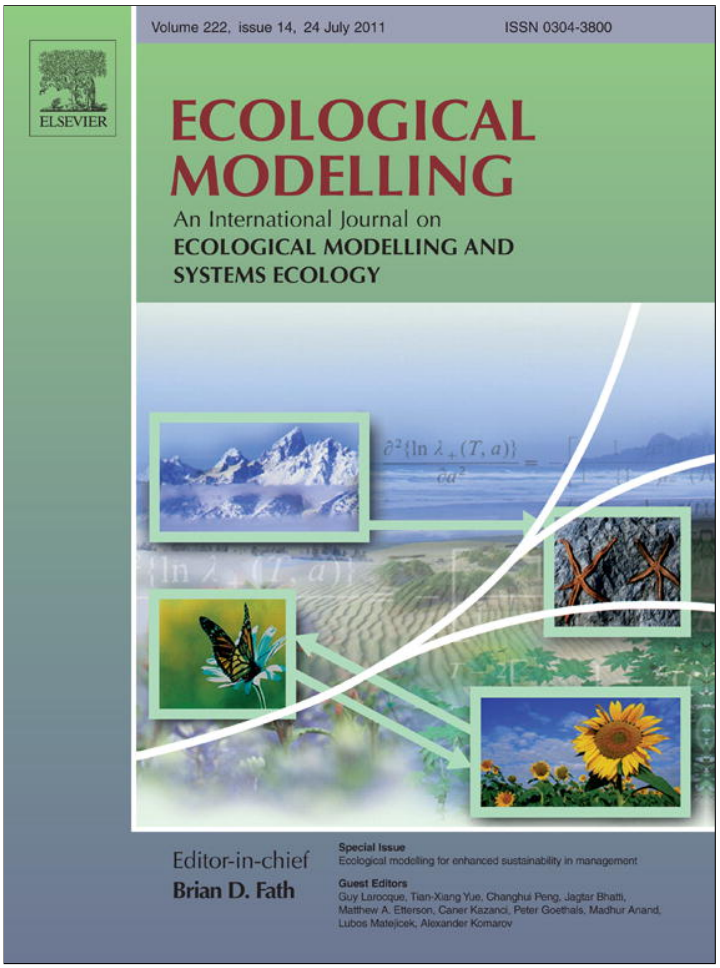

Ecological Modelling 222 (2011) 2524–2529

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Analysis of time series data on agroecosystem vegetation using predictive clustering trees

Marko Debeljak[a,*], Geoffrey R. Squire[b], Dragi Kocev[a], Cathy Hawes[b], Mark W. Young[b], Sašo Džeroski[a]

[a] Jozef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[b] Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK

ARTICLE INFO

Article history:
Available online 23 November 2010

Keywords:
Time series
Agroecology
Weeds
GM
Herbicide tolerance
Oilseed rape
Predictive clustering trees

ABSTRACT

We present an approach to modelling interdependent types of vegetation that support different functions in a managed ecosystem. For optimal management, plants that provide economic output (e.g., crops) and those that support ecological functions (e.g., wild plants or 'weeds') should coexist in an agroecosystem. To make progress with understanding how such plant communities interact over time, we analyse paired time series data about the cover of crop and weed vegetation in oilseed rape fields. The percentage crop and weed cover were measured every 7–14 days at 128 sites in the UK, covering a wide range of localities and management regimes.

To analyze the data, we first cluster the time course profiles of crop cover (using the *k*-medoids clustering algorithm and the dynamic time warping distance between time series). The clustering revealed five typical clusters of crop cover profiles that differed in terms of rate of increase, lag phase and maximum value, but were largely independent of the type of crop (winter/spring oil seed rape) and the weed management regime. Cluster membership for each crop cover profile was used as an additional independent variable (attribute) in the predictive modelling analysis that followed.

We then constructed predictive clustering trees (a generalized form of decision trees) that predict the weed cover profile (time series) from independent (input) variables that include the crop cover cluster, other crops descriptors and environmental variables. The crop cover cluster was more informative in predicting the weed cover profile than any other input variable, including the type of crop and the crop transgenic status (conventional or genetically modified / herbicide tolerant). The approach was successful in identifying the interdependencies between the two types of vegetation. We envisage that it will have plentiful further practical use in relating and interpreting ecological or environmental time series.

