BORNEO SCIENCE 8:57 - 63, DECEMBER, 2000

CLUSTERING ALGORITHM FOR TREE-RING SAMPLES

H.J. Zainodin & Lo Li Lan

School of Science & Technology, Universiti Malaysia Sabah
Locked Bag 2073, 88999 Kota Kinabalu, Sabah, MALAYSIA

ABSTRACT. *Every tree-ring sample needs to be dated correctly before more advanced work or other analyses can be done in developing floating or site chronology of a region. In this paper, a time saving clustering algorithm is successfully used in dendrochronological analyses presented through several data sets. This is presented using a specific similarity measure. The illustration demonstrates the cluster formation using consistent and outlying criteria. So far, with this combined criteria no sample has been incorrectly dated.*

KEYWORDS. Clustering, tree-ring, cross matching, similarity measure.

INTRODUCTION

Useful information of past environment and current exploited environment can be obtained from the remains of houses, ships, bridges and platforms that are in the form of timbers. Having these timbers dated properly and placed at the correct time period does this.

A well-established technique called dendrochronology is used in order to achieve this purpose. It uses the widths of the annual growth rings of trees to date timbers from buildings, waterfront structures, etc. (a cross-section of a timber with visible annual rings is shown in Figure 1). Once every tree-ring sample has been dated correctly, more advanced work or other analyses can be done in developing floating or site chronology of a region. Examination of this material using dendrochronological techniques, in particular, relative (floating) and master chronology building, should enable a clear understanding of the practices used by past communities to capitalise on the materials, coppiced, woodland around them (Cook & Kairiuktis, 1990).

Statistical analysis is complex due to the methods used and the type of wood examined. A few methods have been tried successfully but we would like to get the results faster and more accurately. This paper exposes the effects of similarity measures with proposed simple penalty function in order to gain considerable insight into the intricacies of the data.

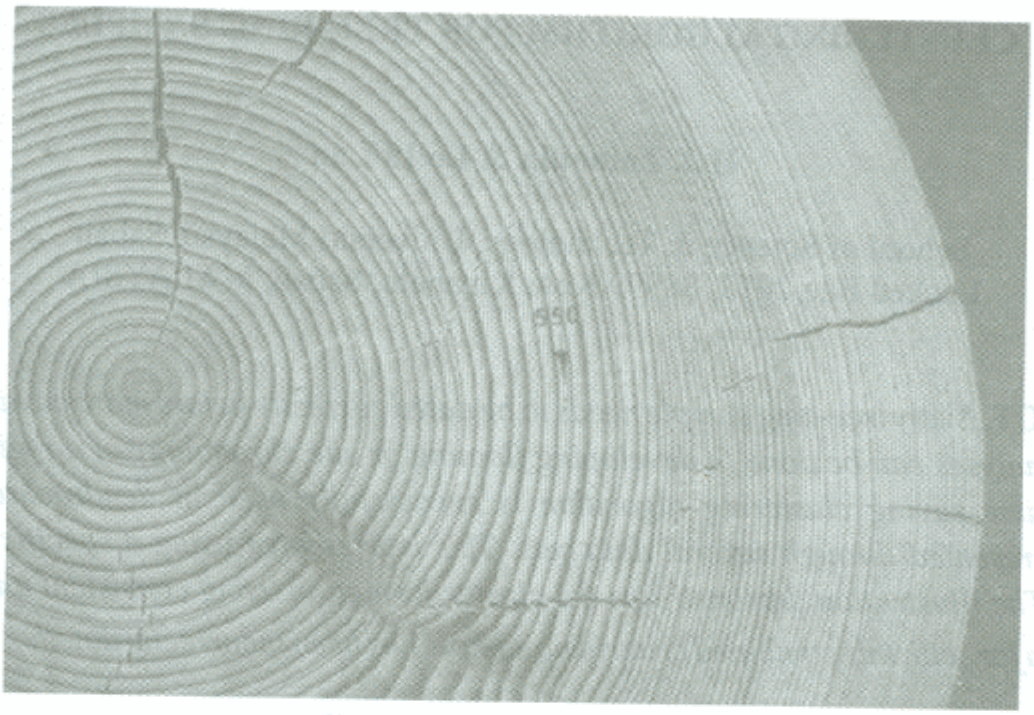

Figure 1: Cross-section of a timber

PROPOSED SIMILARITY MEASURE

To group tree-ring samples into their respective groups, we use a similarity measure in the clustering algorithm. Similarity measures are used to reflect the strength of the match between two sequences at each possible relative position. A few similarity measures have been successfully used previously to cluster the tree-ring samples into groups. It would be better if this could be done in a shorter time (Munro, 1984). Hence, the objective of this paper is to produce a timesaving clustering algorithm that cluster the tree-ring samples into groups.

