\ln \hat{h}(\omega) - \ln h(\omega)\right)^2 d\omega \quad (5)$$

where the $\wedge$ denotes an approximating spectrum. In this way, the ME can also be defined for spectral densities that are not given by time series models. The expectation of ME is for ***unbiased*** models independent of N and of the variance of the signal. Only true and estimated parameter values are required to compute the ME with (4). The asymptotical expectation of the ME of unbiased efficiently estimated ARMA(p', q') models that contain at least all truly nonzero parameters is equal to the number $p' + q'$ of estimated parameters.

III. Probability Density of AR Process

The normal distribution of a variable x with mean μ and variance σ^2 is given by

$$¥(x, \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \quad (6)$$

The joint pdf of N observations $\boldsymbol{X}$, where $\boldsymbol{X}$ is a jointly normally distributed vector stochastic variable

$$X = (x_1\, x_2 \cdots x_N)^T \quad (7)$$

with

$$E\{X\} = \mu_x$$
$$E\left\{(X - \mu_x)(X - \mu_x)^T\right\} = R_{xx} \quad (8)$$

is given by

$$f(X) = \frac{1}{(2\pi)^{N/2}\left|R_{xx}\right|^{1/2}} \cdot exp\left[-\tfrac{1}{2}\left\{(X - \mu_x)^T R_{xx}^{-1}(X - \mu_x)\right\}\right]. \quad (9)$$

This is a general expression that can be applied to AR, MA, or ARMA processes by expressing the Toeplitz matrix R_{xx} in the parameters. It is often preferred in the literature and it has computational advantages to use as much as possible the white noise innovations with diagonal covariance function in the probability density and likelihood functions [2]. The observations $\boldsymbol{X}$ are then considered as filtered innovations. The probability density function of N observations of an AR(p) process with polynomial $A_p(z)$, with normally distributed zero mean innovations ε_n can also be written as a conditional product of the last $N - p$ observations, given the first p

$$f(x_1, x_2, \ldots, x_N) = f(x_{p+1}, \ldots, x_N | x_1, \ldots, x_p) f(x_1, \ldots, x_p). \quad (10)$$

The first part of the right-hand side for the last $N-p$ observations is given by [2, p. 347]

$$f(x_{p+1}, \ldots, x_N|x_1, \ldots, x_p) = \left(\frac{1}{2\pi\sigma_\varepsilon^2}\right)^{(N-p)/2} \exp\left(\frac{1}{2\sigma_\varepsilon^2}\sum_{n=p+1}^{N} A_p(z)x_n\right). \quad (11)$$

The second part describing the first p observations is also known [2, p. 350]. However, a recursive form of that expression is a better starting point for the generation of data. By using conditional densities, it follows that:

$$f(x_1, x_2, \ldots, x_p) = f(x_p|x_1, \ldots, x_{p-1})f(x_1, \ldots, x_{p-1}). \quad (12)$$

Generally, for all observations q, with $q < p$

$$f(x_1, x_2, \ldots, x_q) = f(x_q|x_1, \ldots, x_{q-1})f(x_1, \ldots, x_{q-1}). \quad (13)$$

With those intermediate results for q, (12) can be written as

$$f(x_1, x_2, \ldots, x_p) = \left(\prod_{q=2}^{p} f(x_q|x_1, \ldots, x_{q-1})\right) f(x_1). \quad (14)$$

Elements of the Levinson–Durbin algorithm [6] can be used to evaluate this expression. This algorithm recursively computes parameter sets of increasing order $q < p$ from the AR correlation function, as well as variance expressions. The last parameter of an AR model of order q is always equal to the reflection coefficient k_q. The algorithm starts with

$$\begin{aligned} a_1^1 &= -R(1)/R(0) = k_1 \\ \sigma_1^2 &= R(0)\left(1-k_1^2\right) \\ A^1(z) &= 1 + a_1^1 z^{-1} \end{aligned} \quad (15)$$

where $R(q)$ denotes the covariance of the AR(p) process at lag q. The recursion for $q = 2, 3, \ldots, p$ is given by [7, p. 169]

$$\begin{aligned} a_q^q &= -\left[R(q) + \sum_{i=1}^{q-1} a_i^{q-1} R(q-i)\right] \Big/ \sigma_{q-1}^2 \\ k_q &= a_q^q \\ a_i^q &= a_i^{q-1} + k_q a_{q-i}^{q-1}, \qquad 1 \le i < q \\ \sigma_q^2 &= \sigma_{q-1}^2\left(1-k_q^2\right) \\ A^q(z) &= 1 + a_1^q z^{-1} + \cdots + a_q^q z^{-q}. \end{aligned} \quad (16)$$

