Trees, Dendrograms and Sensitivity

R.D. Braddock

Cooperative Research Centre for Catchment Hydrology, Griffith University, Nathan, Qld 4111, Australia
(r.braddock@mailbox.gu.edu.au)

Abstract: Dendrograms and minimum-weight spanning trees (MWST) are discrete structures which arise in clustering theory, networks, and where strategic choices are made between discrete options. The discrete structure of the solution tree varies discontinuously with respect to changes in the dissimilarity or cost matrix. Tarjan [1982] has obtained bounds on the changes to the elements of the cost matrix, where the structure is not altered. These results are extended to describe the sensitivity of the tree structure with respect to changes in the cost matrix. The sensitivity of the structure to the addition of a node and arcs is solved using potential loops in the tree. This phase of the solution process builds onto the solution to the original MWST. The full sensitivity analysis of a dendrogram to changes in the attribute matrix is very complex, due partly to the corresponding cost matrix being a function of the full attribute matrix. Statistical methods are used to compare changes in the tree structure with changes in the attribute matrix. Analytical results are obviously difficult to obtain. The Kruskal clustering algorithm is used on the similarity or cost matrix to construct both the MWST and the dendrogram. Thus the sensitivity of the dendrogram to the similarity or cost matrix corresponds to the sensitivity of the MWST for this method of clustering.

Keywords: Sensitivity, Discrete Structures, Dendrograms

1. INTRODUCTION

Computer-based decision support is becoming increasingly important and popular, and the underlying mathematical models need to be thoroughly validated as part of the decision support system. An element in model validation is the sensitivity analysis of the model outputs to changes in the model parameters [Castillo et al., 1997]. Sensitivity analysis techniques can be applied in many areas of knowledge and disciplines where models and computer simulations are used; for example physics, chemistry, environmental sciences, economics and many other areas of application. In most cases, the model output is assumed to vary continuously with respect to the input parameters, and there is a wide range of techniques available to handle such problems [Campolongo et al., 2000]. However, there are models where parts of the output are discrete. The optimum solution may involve strategic choices between discrete options. Examples commonly arise in graph theory, and these include Minimum Weight Spanning Trees (MWST), dendrograms and cluster analysis. In such cases, the similarity (dissimilarity) measure, utility function or cost function may depend continuously on the input parameters. However, the structure of the tree, or options for attaining the solution, is discrete and depends discontinuously on the input parameters [Castillo et al., 1997]. Where the choice of options (the strategy) is important, then the sensitivity of this discrete structure needs to be investigated. This paper investigates the sensitivity properties of dendrograms and minimum-weight spanning trees.

2. BACKGROUND

Dendrograms arise in the study of multivariate data and are a commonly used method of studying clusters and groupings [Krzanowski, 1988]. The analysis is usually based on a data matrix of the form

$$X = \begin{array}{c} \\ \text{Individuals} \end{array} \begin{array}{c} \text{Attributes} \\ \begin{bmatrix} x_{11} & x_{12} & & \\ x_{21} & x_{22} & & \\ & & & \\ & & & \\ & & & x_{nm} \end{bmatrix} \end{array} \quad (1)$$

where the rows refer to individuals and the columns refer to attributes of the individuals. The element $x_{i,j}$ is a measure of attribute j of the i^{th} individual. The nature of the data may be very broad, and the data

may be continuous, discrete or dichotomous, or qualitative. The similarity matrix

$$S = \left\{ s_{i_1, i_2} \right\} \tag{2}$$

is calculated from X, and measures the similarity s_{i_1, i_2} between the individuals i_1 and i_2 on the attributes. There are many similarity metrics which can be used to calculate S, a square $n \times n$ matrix, where n is the number of individuals. The one important property with respect to sensitivity analysis is that S_{i_1, i_2} may depend on all elements of X. Thus a change in any element of X may alter all elements of S. The dendrogram or cluster analysis is constructed from the similarity matrix S, and a variety of clustering techniques are available. A major task is combining two individuals to form a cluster, or the combining of two clusters. Part of this task involves calculating the similarity/dissimilarity between the new cluster and other individuals or clusters. Some of the methods in frequent use include nearest (or furthest) neighbour, group average, centroid, median and minimum variance methods [Krzanowski, 1988]. These methods usually result in the reduction in size, or order, of the dissimilarity matrix, and a recomputation of the dissimilarity value between the new cluster and the previous but unaffected clusters. An example of a dendrogram is shown in Figure 1 [Diamond, 1993, page 21]. The left-hand scale on the figure is in terms of percentage difference, or dissimilarity, between the individuals. Generally, dissimilarity and similarity are related inversely and the dendrograms show increasing scales of dissimilarity.

