Yugoslav Journal of Operations Research
12 (2002), Number 2, 185-201

AN ILLUSTRATION OF HARMONIC REGRESSION BASED ON THE RESULTS OF THE FAST FOURIER TRANSFORMATION*

Imre BÁRTFAI

Technical Faculty, University of Pécs, Hungary
bartfai@upisun.jpte.hu

Abstract: The well-known methodology of the Fourier analysis is put against the background in the 2nd half of the century parallel to the development of the time-domain approach in the analysis of mainly economical time series. However, from the author's point of view, the former possesses some hidden analytical advantages which deserve to be re-introduced to the toolbox of analysts.

This paper, through several case studies, reports research results for computer algorithm providing a harmonic model for time series. The starting point of the particular method is a harmonic analysis (Fourier-analysis or Lomb-periodogram). The results are optimized in a multifold manner resulting in a model which is easy to handle and able to forecast the underlying data. The results provided are particularly free from limitations characteristic for that methods. Furthermore, the calculated results are easy to interpret and use for further decisions. Nevertheless, the author intends to enhance the procedure in several ways.

The method shown seems to be very effective and useful in modeling time series consisting of periodic terms. An additional advantage is the easy interpretation of the obtained parameters.

Keywords: Time series, forecasting, regression, Fourier-analysis.

1. INTRODUCTION

The author must confess he has always been fascinated by the analytical capabilities of the Fourier-transformation and its multi-purpose character, rendering it well-usable in a variety of the scientific areas. However, the application of this method is often limited to the technical/engineering area, as the practical usage for a

* An enhanced version of the paper presented at the occasion of EURO XVII, Budapest, 16-19th July, 2000.

statistician or a decision-maker has some disadvantages, even drawbacks. A short list of them follows:

- the result is not a model itself in a philosophical/statistical sense,
- raw data must fulfill certain criteria (e.g. equal distance),
- the resulting spectra are subject to inconvenient limitations,
- output is not easy to interpret and the amount of numbers is clumsy,
- the underlying processes and hidden relations are difficult to identify,
- making a forecast is a boring error-proven job.

Seeing these problems, some researchers have heavily criticized the FT method [18]. Others tried to enhance it to eliminate (at least some of) its disadvantages; a good example for this is the Lomb-periodogram, as will be shown in 2.3.2, or an application of the LSQ estimator to the results of the FT (see [5]). Harmonic regression [10] is also a tool to describe time series of cyclic nature; it can be combined with nonlinear optimization methods to refine the obtained results; here should be mentioned only a brief reference to some of them (see Chapter 10. in [14]; furthermore, Chapter 15 thereof also describes algorithms which need the knowledge of analytical derivatives and therefore in a lot of cases have only restricted applicability). More methods could be put among them such as NELDER-MEAD'S, POWELL'S, and the *simulated annealing method,* which is a variant of that of Nelder-Mead's, perfected for global extrema search.

However, most of the methods mentioned above generally remain limited to particular application areas (such as cardiology or stock market trend analysis).

Having the facts above the author selected the goal of this work to produce a practical *method* characterized by the following properties.

- build a model *based* on the (F)FT (or maybe another spectral method [see below]),
- the resulting model tends to be parsimonious,
- the figures obtained are more or less easy to interpret even for nonprofessionals,
- it is applicable for analyses and forecasts,
- the model is optimal from a given point of view,
- the model can be tested whether it is (or even whether its components are) significant.

In short: the expected outcome of the research is a product: a statistical tool for practical purposes.

The author does not want to declare he discovered something completely new; he rather wants to introduce a new combination of existing methods which provides interesting results and can extend the analytical apparatus in use to analyze time series.

Below we try recapitulate the stages of the development connected with this short theoretical discussion. Afterwards, the current level of the realization will be also shown using several sample data sets.

2. USING THE DFT

I need to mention here two introductory remarks:

1. Discrete Fourier Transformation will be used as the method and its realization works on real finite sample data.
2. DFT and FFT will not be distinguished here as they differ "only" in speed and some other technical details, the results being in theory the same.

Thus, the well-known Fourier Transformation is taken as the starting point. The transformation of y_t, $t = 1,...,N$, having two different ways of writing (see [20]), is

$$y_t = \sum_{i=1}^{q} (a_i \cos 2\pi f_i t + b_i \sin 2\pi f_i t) = \sum_{i=1}^{q} A_i \sin(2\pi f_i t + \varphi_i) \tag{1}$$

The result of the transformation is the set of $[a_i, b_i]$; $i = 1,...,q$ pairs (or that of $[A_i; \varphi_i]s$), where $q = \lfloor (N-1)/2 \rfloor$, and we assume the series y_t has its trend and/or mean removed, i.e. (here intentionally omitted) $a_0 = 0$. Anderson [1] has shown the results obtained are optimal in the least square sense. Computationally,

$$a_i = (2/T) \sum_{t=1}^{N} y_t \cos 2\pi f_i t \tag{2}$$

and

$$b_i = (2/T) \sum_{t=1}^{N} y_t \sin 2\pi f_i t \tag{3}$$

are calculated. Furthermore, $A_i = \sqrt{a_i^2 + b_i^2}$ and $\varphi_i = \arctan(a_i / b_i)$ [20].

