13th World Conference on Earthquake Engineering
Vancouver, B.C., Canada
August 1-6, 2004
Paper No. 2954

REPRESENTATION AND COMPRESSION OF STRUCTURAL VIBRATION MONITORING DATA USING WAVELETS AS A TOOL IN DATA MINING

Maria I. TODOROVSKA[1] and Tzong-Ying HAO[2]

SUMMARY

This paper explores the advantages of expansion in orthonormal wavelet bases—as a preprocessing tool—in analysis of large sets of seismic vibration monitoring data, of ground or structural response, for possible application in data mining. The focus is on the insight that can be gained from the wavelet domain representation, convenience in estimation of energy and correlation, efficiency of representation (data compression property), and dimensionality reduction. Local and global aggregates and average distributions of energy and related quantities (e.g. power, power spectrum density, Fourier amplitude, cross-energy, cross-power, cross power spectrum density, etc.), computed directly in the wavelet domain, are discussed and interpreted as information granules, representative of a frequency interval, of a partition of the phase plane, or of the entire record. The concepts explored are illustrated on a mini database of strong motion records from the 1994 Northridge in a 7-story reinforced concrete building located in the Los Angeles area. Nodal time-frequency distributions of power spectrum density are shown for the ground floor and roof responses. Dimensionality reduction by thresholding is illustrated and compared with sub-sampling. The errors associated with compression are illustrates for a small database of ground response records from six earthquakes recorded in the same building. The results show that the error is very small even for high compression ratios. It is concluded that expansion in orthonormal wavelet series is potentially a very useful preprocessing tool in mining large data sets of ground and structural response vibration data under earthquake excitation. One drawback of the orthonormal wavelet transform is poor resolution at high frequencies, which can be eliminated by using expansion in orthonormal wavelet packets, to which most of the presented concepts are directly applicable. The theory and concepts presented apply directly to datasets of any time series data.

INTRODUCTION

Since first emerged as a consistent theory in the 1980s from the work of French geophysicists Morlet and Grossman (Vetterli and Kovacević [1], Daubechies [2]), wavelet analysis has become a very popular tool for analysis of signals and images in many fields of science and engineering. The wavelet transform is particularly suitable for analysis of transient signals and of time varying systems, because it is localized

[1] Research Associate Professor, Univ. of Southern California, Dept. of Civil and Environmental Eng., Los Angeles, CA 90089-2531
[2] Research Associate, Univ. of Southern California, Dept. of Civil and Environmental Eng., Los Angeles, CA 90089-2531

both in time and frequency. Its widespread use is also due to the existence of orthogonal and bi-orthogonal wavelet bases, the efficiency of representing transient signals in such bases (data compression), and the existence of fast and accurate computational algorithms for signal/image transformation and reconstruction (asymptotically even faster than the fast Fourier transform). The use of wavelets in analyses of mechanical vibrations was first introduced by Newland [3], who proposed the use of wavelet maps as a diagnostics tool for time-varying systems, in particular for detecting and localizing in time hidden details and small perturbations that are practically invisible in the time representation, as well as in providing insight into local correlation of two signals.

The recent advances in sensor, computer, and communication technologies have enabled and encouraged collection of large volumes of scientific data that needs to be efficiently stored, managed and analyzed. These needs have stimulated many advances in the fields of data engineering and digital signal and image processing. In strong motion seismology, the volume of data is increasing rapidly not as much due to the number of new sensors deployed as due to the increasing sensor sensitivity and recorder dynamic range (currently approaching 26 bits or 156 dB), which makes it possible to record, with strong motion instruments, ground and structural response to very small and distant earthquakes as well as to ambient noise (Trifunac and Todorovska [4]). The increasing capability of recording, combined with the dramatic reduction of the cost of digital storage media, lead to lowering the triggering level of recording and even to selective continuous recording. As a part of the Advanced National Seismic System (ANSS) initiative launched by the U.S. Geological Survey (Benz et al. [5]), a large number of instruments will be installed in structures as well, which will further increase the number of recordings.

Seismic and other mechanical vibration data is normally archived by storing the waveform data in the time domain. However, retrieval of such time series data that is of high dimension (i.e. consisting of many data points) from permanent archives on a remote server is slow, especially when, for a specific application, lowering the dimensionality of the data (e.g. by reducing the resolution of the representation, by data compression, or by representing values in intervals by local averages) is permissible or even desirable (e.g. in pattern recognition). A novel approach to store time series or spatial data would be to expand in wavelet series and archive the coefficients of the expansion. This paper summarizes an exploratory analysis of the advantages of wavelet bases representation of strong motion data for mining of large data sets, published in Todorovska and Hao [6].

THEORETICAL BACKGROUND

Wavelets, Wavelet Families, Bases and Orthonormal Wavelet Transform

According to the modern wavelet theory, a wavelet, $\psi(t)$, is a real or complex *zero mean* wiggle on the real line that is localized both in time and in frequency. For the zero mean condition (also called *admissibility* condition) to be satisfied, it must be oscillatory—hence the name wavelet (Vetterli and Kovacević [1], Daubechies [2]). By elementary operations consisting of shifts in time, *b*, and dilation or contraction, a family of wavelets

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \quad a>0,\ b\in \quad ,\ t\in \tag{1}$$

can be constructed from a prototype wavelet, $\psi(t)\in L_2(R)$ (called the "mother wavelet"). In eqn (1), *a* is the *scale* variable, such that $a>1$ corresponds to dilation, and $a<1$ corresponds to contraction. The normalizing constant $1/\sqrt{a}$ is such that all the wavelets in the family have same L_2 norm, usually set to unity. The localization of a wavelet in a family is described by their central time and central frequency, and the spread in time and in frequency. If the mother wavelet is centered at time t_0 and at frequency

ω_0, then $\psi_{a,b}(t)$ is centered at time $\bar{t}(b) = t_0 + b$ and at frequency $\bar{\omega}(a) = \omega_0 / a$. Also, if the mother wavelet has spreads in time and frequency $\sigma_{t,0}$ and $\sigma_{\omega,0}$, the corresponding spreads of $\psi_{a,b}(t)$ are $\sigma_t(a) = a\sigma_{t,0}$ and $\sigma_\omega(a) = \sigma_{\omega,0} / a$, which implies that the larger scale (more dilated) wavelets have larger spread in time but smaller spread in frequency, while the *relative spread* σ_ω / ω is constant. The area of localization in the phase plane (time-frequency plane), $\sigma_t(a)\sigma_\omega(a)$, is also constant for the family, and cannot be made arbitrarily small, as described by the Heisenberg uncertainty principle (Vetterli and Kovacević [1]).