At the final stage p of the recursion, the polynomial $A^p(z)$ is equal to $A(z)$ in (2). The polynomial $1 - A^{q-1}(z)$ is the best linear predictor of order $q-1$, or the best linear combination of $q-1$ previous observations to predict the next observation x_q. Hence, the conditional probability density of x_q, conditional on $q-1$ previous observations has as expectation $[1-A^{q-1}(z)]x_q$ with the variance σ_{q-1}^2. Using this in the conditional expectations (13) gives with (6)

$$f(x_q|x_1, \ldots, x_{q-1}) = \text{¥}\left(x_q, \left(1 - A_{q-1}(z)\right)x_q, \sigma_{q-1}^2\right). \quad (17)$$

The variance of the AR(p) process is given by

$$\sigma_x^2 = R(0) = \frac{\sigma_\varepsilon^2}{\prod_{i=1}^{p}\left(1-k_i^2\right)}. \quad (18)$$

The recursive variance relation for intermediate AR orders can be expressed with increasing or decreasing index, which gives

$$\sigma_q^2 = \frac{\sigma_\varepsilon^2}{\prod_{i=q+1}^{p}\left(1-k_i^2\right)} = \sigma_x^2 \prod_{i=1}^{q}\left(1-k_i^2\right). \quad (19)$$

Now, the conditional density (17) becomes

$$f(x_q|x_1, \ldots, x_{q-1}) = \left(\frac{\prod_{i=q}^{p}\left(1-k_i^2\right)}{2\pi\sigma_\varepsilon^2}\right)^{1/2} \exp\left(-\frac{\prod_{i=q}^{p}\left(1-k_i^2\right)\left(A_{q-1}(z)x_q\right)^2}{2\sigma_\varepsilon^2}\right). \quad (20)$$

The substitution of (20) in (14) and that result together with (11) in (10) is merely an exercise with as result sums of products. However, as all ingredients for the probability density function $f(x_1, x_2, \ldots, x_N)$ are given, the derivation is sufficient to be used in generating data for an AR(p) process.

IV. Data Generation

Data generation is strictly separated in MA and AR generation. For ARMA processes, it is essential that the AR part is used first, because the derivation of Section III used white noise as input, e.g., in (11). Remind that the covariances in (16) are covariances belonging to the AR polynomial only. The principle is for ARMA processes given by

$$\begin{aligned} A(z)v_n &= \varepsilon_n \\ x_n &= B(z)v_n \end{aligned} \quad (21)$$

which together combines to the result of (2)

$$A(z)x_n = A(z)B(z)v_n = B(z)\varepsilon_n. \quad (22)$$

AR(p) Data

The purpose is to generate N observations with the probability density function $f(v_1, v_2, \ldots, v_N)$ like in (10). The first observation v_1 has the only requirement that it has expectation zero and variance $R(0)$ or σ_v^2. That is found with a random number generator with normal or Gaussian density and the prescribed variance. The second observation v_2 follows from (15) and (17) as a normally distributed random variable with mean $-a_1^1 v_1$ and variance σ_1^2. The third observation v_3 uses (16) and (17) with mean $-a_1^2 v_2 - a_2^2 v_1$ and variance σ_2^2. The first p observations are generated in this way. According to (11), all further observations can be generated with the regime

$$v_n + a_1 v_{n-1} + \cdots + a_p v_{n-p} = \varepsilon_n, \qquad n = p+1, \ldots, N \quad (23)$$

This is a filter procedure with a Gaussian random white noise signal as input signal and the first p observations as initial conditions. For AR processes, the p initial observations and the $N-p$ filter results with (23) are together the N observations.