1. Introduction

It is increasingly recognised that production ecosystems have to satisfy several functions or outcomes in the same physical space (Marshall et al., 2003; Hawes et al., 2009). They must balance, for example, the needs of production leading to economic offtake with the provision of sufficient energy, matter and structure to allow essential ecological processes to be sustained. The ecological need is for enough wild plants (weeds) to support an arable food web consisting of a diversity of invertebrate functional types, including herbivores, omnivores, pollinators and predators (Hawes et al., 2009). The economic need is for vigorous growth of the crop together with minimal negative effects on offtake caused by competition from the weeds. Ideally, cropping should not reduce weeds to the point where the essential food web of arable land is impaired, while weeds should not reduce the economic output of the land.

At least a part of this co-habitation between crops and weeds can be achieved by spatially separating them, maintaining areas such as field margins, boundaries or woodlands, where biodiversity is the primary aim, distinct from the disturbed, cultivated centres of fields. Since, however, the cultivated areas tend to be by far the larger in size, and support a distinct flora, temporal co-habitation within the cultivated fields is also essential (Squire et al., 2009). A range of indicators in the soil seedbank, emerged vegetation, and sedentary and mobile invertebrate groups are being developed to assess the relevant ecological and economic processes within the cultivated areas (Hawes et al., 2010). There are

* Corresponding author at: Jozef Stefan Institute, Department of Knowledge Technologies, Jamova 39, 1000 Ljubljana, Slovenia. Tel.: +386 1 477 3124; fax: +386 1 477 3315.
E-mail addresses: Marko.Debeljak@ijs.si (M. Debeljak), Geoff.Squire@scri.ac.uk (G.R. Squire), Dragi.Kocev@ijs.si (D. Kocev), Mark.Young@scri.ac.uk (C. Hawes), Cathy.Hawes@scri.ac.uk (M.W. Young), Saso.Dzeroski@ijs.si (S. Džeroski).

0304-3800/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2010.10.021

major questions, however, as to how these indicators are related to each other. For instance, time profiles or time series of cohabiting groups such as crops and weeds may provide more useful information than occasional measurements. There may be an interaction at one point in the series that has ramifications for the subsequent development of one or more series, while the multiple functions might be satisfied through temporal complementarity — for example, the presence of weeds for only a part of the year might satisfy the ecological functions while having little effect on yield.

If time series for the appropriate indicators could be compared quantitatively, then several questions of practical significance could be addressed. For example, to what extent are the time series for crops and wild plants associated or influenced by each other? Can ideal or near-ideal crop and wild plant time series be identified (and can non-ideal series be identified)? If so, can they be attributed to any causal factors of the site or the agronomy? Can forms of management be devised to achieve the optimal combination of time series? Before such questions can be considered, the feasibility of comparing ecological time series has to be demonstrated and a methodology tested on suitable data.

Quantification and comparison of time series are major problems, however. Typically, the measurements in an ecological time series are taken at irregular time intervals. The time series may have different lengths and differ non-linearly in the intervals between successive time points. Often the data are too variable to allow one series to be related in a simple or direct way to the other. In such a case, it may be possible to look for similarities among time series, i.e., whether the entire set of time series can be partitioned into groups or clusters, such that the time series within a group are more similar to each other than to those in other groups. The cluster, rather than the individual time series, may then be used as an input variable to predict the time series of the second variable.

There are several methods that analyse and cluster time series data from the domain of environmental sciences (Liao, 2005; Shumway and Stoffer, 2006). To model the time series data these approaches typically use hidden Markov models, neural networks, genetic programming, regression base approaches (e.g., autoregressive integrated moving average) or analyse the data in the frequency domain. They are mainly concerned with forecasting weather conditions (e.g. rainfall), predicting river water level (flood protection), analysis of temporal remotely sensed data about land use and land cover (Mari and Le Ber, 2006; Potter et al., 2007; Viovy, 2000; Zhou et al., 2006), etc. For example, Li et al. (2001) perform clustering on time series data using hidden Markov models. The data they analyze concern the ecological effects of mosquito control by changing the drainage patterns in an area south of Brisbane, Australia. However, the aforementioned approaches have some limitations regarding the type of variables that can be used, make assumptions about, prior distributions or missing values and offer limited interpretability. We propose to use predictive clustering trees that have no such prior assumptions and are readily interpretable.

This paper applies several procedures, culminating in predictive clustering trees, to analyse a comprehensive data set consisting primarily of two time series—one for the percentage ground cover of a crop, which is an indicator of the potential production of a crop stand and the other for percentage ground cover of weeds, which is an indicator of both the potential negative effects on the crop and positive effects on the food web. The main aim of the paper is to establish whether an approach to the examination of time series on large agroecological data sets is feasible. In doing this, the paper provides the first application of predictive clustering trees to the analysis of ecological time series in agricultural systems.

