Ultrametric Wavelet Regression of Multivariate Time Series: Application to Colombian Conflict Analysis

Fionn Murtagh
Science Foundation Ireland, Wilton Place, Dublin 2, Ireland, and
Department of Computer Science, Royal Holloway, University of London
Egham TW20 0EX, England
Email: fmurtagh@acm.org

Michael Spagat
Department of Economics, Royal Holloway, University of London
Egham TW20 0EX, England
Email: m.spagat@rhul.ac.uk

Jorge A. Restrepo
Conflict Analysis Resource Center, Carrera 10 No 65-35 Of. 703
Bogotá, Colombia
Email: jorge.restrepo@cerac.org.co

February 17, 2009

Abstract

We first pursue the study of how hierarchy provides a well-adapted tool for the analysis of change. Then, using a time sequence-constrained hierarchical clustering, we develop the practical aspects of a new approach to wavelet regression. This provides a new way to link hierarchical relationships in a multivariate time series data set with external signals. Violence data from the Colombian conflict in the years 1990 to 2004 is used throughout. We conclude with some proposals for further study on the relationship between social violence and market forces, viz. between the Colombian conflict and the US narcotics market.

1 Introduction

In this article we will refer to Appendices 2, 3 and 4 for a basic introduction to Correspondence Analysis, sequence-constrained hierarchical clustering and the

Haar wavelet transform on a hierarchy. In section 1.1 we will first review the terminology that we use.

1.1 Terminology: Ultrametric and Wavelet Regression

By ultrametric we intend a precise definition of "hierarchy" which as such is a loose concept. The ultrametric inequality holds for pairwise distances (for a distance, d, defined for all pairs of points i, j as a symmetric, positive-definite mapping to the reals, the ultrametric inequality holds for all triplets of points, i, j, k: $d(i, k) \leq \max\{d(i, j), d(j, k)\}$). A set of ultrametric distances can be derived from a binary (i.e., each non-terminal node has either one or two child nodes; a terminal node has no child node), ranked (i.e., each node in the hierarchy has a "level" value associated which is typically positive real-valued) rooted tree [2]. Reciprocally, a binary, ranked rooted tree, known as a dendrogram, is a representation of an ultrametric set of distances. We can relax the binary property of this tree but often it is algorithmically convenient to impose this.

By regression, we mean some mapping of independent or explanatory variable x onto a dependent or response variable, y, such that the behavior of y as a function of x can be studied. In this work, x is multidimensional, while y is unidimensional. A novelty is that x is ultrametric, i.e. all triplets of points satisfy the ultrametric inequality; or we avail of a hierarchical structure on x.

Wavelet regression or wavelet smoothing is "potentially highly locally adaptive" [23]. A data signal or set of points is projected into a new basis such that spatial and frequency properties are brought out or exposed. (Consider, for comparative purposes, the following. A sinusoidal basis system used in Fourier analysis helps with frequency properties but not with spatial. A basis of principal components helps with spatial properties but not with frequency properties.) The wavelet transform projection itself is invertible. This means that data can be reconstructed, after carrying out modifications in transform space. (The same principle can be applied in Fourier analysis or principal component analysis.) Modification in wavelet space typically involves setting low-valued wavelet coefficients to zero, leading to a smoother reconstruction. A novelty in our work is that we understand wavelet regression to involve both explanatory and response variables, rather than just wavelet smoothing of erstwhile explanatory variables.

As we will see below, our wavelet transform is a hierarchical one, in the sense of being carried out on a hierarchical clustering or dendrogram. It is therefore a wavelet transform in an ultrametric space, or an ultrametric wavelet transform.

In this context, the effect of wavelet coefficient thresholding leads to piecewise constant function approximation. We will sketch out the explanation for this. The subnodes of any given node are determined in transform space by a positive or a negative detail vector. (The dimensionality of this vector is the dimensionality of the points we are working on. It is possible that these points are single valued.) So left and right subnodes are defined as the parent node's vector, plus or minus (respectively) the detail vector. Setting the detail vector to zero just means that the subnodes, in the reconstructed or wavelet-smoothed data, now become, both identically, equal to the parent node. In other words, within the cluster defined by the parent node, we now have a (constant within the cluster) vector given by the parent node. Wavelet smoothing, in the way that we are implementing it here, furnishes a constant vector within a cluster.

1.2 Time Series Segmentation and the Issue of Normalization

The approach we develop in this article is based on the following.

1. A goal is "subsequence clustering" or segmentation.
2. Normalization and weighting of time series and attributes are addressed.
3. We closely interface the analysis with interpretation. We regress the hierarchical segmentation on external time series.

We will discuss each of these briefly in turn, in a comparative setting. We will also look at statistical changepoint analysis, and a divisive segmentation tree approach.

Unlike in [30] we are concerned with segmentation of a multivariate time series, and not clustering a collection of time series. This distinction is referred to as, resectively, "whole clustering" and "subsequence clustering" by [11]. Singhal et al. [30] build their clustering around principal components analysis, and we also use Correspondence Analysis – appropriate for our frequency count data – to provide an embedding from which distances can be easily defined.

The great deal of recent work on clustering within time series signals, e.g. the incremental and divisive hierarchical algorithm of [29], has been critiqued by [11] on the grounds of meaningfulness. Goldin et al. [7] however point out that use of Euclidean distance is overly simplistic and that data normalization is needed, which they phrase in terms of "distances between cluster shapes rather than between cluster sets". The need to consider normalization issues very attentively, and the close resulting links between distance, shape, and other properties (size, scale, angle, etc.), are all well known in the Correspondence Analysis area. See [17], in particular chapter 4, which presents case studies of analysis of size and shape, and financial time series modeling and forecasting.

In this article we develop a new approach to hierarchical clustering of multivariate time series based on a rigorous treatment of weighting of time series and attributes.

1.3 Short Review of Clusterwise Segmentation

Our innovative approach incorporates piecewise or clusterwise approximation to a time series or one-dimensional signal. Unlike [32], we do not seek to fit patterns directly on the data. Similar type approaches have been pursued in [15, 25, 10, 31]. Instead we fit the piecewise approximation based on a given hierarchical clustering. Furthermore we do not want to determine time series segments in a void, removed from external explanatory and indeed causative discussion. Therefore we are led to regress the hierarchical segmentation, as an explanatory input, onto time series, comprising response variables, that are external to the initial phase of hierarchical segmentation building.

IsoFinder [24] seeks to segment a genome sequence by repeatedly finding a splitting point and looking for a significant difference in the subsequences to the left and right of the splitting point. Significant difference comes from a t-test between means to left and right; and a bootstrap-based setting of significance level. One aspect of our approach that differs from IsoFinder is that we support a fully multivariate input signal. IsoFinder could remain as a divisive approach for multivariate input through use of a quadtree and octtree decomposition, although it would become clumsy relative to our bottom-up tree-building.

Statistical changepoint analysis is exemplified by [6]. A probability model is established for the changepoints. This probability mass function uses the distance between successive changepoints. Within segments linear or polynomial regression is used. This is of unknown order which is to be estimated. Since movement from one changepoint to the next uses transition probabilities, this is to say that a Markov process is used. Positions of changepoints and the regression model orders are sought, using stochastic optimization within a MAP, maximum a posteriori, schema. In [5], piecewise linear regression is used for "regression models where the underlying functional relationship between the response and the explanatory variable is modeled as independent linear regressions on disjoint segments."

In [8] applied to speech and motion tracking, even if called "piecewise constant" it is in our terminology "piecewise linear": the model is local linear regression with "piecewise constant parameters".

In [12], the limits of a clusterwise linear approach are noted. A "piecewise linear model can determine where the data goes up or down and at what rate. Unfortunately, when the data does not follow a linear model, the computation of the local slope creates overfitting." In [12], adaptive polynomial (including a 0 order, or constant) degree of fit in each piecewise region is pursued in the modeling of stock prices and ECG data.

Relative to [6] and, for example, [33], our approach is not parametric. In terms of evaluation therefore we are more concerned to link results with external data, both qualitative (e.g. causes of changepoints of varying degrees of importance) and quantitative (e.g. other univariate time series). Therefore the applications of our work and their context is somewhat different from these parametric modeling approaches.

1.4 Structure of this Article

In this article, we will interchangeably refer to a time series defined on a timeline and a signal with support given by this timeline. Broadly speaking we will also use interchangeably the following terms: hierarchy, tree, dendrogram, ultrametric space. A range of other terms, algorithms and methods are introduced and discussed in the Appendices.

In section 2 the data and problem is discussed, together with aspects relating to methodology that are at issue in this article.

In section 3, the relevance and indeed importance of hierarchy in the data analysis is addressed.

In section 4 the association of a hierarchy with an external signal is at issue. We develop a new way to map the latter onto the hierarchy.

In section 5 we study how the mapping of external signal to the hierarchy can lead to new insights.

2 Background of the Data

The data used in this work related to conflict violence and were produced by CERAC, the Conflict Analysis Resource Center, www.cerac.org.co. We will refer to the data used as the CERAC Colombia Conflict database. It is being grown on an ongoing basis. From [26], we use high frequency micro-data, relating to internal conflict violence in Colombia (population: 44 million) from 1988 to 2004, hence over a period of more than 16 years.

Conflict violence in Colombia is not based primarily on ethnic, religious, or regional differences, as is often the case elsewhere, but it instead has roots in economic and political factors. Economic drivers include, for instance, the narcotics sector and kidnapping. The CERAC data records discrete action types, their intensity, dates, locations and other information. Event types are broken down at the first divide into clashes requiring multilateral engagement, and attacks which are unilateral. Intensity breaks down mainly into killings and injuries. For the data period, the Colombian conflict has seen more than 3000 killings plus injuries per year. One use of the dataset has been to analyze civilian casualties [28].

In [9] distribution of fatalities over time was found to follow a power law. Furthermore the work of these authors points to how remarkably similar power law behavior can be found in conflicts ranging from Iraq to Indonesia. Such power law behavior leads to self-similarity, i.e., burstiness on all aggregation scales. (See examples of application of this in [1] and general discussion in [22].) As a result it is difficult to smooth such data, and hence it is difficult to fit simple parametric models. Under such circumstances it is feasible to seek and find patterns and trends in the data, and it is precisely these goals which are at issue in this work.

