

# A WAVELET-BASED DISCRIMINANT PROCEDURE FOR SIGNAL-PLUS-NOISE PROCESSES

J. Wade Davis, University of Missouri  
Joseph E. Cavanaugh, University of Missouri  
J. Wade Davis, 222 Mathematical Sciences Building, Columbia, MO 65211  
(wdavis@stat.missouri.edu)

**Key Words:** Discriminant analysis, time series analysis, wavelets.

## 1. Introduction

Discrimination and classification are areas of practical as well as theoretical scientific interest and have received much attention in the statistical literature. Aimed at separating and allocating events, the general ideas of discrimination and classification extend to the field of time series analysis. Measurements in time often exhibit patterns that may serve as bases for allocating events of unknown origin into one of several known populations.

Techniques for time series discrimination have been employed in many different scientific areas. Shumway (1982) provides a comprehensive overview of common methodologies. Applications include speech pattern recognition, the categorization of seismic records as originating from either nuclear explosions or earthquakes, and the classification of different states of consciousness based on EEG readings. The problem of distinguishing between aircraft based on radar signatures is particularly relevant, since the underlying series may be viewed as realizations of signal-plus-noise processes (Zyck & Bogner, 1996).

In what follows, we propose and investigate a discriminant procedure for classifying a series of unknown origin that arises as a signal corrupted by noise. The discriminant is formulated using the distributional properties of the wavelet coefficients of the series, and involves the coefficients as well as the parameters representing the underlying populations. The parameters are estimated by beamforming series of known origin within a population, and applying a wavelet transform to the composite series.

We present simulation results that evaluate the performance of the discriminant using the hold-

out cross-validation procedure (Johnson & Wichern, 1998, pg. 654). Several factors are varied in the simulation study: the number of series observed within each population, the length of each series, and the noise level. Results suggest that our method works effectively even when the signal-to-noise ratio is low and the number of realizations available from each population is small. We close with an application in which the procedure is used to classify several Midwestern cities based on their temperature patterns.

## 2. Discriminant Formulation

In the development of our discriminant, we assume a basic understanding of wavelets. For an introductory treatment on wavelets and statistics, see Ogden (1997). Percival and Walden (2000) provide an in-depth look at wavelet-based statistical analyses of time series.

Consider a deterministic signal  $f(x)$  that has been corrupted by noise. Suppose the observed series can be represented as

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n.$$

We will assume for convenience that  $n$  is a power of two, although the methods to be outlined could be easily modified for settings where this does not hold. Also, we require that  $\varepsilon_i \sim iid N(0, \sigma^2)$ .

Suppose we are interested in classifying an unknown realization  $(y_1, \dots, y_n)$  as belonging to one of several possible populations. Without loss of generality, we shall assume that there are only two populations to which the unknown series could belong, Population A and Population B. Population A consists of processes where the underlying signal is  $f^A(x_i)$ , and Population B consists of processes where the underlying signal is  $f^B(x_i)$ ,  $i = 1, \dots, n$ .

Let  $\theta_{i,j}$  denote the  $i$ th population discrete wavelet transform (DWT) coefficient at scale  $2^j$ . The true DWT coefficients for Population A and Population B will be denoted by  $\theta_{i,j}^A$  and  $\theta_{i,j}^B$ , respectively. Let  $w_{i,j}$  be the  $i$ th empirical DWT coefficient at

---

This research was partially funded by a grant from the National Library of Medicine.

scale  $2^j$  for the series we are interested in classifying. The range of  $j$  is determined by the length of the series, the boundary condition, and the type of wavelet used (Bruce & Gao, 1996, pg. 69). We will represent the number of coefficients at scale  $2^j$  by  $n_j$ . Thus  $i = 1, \dots, n_j$  where  $n_j = n/2^j$ .