Of interest in this study are wavelet families such that the scale and time are sampled on a discrete grid

$$a = a_0^m, \quad b = nb_0a_0^m, \qquad m,n \in \quad , \; a_0 \neq 1, \; b_0 \neq 0 \tag{2}$$

where is the set of integers. Then the corresponding grid of central times and frequencies of the family is such that the spacing in frequency is larger at higher frequencies, but is constant on a *logarithmic* frequency axis. The projection of a signal $s(t)$ on such a family is by definition the discrete wavelet transform (DWT) of the signal with respect to that family

$$W_s(m,n) = \left\langle \psi_{m,n}, s \right\rangle_{L_2}, \quad m,n \in \quad , \; a_0 \neq 1, \; b_0 \neq 0 \tag{3}$$

The DWT is a *time-scale* distribution. It is also a *time-frequency* distribution, as scale and frequency are closely related and can be used interchangeably. The frequency corresponding to scale *a* is the central frequency of the wavelet with scale a

$$\omega = \omega_0 / a \tag{4}$$

Most widespread is DWT with $a_0 = 2$ and $b_0 = 1$, referred to as *dyadic*. If the grid is sufficiently dense, the signal can be reconstructed from its DWT. For wavelet families that are *orthonormal*, i.e. $\left\langle \psi_{m,n}, \psi_{m',n'} \right\rangle = \delta_{m=m'} \, \delta_{n=n'}$, the reconstruction formula is

$$\begin{aligned} s(t) &= \sum_m \sum_n \left\langle \psi_{m,n}, s \right\rangle \; \psi_{m,n}(t) \\ &= \sum_m \sum_n W(m,n) \, \psi_{m,n}(t) \end{aligned} \tag{5}$$

Two extreme examples of orthonormal wavelet bases are the Haar basis, such that the prototype wavelet in the time domain is a rectangular up and down pulse, and the sinc basis, such that the prototype wavelet is a rectangular pulse in the Fourier domain (Vetterli and Kovacević [1]). The former basis has finite support in the time domain but infinite support in the frequency domain, while the latter has finite support in the frequency domain, but infinite support in the time domain.

Of particular interest for applications are orthonormal wavelet families that have compact support, but are more "regular" (smoother) than the Haar basis. Such series of families were constructed for the first time by Daubechies in the late 1980s (Daubechies [2]), and are called *doublets* (named after Daubechies), *symlets* (named for their higher degree of symmetry compared to the *daublets*), and *coiflets* (named in honor of R. Coifman). These families are of interest because the analysis and synthesis of a discretely sampled signal can be achieved by using finite impulse response (FIR) filters, using an algorithm—the *pyramid algorithm* (due to Mallat; Vetterli and Kovacević [1]), which for large *N* (the length of the signal) has complexity O(*N*), and hence is faster even than the Fast Fourier Transform. The forward algorithm computes the discrete wavelet transform of the signal (i.e. the coefficients of the expansion in wavelet

series), while the inverse algorithm computes the signal from the coefficients. This algorithm is based on the framework of mutiresolution analysis (postulated by Mallat) and is implemented by filter banks.

Multiresolution Analysis, Subband Decomposition and Their Relation to Wavelet Basis Expansion

The general framework for construction of wavelet bases—multiresolution analysis—was developed in 1986 by Mallat and Meyer (Daubechies [2]), who saw the connection between wavelet analysis and subband theory. This section explains briefly its main features.

Multiresolution analysis consists of a sequence of embedded closed *approximation* subspaces

$$\cdots V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \cdots \tag{6}$$

such that their union covers $L_2(\ )$ and their intersection is the empty set, and they are *scaled versions of each other*. The projections of a signal onto these spaces represent approximations at different resolutions, with larger m corresponding to lower resolution approximation. Let W_m be the orthogonal complement of approximation subspace V_m in approximation subspace V_{m-1}. Then, V_{m-1} is a direct sum of V_m and W_m, and the projection onto W_m contains the detail that has been removed from V_{m-1} to create the lower resolution approximation in V_m. Hence, the subspaces W_m, $m \in \mathbf{Z}$ are called *detail* subspaces. In contrast to the approximation subspaces V_m, $m \in \mathbf{Z}$, which are embedded, subspaces W_m, $m \in \mathbf{Z}$ are disjoint and hence orthogonal to each other, and their union covers $L_2(\ )$. A theorem guarantees the existence of orthonormal bases for each of the approximation and detail subspaces. The basis functions for the detail spaces are wavelets (i.e. zero mean functions), while those for the approximation spaces are scaling function, which are not zero mean, and are, like the wavelet bases, scaled and shifted versions of a prototype scaling function. Let $\{\psi_{m,n}(t)\}_{n \in \mathbf{Z}}$ be a basis for W_m, and $\{\varphi_{m,n}(t)\}_{n \in \mathbf{Z}}$ be a basis for V_m.

Subband theory is an area of digital signal processing, which is based on the design of a set of prototype filters such that divide the signal frequency band in equal parts (subbands). For example, a two-channel filter bank consists of a high-pass and low-pass filter. These filters are applied recursively to the signal, which leads to its division in subbands. The filters applied at each stage have identical properties except for scale. The set of filters is referred to as *filter bank*. The prototype filters are used to construct the wavelet and the scaling function. Hence, subband decomposition becomes equivalent to basis decomposition for discrete time signals. The following example shows how multiresolution approximations of a discrete signal can be obtained using subband decomposition.

All discrete time signals are band-limited, and have maximum frequency $f_{\max} = 1/(2\Delta t)$ Hz (where Δt is the sampling period in seconds), called the Nyquist frequency. This frequency corresponds to circular frequency $\omega_{\max} = \pi$ radians *per sample*, which is the circular frequency in radians per second if $\Delta t = 1$ s. Hence, all discrete signals belong to the space of band-limited functions on $\omega \in [-\pi, \pi]$, which we will call V_0. Let the prototype filters be ideal low- and high-pass filters, as shown in Fig. 1 (top), with impulse responses respectively $h_0[n]$ and $h_1[n]$, and with Fourier transforms $H_0(\omega)$ and $H_1(\omega)$. Filters $H_0(\omega)$ and $H_1(\omega)$ split the signal into two components, S_1 and D_1, such that the former is a smooth approximation of the signal and the latter is the remainder, containing the detail. These two components are respectively band-limited on $\omega \in [-\pi/2, \pi/2]$ and $\omega \in [-\pi, -\pi/2] \cup [\pi/2, \pi]$. Next, the smooth

component, S_1, is split into a low-pass (smooth) and high-pass (detail) components S_2 and D_2, by applying filters $H_0(2\omega)$ and $H_1(2\omega)$. If this procedure is repeated recursively, after *J* steps we have

$$\begin{aligned} s(t) &= D_1(t) + S_1(t) \\ &= D_1(t) + D_2(t) + S_2(t) \\ &= D_1(t) + D_2(t) + D_3(t) + S_3(t) \\ &\vdots \\ &= D_1(t) + D_2(t) + \cdots + D_J(t) + S_J(t) \\ &= \sum_{j=1}^{J} D_j(t) + S_J(t) \end{aligned} \tag{7}$$

