Using the Wavelet Transform for Multivariate Data Analysis and Time Series Forecasting

Fionn Murtagh (1) and Alex Aussem (2)

(1) University of Ulster, Faculty of Informatics, Londonderry BT48 7JL, Northern Ireland. Email fd.murtagh@ulst.ac.uk
(2) Université René Descartes, UFR de Mathématiques et Informatique, 45, rue des Saints-Pères, 75006 Paris, France. Email alex@math-info.univ-paris5.fr

Abstract: We discuss the use of orthogonal wavelet transforms in multivariate data analysis methods such as clustering and dimensionality reduction. Wavelet transforms allow us to introduce multiresolution approximation, and multiscale nonparametric regression or smoothing, in a natural and integrated way into the data analysis. Applications illustrate the powerfulness of this new perspective on data analysis.

Keywords: Wavelet Transform, Additive Decomposition, Partitioning, Clustering, Dimensionality Reduction, Data Display, Exploratory Data Analysis, Principal Components Analysis, Preliminary Data Transformation

1 Introduction

Data analysis, for exploratory purposes, or prediction, is usually preceded by various data transformations and recoding. In fact, we would hazard a guess that 90% of the work involved in analyzing data lies in this initial stage of data preprocessing. This includes: problem demarcation and data capture; selecting non-missing data of fairly homogeneous quality; data coding; and a range of preliminary data transformations.

The wavelet transform offers a particularly appealing data transformation, as a preliminary to data analysis. It offers additionally the possibility of close integration into the analysis procedure as will be seen in this article. The wavelet transform may be used to "open up" the data to de-noising, smoothing, etc., in a natural and integrated way.

2 Some Perspectives on the Wavelet Transform

We can think of our input data as a time-varying signal, e.g. a time series. If discretely sampled (as will almost always be the case in practice), this amounts to considering an input vector of values. The input data may be sampled at discrete wavelength values, yielding a spectrum, or one-dimensional image. A two-dimensional, or more complicated input image, can be fed to the analysis engine as a rasterized data stream. Analysis of such a two-dimensional image may be carried out independently on each dimension, but such an implementation issue will not be of further concern to us here. Even though our motivation arises from the analysis of ordered input data vectors, we will see below that we have no difficulty in using exactly the same approach with (more common) unordered input data vectors.

Wavelets can be introduced in different ways. One point of view on the wavelet transform is by means of filter banks. The filtering of the input signal is some transformation of it, e.g. a low-pass filter, or convolution with a smoothing function. Low-pass and high-pass filters are both considered in the wavelet transform, and their complementary use provides signal analysis and synthesis.

3 The Wavelet Transform Used

The following discussion is based on Strang (1989), Bhatia et al. (1995) and Strang and Nguyen (1996). Our task is to consider the approximation of a vector x at finer and finer scales. The finest scale provides the original data, $x_N = x$, and the approximation at scale m is x_m where usually $m = 2^0, 2^1, \ldots 2^N$. The incremental detail added in going from x_m to x_{m+1}, the detail signal, is yielded by the wavelet transform. If ξ_m is this detail signal, then the following holds:

$$x_{m+1} = H^T(m)x_m + G^T(m)\xi_m \tag{1}$$

where $G(m)$ and $H(m)$ are matrices (linear transformations) depending on the wavelet chosen, and T denotes transpose (adjoint). An intermediate approximation of the original signal is immediately possible by setting detail components $\xi_{m'}$ to zero for $m' \geq m$ (thus, for example, to obtain x_2, we use only x_0, ξ_0 and ξ_1). Alternatively we can de-noise the detail signals before reconstituting x and this has been termed wavelet regression (Bruce and Gao, 1994).

Define ξ as the row-wise juxtaposition of all detail components, $\{\xi_m\}$, and the final smoothed signal, x_0, and consider the wavelet transform W given by

$$Wx = \xi = [\xi_{N-1} \ldots \xi_0 x_0]^T \tag{2}$$

The right-hand side is a concatenation of vectors. Taking $W^T W = I$ (the identity matrix) is a strong condition for exact reconstruction of the input data, and is satisfied by an orthogonal wavelet transform. The important fact that $W^T W = I$ will be used below in our enhancement of multivariate data analysis methods. This permits use of the "prism" (or decomposition in terms of scale and location) of the wavelet transform.