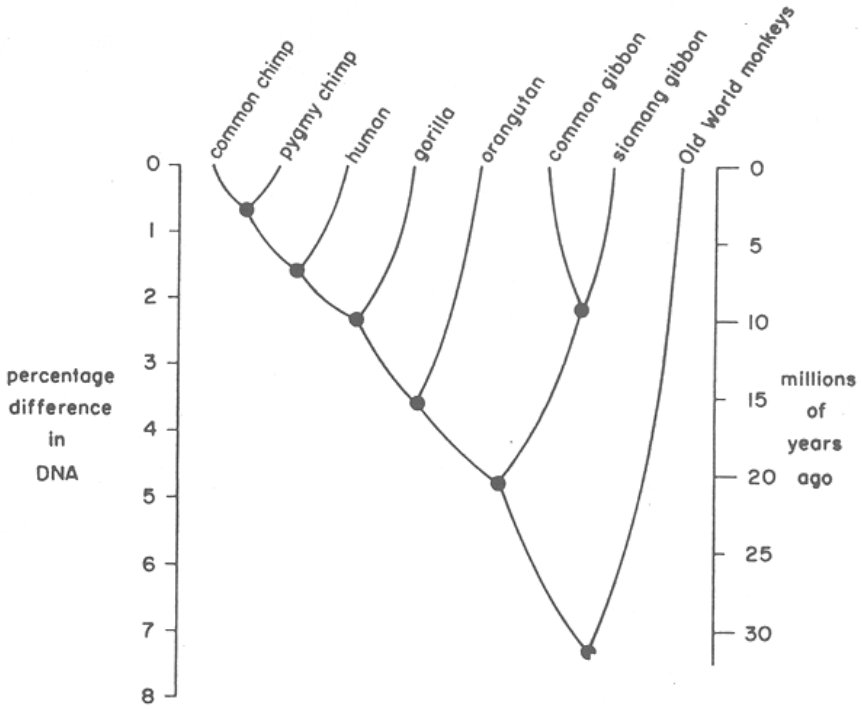

Figure 1. Dendrogram of the DNA of the modern higher primates (After Diamond, 1993). Trace back each pair of modern higher primates to the black dot connecting them. The numbers to the left then give the percentage difference between the DNAs of those modern primates, while the numbers to the right give the estimated number of millions of years ago since they last shared a common ancestor.

The minimum-weight spanning tree (MWST) can be used to obtain a graphical representation of the dissimilarity matrix. The MWST arises in graph theory and plays an important role in the solution strategies of a number of classical operations research problems such as the Travelling Salesman Problem [Bertsekas, 1991]. Consider a set of n nodes and let c_{ij} be the "cost" associated with the edge linking node i to node j. This cost may be the price of flying from node i to node j, as in the Travelling Salesman Problem, or it can be the cost of building or maintaining a physical link between the two nodes. Common examples include communication links or commodity transport [Bertsekas, 1991]. The cost matrix is

$$C = \left\{ c_{ij} \right\}. \tag{3}$$

In the traditional MWST, the costs are usually assessed on economic grounds, and frequently depend on physical features of the terrain along the edge between nodes i and j. There is no analogue to the attribute matrix X, which determines S(X).

A tree is defined as a connected subset of a graph which contains no cycles. A spanning tree contains every node in the network. The MWST is readily computed, and there are several algorithms available [Bertsekas, 1991]. Prim's algorithm builds up the MWST, starting with the cheapest edge, by iteratively adding the edge joining the closest node not yet in the tree. Kruskal's algorithm constructs the MWST, by iteratively choosing the cheapest available edge which does not create a cycle with the edges already chosen. The Kruskal algorithm iteratively selects the edges with shortest cost, but leaves these as separate clusters as it proceeds until the above properties are satisfied.