MA(q) or ARMA(p, q) Data

For MA data, the input signal v_n for the second equation in (21) is a Gaussian white noise ε_n. For ARMA processes, the input to the MA filter will be the output v_n of the AR filter (23). The data x_n are computed with

$$x_n = v_n + b_1 v_{n-1} + \cdots + b_q v_{n-q}. \tag{24}$$

The first q data require negative input index. Therefore, to generate N MA or ARMA observations with this method, the input sequence v_n has to be $N+q$ long. The first q points of the filter output are disregarded.

V. From Spectrum to ARMA Model

Data are always generated with the time series model (1), using white noise of a random generator as input. Therefore, the prescribed spectral density will at the end be defined by the parameters of an ARMA model. If the parameters are already given as characterization for the desired spectral density, they can be used immediately. However, the prescribed spectrum can also be given as an exact continuous function $h(\omega)$, $0 \le \omega \le \pi/T$. Without loss of generality, the sampling time T is taken to be 1. The first step is to determine an approximation for the covariance function. The covariance function $R(k)$ is the inverse integral Fourier transform of the continuous function $h(\omega)$

$$R(k) = \int_{-\pi}^{\pi} h(\omega)\, e^{j\omega k} d\omega. \tag{25}$$

The length of $R(k)$ may be infinite. By defining $h(\omega)$ periodic with period 2π, the integral in (25) can also be taken from 0 to 2π. As ω in (25) is continuous, the usual discrete time inverse Fourier computer transformations can only be approximations for $R(k)$, unless specific assumptions can be made about the interpolated values between the discrete frequency values $h(\omega_k)$ that are used in the computation. Different covariance functions can belong to the same finite sampled discrete version of $h(\omega)$. A unique finite covariance may be found by considering the process as a MA process [6], which has nonzero values for only a finite range of shifts k. Also, considering a finite part of the covariance function as a Fourier series gives an unique interpolation in a continuous function $h(\omega)$ [2, p. 578] with the FFT algorithm. However, the true covariance function of an AR or an ARMA process is infinitely wide, albeit that it will be damping out. Approximating the integral of $h(\omega)$ in (25) by a summation is equivalent with sampling in the frequency domain. This causes the equivalent of aliasing of the covariance function, in the time domain. Therefore, some care in the covariance calculation with inverse Fourier transform is required. Taking the mid-range value of the integrand $h(\omega)$ as representative for an integration interval $\Delta\omega$, a correlation function $R(k)$ can be approximated from discrete values $h(m\Delta\omega)$ as

$$\begin{aligned} R(k) &= \int_0^{2\pi} h(\omega) e^{j\omega k}\, d\omega = \int_{-\Delta\omega/2}^{2\pi-\Delta\omega/2} h(\omega) e^{j\omega k}\, d\omega \\ &\approx \sum_m h(m\Delta\omega) \int_{(m-0.5)\,\Delta\omega}^{(m+0.5)\,\Delta\omega} e^{j\omega k}\, d\omega \\ &= \frac{2}{k\Delta\omega} \sin\left(\frac{k\Delta\omega}{2}\right) \sum_m h(m\Delta\omega) e^{j\Delta\omega\, km}\, \Delta\omega. \end{aligned} \tag{26}$$

For small $\Delta\omega$, this summation is almost equal to the usual summation result in (26) because $sin(y)/y$ has the limiting value 1 for small y. A small value for $\Delta\omega$ is obtained if many equidistant points $h(m\Delta\omega)$ are used. The derivation of (26) for $R(k)$ is preferred above the usual inverse FFT if only a few samples of a given $h(\omega)$ are used in the inverse transformation. It is also possible that the prescribed spectral density is specified for only a small number of frequencies. The integration in (26) defines explicitly what is transformed. However, if $A(z)$ and $B(z)$ are given, it is always more accurate to use the standard time series formulae for the calculation of the covariance function for a finite number of shifts [2]. Integration with (26) or any other approximation of the integral can only provide exact results for the covariance function if the true process is MA.