2. Methods

2.1. Distance-based clustering of time series

In this study, we perform distance-based clustering of ecological time series. To achieve this, we need a distance measure on time series. In the data examined here, measurements of crop cover and height, as well as weed cover, occur at irregular time intervals, while the time series at each location begins and ends at different time points in the year. This has motivated us to use the dynamic time warping (DTW) distance between two time series.

Dynamic time warping (Sakoe and Chiba, 1978) can capture non-linear distortion along the time axis. It accomplishes this by assigning multiple values of one of the time series to a single value of the other. As a result, DTW is suitable if the time series are not properly synchronized, e.g., if one is delayed, or if the two time series are not of the same length, or if the measurements refer to different time points.

Once we have a distance measure, we can cluster the time series, i.e., partition the entire set of time series into groups. We want the time series within a group to be similar to each other and time series in different groups to be different/distant from each other. This means we want to minimize cluster variance, defined as

$$Var(C) = \frac{1}{|C|}\sum_{X \in C} d^2(X, c) \quad (1)$$

where c is the cluster prototype, C is the cluster, X is an example from C and d is the distance measure. The prototype is calculated as

$$c = arg\,\min_q \sum_{X \in C} d^2(X, q) \quad (2)$$

where q ranges over the examples/time series in C. The cluster prototype c as defined above, is called the cluster medoid. The medoid is an example (in this case time series) from the cluster, whose average distance to the other examples in the cluster is minimal (i.e., the example that is 'most centrally' located within the given cluster).

Many different approaches exist to distance-based clustering. Here we discuss two, k-medoids clustering and predictive clustering trees. The latter are of special interest, as they also produce descriptions of the clusters.

2.2. k-Medoids clustering of time series

The k-medoids algorithm is a partition-based clustering algorithm that extends the famous k-means algorithm (which only works for clustering vectors of real numbers). It requires as input the number of clusters (k) that it should produce. The algorithm begins with a random selection of k examples as temporary medoids.

Given a set of (temporary) cluster medoids, each example is associated with the least distant medoid. Next, a non-medoid example is randomly selected. Then, the algorithm checks whether swapping one of the initial medoids with the one that was randomly selected will result in more compact clusters. If more compact clusters are obtained, then the randomly selected example is set as a medoid and the other (non-medoid) examples are again reassigned. The random selection of non-medoid examples is repeated until replacing a medoid with the randomly selected example does not improve the compactness of the clusters.

2.3. Predictive clustering trees for time series

Predictive Clustering Trees (PCTs) (Blockeel et al., 1998) generalize decision trees (Breiman et al., 1984) and can be used for a variety of learning tasks, including different types of prediction and clustering. The PCT framework views a decision tree as a hierarchy of clusters (see Figs. 4 and 5): the top-node of a PCT corresponds to one cluster (group) containing all data, which is recursively partitioned into smaller clusters while moving down the tree. The leaves represent the clusters at the lowest level of the hierarchy and each leaf is labeled with its cluster's prototype (prediction). PCTs can be learned by the system CLUS available at http://www.cs.kuleuven.be/~dtai/clus/.

PCTs are built with a greedy recursive top-down induction (TDI) algorithm, similar to that of C4.5 (Quinlan, 1993) or CART (Breiman et al., 1984). The learning algorithm starts by selecting a test for the root node. Based on this test, the training set is partitioned into subsets according to the test outcome. This is recursively repeated to construct the subtrees. The partitioning process stops when a stopping criterion is satisfied (e.g., the number of records in the induced subsets is smaller than some predefined value; the length of the path from the root to the current subset exceeds some predefined value, etc.). In that case, the prototype is calculated and stored in a leaf.

One of the most important steps in the TDI algorithm is the test selection procedure. For each node, a test is selected by using a heuristic function computed on the training examples. The goal of the heuristic is to guide the algorithm towards small trees with good predictive performance.

The heuristic used in this algorithm for selecting the attribute tests in the internal nodes is intra-cluster variation summed over the subsets induced by the test. Lower intra-subset variance results in more accurate predictions. The cluster variance is calculated as the sum of the squared pairwise distances between the cluster elements, i.e.,

$$Var(C) = \frac{1}{2|C|^2} \sum_{X \in C} \sum_{Y \in C} d^2(X, Y) \quad (3)$$

Note that, no cluster prototypes are required for the computation of variance in this case.