The filter suggested in Laxton, Zainodin & Greig (1996) is applied to all the samples involved in this work. The final sequences of indices are slit to each other and search for the highest similarity value. A similarity function with a simple penalty function is used to present this clustering algorithm and it is successfully used in this paper. We will show that this method can also be used for other tree-ring samples. Below is the proposed similarity measure that is used in this work.

Consider n pairs of overlapping indices (t = 1, 2, 3, …, n) from any two sequences, X and Y (n_i is the length of sequence X, n_j is the length of sequence Y), at relative position p. Let r(i, j, p) be the correlation coefficient between sequences i and j at relative position p, and s(i, j, p) be the similarity value between samples i and j at relative position p. We can find the value of z using equation (1).

$$z = \frac{\sqrt{n-3}}{2} \ln\left(\frac{1+r(i,j,p)}{1-r(i,j,p)}\right) + \frac{\sum_{t=1}^{n-1} D_t}{n-1} \qquad (1)$$

The value of D_t is defined for the following cases.

Case 1:

For sign $(X_{t+1} - X_t) = \text{sign}\,(Y_{t+1} - Y_t)$. We have two possibilities.
If $||X_{t+1} - X_t| - |Y_{t+1} - Y_t|| < 1/2$ then assign $D_t = 1$ or
if $||X_{t+1} - X_t| - |Y_{t+1} - Y_t|| \geq 1/2$ then assign $D_t = 1/2$

Case 2:

For sign $(X_{t+1} - X_t) \neq \text{sign}\,(Y_{t+1} - Y_t)$
If $|X_{t+1} - X_t|$ and $|Y_{t+1} - Y_t| < 1/2$ then assign $D_t = -1/2$.
Otherwise, $D_t = -1$.

The proposed similarity measure use in this work is $s(i, j, p) = 2z_{(1)} - z_{(2)}$ where $z_{(1)}$ is the highest value of z (defined in equation (1)) and $z_{(2)}$ is the second highest value of z.

CLUSTERING ALGORITHM

As mentioned earlier, the raw data need to be transformed so that all the required assumptions are fully satisfied. The indices obtained from the transformation will go through the following steps until all the related samples are grouped into the respective group.

Step 0: Set s^* at a high initial value.

Step 1: Compare samples i and j for $i = 1, 2, \ldots, M \quad 1$; $j = i + 1, \quad i + 2, \ldots, M$ and evaluate similarity values $s(i, j, p)$ for $p = -n_j + q, -n_j + q + 1, \ldots, -1, 0, 1, \ldots, n_i q \quad 1, n_i \quad q$.

Step 2: Sort the current similarity values into order (highest to lowest).

Step 3: Consider the largest similarity value listed in step 2. If this similarity value is less than s^* then stop. Otherwise merge the two corresponding sequences into a group. Let M^* be the number of sequences remaining at this stage. Now if M^* is greater than 1 consider the K-th largest similarity value say $s(i, j)$ in the list, for $K = 2, 3, \ldots, M^*(M^* - 1)/2$.

If this $s(i, j)$ is less than s^*, then proceed to step 4. Otherwise proceed as follows:

3.1 If both the sequence i and j are NOT present in any previously formed groups, merge the two sequences into a group.

3.2 If only one sequence, say i, is present in one of the previously formed groups consisting of k samples labelled $(i_1, i_2, \ldots, i_k)$, and other sequence,

j, is not a member of any group, then sequence j will be merged into the group provided that $p(i, l) = p(i, j) + p(j, l)$ for $l = 1, 2, \ldots, K$ $(l \neq I)$. If this is not the case then sequence j is not included in the group.

3.3 If one sequence is present in any one of the previously formed groups and the other is present in another group then the sequences from both groups are merged into one group provided the relative position of each sequence from one group with respect to each sequence from the other group occurs when the corresponding similarity measure is maximum. In other words provided that $p(i, k) = p(i, j) + p(j, k)$ for all sequences i, j and k in the proposed new group. If this is not the case then these groups remain as they are.