For one-sided prescribed spectra defined for ω between 0 and π, the first step is to sample $h(\omega)$ at intervals $\Delta\omega$, with $\Delta\omega = \pi/L$. The $L+1$ samples $h(m\Delta\omega)$, $m = 0, 1, \ldots, L$ are made to a symmetric spectrum by adding the $L-1$ samples for $m = 1, \ldots, L-1$ in reversed order. Taking $L = 2^k$ gives a convenient length for the FFT. After the inverse Fourier transform of the elongated sampled spectrum, the first L, ($L = \pi/\Delta\omega$) points describe the desired covariance function. L should be chosen high enough, such that the values of the covariance are effectively zero for lags beyond L. If that turns out to be impossible for a prescribed spectrum because the covariance function is too elongated, L should be chosen at least as high as the number of observations N that has to be generated. The larger L, the better the estimated covariance. This is an operational advice for the length L and it should always be validated that the correlation is negligible beyond L, or that $L > N$. The first K lags of the covariance function are transformed to an AR(K) model with the Yule–Walker relations [6]. If the prescribed $h(\omega)$ is exact, taking L and K greater will generally give a better approximation; the improvement may be small.

If the true process is a finite MA(q) process, no problems arise as long as the correlation length L is taken greater than q. All other processes have an infinite long covariance function that generally damps out quickly. The question is if some value M of the AR order, less than L or K, can be considered high enough for an approximation of sufficient accuracy for data generation. It may be advantageous to have a low-order process to generate data. It is still more important to design a low order filter to undo the coloring of some given spectral density to whiten the noise [1]. Any distortion of the true $h(\omega)$ will finally become noticeable if more and more data are generated. Therefore, the minimally required order M depends on the number of observations

N that has to be generated. A natural choice is to allow distortions that are much smaller than the expected statistical uncertainty if the generated data are analyzed. The distortion can be quantified as the bias of the truncated model in the ME measure (4) or in the spectral distortion (5). If the bias contribution of the truncation is smaller than the variance contribution to the ME of estimating a single parameter, the higher order parameters can be omitted without the possibility that the omission can be detected from N generated observations. This is no tight boundary but it is a sensible compromise. This gives an allowed ME of 1 for the truncated AR(M) model with respect to the true AR(∞) process for a given N. With (4), this gives

$$ME\left(\frac{1}{A_M(z)}, \frac{1}{A_\infty(z)}\right) = 1. \tag{27}$$

This is the lowest order M, for which the AR(M) model gives

$$\sigma_M^2 = \frac{\sigma_\varepsilon^2}{\prod_{i=M+1}^{\infty}(1-k_i^2)} = \sigma_x^2 \prod_{i=1}^{M}(1-k_i^2) < \sigma_\varepsilon^2\left(1+\frac{1}{N}\right). \tag{28}$$

In practice, the AR(∞) process in (28) is replaced by the finite order AR(K) process, with K chosen at least so high that the difference between the AR(K) and the AR($K/2$) model has a ME difference smaller than say 0.1 for the given N. This can be verified in practice and K can be chosen higher until this condition is met. Simulations will indicate that orders K and M are often much smaller than N. It is possible to use the AR(K) model as input for a reduced statistics algorithm [4] that computes AR, MA, and ARMA models of various orders. Moreover, it will select the model type and the model order that are closest to the given AR(K) process. Furthermore, the MA model with the lowest order q and the ARMA(r, $r-1$) model with lowest order r can be found that have a ME value less than 1 with respect to the AR(K) model, for a given value of N. In this way, the minimum number of parameters, M or q or $2r-1$, can be determined that is required to meet the accuracy demands. If the true process $h(\omega)$would be a finite order MA(q) or ARMA(p, q) process, this finite order can easily be detected with the reduced statistics estimator. If this MA or ARMA model requires less parameters than the AR(M) model described before, it may be worthwhile to use the most parsimonious time series model to generate or to filter data. However, it should be realized that the MA and ARMA models have long transients in inverse filtering. Using lower order models will give a less accurate approximation of the prescribed spectrum.