This leads to division of the space V_0 into a sequence of detail subspaces W_j, $j = 1,\ldots J$ and a smooth subspace V_J

$$V_0 = W_1 \oplus W_2 \oplus W_3 \oplus \cdots \oplus W_J \oplus V_J \tag{8}$$

shown in Fig. 1 (bottom) for the positive half-band only. Recalling that $D_j \in W_j$, $j = 1,\ldots J$ and $S_J \in V_J$, and that each W_m has basis $\{\psi_{m,n}(t)\}_{n\in\mathbf{Z}}$ and V_J has basis $\{\varphi_{J,n}(t)\}_{n\in\mathbf{Z}}$, it follows that the subband components can be expanded in series of the basis functions as follows

$$D_j(t) = \sum_{k=1}^{N/2^j} d_{j,k}\psi_{j,k}(t) \tag{9}$$

and

$$S_J(t) = \sum_{k=1}^{N/2^J} s_{J,k}\varphi_{J,k}(t) \tag{10}$$

which gives the relationship between subbands and the wavelet bases. The coefficients of the expansion represent the discrete wavelet transform (DWT) of the signal

$$d_{j,k} = \langle \psi_{j,k}, s(t) \rangle \tag{11}$$

$$s_{J,k} = \langle \varphi_{J,k}, s(t) \rangle \tag{12}$$

For dimensionality reduction, discussed in the next section, it is important to note that the number of basis functions for both the approximation and detail subspaces reduces by a factor of 2 at each consecutive level of approximation, and that, for any level of decomposition, *J*, the total number of basis functions, which can represent the signal exactly, is equal to *N*—the length of the signal in the time domain. Hence, the orthonormal wavelet transform is nonredundant.

For ideal low- and high-pass prototype filters and discretely sampled signals, the division of the signal bandwidth into subbands, and the central frequencies $\bar{\omega}$ and bandwidths $\Delta\omega$ of the subbands for some intermediate stage of filtering *j* are

$$D_j: \quad \bar{\omega} = \frac{3\pi}{2^{j+1}}, \quad \Delta\omega = \frac{\pi}{2^j}$$
$$S_j: \quad \bar{\omega} = \frac{\pi}{2^{j+1}}, \quad \Delta\omega = \frac{\pi}{2^j} \tag{13}$$

We recall that ω is the frequency in radians per sample. The corresponding frequency in Hz is $f = \omega/(2\pi\Delta t)$.

The pyramid algorithm does subband decomposition of discrete signals by splitting the signal into a high- and a low-frequency component, followed by sub-sampling by a factor of two, keeping the high-frequency output, while operating further on the low-frequency output by splitting and sub-sampling it in the same fashion. The splitting and sub-sampling of the low-frequency output is applied iteratively in several stages. The final results consist of the outputs of the high frequency channels and the low frequency output of the last channel. The maximum number of iteration stages, called levels of the decomposition, depends on the length of the signal and on the length of the impulse responses of the digital filters used to spit the signal (it is larger for a longer signal and for a shorter filter). The output sequences consist of the coefficients of the Discrete Wavelet Transform of the signal.

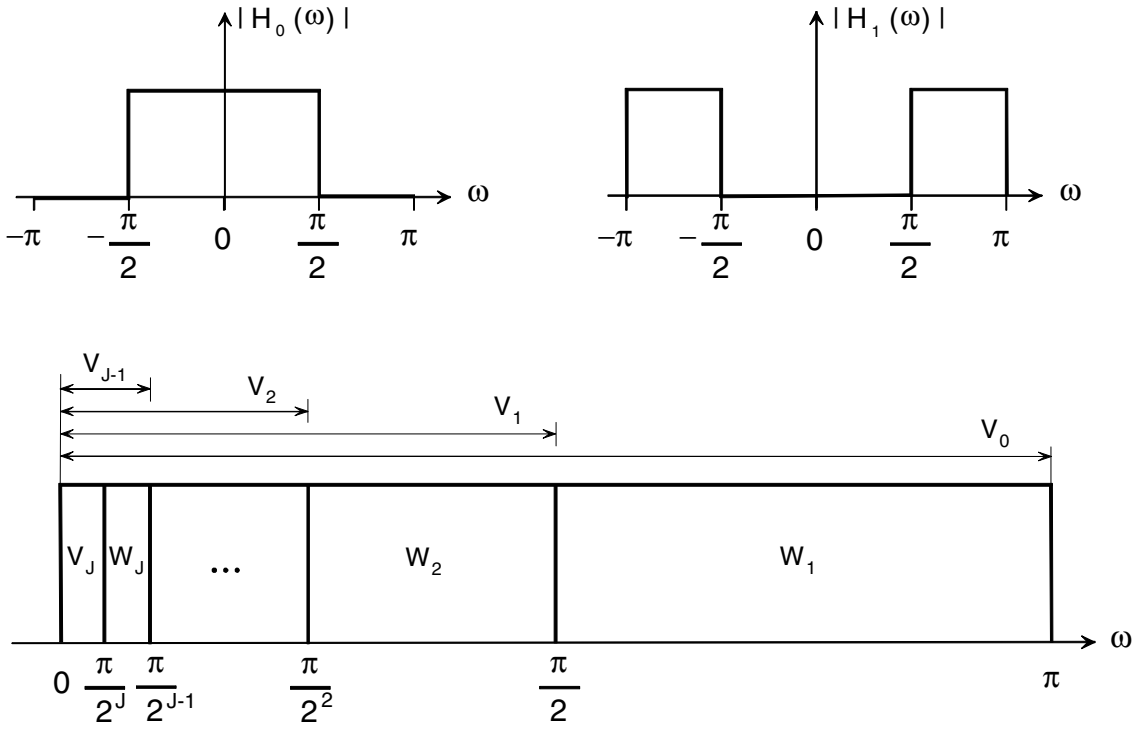

Fig. 1 Ideal low and high pass filters as prototype filters for a two-channel filter bank (top), and an illustration of a *J*-level division of the space of discrete time signals, V_0, by application of the filter bank.

THE DISCRETE WAVELET TRANSFORM AS A DATA MINING TOOL

This section deals with the *theoretical* aspects of dimensionality reduction and information granulation of time series data, using representation in *orthonormal* wavelet bases, for possible application to data mining. Of particular interest to the authors of this paper are time series data of ground and structural vibration response to strong earthquake shaking, as well as to ambient noise, or to forced vibration. The

apparent suitability of the wavelet domain for this purpose is mainly due to the following properties: (1) inherent hierarchical structure, which enables automatically different resolution views of the data, i.e. to zoom in and view the detail (i.e. the trees) or to zoom out and see a coarser view (i.e. the forest); (b) sparse representation compared to that in the time domain of non-stationary and/or band-limited time series data, which leads to high data compression rates with little loss of information; (c) discretely sampled representation, convenient for automated analysis; (d) complete representation, which enables exact reconstruct of the time domain if all the information granules are used, and (e) ability to detect hidden detail in the signal (such as abrupt changes, possibly indicative of damage). Illustrations of these concepts for strong earthquake motion records are presented in the next section.