In this article, we present a novel approach for the study of dynamics. Firstly we look at visualizations using Correspondence Analysis [17]. The output spatial representation is an equiweighted Euclidean one, with embedding of both rows and columns – events, and the attributes used for these events – and this allows cluster analysis and other analyses to bypass any other form of data normalization or standardization.

We then study change versus continuity over time, allowing for gradation in such change. A hierarchical clustering is used, taking account of the timeline, presenting change cluster, breakpoints and timeline resolution-related properties.

Context is offered by the following. In [3], the price of cocaine in the US from 1983 onwards is found to be affected by interdiction effectiveness, and this in turn is found to be beneficial in terms of reduced consumption, at least in the short run.

In our analysis we will look at how narcotics prices might be related to the Colombian violence. Causal mechanisms could take place in different ways. For example a declining world market for narcotics could put great pressure on production, leading to violent reaction when faced with declining returns. Or an expanding market could lead to greater violence because the spoils are greater. A further consideration is that causal mechanisms may work themselves out at different rates, and even in different ways, on the wholesale and retail markets. The retail markets for illegal narcotics are driven by actors who are distributed far and wide relative to the locations of production.

The cocaine drug comes from the *coca* plant. Most initially processed cocaine on the world narcotics markets come from Colombia. A main tool in combating production by the Colombian government has been crop spraying, leading to production being dispersed in jungle regions or national parks. The coca plant is very different from the *cocoa* bean from the *cacao* tree, from which chocolate is produced. *Coffee*, from coffee beans of the coffee plant, is the leading legal agricultural export from Colombia. The latter fact is important because (for example, as just one possible economic mechanism) a declining market for coffee could conceivably lead to product substitution in response.

Opiates come from the opium poppy. Like cocaine, opiates are also narcotic alkaloids. Morphine, derived from opium, is used for medicinal purposes, whereas heroin is a generally banned narcotic. While poppy production has been at times of some significance in Colombia, mostly the leading opiates producer worldwide has been Afghanistan. However for supply of opiates to the US, Colombia and Mexico play a major role. As noted above in regard to the complexity of economic mechanisms, there are intricate linkages between, say, cocaine and opiates, involving many social actors.

3 Hierarchy: Tracking and Prioritizing of Change

3.1 Data

We used 144 numerical attributes relating to killings and injuries coming from 20,288 dated events: see Appendix 1. Aggregating by month the numerical data, 144 attributes used, yielded data for 204 successive months. Further aggregating by year provided us with data for the 15 years, 1990 to 2004, each year having a 144-valued vector of values where we eliminated 1988 and 1989 because we did not have narcotics data for these years (see section 4).

A short periodization of the conflict is as follows [26]: reintensification of conflict in 1986, roughly constant then to 1994, followed by continuous acceleration.

1988–1991, "adjustment period" due to the end of the Cold War; 1992–1996, "stagnation period"; and from 1997 onwards, "upsurge period". In the section to follow there will be further discussion of periodization.

3.2 Considerations on the Analysis

Figure 1 shows a best planar representation, accounting for 62% of the information content of the data as expressed by the combined moments of inertia of the cloud of years (or of the cloud of attributes) about these two axes. Following somewhat clumped characteristics of the years 1990 through 1997, there is a break then with the year 1998. From 1997 an arc is formed, up to 2004. A break between the years 1997 and 1998, (i), is followed by a later break, between 2001 and 2002, (ii).

The 1998 changepoint, (i), was in a period when guerrilla gains from 1996 onwards were being reversed, a process that was seen to be so by 1999. There were high levels of government casualties in 1997–1998, due to guerrilla operations against isolated military and police bases. In particular through airborne weaponry, the government got the upper hand. The "upsurge period" from 1997 onwards also saw a rise of (anti-guerrilla) paramilitary activity whereas before they had been involved in drug trafficking. The paramilitaries started operations around 1997. There was a consolidation of paramilitary groups in 1997, announced publicly in December 1997, and they became active then, having many of their own number killed. It was not until 1999 that paramilitaries began to kill large numbers of guerrillas. Among all of these mutually influencing, reinforcing or retarding, trends and events, our analysis points to a succession of two years where the global change was most intense.

We now come to the 2002 changepoint, (ii). A peak of (paramilitary) casualties brought about by government against paramilitaries was in 2002 due to aerial bombardment of a paramilitary position under attack by guerrillas. It was a major setback for the paramilitaries, who declared a truce following the election that year of President Alvaro Uribe.

Figure 2 shows (i) the caesuras between 1997 (8th year) and 1998 (9th year); and (ii) between 2001 (12th year) and 2002 (13th year). Subtrees before and after these caesuras are clearly distinguishable in the full tree. In line with Figure 1, what Figure 2 indicates is that these are the important caesuras in this data. Note that the same data as seen in the planar projection in Figure 1, viz. the factor 1 and 2 projections, is used for the construction of the hierarchy of Figure 2.

The data used to construct the hierarchy in Figure 2 is of inherent dimension min(20288 events, 204 months, 144 attributes, 15 years) -1 (from section 3.1; see also Appendices 1 and 2). Furthermore, based on Figure 1, we in fact used the first two factors only. So the input data dimensionality on which the hierarchy in Figure 2 is based is 2.

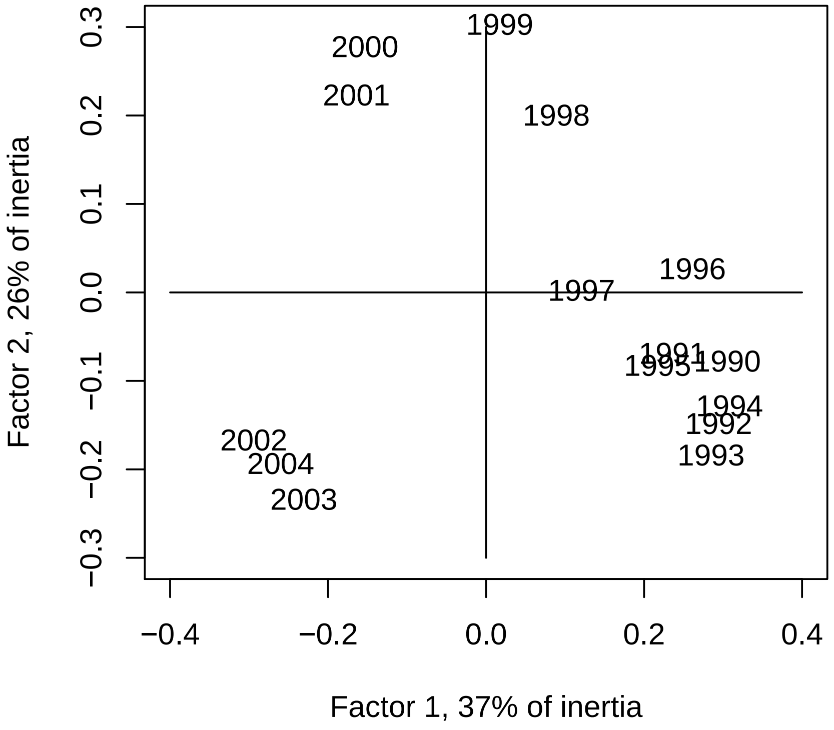

Figure 1: Correspondence Analysis of the annual aggregated data using 144 attributes. Years shown. As discussed in the text, gaps representing the more important changepoints are (i) between 1997 and 1998; and (ii) between 2001 and 2002.

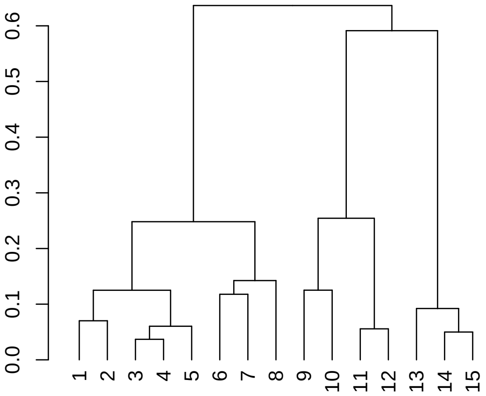

Figure 2: Hierarchical clustering of the fifteen years from 1990 to 2004, using the principal factor values from Figure 1. A sequence constrained complete link agglomeration criterion was used. The main caesuras or changepoints here are (i) in moving from year 8 = 1997 to year 9 = 1998; and (ii) in moving from year 12 = 2001 to year 13 = 2002.

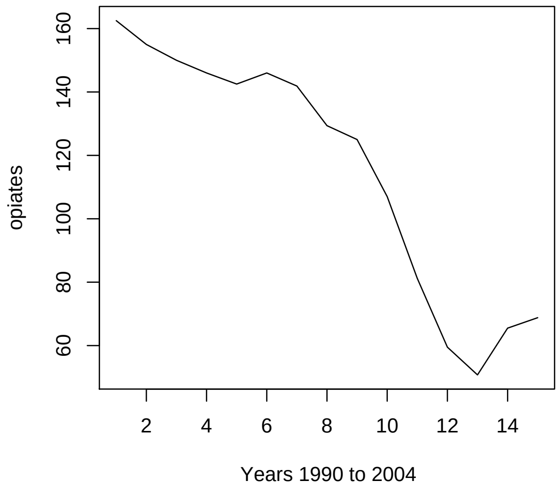

Figure 3: Wholesale street prices of opiates in the US over 15 years, 1990 to 2004. (Source of data: [34].) Ordinate units are US$/kg.

3.3 Novelties in Methodology

Correspondence Analysis (Figure 1) embeds both annually aggregated events and their attributes in the same Euclidean space. As a basis for subsequent analysis, e.g. clustering, weighting and normalization are taken care of as part of the Euclidean embedding algorithm. Appendix 2 presents an overview of Correspondence Analysis.

The hierarchical clustering of Figure 2 is built on the timeline. This means that clusters are contiguous on the yearly timeline. Input to the hierarchical clustering is factor projections from the Correspondence Analysis implying that the 15 yearly aggregated events are equiweighted and endowed with the Euclidean distance. Appendix 3 presents an overview of this sequence-constrained hierarchical clustering.