It is also possible to define a prescribed spectrum with an *estimated* periodogram, instead of with a continuous function $h(\omega)$. In those cases, special care is required, because estimated periodograms contain a lot of spurious details. If N_0 observations are transformed with the FFT and the absolute values are squared, the resulting function is the raw periodogram, say $P(\omega)$. The inverse Fourier transform of $P(\omega)$ should be the covariance function, but it is an aliased version which has as expectation [2]

$$E\left[\hat{R}(s)\right] = \frac{N_0 - s}{N_0} R(s) + \frac{s}{N_0} R(N_0 - s) \tag{29}$$

with $R(s) = E\{x_n x_{n+s}\}$. Using tapers and windows will cause more distortions of the covariance. The use of periodograms cannot be advised. Periodograms as a spectral prescription can be of the same accuracy as some very high-order estimated AR model, if the triangular bias in (29) is negligible [3]. However, a much better solution exists. If one wants to *use measured data* as prototype for the spectrum, it is advised to estimate the time series model for those data with ARMAsel [8]. This automatic computer program finds the best spectral time series model for given data with robust algorithms and criteria, which select only statistically significant details.

Another possibility is that the information in $h(\omega)$ is not exact, at least not for all frequencies. For colored satellite noise, the spectral density at some discrete, not equidistant, frequencies has been given and the continuous function $h(\omega)$ is determined with interpolation [1]. The proposed method to deal with those circumstances is largely the same as dealing with exact $h(\omega)$, but with a completely different critical value for the ME. Variation of the interpolation methods and spline solutions produces different continuous spectra, different covariances, and, hence, different time series models. The difference between those interpolation solutions can be expressed in the ME (4) for a given value of N as the number of data to be generated. If no information about the best interpolation is available, all solutions are equally well obeying the prescriptions. The largest difference found in the ME between the different interpolation results can be used as a critical ME value, instead of the critical value ME $= 1$ that was derived before for a given exact true $h(\omega)$. In this way, lower order models may become accurate enough if the prescription is less precise.

The practical criterion for data generation is always: generate N observations with a time series spectrum that cannot be distinguished from the true given spectrum. It might be possible that the model is good for the generation of N observations, but the difference can, perhaps, be seen if the same model was used to generate 100 N or more observations. Therefore, the generating model may depend on N.

VI. Simulations

To show the possibilities of the data generation, it is applied to an example where broken exponents determine the desired spectral shape. It is some prototype spectrum for turbulence data. The spectrum consists of two declining slopes after a constant at the lowest frequencies. The first slope descends at a rate of $\sim f^{-5/3}$ from 0.01 f_0 and the second slope of $\sim f^{-7}$ starts at 0.1 f_0

$$h(\omega) = \frac{1}{1+\left(\frac{\omega}{0.02\pi}\right)^{5/3}} * \frac{1}{1+\left(\frac{\omega}{0.2\pi}\right)^{16/3}}. \tag{30}$$

Fig. 1 shows no visible difference between the true $h(\omega)$and the AR(1000) approximation. The ME difference between the truncated AR(500) and the AR(1000) model is $1.4 * 10^{-9}N$, which is less than 0.1 for $N < 7 * 10^7$. According to the rules developed in Section V, data can be generated with sufficient accuracy if the ME of the truncated model is less than 1. Table I presents those orders. It is clear that the data generation can be carried out with a very high accuracy. Taking much lower

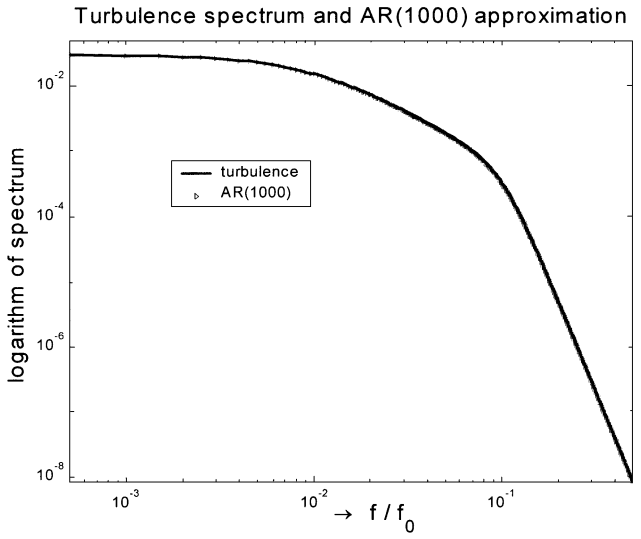


Fig. 1. Turbulence prototype spectrum and approximating AR(1000) model for $L = 2^20$.