The predictive clustering trees approach has a number of desirable properties. No prior assumptions are made on the probability distributions of the dependent and the independent variables. PCTs can handle discrete or continuous independent variables, as well as missing values. In addition, they are tolerant to redundant variables and noise. Furthermore, they are computationally inexpensive and are easily interpretable. Also, from a clustering point of view, the PCTs are unique in the sense that they provide cluster descriptions while constructing the clusters. All in all, PCTs are robust, efficient and interpretable models with satisfactory predictive performance.

3. Data description

The data set consists of paired time series of percentage crop and weed cover from 128 experimental sites throughout the UK. These were the sites that grew oilseed rape in the farm scale evaluations of genetically modified herbicide tolerant (GMHT) crops, about half of them growing winter oilseed rape (WOR) and half spring oilseed rape (SOR). At each site, a split field design (Perry et al., 2003) was used to compare weed management using conventional and GMHT practice (Bohan et al., 2005; Champion et al., 2003). The experiments were run over a wide range of localities and management regimes.

The two crops, conventional and GMHT, in each field were sown and harvested at the same time. They differed only in weed management, all other field operations being the same. In conventional weed management, weeds were treated in the manner usual for the field and crop, generally with a pre-emergence herbicide that targeted weeds around the time the crop itself emerged. In GMHT management, the crop and weeds were allowed to emerge. Later when GM crop plants were still small (no more than a few leaves on each plant), the herbicide glufosinate ammonium was applied to control the weeds. The GM crop plants were little affected by the herbicide, but the weeds were variously set back or killed. Full information on the weed management practices has been published elsewhere (Champion et al., 2003; Bohan et al., 2005).

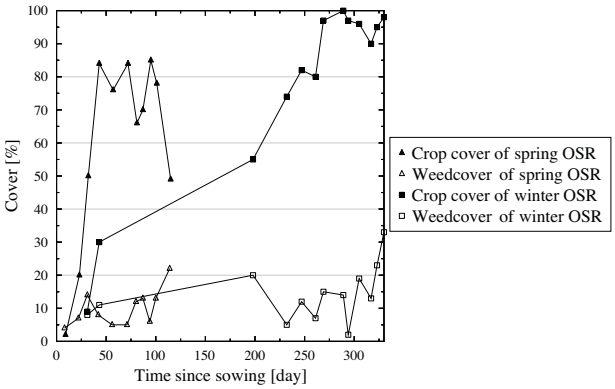

Fig. 1. Examples of time series data of the percentage of ground cover for a winter oilseed rape crop, spring oilseed rape crop, weeds in winter oilseed rape and weeds in spring oilseed rape, all from conventional treatments.

Ground cover and crop height were estimated as described by Champion et al. (2003) at locations throughout each field every 7–14 days during the growing season. The means per treatment (half field) per sample occasion are used subsequently in this analysis. Fig. 1 shows representative examples of the time series data where zero is the time the crop was sown. SOR was sown around day 100 in the harvesting year (mid-April). Measurements began around day 120, continued throughout growth and ceased at typically day 240 (end of August) just before harvest. WOR was sown around 140 days before the beginning of the year of harvest. Measurements continued on typically two to four occasions in that year, ceased over the winter, resumed in early spring and continued until around day 200, just before harvest.

Variation in the time series therefore had the following potential sources: winter vs. spring cropping as indicated in Fig. 1; GMHT vs. conventional treatments in each half field; and differences among sites related to weather, soil and field management. Previous analyses have shown that the mean crop cover did not differ between conventional and GMHT treatments (Champion et al., 2003); the treatments however, had effects on weeds (Bohan et al., 2005; Champion et al., 2003). Notably, the weed cover in the GMHT treatment in SOR was on average about half that of the conventional, while the weed cover in GMHT and conventional treatments in WOR were similar. The variation among sites has not hitherto been systematically examined, but visual inspection of the data indicated that such variation may be greater than that between crops and treatments.