Figure 2 provides a hierarchical understanding of the Colombian conflict violence over the years 1990 to 2004. It ranks change by order of importance. Bigger change is associated with more distinct branches in the tree. A hierarchy defines an ultrametric topology. So we can justifiably characterize Figure 2 as an ultrametric, or tree metric, understanding of the Colombian conflict violence. In the same way the semantics of Colombian society are displayed in a metric, rather than ultrametric, way in Figure 1.

4 Association between Hierarchical Understanding and External Signal

4.1 Regressing Opiates Wholesale Prices Response Variable on Explanatory Hierarchy

We have a structured view in Figure 2 of the socio-dynamics of the situation in Colombia over the years 1990 to 2004. Let us call it $\mathcal{H}$. This is the set of interrelationships expressed in Figure 2 over the years 1990–2004. It is a summarized view of those years. It is a synthetic view, and depends not just on the algorithms applied but – a most important aspect – on the input data used. As a structured view of the socio-dynamics of Colombia, we now seek to associate $\mathcal{H}$ with other external signals defined on the same time period.

As potentially relevant economic sectors outside Colombia we will consider the narcotics sector. We will consider two narcotics signals.

Figure 3 shows the opiates wholesale price, in the US, averaged per year, in US$/kg. Our aim is to look for linkages between the evolving Colombian socio-dynamic expressed in Figure 2 and distant reflections of this, such as the particular narcotics market situation shown in Figure 3. Production is expressed (in a highly complicated way) in the former, Figure 2, while consumption is expressed in the latter, Figure 3.

As noted in section 3.2 there are particular breakpoints easily visible in Figure 2, e.g. in moving from year 8 = 1997 to year 9 = 1998; and in moving from year 12 = 2001 to year 13 = 2002. In Figure 3 something of this is visible too: year 9 is early in a period of precipitous fall (i.e., fall in wholesale price); and year 13 is a clearly visible turning point.

Let us map the opiates data in Figure 3 into the hierarchical structure, $\mathcal{H}$, of Figure 2. One way to do this is to have a set of within-cluster constant approximations to the opiates (Figure 3 data), and furthermore to allow for a range of such constant approximations.

In practice we will use constant approximations of the opiates data, Figure 3, that are governed by the embedded clusters of $\mathcal{H}$, Figure 2. However we will not proceed by cluster level (or dendrogram level from the root) because this would imply different cluster cardinalities at each level – and hence a not very sensible ordering of what are ultimately approximations to the full data set. We will read off approximations to the hierarchy by decreasing *differences* in information between dendrogram levels. We will allow ourselves to be guided by these differences in information between levels. This information is available to us through the dendrogram's Haar wavelet transform. See Appendix 4 for background detail.

In Figure 4, we see the succession of better approximations, quantified by mean squared error (MSE), as given by decreasing absolute values of the dendrogram Haar wavelet detail coefficients.

In regard to use here and later of MSE, note that the size of this goodness of fit value is strongly influenced by the signal's values, – cf. ordinates in displays.

The MSE is not scale-invariant and the ordinate scale impacts on the MSEs.

Note that the indication “thresholded at level 3”, for instance, in Figure 4, is to be understood as at the third, ordered by decreasing value, wavelet detail coefficient. This is not necessarily the third dendrogram level from the root.

4.2 Further Response Variables

We additionally now consider a set of signals, see upper left panel in Figures 4, 5, 6 and 7. There are, including the initial analysis of the previous section:

- Opiates and cocaine retail prices (street price), US$/gram, see [34].
- Opiates and cocaine wholesale price, US$/kg, see [34].

In Figures 4, 5, 6 and 7 exact reconstruction of the input signal, displayed in the upper left, is possible if no thresholding is used.

Figure 5 shows the successive approximations to the opiate retail prices – again in the US, from 1990 to 2004, street prices in US$/gram [34].

Figure 6 shows the successive approximations to the cocaine wholesale prices – in the US, from 1990 to 2004, in US$/kg [34].

Figure 7 shows the successive approximations to the cocaine retail prices – in the US, from 1990 to 2004, street prices in US$/gram [34].

4.3 Novelties in Methodology

We denote the hierarchy in Figure 2 as $\mathcal{H}$ and a signal such as that in Figure 3 as $\int$. Both $\mathcal{H}$ and $\int$ have the same support, viz. the ordered years 1990 to 2004. We are projecting the ultrametric space, $\mathcal{H}$, onto the signal $\int$. We have: $f : \mathcal{H} \longrightarrow \int$. This can be viewed too as an ultrametric regression, i.e. we are regressing $\int$ on $\mathcal{H}$. We estimate our signal from the hierarchy: $\widehat{\int} = f(\mathcal{H})$.

Consider now the segmentation as a clusterwise regression problem. We could approach the clusterwise regression by partition that is read off the hierarchy in Figure 2. Using the Haar wavelet transform instead reads off partitions in order of MSE (mean square error) of approximation of the input signal by the wavelet filtered and reconstructed signal. The MSE provides us with a measure of the approximation of the piecewise constant (which is also linear: but for constant ordinate or response variable) fit to the input signal. Note that each partition of the hierarchy covers the full signal, i.e. each partition is by definition defined on the input's support.

5 Patterns in Data Reflecting the Known Hierarchical Structuring of Events

5.1 Smoothing based on Ultrametric Wavelet Transform

A baseline scenario for what we now seek is shown in Figure 8. In this figure we show a simple and direct clusterwise fit to each of the four signals. We replace

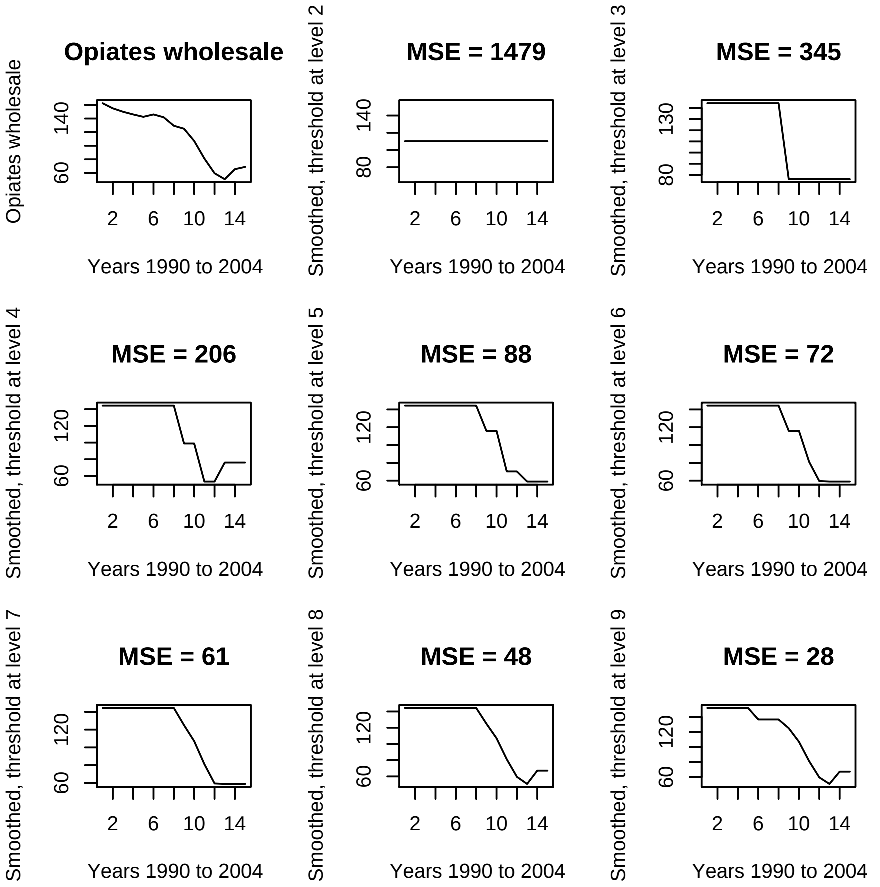

Figure 4: Upper left: opiates, wholesale prices, over the 15 years, 1990 to 2004. Then from left to right, row-wise: reconstruction of the data based on thresholding using decreasing absolute values of the dendrogram wavelet detail coefficients. The MSE, mean square error, indicates the quality of approximation (with a small value being best).

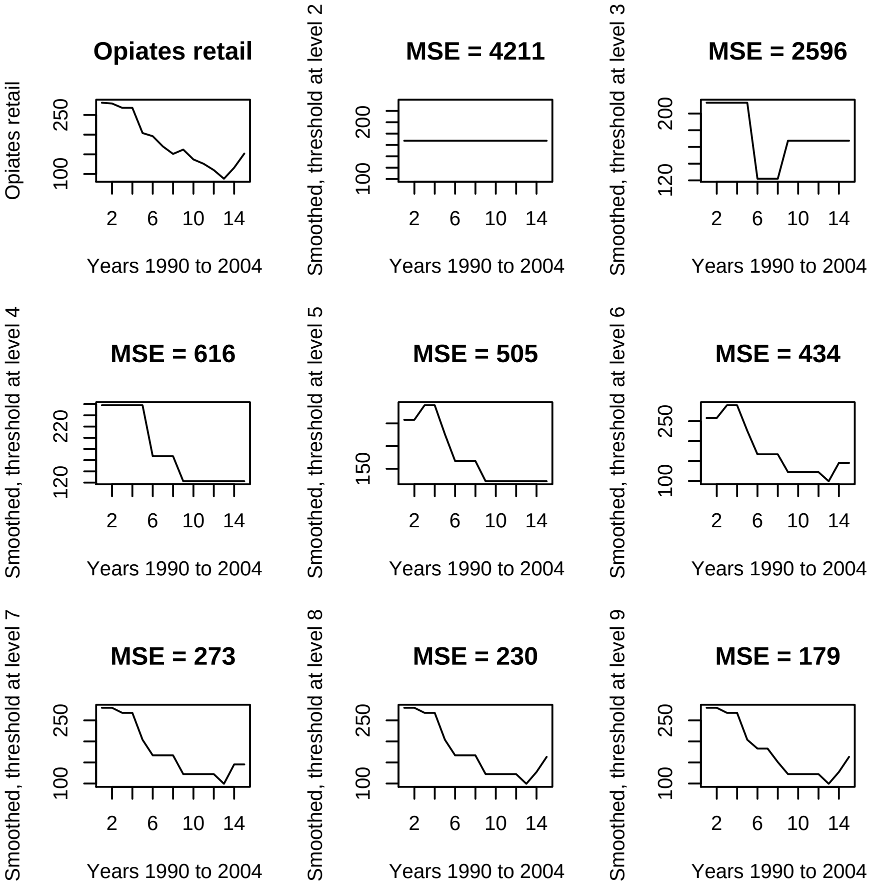

Figure 5: Upper left: opiates, retail prices, over the 15 years, 1990 to 2004. Then from left to right, row-wise: reconstruction of the data based on thresholding using decreasing absolute values of the dendrogram wavelet detail coefficients. The MSE, mean square error, indicates the quality of approximation.