Note that in this approach the $f_i s$ result from the sample directly (see e.g. in [8, 9]): let the time difference for two adjacent observations Δt. Thus, the whole sample spans $T = N \cdot \Delta t$ time. Therefore, the difference between the successive frequencies, called *spectral resolution* can be written as

$$f_i - f_{i-1} = \Delta f = T^{-1} \tag{4}$$

and $f_1 = \Delta f$.

The maximum frequency that can be used is the so-called NYQUIST frequency:

$$f_c = (2\Delta t) \tag{5}$$

Staying here, only domain has been changed, we moved from the time domain to the frequency one. Using the full set, 1 is no model at all. A more or less selective look upon it taken by the investigator already becomes an implicit model building (1), but this is hardly usable yet.

2.1. First step: selecting the DFT result

To obtain a model we consider there are some ith components which are more important than others. Even they can be divided into two groups: the first one represents the parameters of underlying processes which are to be modeled and the second one represents noise of different origins such as measurement and registration error, and disturbances of most various kinds. We refer here to [7]. Of course, there are true or almost completely pure noise processes but we will show that our method seems to be capable of identifying them.

The initial question: how to pick elements which the model itself consists of? The importance of the particular components is proportional to the appropriate power [6]:

$$I_i = (n/2)(a_i^2 + b_i^2)$$

Now the set of i s can be divided into two subsets: the 1st contains the components *only* belonging to the model (we will mark them with j), and the 2nd the other ones. At the very beginning of the research we have defined merely three (partially similar) criteria to determine the model subset:

NUMBER: define the number of components (m) in the model. This approach assumes that only the first m of the sorted I_i s are real contributors of the model, and all the others are random.

PERCENT: the smallest

$$p_j = I_j / \sum_{i=1}^{q} I_i$$

ratio must not be smaller than the predefined limit (e.g. 0.05). The hidden assumption says that all individual components having p_i less than the given limit represent noise.

CUMRATIO: the ratio

$$P_j = \sum_{j=1}^{m} p_j = \sum_{j=1}^{m} I_j / \sum_{i=1}^{q} I_i$$

must exceed the given limit (e.g. 0.8). Virtually, it is the user's guess for the S/N ratio.

All listed selection choices assume that the components are sorted in descending order of their I_i s. Furthermore, it is possible to calculate the GINI-coefficient (g) [19] for both sets using I s in ascending order, and building their cumulative sum. The concentration of the whole spectra (i.e. $g >> 0.5$) means a high contribution of a small number of components to the total variance and such a way promises high odds of building a "good" model, which should have the following properties:

parsimony: $m << q$; this requirement can be evaluated directly.

high power: $EV \to 1$

where EV can be defined in the following way [11]:
let $\hat{y}_t$ s be the estimated time series data, where the estimation itself runs using (1), and the particular components $[A_j;\varphi_j;f_j]$ belong to the model selected in one of the previous ways (of course, the upper limit is now m instead of q, and the order of the components is changed by the sorting procedure described above).
Then,

$$EV = 1 - \underbrace{\sum_{t=1}^{N}(y_t - \hat{y}_t)^2}_{Q_{err}} / \underbrace{\sum_{t=1}^{N}(y_t - \bar{y})^2}_{Q_{tot}} \tag{6}$$

This requirement is virtually identical with $U \to 0$, where U is the THEIL's inequality coefficient [15]. Note, however, that EV could also be negative if the model is extremely wrong (i.e. $Q_{err} > Q_{tot}$); this fact is its possible disadvantage. Otherwise, for linear models, it equals to R^2 called also RSQ.

Statistical significance: the underlying idea is the "good old" ANOVA table with the terms: Model-Error-Total. Remember, FT has variance term: the periodogram I_i. Now, they can be summed up by building the Model (using the j indices) and Error (those not belonging to the model). The total variance is the sum of all I_i s. Then, sloppily expressed, the resulting F should be significant, i.e. $p < 0.05$ (or another predefined limit).

The method up to this point was introduced in [4]. Some remarks should also be mentioned here:

1. The above criterion CUMRATIO prescribes the required minimum for EV directly.
2. ANOVA table can also be calculated using the sum in (6), because the term

$$Q_{tot} = \sum_{i=1}^{q} I_i \tag{7}$$

also counts. A numerical consistency check can be also done comparing Q_{tot} provided by (7) with that of (6).
3. The decrease of g before and after the model building can also be treated as an efficiency indicator. Moreover, the *Lorenz*-curve [19] can be displayed in both cases. One can expect the curve to be near the diagonal line.
4. For large N s, the usage FFT is recommended due to runtime requirements. However, it is not necessary for fast computers and several hundred points.