3. MWST SENSITIVITY ANALYSIS

Consider the undirected graph shown in Figure 2, where the weights along each edge are as shown. The associated MWST is shown by solid lines, called tree edges, and the dashed lines are non-tree edges which are not in the minimal tree. The corresponding cost matrix is

$$C = \begin{bmatrix} 0 & 3 & 5 & a & a & a & a & a \\ 3 & 0 & a & a & a & 7 & a & a \\ 5 & a & 0 & a & a & 9 & 6 & 1 \\ a & a & a & 0 & 2 & a & 4 & 9 \\ a & a & a & 2 & 0 & a & a & a \\ a & 7 & 9 & a & a & 0 & a & a \\ a & a & 6 & 4 & a & a & 0 & a \\ a & a & 1 & 9 & a & a & a & 0 \end{bmatrix} \quad (4)$$

where a is a large positive number that reflects the absence of an edge. Technically, $a = \infty$ to reflect the non-existence of the edge. In practice, a is larger than the cost of any existing link, and the MWST algorithms will operate successfully. The corresponding nearest-neighbour dendrogram is shown in Figure 3, where cost (instead of dissimilarity) is shown on the vertical axis.

The Minimum Weight Spanning Tree structure is sensitive to the costs of both the tree and non-tree edges. Changes to the costs of any edge may lead to discrete changes in the tree as the solution. The analysis is based on the potential cycles in the graphs. Consider the potential cycle of nodes 8, 3, 7 and 4 and the possible range of values of the cost c(4,8) on the non-tree edge 4-8. In this potential cycle, c(4,8) is larger than the maximum value of the other edges in this potential cycle. If c(4,8) = 5, then edge 4-8 will enter the tree and edge 3-7 (with c(3,7) = 6) will be dropped from the tree.

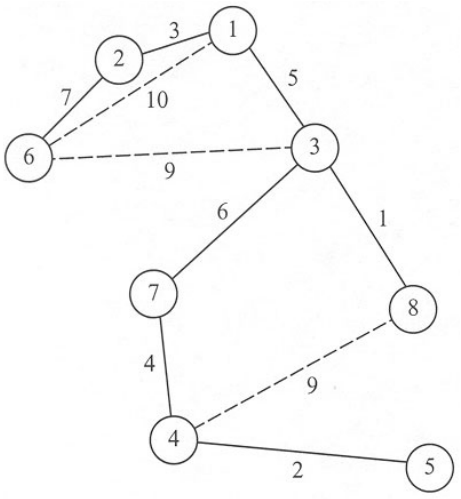

Figure 2. Sample graph and the associated minimum-weight spanning tree (shown as solid lines)

Now consider the tree edge 4-7 and the cost c(4,7); here c(4,8) = 9. Then for c(4,7) < 9, edge 4-7 remains in the tree. However, for c(4,7) > 9, then edge 4-7 is deleted from the tree and edge 4-8 is added. The structure of the tree is altered.

The above changes have been discussed with respect to just one potential cycle in Figure 1. Where more potential cycles are present, i.e. nodes 1, 2, 6 and 3, the above needs appropriate modification. The algorithm developed by Tarjan [1982] uses a transmuter matrix to compute the potential cycles and the relevant bounding values of the edges, for the edges to enter or leave the tree structure.

Now consider the problem of adding an edge to an existing set of nodes. Thus edge 6-7, with cost c(6,7) = 6 (say) is to be added to the graph. In a communications network, this is equivalent to adding a new communication link between two cities (nodes) in the network. This introduces three potential cycles into the graph in Figure 2; the potential cycle 1-2-6-7-3-1 is the important one which needs to be broken. For c(6,7) = 6, then the highest cost edge is 2-6, and the edge 2-6 is removed from the tree. Where c(6,7) is less than the largest cost edge in the cycle, then the tree structure is altered. Where c(6,7) is larger than the largest cost edge in the cycle, then the tree structure is unaltered. This problem is readily handled by performing additions to the transmuter matrix of Tarjan.