Step 4: Calculate the values of the sequences representing each group.

Step 5: Update the entries of the list of similarity values by

(a) deleting all entries involving either sequence i or sequence j,

(b) adding new entries for the similarity values, between the sequences just formed and between them and the remaining sequences.

Step 6: If M^* equals 1 then stop. Otherwise repeat steps 2, 3, 4, 5 and 6.

If no more groups are formed, restart with these groups that have just been formed but now use a lower value of s^* (usually reduced by 0.5). Repeat this procedure gradually reducing the value of s^* until it reaches some pre-set value (usually 3.5).

RESULTS

In this work, 9 samples from Thoresby Estate, England were used for illustration purposes. Detailed of the samples are

Table 1. Samples from Thoresby Estate, England.

SAMPLE	Number of Ring	DATE		OFFSET (Relative Position)
		First Ring	Last Ring	
THO-A01B	167	1810	1976	0
THO-A02A	158	1819	1976	9
THO-A03A	155	1822	1977	12
THO-B01A	158	1820	1977	10
THO-B02B	158	1820	1977	10
THO-B03B	156	1822	1977	12
THO-B04A	144	1834	1977	24
THO-B05A	143	1835	1977	25
THO-O01C	145	1832	1976	22

Three sequences i, j and k are matched with each other as a group. If sequences i and j match "best" at relative position p(i, j), and sequences j and k match "best" at relative position p(j, k), then sequences i and k match best at relative position p(i, j) + p(i, k). Therefore, p(i, k) = p(i, j) + p(j, k). This is shown in Figure 2. The maximum s-values are consistent at the corresponding offsets so that they are strictly coherent.

Using the information in Table 1, group of sequences cross-matched to illustrate the coherent group. To do this, two samples are taken at a time. Altogether there are ${}^{9}C_{2}$ pairs, which are 36 pairs. Consider the dendrogram shown in Figure 3 for the Thoresby data. The grouping process begin with a high s^* and subsequently reduced (usually by 0.5) to the threshold value.

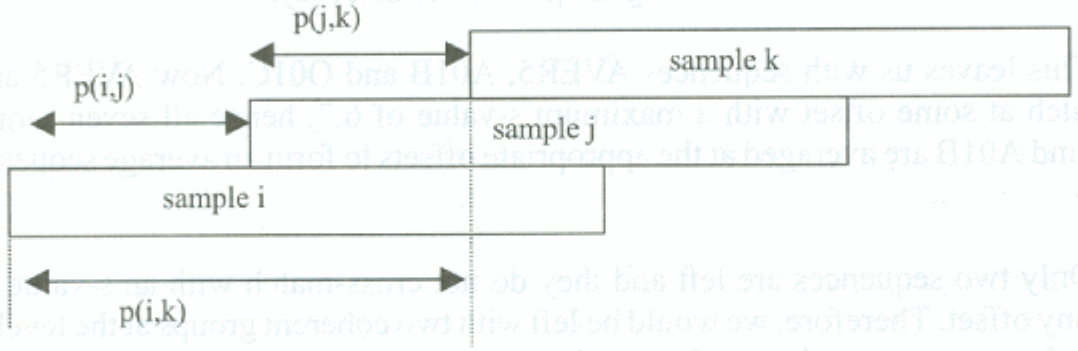

Figure 2. Strictly Coherent Group for the 3 samples.

First, we set $s^* = 7.0$, only A02A and A03A matched and formed a group where an average was taken, call it AVER1. This happened at the offset $p(A02A, A03A) = +3$, where the s-value is 8.1 which is maximum. Eight sequences, AVER1 and all the other sequences except A02A and A03A, are now left to be cross-matched.