2.2. An incidental challenge: the leaking problem

As it is obvious from (4), the set of f_i s are fixed at this stage and they depend only on the sampling process. However, as we mentioned before, there is no warranty whether they coincide with the genuine frequencies of the underlying processes. And, according to Murphy's law, normally this is not the case, especially when observations are few, and therefore the distances of the spectral lines are remarkable. The worst-case scenario is the one where the underlying frequency exactly halves the interval between f_i and its neighbor. This leads to *leakage*, thus the spectral peak becomes blurred, and therefore, the above three criteria can not be fulfilled. To remedy the situation, we have one auxiliary and two essential choices (assuming the sampling conditions are fixed, which is very often the case), respectively:

1. We consider only spectral lines (i.e. I_i s) for the model, where they represent a local maximum, i.e. the following inequality is valid for the unsorted – thus sorted by f_i – sequence:

   $$I_{i-1} < I_i > I_{i+1}$$

   Before applying this condition, the series I_i s could be also smoothed, but it seems to be unnecessary at the moment.
2. We treat every element of the triplets $[A_j, \varphi_j, f_j]$ [1] as variables and minimize Q_{err} as their function changing them appropriately; see below in the Subsection 2.3.1.
3. We try to estimate the triplets with different methodology, as described in Section 2.3.2.

2.3. Suggestion for solution

2.3.1. Optimizing the Results by Nonlinear Minimization

The problem treated in par. 2 above can be seen as a nonlinear optimization one (remember Section 1). More of them were tested extensively, such as NELDER-MEAD'S, POWELL'S, and the *simulated annealing method*.

The third one was dropped soon due to extreme computing times and because it did not converge in all cases. The first method has the advantage of numerical stability, but the results were not perfect in all cases, and the number of iterations was rather high, sometimes excessive. Finally, we have chosen the second one due to its speed and accuracy; the circumstance was utilized the triplets selected by one of the criteria described in 2.1 are *good* initial estimators, so that the iteration step-size must be even reduced.

---

[1] We prefer this notation due to its easy interpretation: A means *amplitude*, φ stands for *phase shift* (at start), and f means *frequency*.

As mentioned before, the goal function to be minimized is Q_{err}. To judge whether the optimization was successful/necessary, we calculated its values before and after the optimization procedure, and tested them by F-statistics, obtaining it by dividing the former by the latter one.

Another interesting and easily interpreted descriptor of a model is the *mean absolute difference* (M.A.D.).

$$M.A.D. = \sum_{t=1}^{N} \underbrace{| y_t - \hat{y}_t |}_{d_t} \tag{8}$$

It is possible to compare the $M.A.D$ values before and after by means of the *Wilcowon*-test [17] (the tail area probability is computed by the formula in [2]), as we treat the d_t series as statistically dependent (like a treatment effect), in such way we apply one-sided test. The ratio of $M.A.D.$ to $\bar{y}$ instantly informs the user about the error of the model.

For the sake of truth it must be mentioned that the speed for longer data sets is rather moderate as the fast version of the inverse Fourier transformation – which is the key part of the evaluation of Q_{err} through $\hat{y}_t$ – has not been implemented yet.

2.3.2. The Lomb Periodogram

A relatively new approach for identification and testing of the individual spectral components (their locations and significance) is the *Lomb Periodogram*; for details see [3, 14]. The method itself shows the following qualities:

- it does not require evenly sampled data (good for e.g. the series containing missing observations)
- it can overcome NYQUIST-frequency limitation (5)
- it can be tuned by the *oversampling parameter* (normally $n \geq 4$)
- it informs about the significance of the resulting – most important – frequency by means of exponential distribution.

However, there are also some drawbacks (partially based on experience):

- the original implementation is slow (which can be partially remedied by the fast version)
- it is assumed that the frequencies are independent (otherwise monotonicity of p can fail)
- the oversampling factor needs some "fingertip-feeling" to hit true frequencies (see PLLFRS[2]) or may not loose itself from a particular peak in the spectrum
- mean should be removed, indeed
- the coefficients are not obtained directly.

---

[2] Power Line Low Frequency Remote Control, see more in 3.2.