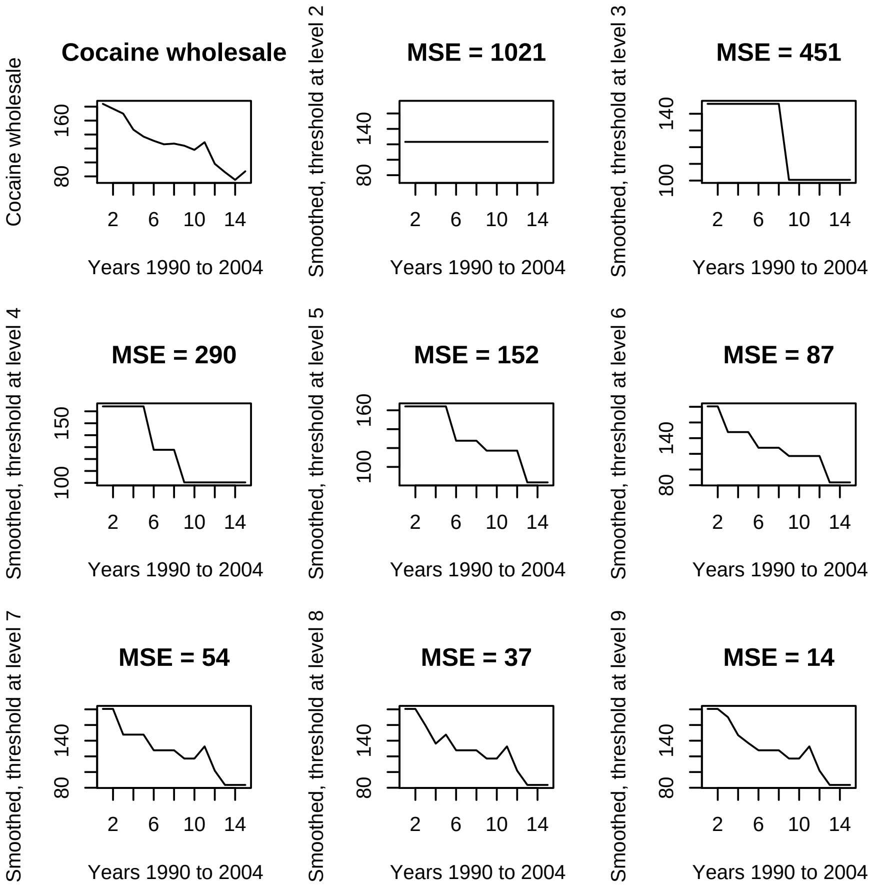

Figure 6: Upper left: cocaine, wholesale prices, over the 15 years, 1990 to 2004. Then from left to right, row-wise: reconstruction of the data based on thresholding using decreasing absolute values of the dendrogram wavelet detail coefficients. The MSE, mean square error, indicates the quality of approximation.

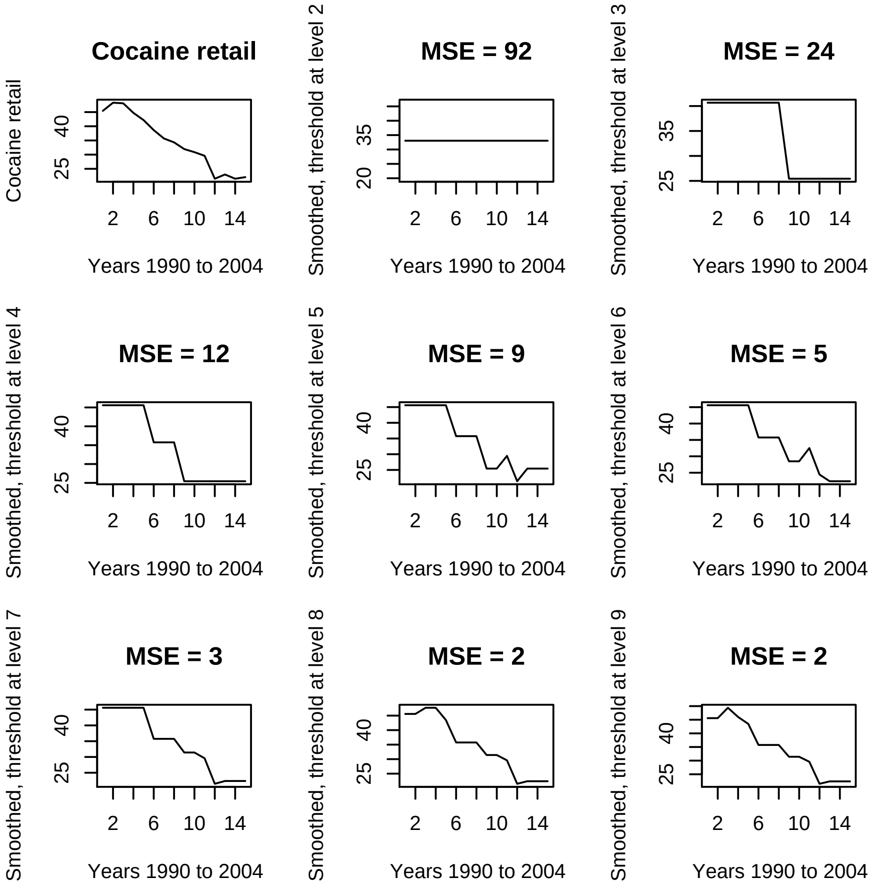

Figure 7: Upper left: cocaine, retail prices, over the 15 years, 1990 to 2004. Then from left to right, row-wise: reconstruction of the data based on thresholding using decreasing absolute values of the dendrogram wavelet detail coefficients. The MSE, mean square error, indicates the quality of approximation.

Crude 3 segments	Fig. 8	1		– 8	9 – 12		13 –	15
Opiates wholesale	Fig. 9	1		– 8	9 – 10	11 – 12	13 –	15
Opiates retail	Fig. 9	1	– 5	6 – 8	9 –			15
Cocaine wholesale	Fig. 10	1	– 5	6 – 8	9 –			15
Cocaine retail	Fig. 10	1	– 5	6 – 8	9 –			15

Table 1: Breakpoints in the Colombian 15-year violence timeline. The breakpoints were read off the figures noted in the Table. We see, for example, that for the cocaine wholesale signal, the breakpoints are such that there is a segment consisting of years 1 to 5; then a segment of years 6 to 8; and finally a segment from year 9 to year 15. From the crude signal segmentation of Figure 8 and the wavelet smoothings of Figures 9 and 10, there is considerable consistency.

the first eight values, in the first cluster, with their mean value; and similarly for the second cluster; and the third. The goodness of fit is given for each quadrant from upper left to lower right as indicated in the figure caption.

Can our wavelet regression perform better? The mean square errors (MSEs), we find, are improved in all four cases. Respectively, as we will see, the MSE of opiates wholesale decreases from 223 to 206. The MSE of opiates retail decreases from 1533 to 616. The MSE of cocaine wholesale decreases from 305 to 290. And the MSE of cocaine retail decreases from 18 to 12.

From Figures 4, 5, 6 and 7 we selected segmentations with at least three segments in each, corresponding to the major breakpoints discussed in section 3.2. The signals used and the clusterwise constant approximated versions are shown in Figures 9 and 10.

Figures 9 and 10 can be viewed as follows. From the hierarchy in Figure 2, with its important caesuras as discussed in section 3.2, take a range of within-cluster signal values as their mean. This implies that we have a piecewise linear approximation of the signals, in the sense of clusterwise constant. Compared to our baseline scenario shown in Figure 8 we intervene on – modify in wavelet transform space – a number of clusters characterized by small changes in the hierarchy.

Our baseline of Figure 8 shows breakpoints over the 15-year time span. The respective breakpoints of Figures 9 and 10 are largely consistent, as summarized in Table 1, while (we may say informally) using the hierarchy as a “key” or “template”.

5.2 Interpretation of Trends and Patterns

Figure 11 shows panels containing (respectively from upper left to lower right) opiates wholesale price and retail price signals; and cocaine wholesale and retail price signals. Over these have been superimposed their approximations that approximate the hierarchy, $\mathcal{H}$.

Note how for both wholesale signals, opiates and cocaine (left up and left

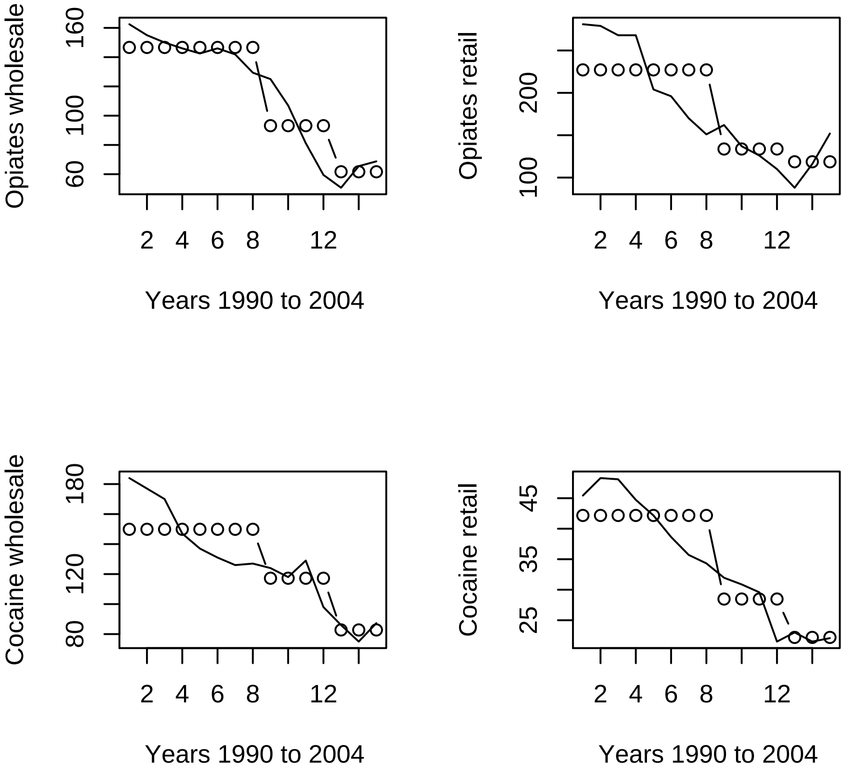

Figure 8: We take the three clearest clusters in Figure 2, viz. from year 1 = 1990 to year 8 = 1997; from year 9 = 1998 to year 12 = 2001; and from year 13 = 2002 to year 15 = 2004. Then we simply trace the piecewise constant fits, in each case, to the four different narcotics market signals. These fits are shown by the small circle piecewise constant curves. The mean square errors are, respectively for opiates wholesale and retail, and cocaine wholesale and retail: 223, 1533, 305 and 18.