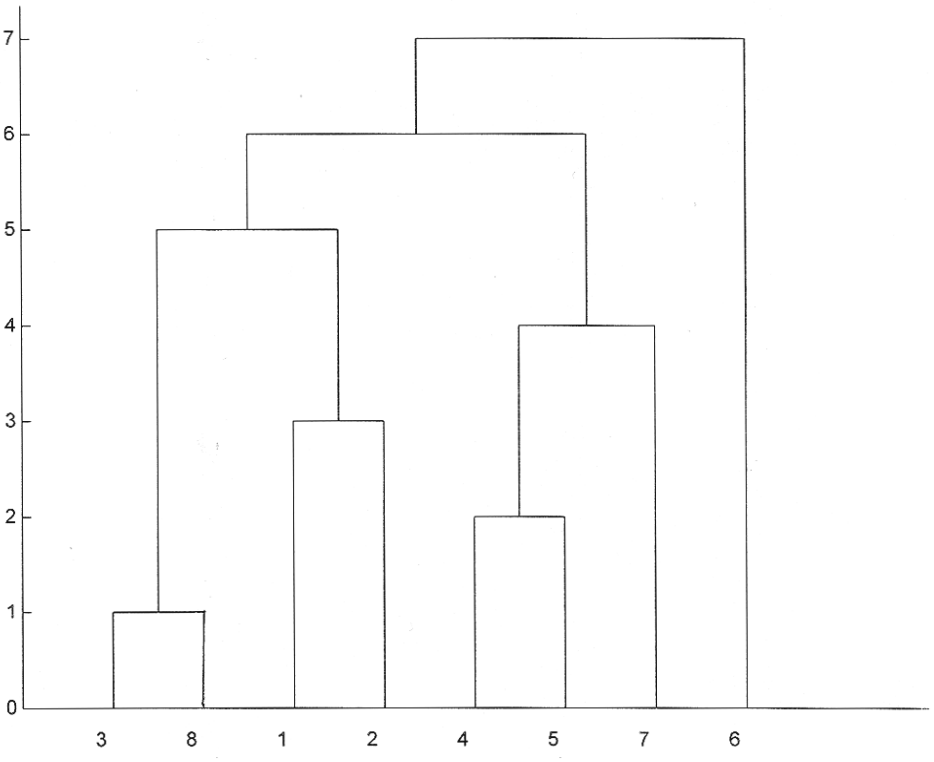

Figure 3. Dendrogram corresponding to the graph in Figure 2 and the dissimilarity matrix C (Equation 4).

The addition of an extra node to the graph can be handled using a two-step process. Consider the graph formed by adding a new node, say node 9, to the graph of Figure 2. The additional edges and edge costs are as shown in Figure 4. The first stage is to attach the new node to the tree, using the lowest cost edge between node 9 and the existing tree in Figure 2. This is essentially the incremental step in the Kruskal algorithm, and adds the edge 4-9 to the tree.

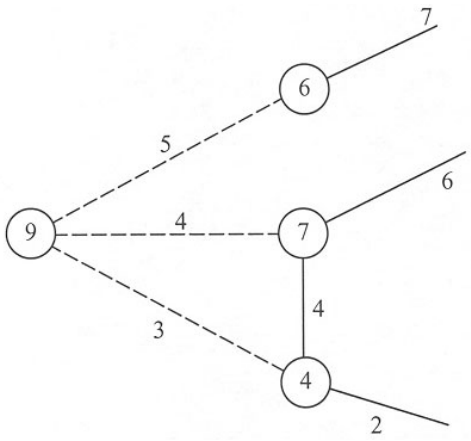

Figure 4. Addition of a node and weighted edges.

The second stage is to consider the resulting potential cycles and adjust the tree to ensure minimal weight. These cycles are

(a) 9, 4, 7, 9;
(b) 9, 4, 7, 3, 1, 2, 6, 9; and
(c) 9, 7, 3, 1, 3, 2, 6, 9.

In cycle (a), the new edge 4-9 is of minimum cost, and the remaining links are of equal cost. Either link can be discarded. In this case, we retain the edge 4-7, and this action also handles case (c). Note that if $c(7,9) < c(7,4)$, then the previous tree edge 4-7 would be removed. In case (b), the edge 2-6 is the highest cost in the cycle. This edge is then deleted from the tree and the edge 6-9 is added. The final spanning tree is shown in Figure 5, and the corresponding dendrogram is shown in Figure 6.

In these simple additions to the graph, the previous optimal solution can be used as a starting point for the extended solution or tree. This is important in practical applications where the graph or pre-existing tree may be extensive.

4. SENSITIVITY OF DENDROGRAMS

The sensitivity analysis of a dendrogram is complex, due to several steps in the construction of the discrete cluster diagram. The attribute matrix, X, is used in the calculation of the similarity matrix S. A variety of metrics can be used in this calculation, including a Euclidean similarity distance measure. In this case, a change in any element of X can alter the

values of all the elements in S. For other distance measures, such as the infinity or maximum element norm, perturbations to an element of X may have no effect on S. The second area of complexity arises in the agglomeration process, where individuals and groups are fused. At the fusion, the current similarity matrix is recalculated and reduced in size. This makes it extremely difficult to follow the effects of changes of the attribute matrix into the dendrogram. The method of agglomeration also affects the structure and properties of the resulting dendrogram, i.e. the centroid method leads to "spherical" clusters with high internal affinity, while the nearest neighbour scheme produces chaining [Krzanowski, 1988]. The third area of complexity arises from the discarding of information by the agglomeration process. Information on nearest neighbour affinity is retained, but information on secondary affinity, or second nearest neighbour, is lost [Krzanowski, 1988; Bertsekas, 1991]. In terms of the MWST , the costs (similarities) of the tree edges are retained, but information on non-tree edges is lost.