In our practice, it is assumed that the model consists of sums of some independent frequency components, and that they will be removed by subtraction from the original series in a successive manner unless no predefined significance criterion is fulfilled. To get the a_i, b_i coefficients, an initial guess as in (2) and (3), respectively, will be generated; here, however, the f_i calculated by the Lomb method is being used. Then an optimization for the $[a_j;b_j;f_j]$ triplet will be done in a similar fashion as in 2.3.1, but the method used is the LEVENBERG-MARQUARDT [14] one, which, in turn, is here a curve fitting task rather than a function minimization problem. It is very fast and quite exact; but requires the providing of analytical derivative functions. Nevertheless, because we fit only a single spectral component at one time, it is possible to provide the said derivatives. By the way, [1] gives an exact formula for the former parameters but I found it rather complicated to implement it compared to applying nonlinear fit using some minimization criteria (here the χ^2). Alternatively, linear fit can also be done for the a_j, b_j pair only, keeping f_j fixed.

The final model consists of the sum of the particular ones. To judge the model, a statistics to the one described previously is used, e.g. (6) and the U-statistic.

The result of linear fit is not free from the potential leakage, but the over-sampling parameter which makes the spectral lines closer together reduces its importance.

Experience has shown that Lomb's method generally seems to be faster and more effective than its DFT counterpart. Seldom a computational problem can occur resulting in peaks at the same frequency in subsequent iterations. A change in procedure and/or sample parameter can remedy this; as a last resort, we can change to DFT.

3. CASE STUDIES

We analyze three data sets to demonstrate the practical implications of the algorithm of this paper. The procedure is currently implemented in Euphoria language [16] because of the following advantages:

- very short development turnaround times,
- interpretative language, but extremly fast in this category,
- no memory limit other than the hardware one,
- multi-platform availability (DOS/Windows/Linux),
- excellent runtime error handling.

Some routines are taken from [14], converted to the target environment, of course.

3.1. White Noise

The data came as sample data set with [12], and consists of 106 data points. The DFT method discovers the random character of the data by the Gini-coefficient: it is 0.4809, thus, below 0.5 (for ordinary data sets it would lie near 1). The Lorenz-curve

(Fig. 1) is smooth. Using the PERCENT criteria (NUMBER 0.05), the method found only 2 components, so the power seems to be spread over the whole frequency range in a relatively uniform manner. The Lomb's method gives a clear unambiguous answer: the 1st iteration failed with $p = 0.344$, which means the data does not contain any significant frequency component. This is an excellent result as it prevents the investigator to waste his/her resources for an unnecessary analysis.

3.2. Artificial Ripple Control Data

Power plants usually control remote load groups by mixing some low frequency signal to the line voltage; the results can be observed as slight ripple on the pure sine wave [13]. This is named: Ripple Control[3]. The appropriate standard prescribes that the amplitude of the control signal should not exceed 0.5% of the carrier and the frequency is 183.3 cps (for 50 cps net). We generated an arbitrary sample from such signal under following conditions:

- the sampling frequency is $0.0025^{-1}\text{sec}^{-1}=400\text{Hz}$,
- the effective sample size is 40 (i.e. 100msec),
- the amplitude of the carrier is $230\,\text{V}\cdot\sqrt{2}$, that of the signal is 0.5% of it,
- the phase shift at the beginning for both signals is zero degree,
- we added random noise to the result exceeding the amplitude of the signal with 12dB.

Obviously, the spectral resolution $\Delta f = 10\,\text{Hz}$, so the frequency of the signal would be misses due to leakage. As a result, the Lomb's method was unable to identify the signal. The DFT with post-optimization produced the following results:

Statistics	Value
M.A.D	1.405641
EV	0.999942
Theil's *U*	0.003816
f_1	50.0032 Hz
A_1	333.02 V
φ_1	-0.1°
f_2	182.7818 Hz
A_2	1.96 V
φ_2	11.8°

One can see that the signal beneath the carrier is clearly identified, and that the huge amount of noise made only a slight trouble, so it seems possible to use the method for special demodulation purposes under noisy circumstances.

---

[3] In the German speaking countries it is named in a different manner: Tonfrequenz-Rundsteuerung (i.e. "tone frequency circle control").

The sample contained in fact 80 points, but the model was fitted on the first 40 only. However, using the model, the second half was forecast and these data were compared with the originals. The fit is virtually perfect ($U = 0.004214$).

3.3. Power Consumption Data

Power consumption of a small town (with industry) was registered in every 30 minutes. Data belonging to the same day were averaged.

For the sake of completeness, we include here a result of the complete run (see Appendix A).

Looking at the listing, we can get an impression of the control language. The model fits good as we can see the M.A.D./$\overline{y}$ ratio, which is about 5%. All components are significant, except that one of them is even more significant (see the asterisks in the fifth column). The interpretation of the frequencies is relatively easy: the first three components show a mixture of sinusoidal period for a year. The 4th and 5th components explain the changes within a month, and the last three are the weekly course. The amplitude is self-explanatory; it expresses the contribution of the particular component, and the phase shift is the horizontal offset from the zero-crossing sinusoidal wave.