Some Basic Concepts in Data Mining and Knowledge Discovery

Sensors collect *data* about a process or an object, but data represent only raw information, which is not useful *per se*. What is useful is the *knowledge* revealed from the information contained in the data, which enables one to draw *opinions* based on the data and take actions accordingly. The purpose of these opinions may be to *understand* the nature of the physical process that generated the data (i.e. understand the past), or further to *predict* future events, or to serve as a basis for *decision-making* (i.e. shape the future). This section reviews some basic concepts in data mining and knowledge discovery, with the purpose of providing a context for the application of the discrete wavelet transform, which follows. The definitions of these concepts closely follow Cios et al. [7].

Data mining methods are tools used in *knowledge discovery* to reveal new pieces of knowledge from large data sets. The terms knowledge discovery and data mining first appeared in the late 1980s, and were defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Here, a *pattern* is an entity representing (characterizing, describing) an abstract concept or a physical object, and may describe relationships, correlations, trends, etc. It is a collection of individual *attributes* (features) of the data, and hence is a vector quantity. The number of attributes defines the dimensionality of a pattern in the pattern space. One may search for patterns in the data and classify the data into classes according to the nature of the patterns discovered. A class can be thought of as a state of nature that governs the pattern generation. For example, in data mining of structural vibration data, the concept of interest can be *damage*, and a pattern characterizing damage can be a collection of attributes such as visible fractures and cracks in the structure, reduction of the natural frequency, specific changes of the mode-shapes of vibration, reduced wave travel times across a fractured structural element, etc. According to this pattern, the data can be classified into two groups (classes), one indicating possible damage and the other one—no damage in the structure. In data mining of ground vibration data, the concept of interest can be hazardous ground shaking, with respect to some event such as structural failure of a particular structure at the site where the ground motion was observed, or liquefaction at the site which may initiate foundation failure, which can be characterized by different patterns, depending on the criticality condition adopted. For example, a pattern indicating hazardous ground shaking can be the peak acceleration or the spectral acceleration exceeding some specified (design) level, or the total energy of ground shaking exceeding some level (which depends on a pattern of velocity and duration of shaking, or on a pattern of the Fourier amplitude spectrum). Then according to these patterns, the ground shaking can be classified as hazardous or not. It should be noted here that the classification of the data could be "hard," i.e. with deterministically defined boundaries and exclusive membership in a class, or "soft," i.e. with fuzzy boundaries between the classes and membership defined by a probability distribution function.

The examples mentioned earlier in this section already indicate that the patterns characterizing some concept or object may be more naturally defined in some *transformed domain*, rather than in the domain where the data was recorded. In the case of mechanical vibrations, the data is typically recorded in the

time domain, but is commonly Fourier transformed and analyzed in the frequency domain. In applications that require preserving the time localization, the data is transformed by some time-frequency distribution, e.g. by the wavelet transform or the Gabor transform. Transformation of the data is part of the preprocessing stage.

Another important element of preprocessing is *information granulation*, which refers to reducing the dimensionality of the data by encapsulating numeric data, e.g. in an interval, into a single conceptual entity. Information granulation is necessary for several reasons. Firstly, while data is usually recorded as some long sequence of numbers, it is the nature of the human mind to process information by reducing this longer sequence of numbers into fewer groups or intervals, and associating to each interval some concept, which represents some higher level of abstraction but which is more naturally related to the phenomenon of interest. For example, the sequence of amplitudes of recorded vibrations can be granulated by classification into intervals, which, at a higher level of abstraction, are viewed as representing small, moderate or large response. Information granulation also often consists of aggregating the information by computing local and global averages, for example, representative of a segment of the domain or of the entire domain. For example, one such aggregate granule of information may be the average amplitude of the signal, the root mean square amplitude, which is related to the total energy of the signal, or energy of the signal contained in different frequency subbands. Information granulation is also a practical necessity to avoid a combinatorial explosion in analysis of large sets of data, but may be also convenient in comparing different patterns, i.e. in defining *distance between patterns*. This report explores the use of the discrete wavelet transform as a means for information granulation and definition of patterns in data mining through large set of mechanical vibration data, such as earthquake response data. As it will be seen later in this chapter, the orthonormal discrete wavelet transform appears to be a very convenient tool for representing the signal energy, both on a local and on a global scale.

Wavelet Transformed Database

This sections describe how the discrete wavelet transform of a time series data can be used as a tool in data mining, in particular for reducing the dimensionality of the data, as well as for particular characterization of the data (feature selection) and pattern recognition. We assume that the database has been transformed, and are interested, in particular, in features that can be obtained directly in the wavelet domain, by simple manipulation of the wavelet coefficients, such that can be done at a database level. The record stored in the database for each discrete time series $s[n], n=1,...,N$ is the set of N coefficients of orthonormal wavelet expansion, $\{d_{j,k}\}_{\substack{k=1,...,N/2^j \\ j=1,...,J}}$ and $\{s_{J,k}\}_{k=1}^{N/2^J}$, of the signal in a chosen orthonormal wavelet basis $\{\psi_{j,k}[n]\}_{\substack{k=1,...,N/2^j \\ j=1,...,J}} \cup \{\varphi_{J,k}[n]\}_{k=1,...,N/2^J}$ for agreed level of expansion J, such that

$$s[n]=\sum_{j=1}^{J}\sum_{k=1}^{N/2^j} d_{j,k}\psi_{j,k}[n]+\sum_{k=1}^{N/2^J} s_{J,k}\varphi_{J,k}[n], \quad j=1,...,J \tag{14}$$

It is understood that the discrete time series $s[n], n=1,...,N$ is a sampled version of a continuous time signal $s(t)$ at sampling interval Δt. Each coefficient of the expansion is the orthogonal projection of the signal onto the corresponding basis function, and hence is a measure of how similar the signal is to that basis function. The basis functions are localized both in time and in frequency, *effectively* within a tile of the time-frequency plane. Figure 2 shows a multiresolution partition of the time-frequency plane in non-overlapping tiles. In reality, and even theoretically, the time-frequency distributions of the basis functions extend beyond the boundaries of the corresponding tiles, but most of their energy is within the tile. The

leakage of energy in the neighboring tiles depends on the nature of the basis. Extreme cases are the Haar basis, for which there is no leakage of energy in the neighboring tiles along the time axis but there is leakage along the frequency axis, and the sinc basis, for which there is no leakage of energy in the neighboring tiles along the frequency axis but there is leakage in the neighboring tiles along the time axis. The compactly supported wavelets, such as the s8 wavelet used in the examples in this report, are localized reasonably well both in time and in frequency.