If the realization belongs to Population A, then  $(w_{i,j} - \theta_{i,j}^A) \sim iid N(0, \sigma_{A,j}^2)$ , where  $\sigma_{A,j}^2$  denotes the population variance for the scale  $2^j$  DWT coefficients of Population A (Ogden, 1997 pg. 122). Similarly, if the realization belongs to Population B, then  $(w_{i,j} - \theta_{i,j}^B) \sim iid N(0, \sigma_{B,j}^2)$ . Implicitly, we assume homogeneous variances within each level.

With the preceding distributional results, we can apply traditional discrimination techniques (Johnson & Wichern, 1998, Chp. 11) for classifying the series at each of the  $2^j$  scale levels. We define two measures of discrepancy,  $d_j^A$  and  $d_j^B$ , in terms of the negative log likelihood:

$$d_j^A = \frac{n_j}{2} \ln \sigma_{A,j}^2 + \frac{1}{2} \sum_{i=1}^{n_j} \frac{(w_{i,j} - \theta_{i,j}^A)^2}{\sigma_{A,j}^2},$$

$$d_j^B = \frac{n_j}{2} \ln \sigma_{B,j}^2 + \frac{1}{2} \sum_{i=1}^{n_j} \frac{(w_{i,j} - \theta_{i,j}^B)^2}{\sigma_{B,j}^2}.$$

For a given scale level  $2^j$ ,  $d_j^A$  and  $d_j^B$  may be viewed as measures which assess how effectively the unknown realization matches the attributes of each population.

Defining the vectors  $\mathbf{w}_j$ ,  $\boldsymbol{\theta}_j^A$ , and  $\boldsymbol{\theta}_j^B$  as  $\mathbf{w}_j = [w_{1,j} \dots w_{n_j,j}]'$ ,  $\boldsymbol{\theta}_j^A = [\theta_{1,j}^A \dots \theta_{n_j,j}^A]'$ , and  $\boldsymbol{\theta}_j^B = [\theta_{1,j}^B \dots \theta_{n_j,j}^B]'$ , we can express the discrepancy measures in vector form as

$$d_j^A = \frac{n_j}{2} \ln \sigma_{A,j}^2 + \frac{1}{2\sigma_{A,j}^2} (\mathbf{w}_j - \boldsymbol{\theta}_j^A)' (\mathbf{w}_j - \boldsymbol{\theta}_j^A),$$

$$d_j^B = \frac{n_j}{2} \ln \sigma_{B,j}^2 + \frac{1}{2\sigma_{B,j}^2} (\mathbf{w}_j - \boldsymbol{\theta}_j^B)' (\mathbf{w}_j - \boldsymbol{\theta}_j^B).$$

It can be shown that

$$d_j^A - d_j^B = \frac{n_j}{2} \ln \frac{\sigma_{A,j}^2}{\sigma_{B,j}^2}$$

$$+ \frac{1}{2} \mathbf{w}_j' \left( \frac{1}{\sigma_{A,j}^2} - \frac{1}{\sigma_{B,j}^2} \right) \mathbf{w}_j$$

$$- \left( \frac{1}{\sigma_{A,j}^2} \boldsymbol{\theta}_j^{A'} - \frac{1}{\sigma_{B,j}^2} \boldsymbol{\theta}_j^{B'} \right) \mathbf{w}_j$$

$$+ \frac{1}{2} \left( \frac{1}{\sigma_{A,j}^2} \boldsymbol{\theta}_j^{A'} \boldsymbol{\theta}_j^A - \frac{1}{\sigma_{B,j}^2} \boldsymbol{\theta}_j^{B'} \boldsymbol{\theta}_j^B \right).$$

We call the preceding a level discriminant. Note that this is a quadratic discriminant since it is quadratic in the data.