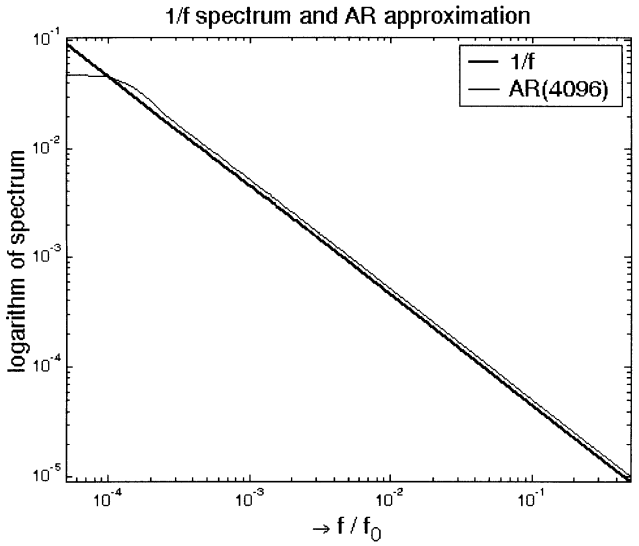


Fig. 2. $1/f$ prototype spectrum and approximating AR(4096) model for $L = K = 4096$.

TABLE I
LOWEST ALLOWED TRUNCATED AR ORDER M WITH THE ME LESS THAN 1 FOR THE GENERATION OF N TURBULENCE OBSERVATIONS, AS A FUNCTION OF N; $L = 2^{20}$ AND $K = 1000$

$\Rightarrow N$	10^2	10^3	10^4	10^5	10^6	10^7
M	7	11	19	37	71	143

values for L, the length of the inverse Fourier transform, has hardly any influence on the minimum order M. The values for M obtained with $L = 2^{12}$ and $K = 500$ are equal to those in Table I, except the last one that becomes 141. This demonstrates that this turbulence covariance function is already damped out sufficiently at lag 500. The sampling distance in the frequency domain hardly influences the estimated covariance function and the AR parameters if the length K of the AR model is greater than 500.

An interesting spectral density in physics is the $1/f$ spectrum in Fig. 2. The normalized true spectrum and an approximation are shown. Both have the same sum of the total power in the normalization. Due to the singularity at $f = 0$, it is difficult or even impossible to generate data that obey this spectrum. The length of the approximating covariance function is strongly influenced by the value of L. Particularly, the approximating finite value of the spectrum taken for $\omega = 0$ and the choice for L, determining the smallest nonzero sampled frequency at $\omega = \pi/L$ or $f = f_0/2L$, have an enormous impact on the length of the correlation. On the other hand, it is impossible to recognize or to verify the $1/f$ shape from N_0 observations at frequencies lower than say $1/N_0$. This is visible in Fig. 2 because the approximating spectrum becomes inaccurate for f less than say 0.0001 f_0. In approximating this true $1/f$ spectrum, a choice had to be made for $f = 0$, because the true value of the spectrum becomes infinite and leads to numerical problems. Therefore, some extrapolation from the first two nonzero sampled values for $h(\omega)$ have been used. Those are at the frequencies $f_0/2L$ and f_0/L. A parabola through the spectrum of those two points,

TABLE II
LOWEST ALLOWED TRUNCATED AR ORDER M WITH THE ME LESS THAN 1 FOR THE GENERATION OF N FROM 100 TO 100 000 OBSERVATIONS, AS A FUNCTION OF L AND K, FOR $1/f$ AS PRESCRIBED SPECTRAL DENSITY. N_0 IS THE NUMBER FOR WHICH THE AR(K) AND THE AR($K/2$) MODEL HAVE 0.1 AS THEIR ME DIFFERENCE