It can be seen from Fig. 2 that, for smaller frequencies, the tiles become wider and shorter, while the area of each tile remains *always* constant and equal to π. For the detail coefficient $d_{j,k}$, the tile is a rectangle with width $\Delta n = 2^j$ and height $\Delta\omega = \pi/2^j$, and centered at time $n = 2^j\left(k+1/2\right)$ and at frequency $\omega = 3\pi/2^{j+1}$. For the smooth coefficient $s_{J,k}$, the tile is also a rectangle with width $\Delta n = 2^J$ and height $\Delta\omega = \pi/2^J$, and centered at time $n = 2^J\left(k+1/2\right)$ and at frequency $\omega = \pi/2^{J+1}$. Note that the half plane is sufficient to represent the signal.

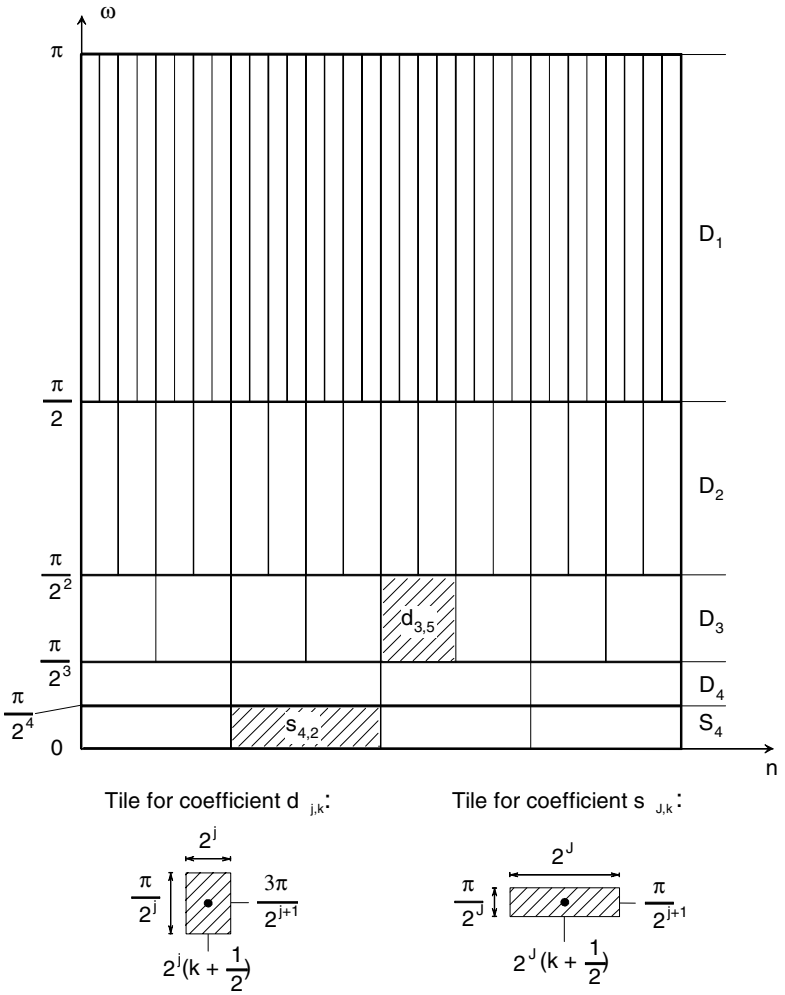

Fig. 2 Multiresolution division of the time-frequency plane for level J=4 expansion.

The N coefficients of orthonormal wavelet expansion represent *all* the information about the signal, and the signal can be *exactly* reconstructed (i.e. up to the precision of the machine) using the inverse wavelet

transform, which can be practically implemented for compactly supported wavelets by the pyramid algorithm. Hence, the set of all of the wavelet coefficients represents the signal at the finest information granulation level.

While the coefficients for fixed j represent a sub-sampled version of the subband components, the subband components as functions of time can be reconstructed at the initial sampling rate of the signal by synthesis of the corresponding coefficients

$$D_j[n] = \sum_{k=1}^{N/2^j} d_{j,k}\psi_{j,k}[n], \quad j = 1,...,J$$
$$S_j[n] = \sum_{k=1}^{N/2^J} s_{j,k}\varphi_{j,k}[n] \tag{15}$$

which can be also accomplished by the inverse pyramid algorithm, by setting to zero all other coefficients.

Reducing Data Dimensionality

Data compression is reduction of the size of a signal while preserving its significant features or most of the energy. The wavelet basis expansion and its relative—the wavelet packet expansion are very efficient for compression of non-stationary time series, such as strong motion earthquake records. As it will be seen in the illustrations in the next section, especially for structural response records, dramatic compression rates are achieved (i.e. reduction of size) by preserving most of the energy.

Data compression is accomplished by expanding the signal in a wavelet basis, and dropping the coefficients of the expansion that are not *significant*, based on an adopted rule, depending on the specific application. One rule may consist of creating a lower resolution approximation of the signal, by dropping the detail coefficients that correspond to the smaller scales, considering them as *noise*. This may be desirable in applications in which the features at the coarser scales are considered as those defining the signal, while the details are considered not so important or contaminated with noise. Another rule is the one based on thresholding, which consists of defining some application specific or universal threshold level, and dropping the coefficients that have magnitude smaller than that level. This rule may be desirable when the signal may have some significant higher resolution features (e.g. some singularities) that need to be preserved. Then this method will preserve the high-resolution features associated with a large wavelet coefficient, while still dropping the less significant features. The threshold level can be a specified as a single value (hard thresholding) or as a range, such that below that range, all coefficients are dropped, while within the range, the coefficients are gradually "shrunk" (soft thresholding), which may reduce some artifacts resulting from the compression. Depending on specific applications, other thresholding schemes can be defined, e.g. such that have subband-specific threshold levels. Data compression is closely related to nonparametric estimation and noise removal using wavelets (Bruce and Gao [8]), as both are based on the same principle of "shrinking" the less significant coefficients.