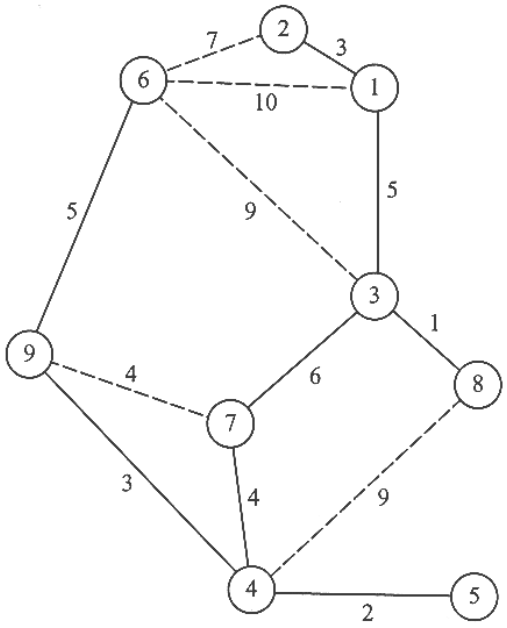

Figure 5. Final spanning tree for the extended graph.

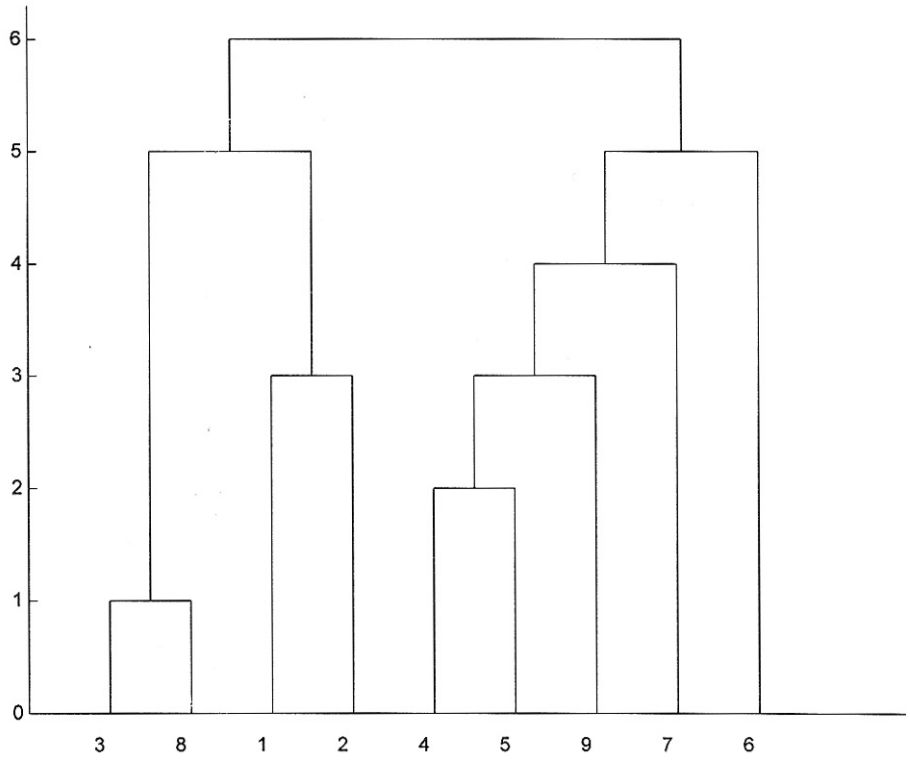

Figure 6. The dendrogram corresponding to the nine-node spanning tree.