Examples of these orthogonal wavelets, i.e. the operators G and H, are the Daubechies family, and the Haar wavelet transform (Press et al., 1992; Daubechies, 1992). For the Daubechies D_4 wavelet transform, H is given by

$$(0.4829629131, 0.8365163037, 0.2241438680, -0.1294095226)$$

and G is given by

$$(-0.1294095226, -0.2241438680, 0.8365163037, -0.4829629131).$$

Implementation is by decimating the signal by two at each level and convolving with G and H: therefore the number of operations is proportional to $n + n/2 + n/4 + \ldots = O(n)$. Wrap-around (or "mirroring") is used by the convolution at the extremities of the signal.

4 Wavelet-Based Multivariate Data Analysis: Basis

We consider the wavelet transform of x, Wx. Consider two vectors, x and y. The squared Euclidean distance between these is $\|x - y\|^2 = (x - y)^T(x - y)$. The squared Euclidean distance between the wavelet transformed vectors is $\|Wx - Wy\|^2 = W^T W(x - y)^T(x - y)$, and hence identical to the distance squared between the original vectors. For use of the Euclidean distance, the wavelet transform can replace the original data in the data analysis. The analysis can be carried out in wavelet space rather than direct space. This in turn allows us to directly manipulate the wavelet transform values, using any of the approaches found useful in other areas. The results based on the orthogonal wavelet transform exclusively imply use of the Euclidean metric, which nonetheless covers a considerable area of current data analysis practice.

Note that the wavelet basis is an orthogonal one, but is not a principal axis one (which is orthogonal, but also optimal in terms of least squares projections). Wickerhauser (1994) proposed a method to find an approximate principal component basis by determining a large number of (efficiently-calculated) wavelet bases, and keeping the one closest to the desired Karhunen-Loève basis. If we keep, say, an approximate representation allowing reconstitution of the original n components by n' components (due to the dyadic analysis, $n' \in \{n/2, n/4, \ldots\}$), then we see that the space spanned by these n' components will not be the same as that spanned by the n' first principal components.

5 Wavelet Filtering or Wavelet Regression

Foremost among modifications of the wavelet transform coefficients is to approximate the data, progressing from coarse representation to fine representation, but stopping at some resolution level m. As noted above, this implies setting wavelet coefficients $\xi_{m'}$ to zero when $m' \geq m$.

Filtering or non-linear regression of the data can be carried out by deleting insignificant wavelet coefficients at each resolution level (noise filtering), or by "shrinking" them (data smoothing). Reconstitution of the data then provides a cleaned data set. A practical overview of such approaches to data filtering (arising from work by Donoho and Johnstone at Stanford University) can be found in Bruce and Gao (1994, chapter 7). For other model-based work see Starck et al. (1995).

6 Examples of Multivariate Data Analysis in Wavelet Space

We used a set of 45 astronomical spectra. These were of the complex AGN (active galactic nucleus) object, NGC 4151, and were taken with the small but very successful

IUE (International Ultraviolet Explorer) satellite which was still active in 1996 after nearly two decades of operation. We chose a set of 45 spectra observed with the SWP spectral camera, with wavelengths from 1191.2 Å to approximately 1794.4 Å, with values at 512 interval steps. There were some minor discrepancies in the wavelength values, which we discounted: an alternative would have been to interpolate flux values (vertical axis, y) in order to have values at identical wavelength values (horizontal axis, x), but we did not do this since the infrequent discrepancies were fractional parts of the most common regular interval widths. Fig. 1 shows a sample of 20 of these spectra. A wavelet transform (Daubechies 4 wavelet used) version of these spectra was generated, with a number of scales generated which was allowed by dyadic decomposition. An overall $0.1\ \sigma$ (standard deviation, calculated on all wavelet coefficients) was used as a threshold, and coefficient values below this were set to zero. Spectra which were apparently more noisy had relatively few coefficient values set to zero, e.g. 31%. More smooth spectra had up to over 91% of their coefficients set to zero. On average, 76% of the wavelet coefficients were zeroed in this way. Fig. 2 shows the relatively high quality spectra re-formed, following zeroing of wavelet coefficient values.