Thus far in the development, we have not assumed that the variances  $\sigma_{A,j}^2$  and  $\sigma_{B,j}^2$  are the same for each  $j$ . However, such an assumption is quite realistic. In fact, if the variances differed substantially, the qualitative differences in population realizations would be such that discrimination could be subjectively accomplished. If we assume  $\sigma_{A,j}^2 = \sigma_{B,j}^2 = \sigma_j^2$  for each  $j$ , the preceding quadratic level discriminant becomes a linear level discriminant and we obtain

$$d_j^A - d_j^B =$$

$$- \frac{1}{\sigma_j^2} (\boldsymbol{\theta}_j^A - \boldsymbol{\theta}_j^B)' \mathbf{w}_j$$

$$+ \frac{1}{2\sigma_j^2} (\boldsymbol{\theta}_j^A - \boldsymbol{\theta}_j^B)' (\boldsymbol{\theta}_j^A + \boldsymbol{\theta}_j^B).$$

The overall discriminant which leads to our classification rule is a weighted sum of the level discriminants. Let  $D = \mathbf{k}'\mathbf{d}$ , where  $\mathbf{k}$  represents a vector of weights and  $\mathbf{d}$  is the vector of level discriminants. We assign an unclassified realization to Population A if  $D < 0$ , and to Population B otherwise.

The weight vector  $\mathbf{k}$  could be chosen several ways. The simplest weight vector would consist of ones, thereby assigning each level discriminant equal weight. However, due to the nature of the DWT, successive levels consist of only half as many coefficients as the previous level. Since each level discriminant is a sum of  $n_j$  terms, and the  $n_j$  are halved as  $j$  increases, the relative magnitude of the level discriminants varies greatly. This complicates direct comparison of different level discriminants, so a unit weight vector is less than ideal.

We propose a weight vector that adjusts for the differences in magnitude by giving each successive level discriminant twice as much weight as the previous level. Thus the components of the weight vector ascend in powers of two. Our intuitive justification for this scheme arises from the pyramid algorithm (also known as the cascade algorithm) (Ogden, 1997, pg. 63). Recall that the  $n_j$  coefficients for scale  $2^j$  can be used to compute the  $n_{j+1}$  coefficients for scale  $2^{j+1}$  and that  $(n_j/n_{j+1}) = 2$ . Although there are half as many coefficients at scale  $2^{j+1}$ , each coefficient at scale  $2^{j+1}$  is calculated using two coefficients at scale  $2^j$ . Thus, the coefficients at scale  $2^j$  represent a higher level of resolution than those at scale  $2^{j+1}$ . By giving scale  $2^{j+1}$  twice as much weight as scale  $2^j$ , we implicitly assume that as much discrimination information is provided by the  $n_{j+1}$  coefficients at scale  $2^{j+1}$  as by the  $n_j$  coefficients at scale  $2^j$ . This weighting scheme puts all of the level discriminants on the same caliber, and places

a greater emphasis on the coefficients corresponding to the coarser detail levels where the signal is best represented. Simulation results verify that such a weighting scheme is superior to the unit weights scheme.

For situations where some levels are assumed to have equal variances while others do not, the overall discriminant  $D$  would be a weighted linear combination of the quadratic and linear level discriminants  $d_j^A - d_j^B$ . For certain settings, Percival and Walden (2000, pg. 380) discuss methods for testing the homogeneity of variance assumption.

We now discuss estimation of the population DWT coefficients  $\theta_j^A$  and  $\theta_j^B$  as well as the population variances  $\sigma_{A,j}^2$  and  $\sigma_{B,j}^2$ . The parameters  $\theta_j^A$  and  $\theta_j^B$  are estimated by beamforming series of known origin within a population. The series from a given population are first aligned and then averaged together to form a “pure” signal that represents the population. The DWT coefficients from the beamformed series comprise  $\theta_j^A$  and  $\theta_j^B$ . We estimate each population variance by the sample variance of the scale  $2^j$  empirical wavelet coefficients of the  $N$  training series. Thus, our estimates are of the form

$$s_{A,j}^2 = \frac{1}{(N n_j)} \sum_{k=1}^N \sum_{i=1}^{n_j} (w_{i,j,k} - \theta_{i,j}^A)^2$$

$$s_{B,j}^2 = \frac{1}{(N n_j)} \sum_{k=1}^N \sum_{i=1}^{n_j} (w_{i,j,k} - \theta_{i,j}^B)^2.$$