The Gini coefficient is nice as the spectral contribution in column I is relatively smooth (see Fig. 2). The number of components (8) out of possible 182 stresses the parsimony of the model, which is, in turn, according to the ANOVA, highly significant. The *EV* ratio could be higher, but for practical data as high as 0.71 is fairly acceptable. The U is small enough to accept the model, too. (Look at Fig. 3 to see the goodness-of-it).

In the DFT approach we have intentionally chosen as many components as the previous analysis had produced. However, we needed to exclude some consecutive spectral components by allowing only local maxima to take part in the analysis because a run not shown here has provided several components with identical frequencies. Therefore, the initial *EV* before the optimization is only 0.45. The nonlinear optimization produced a significant decrease both in the *SSQ* and in the *M.A.D.*, however, they are not as good as those in Lomb. The rest of output can be analyzed in a similar way as we did before.

4. DISCUSSION AND CONCLUSIONS

We have seen that the suggested methodology can serve for several purposes, especially for those mentioned in Section 1. We point out the following properties of it:

- to check whether the investigated process has a white-noise character, applying also such interesting approaches like the GINI-coefficient,
- to identify relevant harmonic component estimations out of the set generated by FT,
- the ability to discover hidden periodicities,
- to build a parsimonious model able to forecast the underlying process data,

- the goodness-of-it is checked in a multifold manner using an abundance of different statistics, from the simple R^2 to THEIL's coefficients and *M.A.D.*,
- nevertheless, the leakage problem is eliminated in a way which identifies the "true" underlying spectral contributors even if they are not present in the intermediate analysis results.

Especially the results shown in 3.2 have stressed the advantages of the method, so again we refer back to them.

We suggest to treate the method shown as a kind of enhancement of the harmonic regression model which now contains a lot of statistics to test the goodness-of-it, and being able to detect such hidden components which remain undiscovered or hard to identify by other means. Nonetheless, the exactness of the obtained descriptors compared to those of the underlying process are surprisingly good and make us have high hopes regarding the practical usefulness and further expandability of the introduced technology.

5. FUTURE PLANS

To make the program to multi-purpose tool, there is a need for enhancements in many ways. Here is a short review of them:

5.1. Speed

There is a need to improve the speed of the program if one wants to use it on a longer data set. The following possibilities have been recently taken into account:

- switching to FFT, especially in the case of inverse transformations
  - for particular – e.g. embedded – applications even integer FFT can be considered,
- a compiled programming language could also make things faster; here Pascal is the candidate number one, due to the runtime error checking and dynamic memory handling capabilities.

5.2. Services

Dynamizing: the time series to be analyzed could be divided into smaller parts, and the analysis will be made on every section, even by sliding a short window on the data. Consequently, the changes in the spectral nature of processes can be discovered, too. The window can be given a particular shape (HAMMING, HANNING, PARZEN, etc.) to eliminate eventual side-lobes of the spectra.

Filtering: some peaks on the data may be treated as outliers and should not be considered. An appropriate input filtering procedure (e.g. SAVITZKY-GOLAY) could eliminate them allowing a more precise estimation of the underlying process. The power spectrum could also be smoothed making the identification of the true peaks easier.

More information: The program's output can be extended by other statistical descriptors. There are also some significance checks for the spectra in the literature which could also be introduced if needed. The Fourier method could

also be applied in a stepwise manner as well, so the improvement of the whole model could be tested between the subsequent steps.

Interactivity: the user may want to intervene in the component selection and/or in the iteration process.

Remote usage: a web page is to be set up where the visitor should be able to at least edit a job and make the analysis remotely, obtaining the result on his/her screen.

6. SUMMARY

The reader got acquainted with a method extending the application range of the traditional spectral analysis technique, which makes it possible to set up true spectral models on time series. The different variants of the method should not be treated as competitors, rather they complement each other. The author showed sample analysis and future extensions are briefly discussed, too. They show that the development process has not been finished yet, but the effort seems to be worthy to continue the task.