For discrete sequences, data compression or nonparametric estimation can be formally defined as creating and approximation $\tilde{s}[n]$ of the signal $s[n]$ such that

$$\tilde{s}[n] = \sum_{j=1}^{J}\sum_{k=1}^{2^j}\tilde{d}_{j,k}\psi_{j,k}[n] + \sum_{k=1}^{2^j}\tilde{s}_{J,k}\varphi_{J,k}[n] \tag{16}$$

where the new coefficients of the expansion $\tilde{d}_{j,k}$ and $\tilde{s}_{J,k}$ are

$$\tilde{d}_{j,k} = \delta\left(d_{j,k}\right)$$
$$\tilde{s}_{J,k} = \delta\left(s_{J,k}\right) \tag{17}$$

where $\delta(x)$ is a threshold function, which depends on the adopted threshold scheme. The degree of compression can be measured by the compression rate, herein defined as

$$\text{Compression rate } = 1 - \frac{\tilde{N}}{N} \tag{18}$$

where $\tilde{N}$ is the number of nontrivial reals used to represent the approximation (i.e. the coefficients in the wavelet expansion that are different from zero), and N is the dimension of the original signal (i.e. the number of reals that represent is exactly). In digital signal and image processing, the compression ratio is defined via the ratio of the number of bits used to represent the original signal and the number of bits used to represent the compressed and quantized and coded signal. Quantization consists of dividing the range of possible values of the coefficients in discrete levels, and assigning the coefficient to one of these discrete levels, and coding consists of assigning each level a unique string of bits, with length that it is shorter for the intervals occurring most often. Quantization and coding are out of the scope of this analysis. Dimensionality reduction of mechanical vibration data by compression for feature selection has been previously considered by Staszewski [9].

Estimation of Energy, Correlation and Related Quantities—Local and Global Aggregates and Averages

The estimation of these quantities is based on the Parseval's relation for the orthonormal DWT, also called wavelet Placherel formula. Let $s[n]$ and $g[n]$ be two signals of equal length, and $d_{j,k}^{(s)}$ and $s_{J,k}^{(s)}$, and $d_{j,k}^{(g)}$ and $s_{J,k}^{(g)}$, respectively, be their coefficients of expansion in orthonormal wavelet basis. The Parseval's equality for these two signals is

$$\langle s, g \rangle = \sum_{n=-\infty}^{\infty} s[n]\, g[n] = \sum_{j=1}^{J} \sum_{k=1}^{N/2^j} d_{j,k}^{(s)}\, d_{j,k}^{(g)} + \sum_{k=1}^{N/2^J} s_{J,k}^{(s)}\, s_{J,k}^{(g)} \tag{19}$$

and if the two sequences are equal

$$\|s\|^2 = \sum_{n=-\infty}^{\infty} \left|s[n]\right|^2 = \sum_{j=1}^{J} \sum_{k=1}^{N/2^j} \left|d_{j,k}\right|^2 + \sum_{k=1}^{N/2^J} \left|s_{J,k}\right|^2 \tag{20}$$

Physically, the inner product of two signals is their correlation or cross-energy, and the L_2 norm squared is the energy of the signal, and both represent single aggregate quantities (single granules of information) representing the individual signals an their relationship, i.e. what they have in common. Division by the length of the signals gives the average power and average cross-power, which are single value average distributions of the energy and cross-energy. The sums on the right hand side of equations (19) and (20) allow convenient deaggregation of the energy and cross-energy, into local aggregates and also local averages for the subbands, and further into local aggregates and averages for each bin of the time-frequency plane (Fig. 2), which represent the next two levels of granulation. For all of these levels of granulation, one can compute average spectral density of energy/power for the entire signal, for the subbands and for the bins of the time-frequency plane. Similarly, one can compute average cross-energy/cross power spectrum density at each granulation level. These granules of information then can be used to plot *nodal spectra* and time *nodal time-frequency distributions* of (cross) energy, (cross) power, (cross) power spectrum density, Fourier spectra, etc. Detailed expressions for these quantities, as

well as for average Fourier spectra, are not shown here due to lack of space, and can be found in Todorovska and Hao [6]. It is only noted that the square of the magnitude of each coefficient of the expansion is equal to the energy in the appropriate bin of the time-frequency plane, and also the average power spectrum density in the bin. The convenience offered by the orthonormal wavelet transform to study energy of earthquake motion has been previously recognized by Iyama and Kuwamura [10] who studied the cumulative energy and the rate of energy input for earthquake ground motions.

RESULTS

Few of the concepts discussed in the preceding section are illustrated on acceleration records in a 7-story reinforced building in the city of Van Nuys of the Los Angeles metropolitan area (Trifunac et al. [12]; data was provided by California Div. of Mines and Geology). This building was severely damaged by the 1994 Northridge earthquake and aftershocks. More detailed illustrations of all of the concepts discussed can be found in Todorovska and Hao [6]. These results were computed using the S-Plus wavelet toolbox and the s8 wavelet (Bruce and Gao [8]). Figure 3 illustrates the efficiency of the representation in the (discrete) wavelet domain. It shows the fraction of the total energy that is represented by the top coefficients as function of their number, for the EW accelerations of the Northridge building recorded at the ground floor and at the roof (sampled at 0.02 s). It can be seen that very few of the top coefficients represent most of the energy (3.3% of the coefficients for the roof record and 5% for the ground floor record account for 90% of the total energy). Figure 4 shows, for the same two records, the effect of reduction of dimensionality by thresholding and by subsampling. It is seen that, for the same reduction factor, compression by thresholding preserves the high frequency content where it is significant, while subsampling does not. The error due to the compression for earthquake records that are transient in nature is small, and its value depends on the nature of the excitation. Fig. 5 shows the normalized root mean square error as function of the compression rate, defined in eqn (18), for the EW component of the ground motion records from six earthquakes. The error is the largest for the smallest earthquakes (Montebello and Malibu), which caused shaking that was richer in high frequencies and more "stationary like," and is the smallest for the earthquakes that were the closest to the site Northridge and San Fernando), which caused the largest amplitude shaking. Finally, Fig. 6 shows nodal time-frequency distribution of normalized power spectrum density (by the maximum value for each subband), for the EW displacements at the ground floor and roof during the 1994 Northridge earthquake.

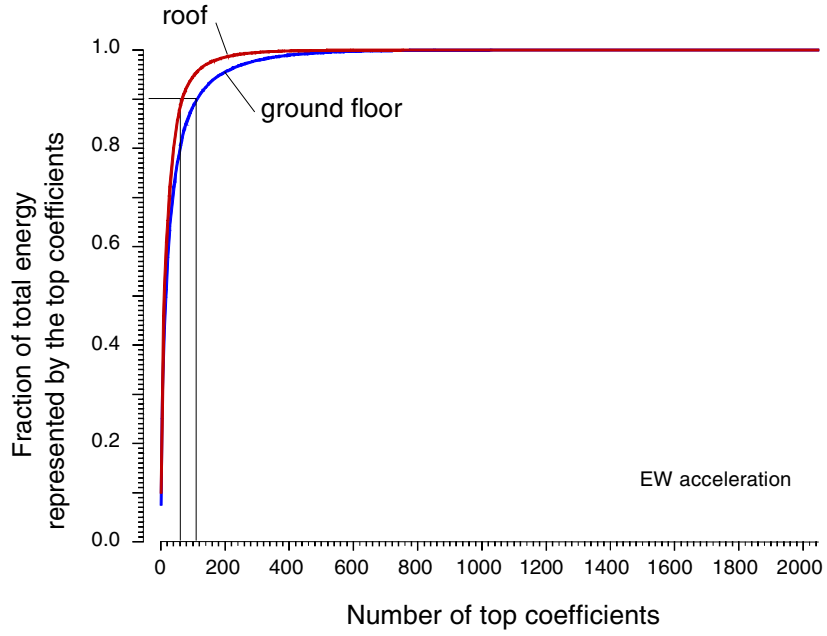

Fig. 3 Fraction of the total energy represented by the top coefficients versus their number for the EW acceleration records at the ground floor and at the roof from the 1994 Northridge earthquake.