### 3. Simulations

To evaluate our discriminant procedure, we generated training samples based on two pairs of competing signals corrupted with varying levels of noise. For our first pair of signals, we let  $f^A(x) = \sin(6x)$  and  $f^B(x) = \sin(6x - 1)$ . To create the realizations, the functions were sampled  $n$  times on the interval  $[0, 2\pi]$  and corrupted with three different levels of Gaussian white noise. The noise levels were such that the signal-to-noise ratio (SNR) was -6 dB, -9 dB, or -12 dB, where the signal-to-noise ratio is defined as

$$\text{SNR} = 10 \log_{10} \left( \frac{\sum_{i=1}^n f(x_i)^2}{n\sigma^2} \right).$$

Here, the  $f(x_i)$ 's represent the “pure” signal and  $\sigma^2$  is the variance of the noise component.

The wavelet used in the decomposition belongs to the least asymmetric (LA) family (Percival & Walden, 2000). Also known as a symmetlet, this

wavelet is orthogonal, smooth, non-zero on a relatively short interval, and nearly symmetric (hence the name least asymmetric). Although there are no definitive rules for choosing a wavelet, the preceding properties make the LA(8) wavelet a good overall choice for many applications (Bruce & Gao, 1996, pg. 69). Of course, the choice of a wavelet ultimately depends on the characteristics of the signal being analyzed.

Figure 1 depicts the two “pure” signals  $f^A(x) = \sin(6x)$  and  $f^B(x) = \sin(6x - 1)$ . Note that the two signals are identical except for a slight phase shift. Figure 2 shows corrupted versions of  $f^A(x) = \sin(6x)$  with SNR's -6 dB, -9 dB, and -12 dB. The three graphs demonstrate the impact of the noise which is typical in these simulations.

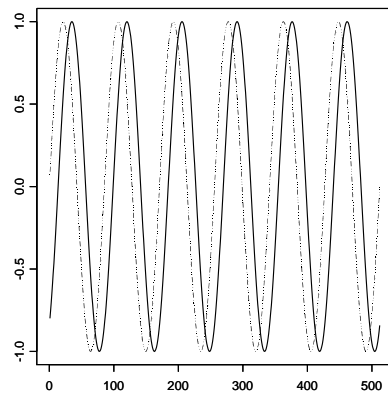


Figure 1:  $f^A(x)$  and  $f^B(x)$  for the first simulation experiment.

Our linear discriminant was evaluated using a cross-validation holdout procedure. The apparent error rate (APER) for each trial was calculated. The APER is nearly an unbiased estimate of the expected actual error rate,  $E(\text{AER})$  (Johnson & Wichern, 1998, pg. 654). Table 1 contains the average APER based on 100 trials for each set of simulation conditions. The size of the training set was either four or eight, and the number of points sampled was 64, 256, or 512.

As expected, increasing the size of the training set decreases the APER. This is due to improved beamformed estimates of  $\theta_{i,j}^A$  and  $\theta_{i,j}^B$ . For both training set sizes, the results are quite satisfactory considering the small number of realizations employed. The number of sample points has a strong impact on the error rate as well. This is not surprising, since the larger the number of sample points, the better the resolution of the signal. On a practical note, series should be collected at the highest sampling rate