The detailed sensitivity analysis of the dendrogram is not generally possible. The Kruskal algorithm can be used on the similarity matrix to generate both the MWST and the corresponding dendrogram. This would also require that the corresponding similarity matrix is not recalculated at each fusion, and thus that non-tree-edge information is preserved. The Kruskal algorithm proceeds by selecting the lowest cost arc in the network and including it in the potential tree [Bertsekas, 1991]. Further arcs, not yet in the potential tree, are selected on the basis of minimum cost, and are added to the potential tree. Cycles are not

permitted, and the algorithm proceeds until all nodes are connected. The arcs may form isolated clusters at each iteration except the last. There is a one-to-one correspondence between the dendrogram with nearest neighbour clustering and the isolated clusters in the iterative stages of the Kruskal theorem. Consider the tree in Figure 2 and the nearest neighbour dendrogram in Figure 3. The lowest cost arc in the tree construction is the arc 3-8. This is the arc selected at the first iteration of Kruskal's algorithm; it is also the first fusion point in the dendrogram. The second fusion is between 4 and 5 on the dendrogram, corresponding to arc 4-5 on Figure 2. Drawing any horizontal line across the dendrogram in Figure 2 gives, under the line, the fusions of the dendrogram corresponding to the isolated clusters of the Kruskal algorithm. Thus the sensitivity analysis of the MWST can be immediately applied to the nearest neighbour dendrogram.

The Prim algorithm can also be used to estimate the MWST, but with the observation that the full MWST or dendrogram needs to be computed first, before the development of the clusters can be considered.

Addition of an individual to X will also affect the dendrogram structure. The effects of adding the ninth node in Figure 4 are illustrated by the dendrograms in Figures 3 and 6. The cluster of nodes 1, 2, 3 and 8 are not affected by the addition of node 9. However, the addition of node 9 does affect the right-hand cluster of nodes 4, 5, 6 and 7. Node 6 is brought into the right-hand cluster, the topology of the tree is changed, and some of the fusions occur at lower values of the dissimilarity.

The difficulties in dealing with the dependence $S = S(X)$, as well as the problems discussed above, have led to the comparison of the dendrograms derived from the original data X and the perturbed data $X + \Delta X$. Unfortunately, a change in any element of X may affect all of S, the similarity matrix, and hence may alter the full structure of the MWST and the dendrogram. This leads to the problem of comparing and contrasting two dendrograms (or minimum weight spanning trees). This is a difficult and mainly unsolved problem. Most of the work on comparing dendrograms is rooted in the biological literature [Lapointe and Legendre, 1995; Berntson, 1995; van Pelt, 1997]. Monte Carlo methods are often used to derive probability distributions against which the statistical significance of variations in dendrograms are assessed. The reader is referred to the literature cited above for details.

5. CONCLUSION

The sensitivity analysis of a MWST is easily handled using the graph theoretic properties which need to be satisfied by the spanning tree. The sensitivity properties can be calculated using the concept of the potential cycle in the graph. Larger scale problems are readily handled using the transmuter matrix approach of Tarjan [1982] in considering one-at-a-time changes to the graph. Addition of nodes or arcs can also be handled using the same potential cycle approach. Here, the current MWST can be used as the starting point for the handling additions to the graph. This implies an economy of calculation for small additions (or subtractions) to the graph. In these cases, the transmuter matrix approach is less helpful as it needs to be constructed anew when new nodes are added. The sensitivity of a dendrogram may draw on the MWST results in a limited set of circumstances, i.e. nearest-neighbour agglomeration. In general, the variations to the attribute matrix X need to be considered by totally recomputing the dendrogram for each data perturbation. Some statistical work has been done, but the results are not conclusive.

6. REFERENCES

Berntson, G.M., The characterisation of topology: a comparison of four topological indices for rooted binary trees, *J. Theoretical Biology, 177*, 271-281, 1995.

Bertsekas, D.P., *Linear Network Optimisation, Algorithms and Codes*, MIT Press, Cambridge, Massachusetts, 1991.

Campolongo, F., A. Saltelli, T. Sorensen, and S. Tarantola, Hitchhiker's Guide to Sensitivity Analysis, In Saltelli, A., K. Chan, and M. Scott (eds) *Sensitivity Analysis*, Wiley, pp. 15-47, 2000.

Castillo, E., J.M. Gutierrez, and A.S. Hadi, Sensitivity analysis in discrete Bayesian networks, *IEEE Transactions on Systems, Man and Cybernetics, 26*(7), 412-423, 1997.

Diamond, J., *The Third Chimpanzee*, Harper Collins, 1993.

Krzanowski, W.J., *Principles of Multivariate Analysis; a User's Perspective*, Oxford University Press, 1988.

Lapointe, F.J., and P. Legendre, Comparison tests for dendrograms: a comparative evaluation, *Journal of Classification, 12*, 265-282, 1995.

Tarjan, R.E., Sensitivity analysis of minimum spanning trees and shortest path trees. *Information Processing Letters, 14*, 30-33, 1982.

van Pelt, J., Effect of pruning in dendritic tree topology. *Journal of Theoretical Biology*, *186*, 17-32, 1997.