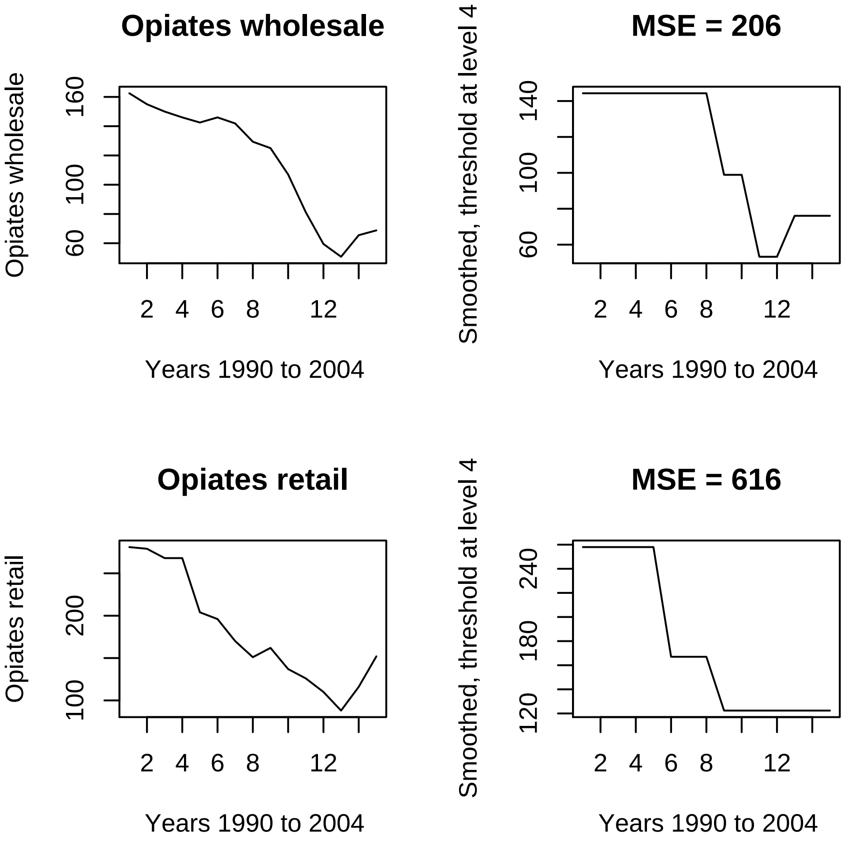

Figure 9: Left: opiates, wholesale and retail prices, over the 15 years, 1990 to 2004. Right: corresponding reconstruction of the signal based on wavelet detail thresholding. (The two upper panels are identical to panels shown in Figure 4; and the two lower panels are identical to panels shown in Figure 5.)

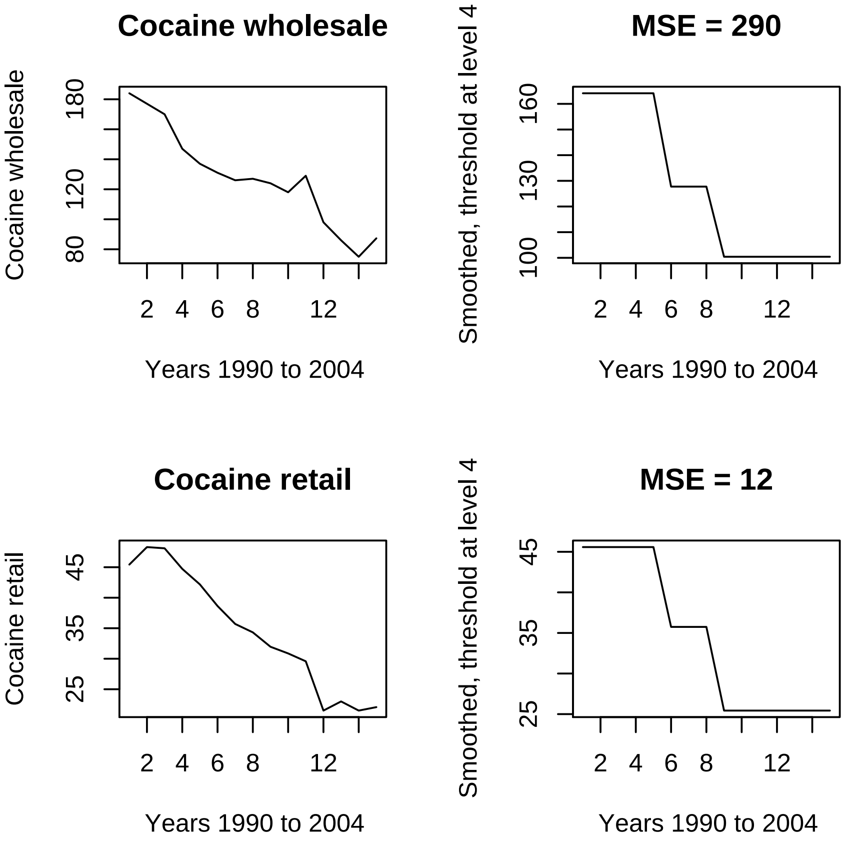

Figure 10: Left: cocaine, wholesale and retail prices, over the 15 years, 1990 to 2004. Right: corresponding reconstruction of the signal based on wavelet detail thresholding. (The two upper panels are identical to panels shown in Figure 6; and the two lower panels are identical to panels shown in Figure 7.)

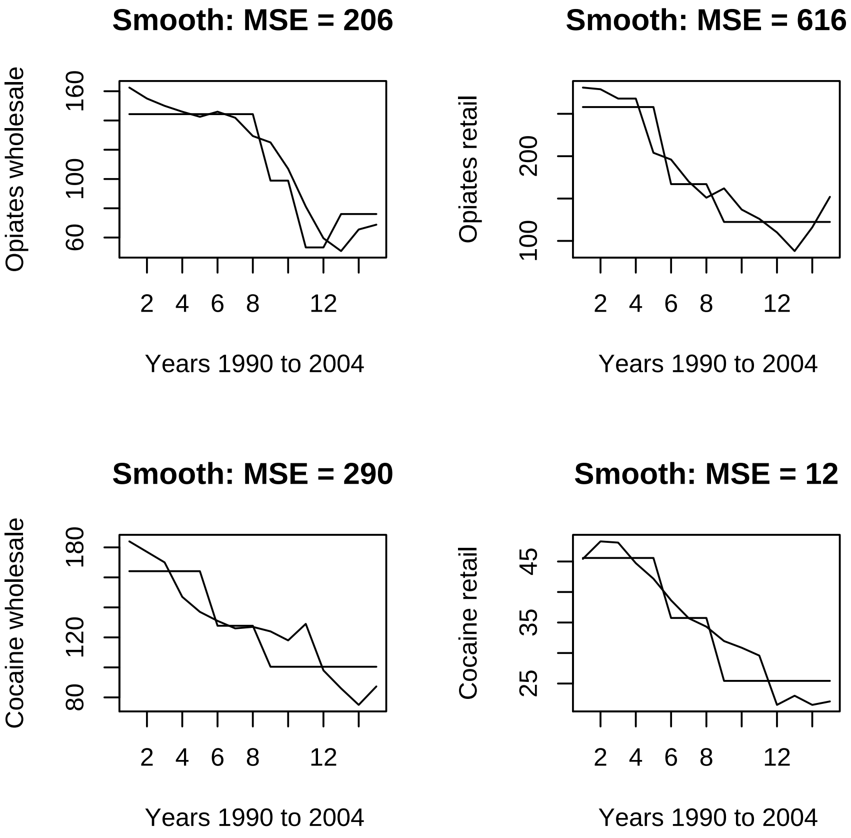

Figure 11: Shown are the input signals with one selected smooth from, respectively, Figures 4, 5, 6 and 7. The same plots are shown separately in Figures 9 and 10. The smooth referred to here is the piecewise linear smooth. The goodness of fit between piecewise linear (more strictly, clusterwise constant) fit and original signal is given by MSE, mean square error.

down panels in Figure 11), the hierarchy-representing signal is to the left of or equal to the wholesale price signal from, and including, years 8 = 1997 to 12 = 2001. In those years the information from the hierarchy, expressing the Colombian conflict, was lower than the wholesale price. Now, the wholesale price itself was falling. So we conclude that this points to the Colombian internal conflict as having possibly played a causative role in the wholesale price movement.

The corresponding situation for the retail price signals (right up and right down in Figure 11) is different in this regard. The analogous influence by the Colombian internal conflict on the retail price movement is taking place from year 6 = 1995 to year 7 = 1996, and again from year 8 = 1997 to year 11 = 2000.

From a visual point of view there is a tighter fit between the opiates (top two panels in Figure 11) compared to cocaine (bottom two panels in Figure 11). Note that the opiates wholesale hierarchy fit is the only one with 4 segments; all others have 3 segments. This was due to our selection of partition. It is interesting to note the similarities of our findings, i.e. the same years being indicated for the two wholesale signals, and again the same years being indicated for the two retail signals.

Obviously we are dealing with one possible mechanism relating to price formation here and there are other causal factors.