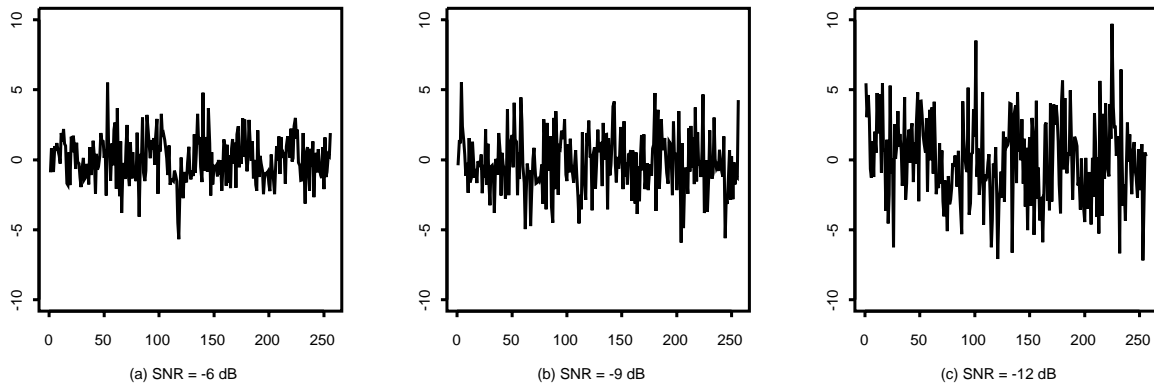


Figure 2:  $f^A(x) = \sin(6x)$  corrupted with varying degrees of noise.

Table 1: Cross-Validation Error Rates for  $f^A(x) = \sin(6x)$  vs.  $f^B(x) = \sin(6x - 1)$ .

| Sample Points | Training Set Size = 4      |          |          | Training Set Size = 8      |          |          |
|---------------|----------------------------|----------|----------|----------------------------|----------|----------|
|               | Signal-to-Noise Ratio (dB) |          |          | Signal-to-Noise Ratio (dB) |          |          |
|               | -6                         | -9       | -12      | -6                         | -9       | -12      |
| 64            | 0.200000                   | 0.343750 | 0.491250 | 0.110625                   | 0.232500 | 0.377500 |
| 256           | 0.005000                   | 0.076250 | 0.258750 | 0.001250                   | 0.021250 | 0.106250 |
| 512           | 0.000000                   | 0.003750 | 0.097500 | 0.000000                   | 0.000000 | 0.024375 |

Table 2: Cross-Validation Error Rates for  $f^A(x) = \cos(x)$  vs.  $f^B(x) = \cos(x) + (\frac{3}{2}\sin(5x))^3$ .

| Sample Points | Training Set Size = 4      |          |          | Training Set Size = 8      |          |          |
|---------------|----------------------------|----------|----------|----------------------------|----------|----------|
|               | Signal-to-Noise Ratio (dB) |          |          | Signal-to-Noise Ratio (dB) |          |          |
|               | -6                         | -9       | -12      | -6                         | -9       | -12      |
| 64            | 0.236250                   | 0.337500 | 0.422500 | 0.100625                   | 0.200625 | 0.297500 |
| 256           | 0.018750                   | 0.118750 | 0.293750 | 0.003125                   | 0.043750 | 0.130000 |
| 512           | 0.000000                   | 0.002500 | 0.083750 | 0.000000                   | 0.000000 | 0.030625 |

available, or at least at a sampling rate high enough to recover all of the pertinent features in the signals.

Excellent results were obtained when 512 sample points were collected, regardless of the noise level or the population size. The APER for these cases ranged from 0% to less than 10%.

For our second simulation experiment, we let  $f^A(x) = \cos(x)$  and  $f^B(x) = \cos(x) + (\frac{3}{4}(\sin(5x)))^3$ . Figure 3 shows  $f^A(x)$  and  $f^B(x)$  superimposed with no added noise. The signals have similar low-frequency behavior, but different higher-frequency behavior. Table 2 summarizes the simulation results. The size of the training set, the number of sample points, and the noise level have the same effect on the APER as before. The highest sampling rate gives the best results: the APER for these cases

ranged from 0% to approximately 8%. A high sampling rate is advantageous due to the nature of the frequencies comprising  $f^B(x)$ .