The Kohonen "self-organizing feature map" (SOFM; Murtagh and Hernández-Pajares, 1995) was applied to this data. A 5×6 output representational grid was used. In wavelet space or in direct space, the assignment results obtained were identical. With 76% of the wavelet coefficients zeroed, the result was very similar, indicating that redundant information had been successfully removed. This approach to SOFM construction leads to the following possibilities:

1. Efficient implementation: a good approximation can be obtained by zeroing most wavelet coefficients, which opens the way to more appropriate storage (e.g. offsets of non-zero values) and distance calculations (e.g. implementation loops driven by the stored non-zero values). Similarly, compression of large datasets can be carried out. Finally, calculations in a high-dimensional space, $I\!R^m$, can be carried out more efficiently since, as seen above, the number of non-zero coefficients may well be $m^{''} << m$ with very little loss of useful information.

2. Data "cleaning" or filtering is a much more integral part of the data analysis processing. If a noise model is available for the input data, then the data can be de-noised at multiple scales. By suppressing wavelet coefficients at certain scales, high-frequency (perhaps stochastic or instrumental noise) or low-frequency (perhaps "background") information can be removed. Part of the data coding phase, prior to the analysis phase, can be dealt with more naturally in this new integrated approach.

A number of runs of the k-means partitioning algorithm were made. The exchange method, described in Späth (1985) was used. Four, or two, clusters were requested. Identical results were obtained for both data sets, which is not surprising given that this partitioning method is based on the Euclidean distance. For the 4-cluster, and 2-cluster, solutions we obtained respectively these assignments:

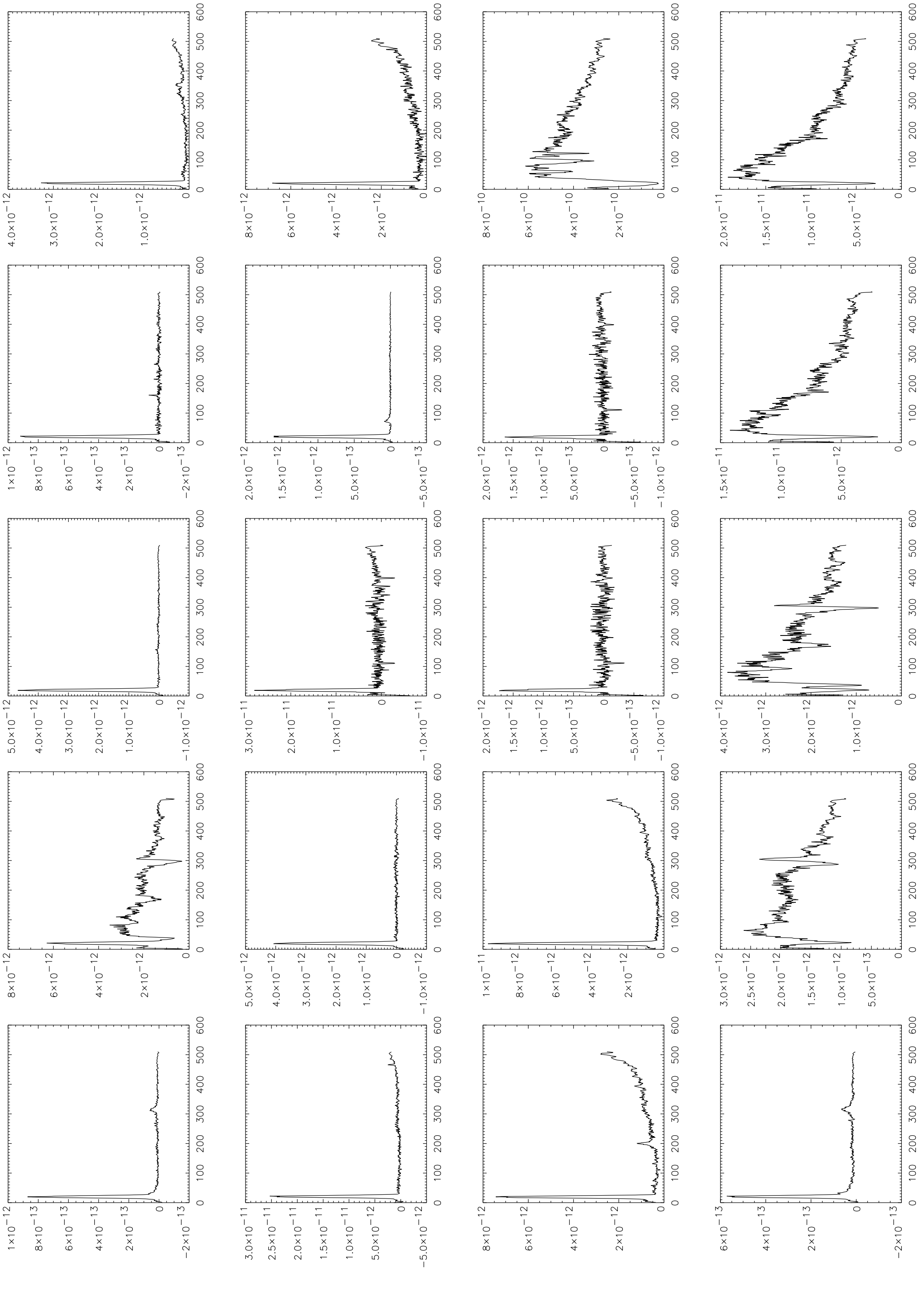

Figure 1: Sample of 20 spectra (from 45 used) with original flux measurements plotted on the y-axis.