While we use 144 attributes in this work for tracking overall violence levels, and while we find that breakpoints in the conflict occur at consistent times relative to big movements in narcotics prices, we note that these narcotics prices are largely falling in the period considered. So it is not feasible to seek a link between prices and increasing or decreasing violence. Ultimately we are looking for more subtle associations.

An unanticipated finding that comes out of this analysis is as follows. One could well expect the Colombian conflict to be a driver of US cocaine prices. What we have found is that the Colombian conflict could also be influential in regard to US opiate prices. We note this as a potentially useful lead.

5.3 Novelties in Methodology

Figure 11 shows inter-related breakpoints arising from the hierarchical interpretation of the Colombian conflict violence from 1990 to 2004. Each wavelet filtering and reconstructing of the data provides one set of breakpoints. The simplest way to approximate any other signal – for example, the annually averaged wholesale price of opiates in the US – is to use a clusterwise constant fit to the signal *with the breakpoints taken into account.* This is what is done in Figure 8. What is innovative about the wavelet-based smoothing is how the breakpoints, defining the piecewise linear approximation, are defined. They are defined from the *partial order* specified by $\mathcal{H}$.

6 Conclusion

We have described how hierarchy expresses changes at varying scales; how hierarchy can be induced on a time-varying data set; and then how other data sets defined on the same timeline can be "folded over" the hierarchically structured data. Our aim is initial screening of relationships which may lead to later stages of modeling, as exemplified by the following.

Taking the important products, coffee and oil, for the Colombian economy, with the former being labor-intensive and the latter capital-intensive, [4] studied their relationship on the Colombian conflict at a low (fine granularity) level: "Exploiting exogenous price shocks in international markets, we find that a drop in the price of coffee during the late 1990s increased civil war violence in municipalities that cultivate coffee more intensively. In contrast, a rise in oil prices during the same period induced greater conflict in the oil municipalities relative to the non-oil region. Lower coffee prices exacerbated conflict by reducing wages and the opportunity cost of recruiting fighters into armed groups, while higher oil prices fuelled conflict by increasing contestable revenue in local governments, which invited predation by groups that steal these resources."

Compared to this we have sought to preserve a wide range of levels of aggregation. We do this because of our view that causative factors are often on a range of resolution levels. Scale, in this case temporal, is important. So rather than the simple smoothing of Figure 8 we instead wanted to take more of the hierarchical information into account. Apart from the goodness of fit being improved there were also a few intriguing findings – to be further studied with more detailed data – relating to the relationship between market processes and conflict violence.

Appendix 1: Attributes

144 attributes were used for all events. Organisational acronyms used here: FARC, Armed Revolutionary Forces of Colombia; ELN, National Liberation Army; ERP, Popular Revolutionary Army; CGSB, Simon Bolivar Guerrilla Coordination; DAS, Administrative Security Department; M-19, 19th of April Movement.

K0, Number of civilians killed

K1, Number of military forces members killed

K2, Number of police members killed

K3, Number of paramilitaries killed

K4, Number of ELN members killed

K5, Number of FARC members killed

K6, Number of EPL members killed

K7, Number of ERP members killed

K8, Number of other guerrilla group members killed

K9, Number of other group members killed

K10, Number of other non-identified group members killed

K11, Number of CGSB members killed

K12, Number of DAS members killed

K13, Number of M19 members killed

I0, Number of civilians injured

I1, Number of military forces members injured

I2, Number of police members injured

I3, Number of paramilitaries injured

I4, Number of ELN members injured

I5, Number of FARC members injured

I6, Number of EPL members injured

I7, Number of ERP members injured

I8, Number of other guerrilla group members injured

I9, Number of other group members injured

I10, Number of other non-identified group members injured

I11, Number of CGSB members injured

I12, Number of DAS members injured

I13, Number of M19 members injured

KGue, Number of guerrilla members killed

KPar, Number of paramilitary members killed

KGob, Number of government members killed

KCiv, Number of civilians killed

K, Number of killings

IGue, Number of guerrilla members injured

IPar, Number of paramilitary members injured

IGob, Number of government members injured

ICiv, Number of civilians injured

I, Number of injured

KIGue, Number of guerrilla members casualties

KIPar, Number of paramilitary members casualties

KIGob, Number of government members casualties

KICiv, Number of civilians casualties

KI, Number of casualties

KClAtCivGue, Number of civilians killed in guerrilla events

KClAtCivPar, Number of civilians killed in paramiltary events

KClAtCivGob, Number of civilians killed in government forces events

KICivELN, Number of civilian casualties in ELN events

KICivFARC, Number of civilian casualties in FARC events

KCivGueMas, Number of civilians killed in guerrilla massacres

KCivGueInc, Number of civilians killed in guerrilla incursions

KCivGueBom, Number of civilians killed in guerrilla bombings

KGueGue, Number of guerrillas killed in guerrilla unilateral attacks

KParGue, Number of paramilitaries killed in guerrilla unilateral attacks

KGobGue, Number of government members killed in guerrilla unilateral attacks

KCivGue, Number of civilians killed in guerrilla unilateral attacks

KTotGue, Total number of killings in guerrilla unilateral attacks

IGueGue, Number of guerrillas injured in guerrilla unilateral attacks

IParGue, Number of paramilitaries injured in guerrilla unilateral attacks
IGobGue, Number of government members injured in guerrilla unilateral attacks
ICivGue, Number of civilians injured in guerrilla unilateral attacks
ITotGue, Total number of injured in guerrilla unilateral attacks
KIGueGue, Number of guerrilla casualties in guerrilla unilateral attacks
KIParGue, Number of paramilitary casualties in guerrilla unilateral attacks
KIGobGue, Number of government member casualties in guerrilla unilateral attacks
KICivGue, Number of civilian casualties in guerrilla unilateral attacks
KITotGue, Total number of casualties in guerrilla unilateral attacks
KGuePar, Number of guerrillas killed in paramilitary unilateral attacks
KParPar, Number of paramilitaries killed in paramilitary unilateral attacks
KGobPar, Number of government members killed in paramilitary unilateral attacks
KCivPar, Number of civilians killed in paramilitary unilateral attacks
KTotPar, Total number of killings in paramilitary unilateral attacks
IGuePar, Number of guerrillas injured in paramilitary unilateral attacks
IParPar, Number of paramilitaries injured in paramilitary unilateral attacks
IGobPar, Number of government members injured in paramilitary unilateral attacks
ICivPar, Number of civilians injured in paramilitary unilateral attacks
ITotPar, Total number of injured in paramilitary unilateral attacks
KIGuePar, Number of guerrilla casualties in paramilitary unilateral attacks
KIParPar, Number of paramilitary casualties in paramilitary unilateral attacks
KIGobPar, Number of government member casualties in paramilitary unilateral attacks
KICivPar, Number of civilian casualties in paramilitary unilateral attacks
KITotPar, Total number of casualties in paramilitary unilateral attacks
KGueGob, Number of guerrillas killed in government unilateral attacks
KParGob, Number of paramilitaries killed in government unilateral attacks
KGobGob, Number of government members killed in government unilateral attacks
KCivGob, Number of civilians killed in government unilateral attacks
KTotGob, Total number of killings in government unilateral attacks
IGueGob, Number of guerrillas injured in government unilateral attacks
IParGob, Number of paramilitaries injured in government unilateral attacks
IGobGob, Number of government members injured in government unilateral attacks
ICivGob, Number of civilians injured in government unilateral attacks
ITotGob, Total number of injured in government unilateral attacks
KIGueGob, Number of guerrilla casualties in government unilateral attacks
KIParGob, Number of paramilitary casualties in government unilateral attacks
KIGobGob, Number of government member casualties in government unilateral attacks
KICivGob, Number of civilian casualties in government unilateral attacks
KITotGob, Total number of casualties in government unilateral attacks
KCivOth, Number of civilians killed in other group unilateral attacks
ICivOth, Number of civilians injured in other group unilateral attacks
KICivOth, Number of civilian casualties in other group unilateral attacks
KGueClGobGue, Number of guerrillas killed in government-guerrilla clashes
KParClGobGue, Number of paramilitaries killed in government-guerrilla clashes
KGobClGobGue, Number of government members killed in government-guerrilla clashes
KCivClGobGue, Number of civilians killed in government-guerrilla clashes

KTotClGobGue, Total number of killings in government-guerrilla clashes
IGueClGobGue, Number of guerrillas injured in government-guerrilla clashes
IParClGobGue, Number of paramilitaries injured in government-guerrilla clashes
IGobClGobGue, Number of government members injured in government-guerrilla clashes
ICivClGobGue, Number of civilians injured in government-guerrilla clashes
ITotClGobGue, Total number of injured in government-guerrilla clashes
KIGueClGobGue, Number of guerrilla casualties in government-guerrilla clashes
KIParClGobGue, Number of paramilitary casualties in government-guerrilla clashes
KIGobClGobGue, Number of government member casualties in government-guerrilla clashes

KICivClGobGue, Number of civilian casualties in government-guerrilla clashes
KITotClGobGue, Total number of casualties in government-guerrilla clashes
KGueClGobPar, Number of guerrillas killed in government-paramilitary clashes
KParClGobPar, Number of paramilitaries killed in government-paramilitary clashes
KGobClGobPar, Number of government members killed in government-paramilitary clashes

KCivClGobPar, Number of civilians killed in government-paramilitary clashes
KTotClGobPar, Total number of killings in government-paramilitary clashes
IGueClGobPar, Number of guerrillas injured in government-paramilitary clashes
IParClGobPar, Number of paramilitaries injured in government-paramilitary clashes
IGobClGobPar, Number of government members injured in government-paramilitary clashes

ICivClGobPar, Number of civilians injured in government-paramilitary clashes
ITotClGobPar, Total number of injured in government-paramilitary clashes
KIGueClGobPar, Number of guerrilla casualties in government-paramilitary clashes
KIParClGobPar, Number of paramilitary casualties in government-paramilitary clashes
IGobClGobPar, Number of government members injured in government-paramilitary clashes