We also tested the discriminant on other pairs of functions. Simulation results were virtually the same as those shown in Table 1 and Table 2. For the sake of brevity, we have omitted these results.

The simulation results in Table 1 and Table 2 are based on the linear discriminant. Very similar error rates were observed for the quadratic discriminant, which are therefore not reported.

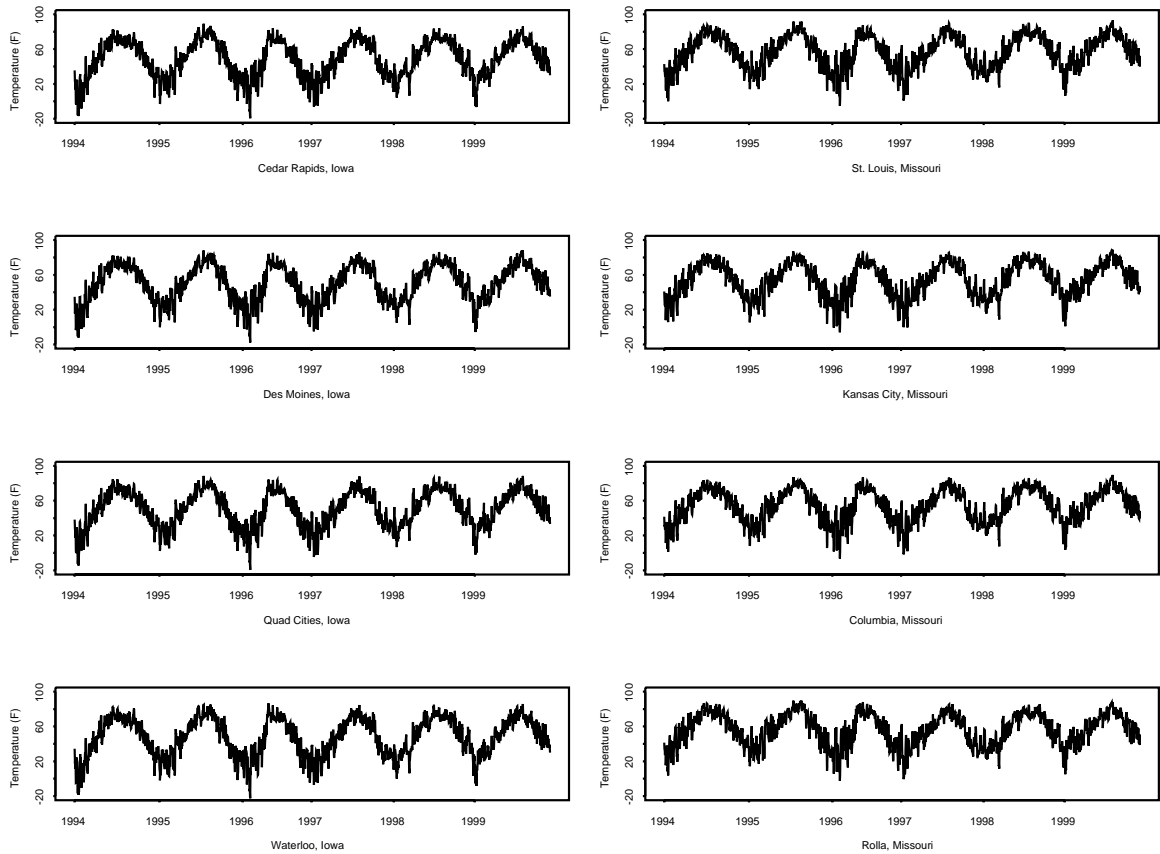


Figure 4: Daily mean temperatures for eight Midwestern U.S. cities

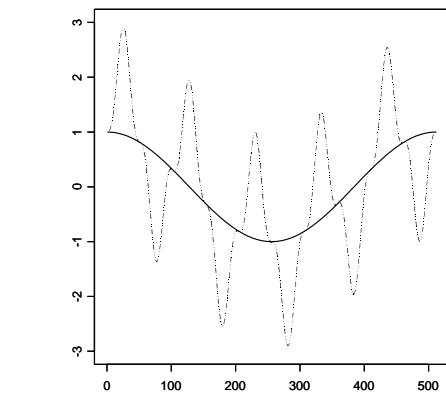


Figure 3:  $f^A(x)$  and  $f^B(x)$  for the second simulation experiment.