L	K	N_0	$\Rightarrow N$	10^2	10^3	10^4	10^5
2^{10}	2^{10}	$2.1\ 10^4$		8	38	109	223
2^{11}	2^{10}	$8.6\ 10^4$		10	50	165	339
2^{12}	2^{10}	$7.2\ 10^3$		11	63	242	560
2^{12}	2^{12}	$8.3\ 10^4$		11	63	244	589
2^{14}	2^{12}	$2.9\ 10^4$		13	89	455	1429
2^{16}	2^{12}	$6.6\ 10^3$		15	112	688	2410
2^{14}	2^{14}	$3.3\ 10^5$		13	89	457	1454
2^{16}	2^{14}	$1.2\ 10^5$		15	113	716	3057
2^{16}	2^{16}	$1.3\ 10^6$		15	113	716	3075

with the additional demand that the derivative of the spectrum at $f = 0$ equals 0 yields: $h(0) = 7/6 * h(f_0/2L)$.

Table II gives some results for the prescribed $1/f$ spectrum. The value for M depends much more on L than on K because the length of the approximating covariance function depends strongly on L. If K is chosen long enough, the actual value of K has almost no influence. The necessary AR orders are much higher than for the turbulence example. Despite that, the accuracy at very low frequencies is still not good. The accuracy of AR models of different orders is shown in Fig. 3. The low order models in Fig. 3 are still reasonable for higher frequencies. AR(10) is reasonable for $f > 0.03 f_0$, AR(100) is reasonable for $f > 0.003 f_0$, and AR(1000) is reasonable for $f > 0.000\,05 f_0$. The full length AR(4096) model is shown in Fig. 2. This demonstrates that it is possible to find a satisfactory time series model for any N, by selecting L at least $N/2$. However, in cases with a singularity at $f = 0$, the required minimum model order M is rather high. It can be advised to always compare visually the truncated AR spectra with the prescription, like in Fig. 3. The truncation value of M in Table II is a minimum; using the full length K gives always a better approximation.

The first example shows that smooth spectral shapes can be modeled easily with time series models. The second example

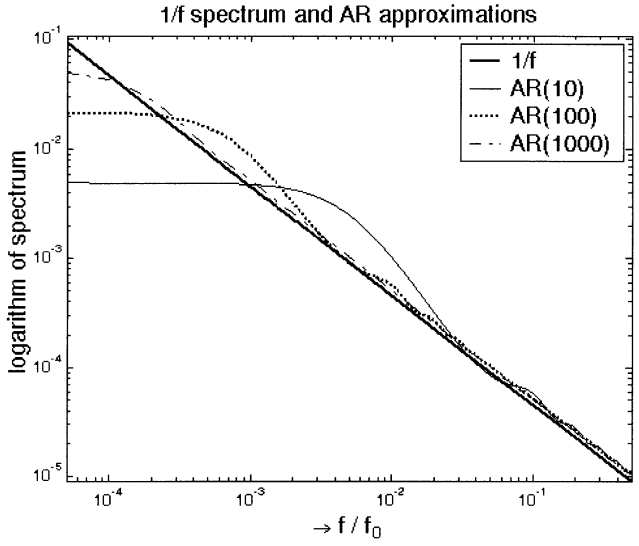


Fig. 3. $1/f$ prototype spectrum and three AR approximations for $L = K = 4096$.

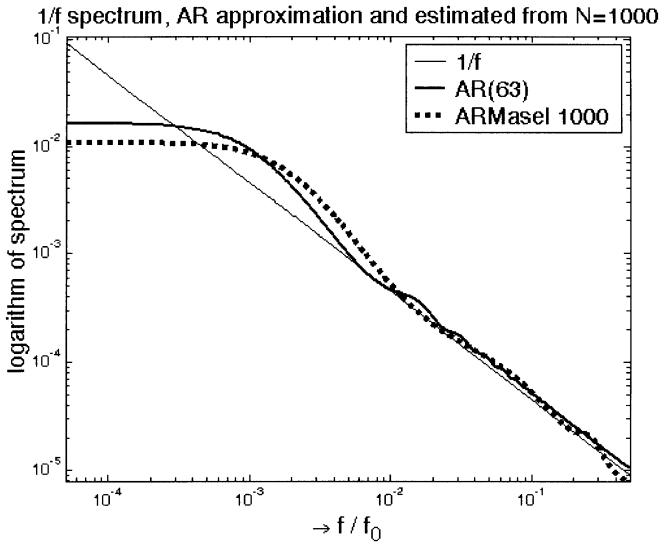


Fig. 4. $1/f$ prototype spectrum, the AR(63) approximation for $L = K = 2^{12}$ and the ARMAsel spectrum selected from 1000 observations that had been generated with the AR(63) process.