ICivClGobPar, Number of civilians injured in government-paramilitary clashes
ITotClGobPar, Total number of injured in government-paramilitary clashes
KIGueClGobPar, Number of guerrilla casualties in government-paramilitary clashes
KIParClGobPar, Number of paramilitary casualties in government-paramilitary clashes
KIGobClGobPar, Number of government casualties in government-paramilitary clashes
KICivClGobPar, Number of civilian casualties in government-paramilitary clashes
KITotClGobPar, Total number of casualties in government-paramilitary clashes
KGueClGuePar, Number of guerrillas killed in guerrilla-paramilitary clashes
KParClGuePar, Number of paramilitaries killed in guerrilla-paramilitary clashes
KGobClGuePar, Number of government members killed in guerrilla-paramilitary clashes
KCivClGuePar, Number of civilians killed in guerrilla-paramilitary clashes
KTotClGuePar, Total number of killings in guerrilla-paramilitary clashes
IGueClGuePar, Number of guerrillas injured in guerrilla-paramilitary clashes
IParClGuePar, Number of paramilitaries injured in guerrilla-paramilitary clashes
IGobClGuePar, Number of government members injured in guerrilla-paramilitary clashes
ICivClGuePar, Number of civilians injured in guerrilla-paramilitary clashes
ITotClGuePar, Total number of injured in guerrilla-paramilitary clashes
KIGueClGuePar, Number of guerrilla casualties in guerrilla-paramilitary clashes

KIParClGuePar, Number of paramilitary casualties in guerrilla-paramilitary clashes
KIGobClGuePar, Number of government member casualties in guerrilla-paramilitary clashes

KICivClGuePar, Number of civilian casualties in guerrilla-paramilitary clashes
KITotClGuePar, Total number of casualties in guerrilla-paramilitary clashes

Appendix 2: the Correspondence Analysis Platform

This Appendix and the next introduce important aspects of Correspondence Analysis and hierarchical clustering. Further reading is to be found in [2] and [17].

Analysis Chain

1. Starting point: a matrix that cross-tabulates the dependencies, e.g. frequencies of joint occurrence, of an observations crossed by attributes matrix.

2. By endowing the cross-tabulation matrix with the χ^2 metric on both observation set (rows) and attribute set (columns), we can map observations and attributes into the same space, endowed with the Euclidean metric.

3. A hierarchical clustering is induced on the Euclidean space, the factor space.

4. Interpretation is through projections of observations, attributes or clusters onto factors. The factors are ordered by decreasing importance.

Various aspects of Correspondence Analysis follow on from this, such as Multiple Correspondence Analysis, different ways that one can encode input data, and mutual description of clusters in terms of factors and vice versa. In the following we use elements of the Einstein tensor notation of [2]. This often reduces to common vector notation.

Correspondence Analysis: Mapping χ^2 Distances into Euclidean Distances

- The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$.

- I is the set of observation indexes, and J is the set of attribute indexes. We have $k(i) = \sum_{j \in J} k(i,j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i,j)$.

- Relative frequencies: $f_{IJ} = \{f_{ij} = k(i,j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly f_I is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and f_J analogously.

- The conditional distribution of f_J knowing $i \in I$, also termed the jth *profile* with coordinates indexed by the elements of I, is:

$$f_J^i = \{f_j^i = f_{ij}/f_i = (k_{ij}/k)/(k_i/k); f_i > 0; j \in J\}$$

  and likewise for f_I^j.

- What is discussed in terms of information focusing in the text is underpinned by the *principle of distributional equivalence.* This means that if two or more profiles are aggregated by simple element-wise summation, then the χ^2 distances relating to other profiles are not effected.

Input: Cloud of Points Endowed with the Chi Squared Metric

- The cloud of points consists of the couples: (multidimensional) profile coordinate and (scalar) mass. We have $N_J(I) = \{(f_J^i, f_i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.

- Included in this expression is the fact that the cloud of observations, $N_J(I)$, is a subset of the real space of dimensionality $|J|$ where $|.|$ denotes cardinality of the attribute set, J.

- The overall inertia is as follows:

$$M^2(N_J(I)) = M^2(N_I(J)) = \|f_{IJ} - f_I f_J\|^2_{f_I f_J}$$

$$= \sum_{i \in I, j \in J} (f_{ij} - f_i f_j)^2 / f_i f_j$$

- The term $\|f_{IJ} - f_I f_J\|^2_{f_I f_J}$ is the χ^2 metric between the probability distribution f_{IJ} and the product of marginal distributions $f_I f_J$, with as center of the metric the product $f_I f_J$.

- Decomposing the moment of inertia of the cloud $N_J(I)$ – or of $N_I(J)$ since both analyses are inherently related – furnishes the principal axes of inertia, defined from a singular value decomposition.

Output: Cloud of Points Endowed with the Euclidean Metric in Factor Space

- The χ^2 distance with center f_J between observations i and i' is written as follows in two different notations:

$$d(i, i'^i_J - f_J^{i'}\|^2_{f_J} = \sum_j \frac{1}{f_j} \left(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}} \right)^2$$

- In the factor space this pairwise distance is identical. The coordinate system and the metric change.

- For factors indexed by α and for total dimensionality N ($N =$ min $\{|I| - 1, |J| - 1\}$; the subtraction of 1 is since the χ^2 distance is centered and hence there is a linear dependency which reduces the inherent dimensionality by 1) we have the projection of observation i on the αth factor, F_α, given by $F_\alpha(i)$:

$$d(i, i') = \sum_{\alpha=1..N} \left(F_\alpha(i) - F_\alpha(i')\right)^2 \qquad (1)$$

- In Correspondence Analysis the factors are ordered by decreasing moments of inertia.

- The factors are closely related, mathematically, in the decomposition of the overall cloud, $N_J(I)$ and $N_I(J)$, inertias.

- The eigenvalues associated with the factors, identically in the space of observations indexed by set I, and in the space of attributes indexed by set J, are given by the eigenvalues associated with the decomposition of the inertia.

- The decomposition of the inertia is a principal axis decomposition, which is arrived at through a singular value decomposition.

Dual Spaces and Transition Formulas

- $$F_\alpha(i) = \lambda_\alpha^{-\frac{1}{2}} \sum_{j \in J} f_j^i G_\alpha(j) \text{ for } \alpha = 1, 2, \ldots N; i \in I$$

$$G_\alpha(j) = \lambda_\alpha^{-\frac{1}{2}} \sum_{i \in I} f_i^j F_\alpha(i) \text{ for } \alpha = 1, 2, \ldots N; j \in J$$

- *Transition formulas*: The coordinate of element $i \in I$ is the barycenter of the coordinates of the elements $j \in J$, with associated masses of value given by the coordinates of f_j^i of the profile f_J^i. This is all to within the $\lambda_\alpha^{-\frac{1}{2}}$ constant.

- In the output display, the barycentric principle comes into play: this allows us to simultaneously view and interpret observations and attributes.

Appendix 3: Hierarchical Clustering

Sequence-Constrained Hierarchical Clustering

Background on hierarchical clustering in general, and the particular algorithm used here, can be found in [16]. The sequence constraint considered here is, for example, the total order involved in a time series.

Consider the projection of observation i onto the set of all factors indexed by α, $\{F_\alpha(i)\}$ for all α, which defines the observation i in the new coordinate frame. This new factor space is endowed with the (unweighted) Euclidean distance, d. We seek a hierarchical clustering that takes into account the observation sequence, i.e. observation i precedes observation i' for all $i, i' \in I$. We use the linear order on the observation set.

Agglomerative hierarchical clustering algorithm:

1. Consider each observation in the sequence as constituting a singleton cluster. Determine the closest pair of adjacent observations, and define a cluster from them.
2. Determine and merge the closest pair of adjacent clusters, c_1 and c_2, where closeness is defined by $d(c_1, c_2) = \max \{d_{ii'} \text{ such that } i \in c_1, i' \in c_2\}$.
3. Repeat the second step until only one cluster remains.

This is a sequence-constrained complete link agglomeration criterion. The cluster proximity at each agglomeration is strictly non-decreasing.

Recent application of this method can be found in [21], relating to the sequence of scenes in a film script (further discussed in [13]).

How a Hierarchy Expresses Change

Consider Figure 12. The schematic representation, Figure 12 left, is of document retrieval where the query is in the upper right. In Figure 12 right the ultrametric distance is illustrated.

Further discussion of how a hierarchy expresses the semantics of change and distinction, themes that are central in this article, can be found in [20].

The hierarchies that we induce on given data are based on an embedded set of clusters. Through a series of $n - 1$ agglomerations starting from n terminal nodes, the hierarchical clustering is constructed. The dendrogram tree has, by construction, two-way branchings at each node.

Appendix 4: Haar Wavelet Transform of a Dendrogram

The discrete wavelet transform is a decomposition of data into spatial and frequency components. In terms of a dendrogram these components are with re-

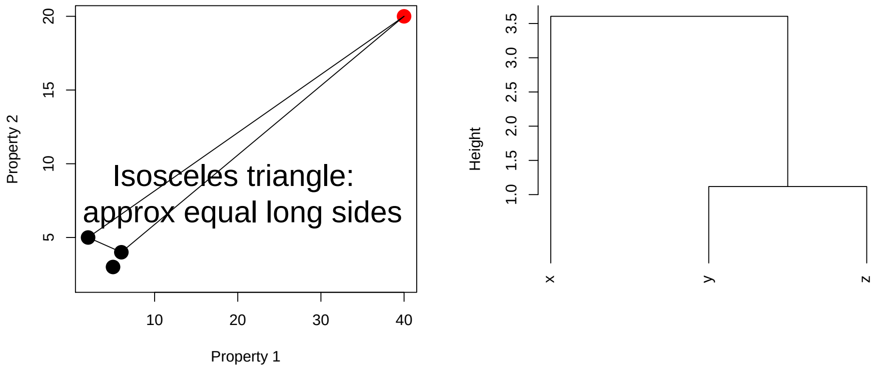

Figure 12: Left: The query is on the far right. While we can easily determine the closest target (among the three objects represented by the dots on the left), is the closest really that much different from the alternatives? Right: The strong triangular inequality defines an ultrametric: every triplet of points satisfies the relationship: $d(d,z) \leq \max\{d(x,y), d(y,z)\}$ for distance d. Cf. by reading off the hierarchy, how this is verified for all x, y, z: $d(x,z) = 3.5; d(x,y) = 3.5; d(y,z) = 1.0$. In addition the symmetry and positive definiteness conditions hold for any pair of points.

spect to, respectively, within and between clusters of successive partitions. We show how this works taking the data of Table 2.