#### 4. Application

Climatologists are often interested in determining regions which comprise a climate regime. However, once such regions are defined, the classification of areas which lie on the geographic boundary between regimes is problematic. We seek an objective procedure to facilitate such classification. The discrimi-

nation method developed in this paper could be used for this purpose. The relevant regimes could be regarded as the populations, the mean temperature pattern over a regime could be considered a signal, and the temperature pattern for the boundary location could be viewed as the corrupted signal to be classified.

We evaluated our method by applying it to data obtained from the National Oceanic & Atmospheric Administration’s (NOAA) National Climatic Data Center. Daily mean temperatures were collected from eight different cities in the Midwestern United States – four from Missouri and four from Iowa. The cities from each region cover roughly the same longitudes, but the cities in Iowa are approximately 250 miles north of the cities in Missouri. Thus, the competing populations are the “Missouri area climate regime” and the “Iowa area climate regime”. The “signals” in this setting are the daily mean temperatures for the two regimes over a five year period. The temperature series for the eight cities are shown in Figure 4.

As a preliminary performance check of the discriminant, the holdout cross-validation procedure

was conducted for the eight cities comprising the training samples. Due to the nature of the signals, the linear discriminant was used. All of the cities were correctly classified as either belonging to Missouri or Iowa.

We were also interested in classifying a city that is located between the two groups of cities. Kirksville, Missouri was chosen as our test city. Since Kirksville is only 20 miles south of the Iowa border, it was an ideal choice. The Kirksville temperature series is shown in Figure 5. We applied our discriminant to this series. The temperature pattern of Kirksville between 1994 and 1999 was matched to the pattern of the Missouri area regime.

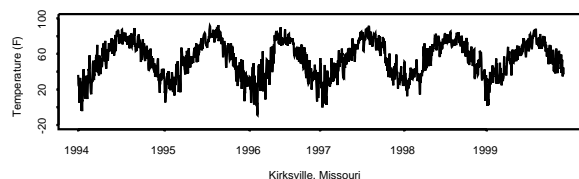


Figure 5: Daily mean temperatures for unclassified city.

## 5. Conclusion

We have developed an approach for discriminating time series originating from signal-plus-noise processes. Although traditional linear and quadratic discriminants form the bases of our method, the use of wavelet coefficients as well as level discriminants makes the method novel. Simulations demonstrate the procedure performs well under adverse noise levels and with as few as four realizations per population as training samples. Applying the method to temperature data also gave favorable results.

## 6. Bibliography

Bruce, A.G. and Gao, H.-Y. (1996). *Applied Wavelet Analysis with S-PLUS*. New York: Springer.

Johnson, R.A. and Wichern, D. (1998). *Applied Multivariate Analysis (4th Ed)*. Englewood Cliffs, N.J.: Prentice-Hall.

Ogden, R.T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Boston: Birkhäuser.

Percival, D.B. and Walden, A.T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge: Cambridge University Press.

Shumway, R.H. (1982). Discriminant Analysis for Time Series. In *Handbook of Statistics, Vol. 2, Pattern Recognition and Reduction of Dimensionality*, ed. P.R. Krishnaiah, 1-43. Amsterdam: North-Holland.

Zyweck, A. and Bogner, R. (1996). Radar Target Classification of Commercial Aircraft. *IEEE Transactions in Aerospace and Electronic Systems*, **32**, 598-606.