REFERENCES

[1] Anderson, T.W., *The Statistical Analysis of Time Series*, John Wiley & Sons, Inc., 1971.
[2] Casio, *Casio Programmbibliothek FX-601P/FX-602P*.
[3] Castiglioni,, P., *Lomb Periodogram*, http://www.cbi.polimit.it/glossary/Lomb.html
[4] Bártfai, I., "On prediction of time series using DFT", *Procc. of XIV International Conference on Mathematical Programming*, Mátraháza, Hungary, March 1990.
[5] Dynacomp, *DYNACOMP, Inc: Fourier Analysis Forecaster*.
[6] Fuller, W.A., *Introduction to Statistical Time Series*, John Wiley & Sons, Inc., 1995.
[7] Granger, C.W.J., "The typical spectral shape of an economic variable", *Econometrica*, 34 (1) (1966) 150-161.
[8] Hesselmann, N., *Digitale Signalverarbeitung*, Vogel-Verlag, Würzburg, 1983.
[9] Hesselmann, N., *Digitális Jelfeldolgozás*, Müszaki Könyvkiadó, Budapest, 1985.
[10] Hintze, J.L., *Time Series Analysis and Forecasting*, BMDP Statistical Software, Inc., 1991.
[11] Máhr, J., and Varga-Haszonits, Z., *Az Idöjárás Elörejelzése és a Mindennapi élet*, Gondolat, Budapest, 1978.
[12] Newton, H.J., *Timeslab*, http://stat.tamu.edu/~jnewton.
[13] Osvald, K., Nagy, G., and Vimi, J., *Hangfrekvenciás Központi Verzérlés*, Müszaki Könyvkiadó, 1981.
[14] Press, H.W., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in Fortran 77, Volume 1*, Cambridge University Press, Second Edition, 1992.
[15] Qi, Y., "A simulation laboratory to evaluation of dynamic traffic management system", PhD thesis, Center of Transportation Studies in Massachusetts Institute of Technology, 1997.
[16] Rapid Deployment Software, The Euphoria Programming Language, http://www.RapidEuphoria.com.
[17] Sachs, L., *Angewandte Statistik*, Springer-Verlag, 4. Edition, 1974.
[18] Schlittgen, R., and Streitberg, B.H.J., *Zeitreihenanalyse,* R. Oldenbourg, 1995.
[19] Smith, L., *NCEPH Seminar*, *Chapter Lorenz Curve and Gini Coefficient*, http://www.ann.edu.au/nceph/inequalities/nceph-presentation, February 1998.
[20] Stöcker, H., *Taschenbuch Mathematischer Formeln und Moderner Verfahren*, Verlag Harri Deutsch, 1995.

APPENDIX A

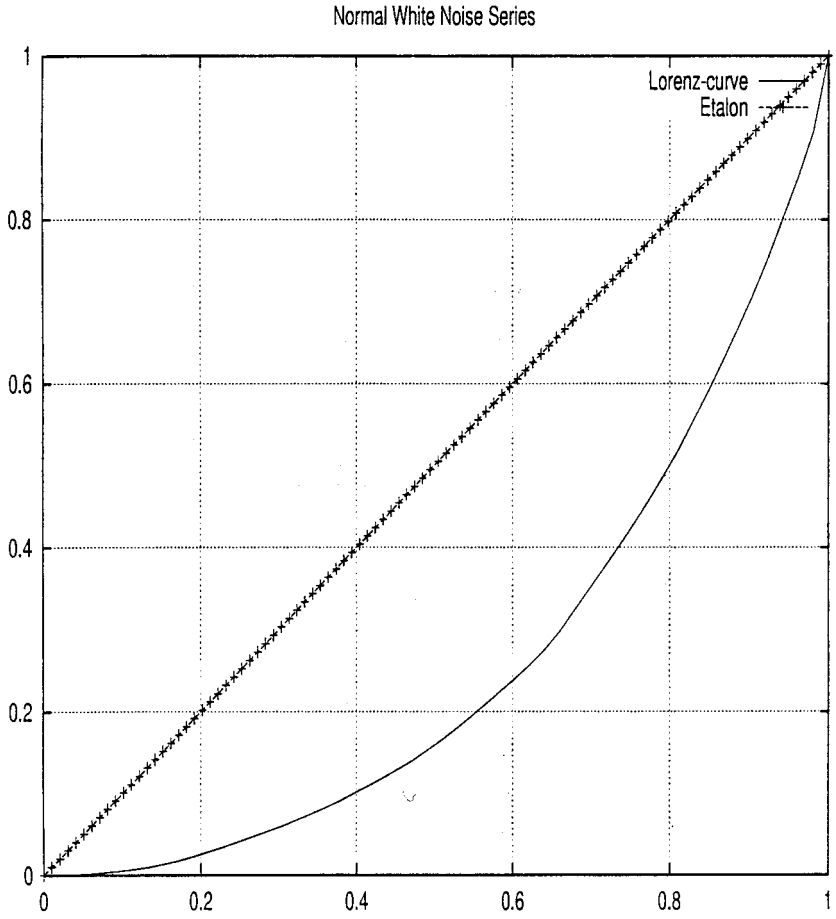

Figure 1: Lorenz-curve of normal white noise series

```
***PROGRAM DECK***

1:     TITLE "Daily Average Load - Lomb method"
2:     FILE/DOSC/temp/rau/stadt2.txt
3:     VAR 1 # 1st column
4:     DETREND Mean # remove mean
5:     METHOD Lomb
6:     NUMBER 0 # take as many as necessary
7:     TNAME  year # time unit
8:     DELTA  0.002793296 # 1/365
9:     REFINE # frequency can be changed
10:    PLOT   stadt21.glp # GNU Plot Output
11:    END # end of job
```