The hierarchy built on the 8 observations of Table 2 is shown in Figure 13.

Something more is shown in Figure 13, namely the detail signals (denoted $\pm d$) and overall smooth (denoted s), which are determined in carrying out the wavelet transform, the so-called forward transform.

The inverse transform is then determined from Figure 13 in the following way. Consider the observation vector x_2. Then this vector is reconstructed exactly by reading the tree from the root: $s_7 + d_7 = x_2$. Similarly a path from root to terminal is used to reconstruct any other observation. If x_2 is a vector of dimensionality m, then so also are s_7 and d_7, as well as all other detail signals.

This procedure is the same as the Haar wavelet transform, only applied to the dendrogram and using the input data.

A complete specification of this wavelet transform for the data in Table 2 is shown in Table 3.

The principle of "folding" the hierarchy onto an external signal is as follows. The wavelet transform codifies the hierarchy. Having that, we apply the "codification" of the hierarchy with the new, external signal as input.

Wavelet regression entails setting small and hence unimportant detail coefficients to 0 before applying the inverse wavelet transform.

	Sepal.L	Sepal.W	Petal.L	Petal.W
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2

Table 2: First 8 observations of Fisher's iris data. L and W refer to length and width.

	s7	d7	d6	d5	d4	d3	d2	d1
Sepal.L	5.146875	0.253125	0.13125	0.1375	−0.025	0.05	−0.025	0.05
Sepal.W	3.603125	0.296875	0.16875	−0.1375	0.125	0.05	−0.075	−0.05
Petal.L	1.562500	0.137500	0.02500	0.0000	0.000	−0.10	0.050	0.00
Petal.W	0.306250	0.093750	−0.01250	−0.0250	0.050	0.00	0.000	0.00

Table 3: The hierarchical Haar wavelet transform resulting from use of the first 8 observations of Fisher's iris data shown in Table 2. Wavelet coefficient levels are denoted d1 through d7, and the continuum or smooth component is denoted s7.

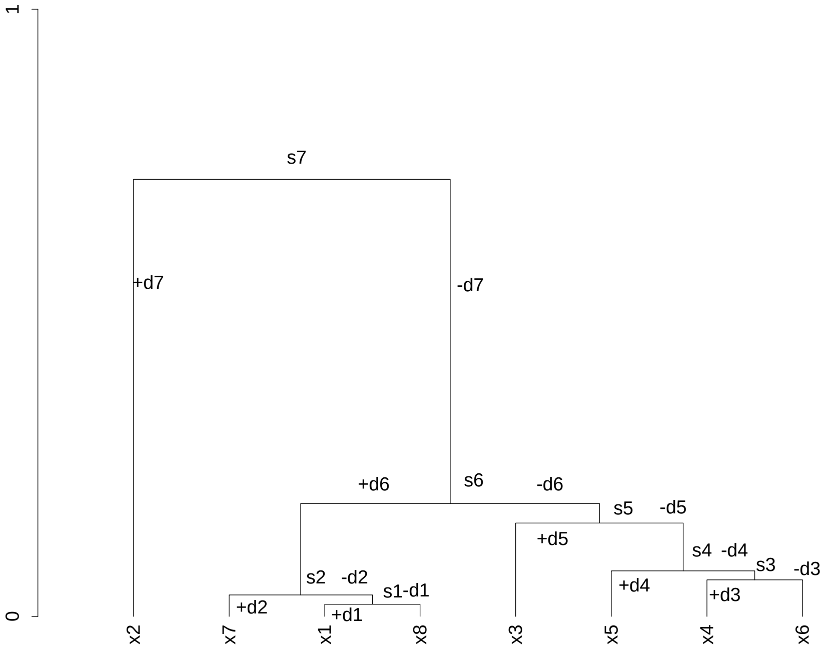

Figure 13: Dendrogram on 8 terminal nodes constructed from first 8 values of Fisher iris data. (Median agglomerative method used in this case.) Detail or wavelet coefficients are denoted by d, and data smooths are denoted by s. The observation vectors are denoted by x and are associated with the terminal nodes. Each *signal smooth*, s, is a vector. The (positive or negative) *detail signals*, d, are also vectors. All these vectors are of the same dimensionality.

More discussion can be found in [18].

References

[1] A. Aussem and F. Murtagh, “A neuro-wavelet strategy for Web traffic forecasting”, Journal of Official Statistics, 1, 65–87, 1998.

[2] J.-P. Benzécri, L'Analyse des Données, Tome I Taxinomie, Tome II Correspondances, 2nd ed. Dunod, Paris, 1979.

[3] B.D. Crane, A. Rex Rovolo and G.C. Comfort, An Empirical Examination of Counterdrug Interdiction Program Effectiveness, IDA Paper P-3219, Institute of Defense Analysis, Alexandria, VA, pp. 104, Jan. 1997.

[4] O. Dube and J.F. Vargas, “Commodity price shocks and civil conflict: evidence from Colombia”, Documentos de CERAC, No. 2, 2006, pp. 62.

[5] P. Fearnhead, “Exact Bayesian curve fitting and signal segmentation”, IEEE Transactions on Signal Processing, 53, 2160–2166, 2005.

[6] P. Fearnhead and Zhen Liu, “On-line inference for multiple changepoint problems”, Journal of the Royal Statistical Society, Series B, 69, 589–605, 2007.

[7] D. Goldin, R. Mardales and G. Nagy, “In search of meaning for time series subsequence clustering: matching algorithms based on a new distance measure”, Proceedings of ACM Conference on Information and Knowledge Management, CIKM (Washington DC, November 2006), ACM Press, New York, 347–356, 2006.

[8] F. Gustafsson, “Segmentation of signals using piecewise constant linear regression models”, Technical report LiTH-ISY-R-1672, Linköpings Universitet, 1994, 29 pp., ftp://ftp.isy.liu.se/pub/rt/Reports/1994/1672.ps.Z

[9] N. Johnson, M. Spagat, J. Restrepo, J. Bohorquez, N. Suarez, E. Restrepo and R. Zarama, “From old wars to new wars and global terrorism”, 2005, http://lanl.arxiv.org/abs/physics/0506213

[10] B. Kamgar-Parsi, B. Kamgar-Parsi and H. Wechsler, “Simultaneous fitting of several planes to point sets using neural networks”, Computer Vision, Graphics, and Image Processing, 52, 341-359, 1990.

[11] E. Keogh and J. Lin, “Clustering of time series subsequences is meaningless: implications for previous and future research”, Knowledge and Information Systems, 8, 154–177, 2005.

[12] D. Lemire, “A better alternative to piecewise linear time series segmentation”, preprint: http://arxiv.org/pdf/cs/0605103v8, in SIAM Data Mining, pp. 545–550, 2007.

[13] Z. Merali, "Here's looking at you, kid. Software promises to identify blockbuster scripts", Nature, 453, 708, 4 June 2008.

[14] G.W. Milligan and M.C. Cooper, "A study of standardization of variables in cluster analysis", Journal of Classification, 5, 181–205, 1988.

[15] F. Murtagh and A.E. Raftery, "Fitting straight lines to point patterns", Pattern Recognition, 17, 479-483, 1984.

[16] F. Murtagh, Multidimensional Clustering Algorithms, Physica-Verlag, Würzburg, 1985.

[17] F. Murtagh, Correspondence Analysis and Data Coding with R and Java, Chapman & Hall/CRC, 2005.

[18] F. Murtagh, "The Haar wavelet transform of a dendrogram", Journal of Classification, 24, 3-32, 2007.

[19] F. Murtagh, G. Downs and P. Contreras, "Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding", SIAM Journal on Scientific Computing, 30, 707–730, 2008.

[20] F. Murtagh, "The Correspondence Analysis platform for uncovering deep structure in data and information", Sixth Annual Boole Lecture in Informatics, Computer Journal, 2008, in press. Advance Access 9 Sept. 2008 (doi:10.1093/comjnl/bxn045).

[21] F. Murtagh, A. Ganz, and S. McKie, "The structure of narrative: the case of film scripts", Pattern Recognition, 42, 302–312, 2009.

[22] M.E.J. Newman, "Power laws, Pareto distributions and Zipf's law", Contemporary Physics, 46, 323–351, 2005.

[23] G.P. Nason and B.W. Silverman, "Wavelets for regression and other statistical problems", In M.G. Schimek, Ed., Smoothing and Regression: Approaches, Computation and Application. New York: Wiley, 159–192, 2000.

[24] J.L. Oliver, P. Carpena, M. Hackenberg and P. Benaola-Galván, "IsoFinder: computation prediction of isochores in genome sequences", Nucleic Acids Research, 32, 287–292, 2004.

[25] T.-Y. Phillips and A. Rosenfeld, "An ISODATA algorithm for straight line fitting", Pattern Recognition Letters, 7, 291-297, 1988.

[26] J. Restrepo, M. Spagat and J.F. Vargas, "The dynamics of the Colombian civil conflict: a new data set", Homo Oeconomicus, 21, 396–428, 2004.

[27] J. Restrepo, and M. Spagat, "Colombia's tipping point?", Survival, 47, 131–152, 2004.

[28] J. Restrepo and M. Spagat, "Civilian casualties in the Colombian conflict: a new approach to human security", preprint, 2004.

[29] P.P. Rodrigues, J. Gama and J.P. Pedroso, "ODAC: hierarchical clustering of time series data streams", SDM 2006, Prof. Sixth SIAM International Conference on Data Mining, 499–503, 2006.

[30] A. Singhal and D.E. Seborg, "Clustering multivariate time-series data", Journal of Chemometrics, 19, 427–438, 2005.

[31] H. Späth, "A fast algorithm for clusterwise linear regression", Computing, 29, 175-181, 1982.

[32] H. Späth, Cluster Dissection and Analysis, Ellis Horwood, Chichester, 1985.

[33] D.A. Stephens, "Bayesian retrospective multiple-changepoint identification", Applied Statistics, 43, 159–178, 1994.

[34] United Nations Office on Drugs and Crime, World Drug Report – Global Illicit Drug Trends, 2008. https://www.unodc.org/unodc/en/data-and-analysis/WDR.html