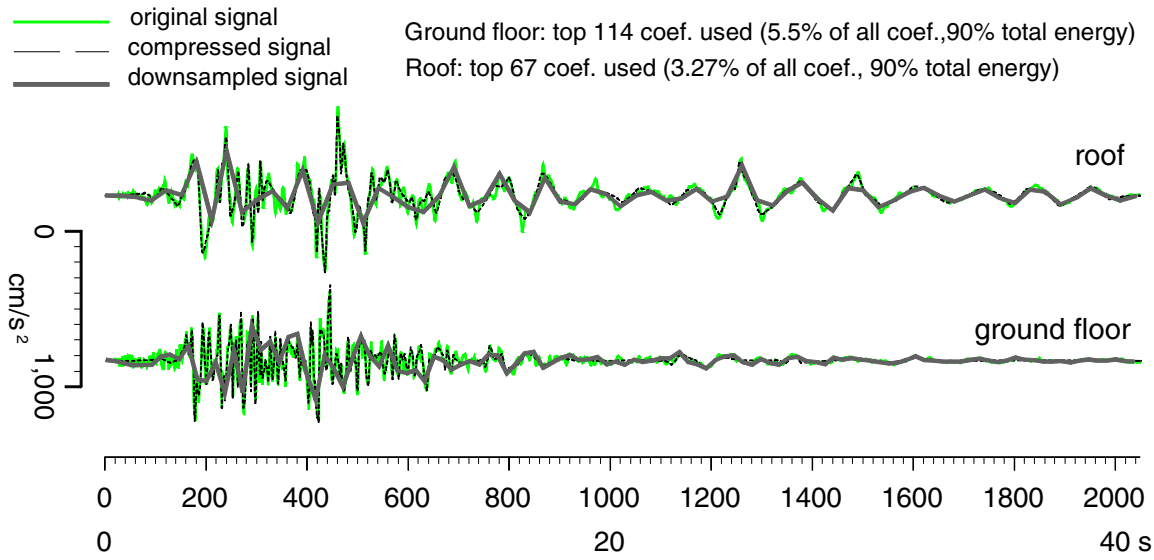

Fig. 4 Comparison of results of reduction of dimensionality by shrinkage of the less significant wavelet coefficients and by sub-sampling, for the EW accelerations from the 1994 Northridge earthquake.

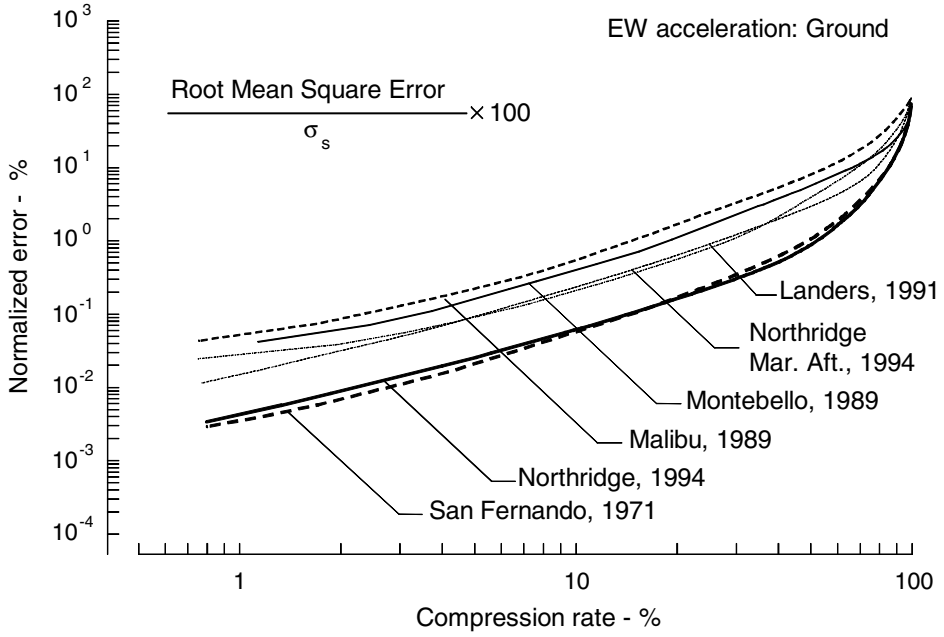

Fig. 5 Root mean square error, normalized by the standard deviation of the signal for EW accelerations recorded at the ground floor during six earthquakes.

The central frequency of the subbands is shown on the right-hand-side of the plot, and the plot on the bottom shows the instantaneous soil-structure system frequency for EW rocking, determined by the complex continuous wavelet transform using the Morlet wavelet (Todorovska [11]). This plots shows significant reduction of this frequency at about 10 s from the beginning of shaking. The large spikes in the highest frequency subband of the roof record indicate (otherwise hidden) abrupt changes in this signal, possibly due to rapid loss of stiffness as a result if damage (Hou et al. [13], Rezai et al. [14]). The discrete wavelet transform is a promising method for detection and localization of damage in a structure. However, more research is needed to determine the reliability of this method when applied to *actual* earthquake records.