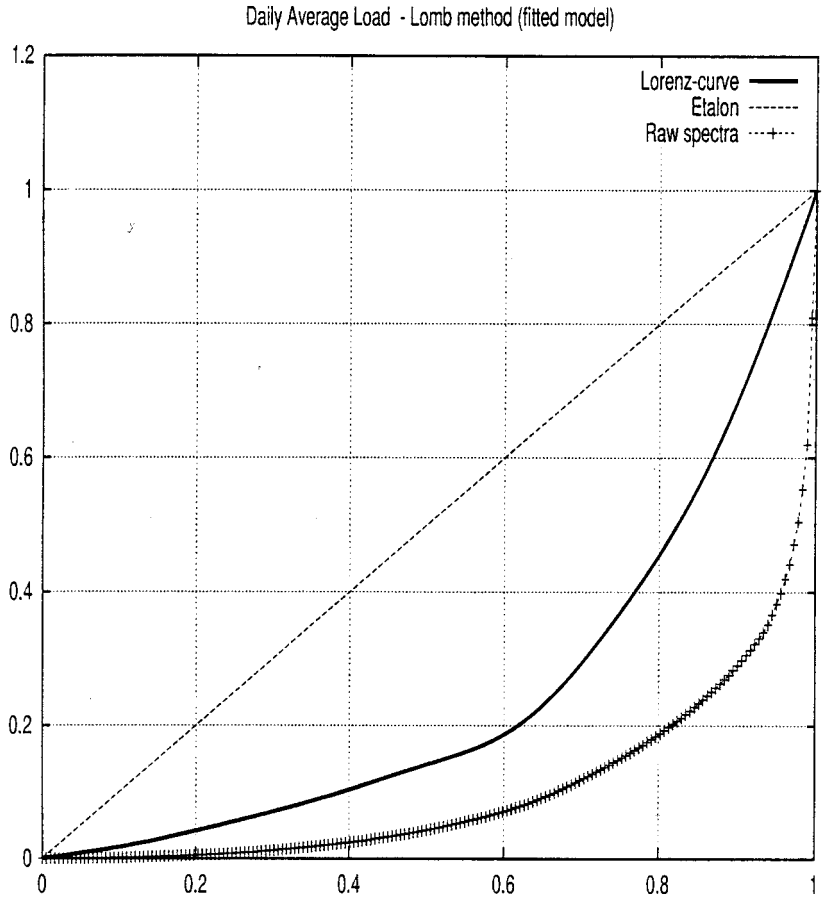

Figure 2: Lorenz-curves for the daily average data

```
***EXECUTION***

Time points read: 365
Time points used: 365

Mean subtracted:           426.2327

Significance level leaving iteration loop: 0.097947
      M.A.D.               21.451474

Frequency analysis results (units in cycle/year)
Lomb periodogram used
Note: only terms used in the model are printed.
```

f	a	b	I		A	fi(∅)
1.0136	22.94	19.58	166040.30	***	30.20	49.5
2.9328	-11.59	2.82	25981.48	***	11.95	-76.4
8.0384	-5.08	-12.15	31639.20	***	13.18	22.7
11.3747	6.00	6.91	15269.22	*	9.16	41.0
20.8660	-10.81	2.47	22442.65	***	11.10	-77.1
51.2001	24.44	-28.71	259434.73	***	37.76	-40.4
102.3275	0.84	24.60	110551.25	***	24.65	2.0
153.4132	-8.72	-12.83	43915.89	***	15.53	34.2

Gini coefficient: 0.4985 (8)