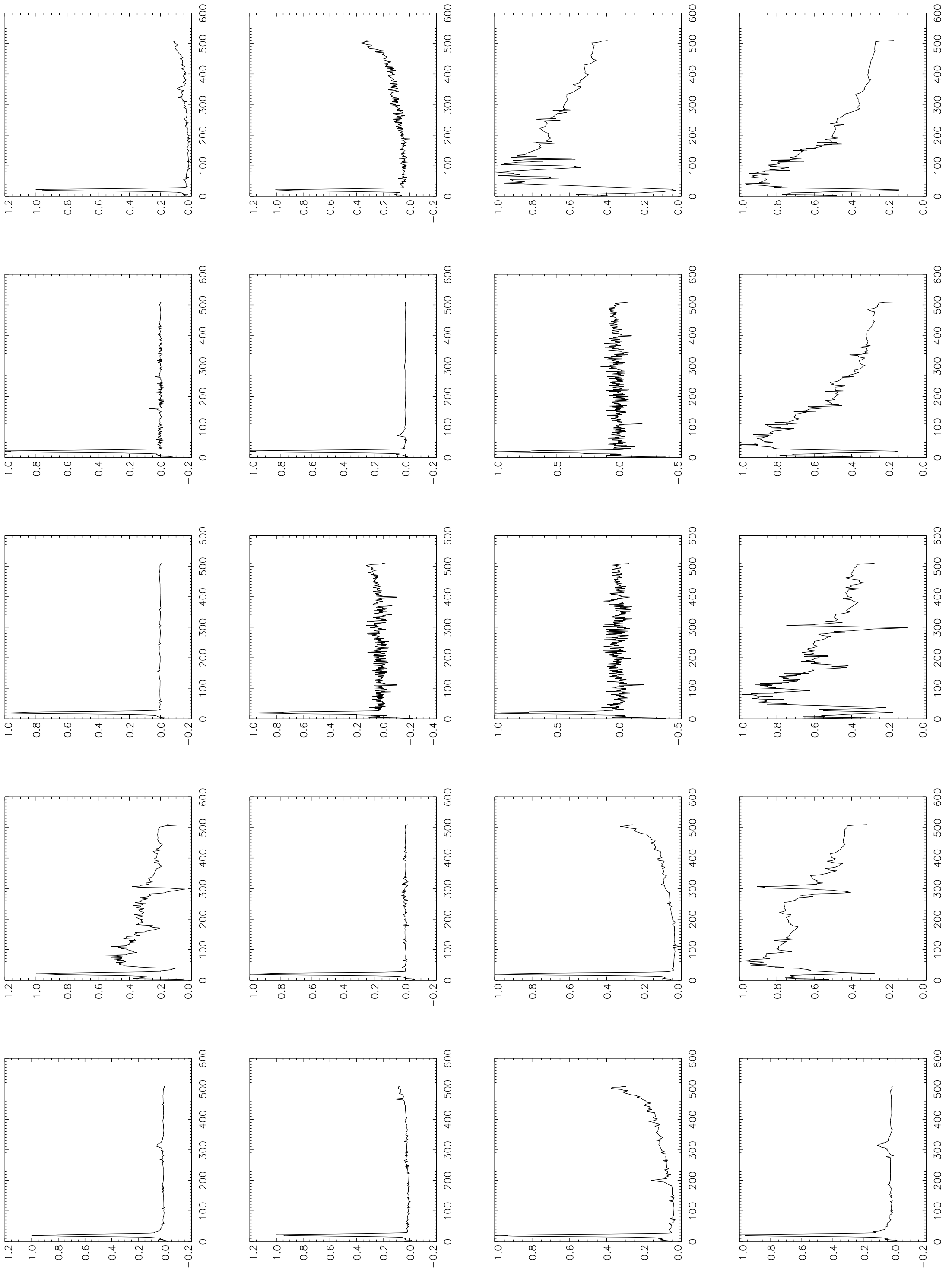

Figure 2: Sample of 20 spectra (as in previous Fig.), each normalized to unit maximum value, then wavelet transformed, approximately 75% of wavelet coefficients set to zero, and reconstituted.

```
123213114441114311343133141121412222222121114

122211111111111111111111111121112222222121111
```

The case of principal components analysis was very interesting. We know that the basic PCA method uses Euclidean scalar products to define the new set of axes. Often PCA is used on a variance-covariance input matrix (i.e. the input vectors are centered); or on a correlation input matrix (i.e. the input vectors are rescaled to zero mean and unit variance). These two transformations destroy the Euclidean metric properties vis-à-vis the raw data. Therefore we used PCA on the unprocessed input data. We obtained identical eigenvalues and eigenvectors for the two input data sets.

The eigenvalues are similar up to numerical precision:

```
1911.217163  210.355377   92.042099   13.908587    7.481989
   2.722113    2.304520

1911.220703  210.355392   92.042336   13.908703    7.481917
   2.722145    2.304524
```

The eigenvectors are similarly identical. The actual projection values are entirely different. This is simply due to the fact that the principal components in wavelet space are themselves inverse-transformable to provide principal components of the initial data.

Various aspects of this relationship between original and wavelet space remain to be investigated. We have argued for the importance of this, in the framework of data coding and preliminary processing. We have also noted that if most values can be set to zero with limited (and maybe beneficial) effect, then there is considerable scope for computational gain also. The processing of sparse data can be based on an "inverted file" data-structure which maps non-zero data entries to their values. The inverted file data-structure is then used to drive the distance and other calculations. Murtagh (1985, pp. 51–54 in particular) discusses various algorithms of this sort.

7 An Isotropic Redundant Wavelet Transform

It is common in pattern recognition to speak of "features" when what is intended are small density perturbations in feature space, small glitches in time series, etc. Such "features" may include sharp (edge-like) phenomena which can be demarcated using wavelet transforms like the orthogonal ones described above. Sometimes the glitches which are of interest are symmetric or isotropic. If so, a symmetric wavelet may be more useful. The danger with an asymmetric wavelet is that the wavelet itself may impose artifacts.

The "à trous" (with holes) algorithm is such an isotropic wavelet transform. It does not have the orthogonality property of the transform described earlier. The French term is commonly used, and arises from an interlaced convolution which is used instead of the usual convolution (see Shensa, 1992; Holschneider et al., 1989; see also Starck and

Bijaoui, 1994; and Bijaoui et al., 1994). The algorithm can be described as follows: (i) smoothing p times with a B_3 spline – hence Gaussian-like, but of compact support; (ii) the wavelet coefficients are given by the differences between successive smoothed versions of the signal. The latter provide the detail signal, which (we hope) in practice will capture small "features" of interpretational value in the data. The following attractive additive decomposition of the data follows immediately from the design of the above scheme:

$$c_0(k) = c_p(k) + \sum_{i=1}^{p} w_i(k) \tag{3}$$

The set of values provided by c_p provide a "residual" or "continuum" or "background". Adding w_i values to this, for $i = p, p-1, \ldots$ gives increasingly more accurate approximations of the original signal. Note that no decimation is carried out here, which implies that the size or dimension of w_i is the same as that of c_0. This may be convenient in practice: cf. next section. It is readily seen that the computational complexity of the above algorithm is $O(n)$ for an n-valued input, and the storage complexity is $O(n^2)$.