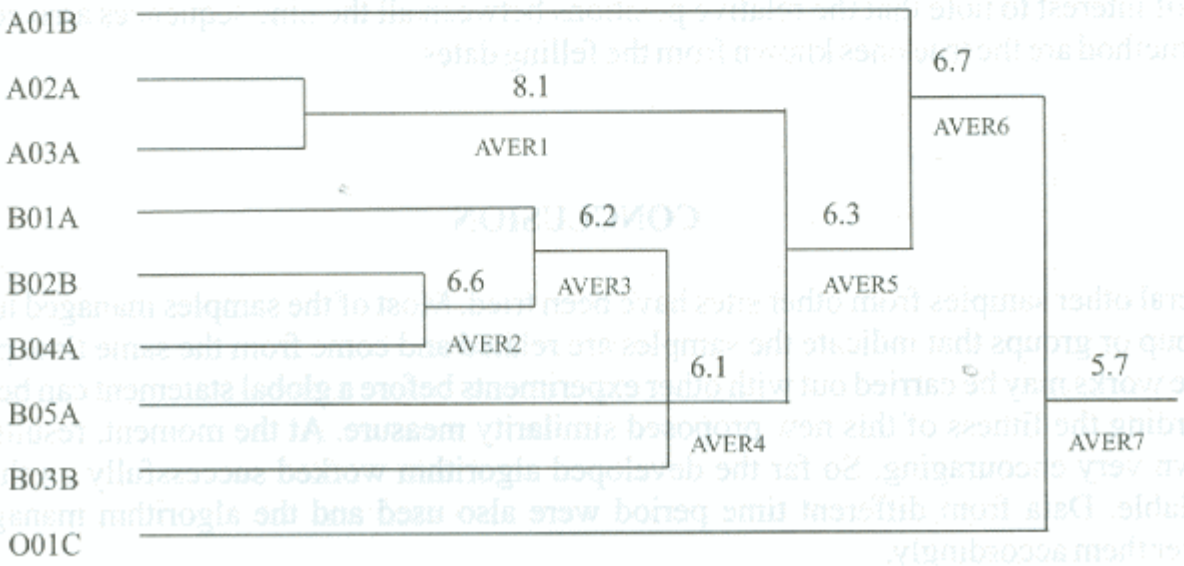

Figure 3. Final Group of the Samples.

Now put $s^* = 6.0$. First, only B02B and B04A cross-match at this level with a maximum s-value of 6.6 and formed an average, call it AVER2. Then, B01A cross-matches AVER2 at some offset (with a maximum s-value of 6.2, see Figure 3) so the two sequences, AVER2 and B01A, are matched and averaged at the appropriate offsets to form a new sequence, say AVER3. So at this matching level of $s^* = 6.0$, the sequences are reduced to six, which are AVER1, AVER3, A01B, B05A, B03B and O01C.

Cross matching between sequences continue at the level $s^* = 6.0$. The five sequences B01A, B02B, B04A (in AVER3), B05A and B03B are averaged at the appropriate offsets to form an average, AVER4, say. Now, we are left with sequences AVER1, AVER4, A01B and O01C. Then, AVER1 and AVER4 cross-match at some offset with a maximum s-value of 6.3 so the seven component sequences in AVER1 and AVER4 are now averaged at the appropriate offsets to form a new average sequence, AVER5, say.

This leaves us with sequences AVER5, A01B and O01C. Now AVER5 and A01A cross-match at some offset with a maximum s-value of 6.7, hence all seven sequences in AVER5 and A01B are averaged at the appropriate offsets to form an average sequence, call it AVER6.

Only two sequences are left and they do not cross-match with an s-value of 6.0 or more at any offset. Therefore, we would be left with two coherent groups at the level $s^* = 6.0$, one of eight sequences and one of one only (notice that the coherent group of eight is not strictly coherent, see Figure 3).

Finally, s^* is decreased to 5.0 where AVER6 and O01C will cross-match at some offset with a maximum s-value of 5.7. One coherent group consisting of all nine sequences are formed and a site sequence has been produced.

We have shown that the nine Thoresby sequences as in Table 1 form a coherent group. It is of interest to note that the relative positions between all the nine sequences arrived at by this method are the true ones known from the felling dates.

CONCLUSION

Several other samples from other sites have been tried. Most of the samples managed to form a group or groups that indicate the samples are related and come from the same time period. More works may be carried out with other experiments before a global statement can be made regarding the fitness of this new proposed similarity measure. At the moment, results have shown very encouraging. So far the developed algorithm worked successfully on the data available. Data from different time period were also used and the algorithm managed to cluster them accordingly.

REFERENCES

Cook, E.R. & Kairiuktis, L.A. 1992. *Methods in dendrochronology: Applications in the Environmental Sciences*. Dorderecht: Kluwer Academic Publishers.

Laxton, R.R., Zainodin, H.J. & Greig, B.J.W. 1996. Model for the decline & dieback of oaks: Based on an analysis of ring widths. *Radiocarbon*: 427-436.

Munro, M.A.R. 1984. An improved algorithm for cross-dating tree-ring series. *Tree-ring Bulletin*. **44**: 17-27.