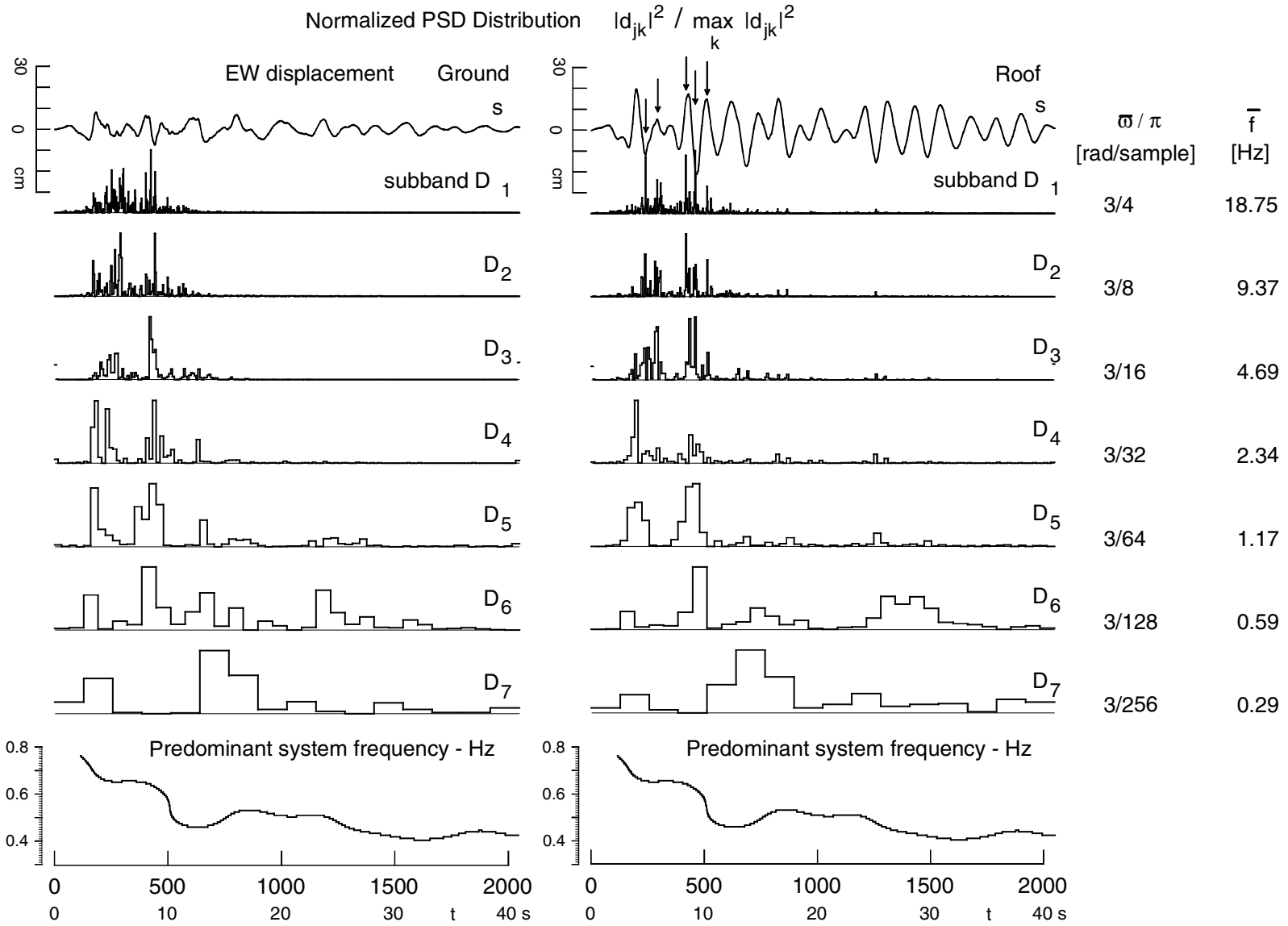

Fig. 5.7.5 Time-frequency distribution of average power spectrum density (PSD), normalized to unit amplitude for each subband, of the EW absolute displacement at the ground floor and on the roof during the 1994 Northridge earthquake.

DISCUSSION AND CONCLUSIONS

From the presented theory and illustrations, it can be concluded that expansion in orthonormal wavelet series is potentially a very useful preprocessing tool in mining large data sets of ground and structural response vibration data under earthquake excitation. The beneficial properties of the transform are: linearity, orthogonality (convenient for estimation of energy), time-frequency localization (convenient for analysis of time varying systems), ability to detect and localize in time abrupt changes in the signal (promising for application to structural health and monitoring damage detection), information granulation (i.e. encapsulation of information about an interval into a single value, hence providing more robust estimation than single values), data compression (enabling a very high degree of dimensionality reduction while preserving the significant features at all scales), multi-resolution structure (enabling separation of features that appear at different scales, as well as dimensionality reduction by lower resolution approximation), and discrete structure (convenient for automated analysis in data mining). One drawback of the orthonormal wavelet transform, as pointed out by Todorovska and Hao [6], is poor resolution at high frequencies. This drawback can be eliminated in principle by further expansion of the subbands that require better frequency resolution in wavelet packets.

REFERENCES

1. Vetterli M, Kovacević J. "Wavelets and Subband Coding", *Prentice Hall*, 1995.
2. Daubechies I. "Ten lectures on wavelets," Philadelphia: *Society for Industrial Application of Mathematics* (SIAM), 1992
3. Newland DE. "Wavelet Analysis of Vibration, Parts I and II", ASME: *J. of Vibration and Acoustics* 1994; **116**, 409-416, and 417-425.
4. Trifunac MD, Todorovska MI. "Evolution of accelerographs, data processing, strong motion arrays and amplitude and spatial resolution in recording strong earthquake motion," *Soil Dynamics and Earthquake Engineering* 2001; **21**, 537-555.
5. Benz, H, Buland R, Filson J, Frankel A, Sheldok K. "Advanced National Seismic System," *Seism. Res. Letters* 2001; **72**(1), 70-75.
6. Todorovska MI, Hao TY. "Information granulation and dimensionality reduction of seismic vibration monitoring data using orthonormal discrete wavelet transform for possible application to data mining," Dept. of Civil Eng., U. Southern California: Report CE 03-02, 2003.
7. Cios KJ, Pedrycz W, Swiniarski R. "Data mining methods for knowledge discovery," Boston: *Kluwer Academic Publishers*, 1998.
8. Bruce, A, Gao HY. "Applied Wavelet Analysis with S-Plus", New York: *Springer-Verlag*, 1996.
9. Staszewski WJ. "Wavelet based compression and feature selection for vibration analysis," *J. of Sound and Vibration* 1998; **211**(5), 735-760.
10. Iyama J, Kuwamura H. "Application of wavelets to analysis and simulation of earthquake motions," *Earthquake Engineering and Structural Dynamics* 1999; **28**, 255-272.
11. Todorovska MI. "Estimation of instantaneous frequency of signals using the continuous wavelet transform," Dept. of Civil Eng., U. Southern California: Report CE 01-07, 2001.
12. Trifunac MD, Ivanović SS, Todorovska MI. "Instrumented 7-storey reinforced concrete building in Van Nuys, California: description of damage from the 1994 Northridge earthquake and strong motion data," Dept. of Civil Engrg., U. Southern California: Report CE 99-02, 1999.
13. Hou Z, Noori M, St. Amand R. "Wavelet based approach for structural damage detection," ASCE: *J. of Engineering Mechanics* 2000; **126**(7), 677-683.
14. Rezai M, Rahmatian P, Ventura CE. "Seismic data analysis of a seven-storey building using frequency response function and wavelet transform," CUREE: *Proc. NEHRP Conf. and Workshop on Research on the Northridge, California Earthquake of January 17, 1994*, 1996, Vol. III, 421-428.