8 Wavelet-Based Forecasting

In experiments carried out on the sunspots benchmark dataset (yearly averages from 1720 to 1979, with forecasts carried out on the period 1921 to 1979: see, e.g., Tong, 1990), a wavelet transform was used for values k up to a time-point k_0. One-step-ahead forecasts were carried out independently at each w_i. These were summed to produce the overall forecast (cf. the additive decomposition of the original data, provided by the wavelet transform). An interesting variant on this was also investigated: this variant was that there was no need to use the same forecasting method at each level, i. We ran autoregressive, multilayer perceptron and recurrent connectionist networks in parallel, and kept the best results indicated by a cross-validation on withheld data at that level. We found the overall result to be superior to working with the original data alone, or with one forecasting engine alone. Details of this work can be found in Aussem and Murtagh (1996).

9 Conclusion

The results described here, from the multivariate data analysis perspective, are very exciting. They not only open up the possibility of computational advances but also provide a new approach in the area of data coding and preliminary processing.

The chief advantage of these wavelet methods is that they provide a multiscale decomposition of the data, which can be directly used by multivariate data analysis methods, or which can be complementary to them.

A major element of this work is to show the practical relevance of doing this. It has been the aim of this paper to do precisely this in a few cases. Finding a symbiosis between what are, at first sight, methods with quite different bases and quite different objectives, requires new insights. Wedding the wavelet transform to multivariate data analysis no doubt leaves many further avenues to be explored.

Further details of the experimentation described in this paper, details of code used, and further information, can be found in Murtagh (1996).

References

AUSSEM, A. and MURTAGH, F. (1996), "Combining neural network forecasts on wavelet-transformed time series", *Connection Science*, submitted.

BHATIA, M., KARL, W.C. and WILLSKY, A.S. (1995), "A wavelet-based method for multiscale tomographic reconstruction", *IEEE Transactions on Medical Imaging*, submitted, MIT Technical Report LIDS-P-2182.

BIJAOUI, A., STARCK, J.-L. and MURTAGH, F. (1994), "Restauration des images multi-échelles par l'algorithme à trous", *Traitement du Signal*, 11, 229–243.

BRUCE, A. and GAO, H.-Y. (1994), *S+Wavelets User's Manual*, Version 1.0, Seattle, WA: StatSci Division, MathSoft Inc.

DAUBECHIES. I. (1992), *Ten Lectures on Wavelets*, Philadelphia: SIAM.

HOLSCHNEIDER, M., KRONLAND-MARTINET, R., MORLET, J. and TCHAMITCHIAN, Ph. (1989), "A real-time algorithm for signal analysis with the help of the wavelet transform", in J.M. Combes, A. Grossmann and Ph. Tchamitchian (eds.), *Wavelets: Time-Frequency Methods and Phase Space*, Berlin: Springer-Verlag, 286–297.

MURTAGH, F. (1985), *Clustering Algorithms*, Würzburg: Physica-Verlag.

MURTAGH, F. and HERNÁNDEZ-PAJARES, M. (1995), "The Kohonen self-organizing feature map method: an assessment", *Journal of Classification*, 12, 165–190.

MURTAGH, F. (1996), "Wedding the wavelet transform and multivariate data analysis", *Journal of Classification*, submitted.

PRESS, W.H., TEUKOLSKY, S.A., VETTERLING, W.T. and FLANNERY, B.P. (1992), *Numerical Recipes*, 2nd ed., Chapter 13, New York: Cambridge University Press.

SHENSA, M.J. (1992), "The discrete wavelet transform: wedding the à trous and Mallat algorithms", *IEEE Transactions on Signal Processing*, 40, 2464–2482.

SPÄTH, H. (1985), *Cluster Dissection and Analysis*, Chichester: Ellis Horwood.

STARCK, J.-L. and BIJAOUI, A. (1994), "Filtering and deconvolution by the wavelet transform", *Signal Processing*, 35, 195–211.

STARCK, J.-L., BIJAOUI, A. and MURTAGH, F. (1995), "Multiresolution support applied to image filtering and deconvolution", *Graphical Models and Image Processing*, 57, 420–431.

STRANG, G. (1989), "Wavelets and dilation equations: a brief introduction", *SIAM Review*, 31, 614–627.

STRANG, G. and NGUYEN, T. (1996), *Wavelets and Filter Banks*, Wellesley, MA: Wellesley-Cambridge Press.

TONG, H. (1990), *Non Linear Time Series*, Oxford: Clarendon Press.

WICKERHAUSER, M.V. (1994), *Adapted Wavelet Analysis from Theory to Practice*, Wellesley, MA: A.K. Peters.