deals with the problem that the estimated covariance will always depend on the arbitrary length L. That determines the sampling distance in the frequency domain and, therefore, the lowest nonzero frequency, which is taken into account in the transformation of the true spectrum. Taking $L > N/2$ gives some guarantee that the differences between the generating process and the prescribed spectrum are not noticeable from N observations.

This is demonstrated in Fig. 4 where the ARMA(5,4) spectrum is shown that was selected with the ARMAsel algorithm [8] from 1000 observations that were generated with the AR(63) model obtained with $L = 2^{12}$. The resulting ARMAsel spectrum is good for $f > 0.01f_0$, reasonable for $f > 0.0002f_0$ and poor for still lower frequencies. It closely follows the AR(63) spectrum that generated the observations. Taking a higher AR order for the generating process gives ARMAsel results that are closer to the $1/f$ spectrum for a larger part of the frequency range.

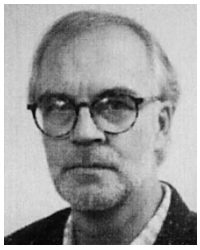
VII. Conclusion

Data with a prescribed spectral density can be generated in a simple way with time series models. The first step is to determine a finite-order AR model that has a spectrum close enough to the prescribed spectrum. Several tools are available to facilitate the search for a parsimonious time series model of other types. Use of higher orders always improves the accuracy, unless the spectral prescriptions were specified as a finite-order time series model.

By using an exact description of the probability density function of N autoregressive observations as a product of conditional probabilities, it is possible to efficiently generate autoregressive data. The generation of MA or ARMA data is solved with a simple filter operation.

References

[1] R. Klees, P. Ditmar, and P. M. T. Broersen, "How to handle colored observation noise in large least-squares problems?," *J. Geodesy*, vol. 76, pp. 629–640, 2003.
[2] M. B. Priestley, *Spectral Analysis and Time Series*. London, U.K.: Academic, 1981.
[3] P. M. T. Broersen, "Automatic spectral analysis with time series models," *IEEE Trans. Instrum. Meas.*, vol. 51, pp. 211–216, Apr. 2002.
[4] P. M. T. Broersen and S. de Waele, "Selection of order and type of time series models from reduced statistics," in *Proc. IEEE/IMTC Conf.*, Anchorage, AK, May 2002, pp. 1309–1314.
[5] P. M. T. Broersen, "The quality of models for ARMA processes," *IEEE Trans. Signal Processing*, vol. 46, pp. 1749–1752, June 1998.
[6] S. M. Kay and S. L. Marple, "Spectrum analysis—A modern perspective," *Proc. IEEE*, vol. 69, pp. 1380–1419, Nov. 1981.
[7] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York: Springer-Verlag, 1987.
[8] P. M. T. Broersen. ARMASA Toolbox. [Online]. Available: http://www.tn.tudelft.nl/mmr/downloads.



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 the Delft University of Technology (DUT), Delft, The Netherlands, in 1968 and 1976, respectively.

He is currently with the Department of Applied Physics, DUT. His main research interest is in automatic identification on statistical grounds. He found a solution for the selection of order and type of time series models for stochastic processes and the application to spectral analysis, model building, and feature extraction.

Stijn de Waele was born in Eindhoven, The Netherlands, in 1973. He received the M.Sc. degree in applied physics and the Ph.D. degree, with a thesis entitled "Automatic model inference from finite time observations of stationary stochastic processes," from the Delft University of Technology, Delft, The Netherlands, in 1998 and 2003, respectively.

Currently, he is a Research Scientist at Philips Research Laboratories, Eindhoven, The Netherlands, where he works in the field of digital video compression.