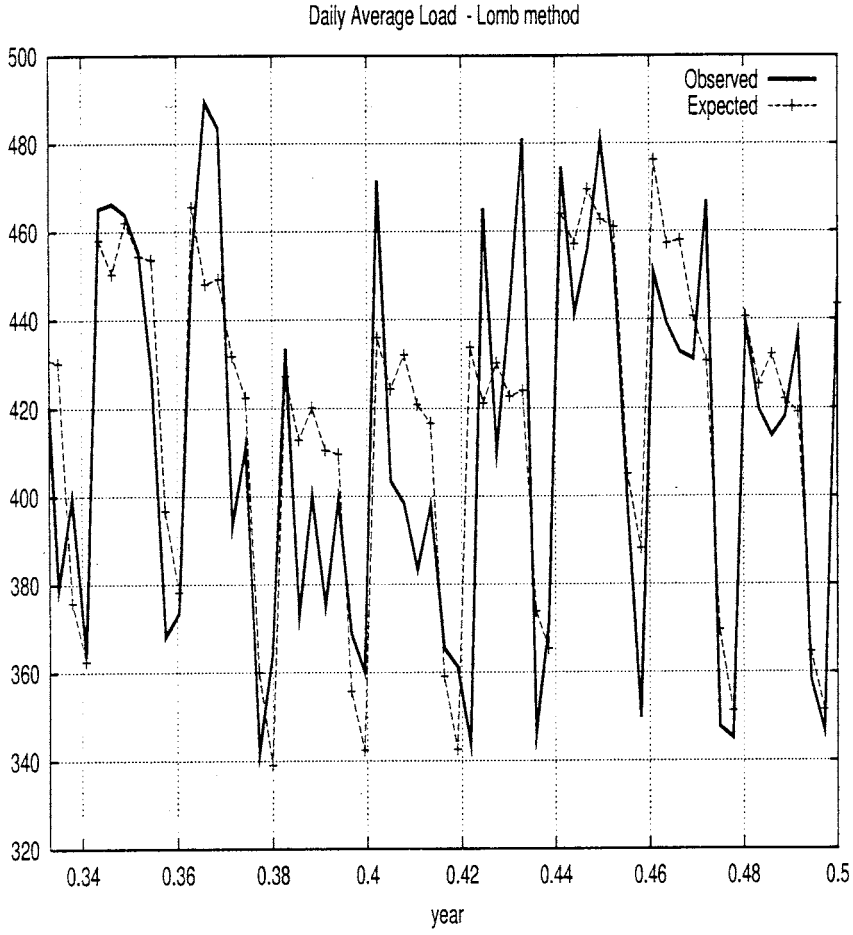

Figure 3: Original and fitted data plot for May and June

```
Comparison of Observed and Expected Values
==========================================
(Current Analysis)

***ANOVA***

   Source        SSQ       D.F.    Var.         F        p
   Model     676433.904     24   28184.746    34.909   0.000
   Error     274509.284    340     807.380
   Total     950943.188    364    2612.481

EV ratio.............................. 0.711329
Correlation coefficient................. 0.843524

Original data: mean ..................... 426.2327
      standard deviation .................  51.0424
Theil's inequality coefficient (U) ......   0.31948

Decomposition into proportions of inequality
      Um (Bias proportion) ..............   0.000702
      Us (Variance proportion) ..........   0.084877
      Uc (Covariance proportion) ........   0.914421
==================================================

Elapsed time: 17.40 sec; that of analysis: 17.36 sec

***PROGRAM DECK***

12:    #
13:    TITLE  "daily Average Load - with Fourier"
14:    FILE/DOSC/temp/rau/stadt2.txt
15:    VAR 1
16:    METHOD Number # specify number of components directly
17:    NUMBER 8 # as many as in Lomb
18:    TNAME  year
19:    DELTA  0.002793296
20:    PEAK   # only local maxima
21:    REFINE # Powell's method set
22:    END    # end of job

***EXECUTION***

Time points read: 365
Time points used: 365

Cumulative variance ratio of 8 terms used out of 182: 0.45133

Optimization results
   Initial SSQ:     521758.585792
   M.A.D.               30.920560
   Iterations in POWELL: 9  (2318 function evaluations)
   Final SSQ:       413411.677740
   M.A.D.               27.609542
```

```
   Wilcoxon-test for M.A.D.
      z= 3.692075     p=0.000111
   F test for SSQ decrease
   F statistics:    1.262080     p=0.013333

Frequency analysis results (units in cycle/year)
FFT method used
Note: only terms used in the model are printed.

     f       a       b          I                  A          fi(∅)
  0        426.23           66311128.65      *
  2.6497    -8.95   13.72     48977.71      *    16.40        -33.1
  4.7867    -6.92   -5.72     14699.05      *     8.99         50.4
  7.8190   -10.70   -4.89     25247.26      *    11.78         65.4
 14.6901    -4.74    7.35     13953.00      *     8.76        -32.8
 20.8408    -9.81    3.75     20125.11      *    10.52        -69.1
 51.1920    23.97  -29.17    260083.01      *    37.80        -39.4
102.3019     3.15   24.46    111013.99      *    24.70          7.3
153.3875    -9.41  -12.12     42938.15      *    15.36         37.8

Gini coefficient:   0.8221 (182) -> 0.5374 (8)

Comparison of Observed and Expected Values
==========================================
(Current Analysis)

***ANOVA***

   Source        SSQ       D.F.    Var.         F        p
   Model     537531.510     17      31619    26.540   0.000
   Error     413411.678    347    1191.388
   Total     950943.188    364    2612.481

EV ratio.............................. 0.565261
Correlation coefficient................. 0.752571

Original data: mean ..................... 426.2327
      standard deviation ................  51.0424
Theil's inequality coefficient (U) ......   0.39183

Decomposition into proportions of inequality
      Um (Bias proportion) ..............   0.002529
      Us (Variance proportion) ..........   0.142455
      Uc (Covariance proportion) ........   0.855016
==================================================

Elapsed time: 33.03 sec; that of analysis: 0.23 sec

***PROGRAM DECK***

23:    FINISH # end of file

Program terminated normally.
```