In total, the dataset consisted of 2665 individual measurements of percentage cover (1322 conventional and 1333 GMHT) which were then converted to 254 time series. Accompanying these measurements of crop and weed cover for each half-field were data describing the site (e.g., field size, soil texture, soil carbon and nitrogen) and the management (e.g., soil cultivation, herbicide, fertiliser, harvesting). For illustration, these data were used as inputs to some of the analyses presented here, but understanding their effects on percentage cover is not the primary aim of the paper. Examples of

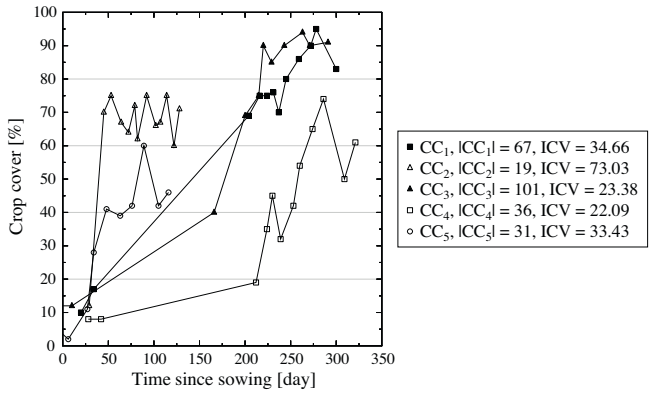

Fig. 2. The medoids of the five clusters for crop cover. The sizes of the clusters ($|CC_i|$) and the intra-cluster variances (ICV) are given in the legend.

the use of such data in another application are given by Debeljak et al. (2008).

4. Results

4.1. Clustering of crop cover and crop height time series

Time series data for crop height and cover were clustered to include the information about crop height (CH) and crop cover (CC) in the induction process. We applied the *k*-medoids algorithm as described in the previous section.

The *k*-medoids clustering algorithm requires as input *k*, the number of clusters that the algorithm should output. This value was set to 3, 4, 5, 6, 7, 8 and 9 for both crop cover and crop height. The clusters were inspected for the intra cluster variance (how homogenous the clusters were). The clusterings with least variance were those of 5 clusters for crop cover (CC_1, CC_2, CC_3, CC_4 and CC_5) and 6 for crop height (CH_1, CH_2, CH_3, CH_4, CH_5 and CH_6). The cluster medoids for each of the two clusterings are presented in Figs. 2 and 3.

The clusters of crop cover time series differed largely in the shape of the profiles as affected by the length of the period of slow growth at the beginning of the season (causing an offset rightwards of the period of rapid growth), the steepness of slope of the rapid growth phase, the duration of this early growth and the overall period of growth and the maximum cover. The obtained clusters present an interesting finding in that each of the clusters included both WOR and SOR as well as GM and conventional, suggesting that the shape of the crop cover profile varied systematically and independently of the crop unit (Table 1).

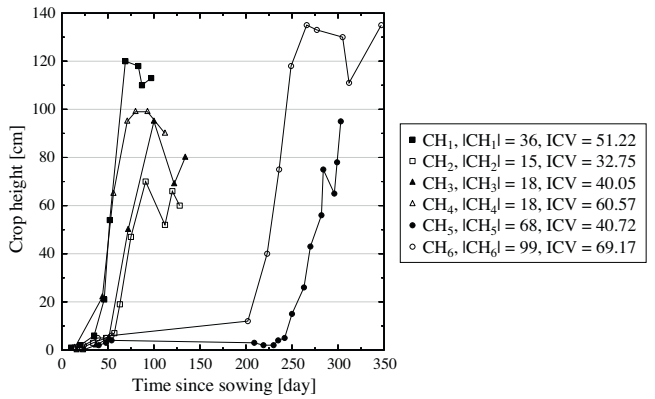

Fig. 3. The medoids of the six clusters for crop height. The sizes of the clusters ($|CH_i|$) and the intra-cluster variances (ICV) are given in the legend.

Table 1
The composition of the five clusters for crop cover (Fig. 2). For each cluster, the total number of sites is shown first, and then is broken down into winter oilseed rape (WOR) or spring oilseed rape (SOR) and conventional (C) or GM.

Cluster	Total	WOR	SOR	C	GM
1	67	37	30	34	33
2	19	3	16	9	10
3	101	57	44	53	48
4	36	15	21	20	16
5	31	13	18	12	19

Each time series in the dataset is therefore assigned two more explanatory variables, one for each of the CC and CH clusterings, that indicate the cluster memberships for a given sample. For example, if a sample has values CC_4 and CH_5 for the CC and CH variables, this means that this sample belongs to the 2nd cluster of CC and 5th cluster of CH.

4.2. Predictive clustering trees for weed cover

Two different predictive clustering trees were built to cluster the weed cover profiles (time series). In the first, the input attributes were crop cover, crop unit (WOR, SOR, C, GM), and the membership in the crop cover and crop height clusters. To increase the efficiency of the clustering process, and to prevent the tree from forming many small clusters, each leaf of the PCTs was required to contain a minimum number of 32 instances (i.e., time series). The tree and the cluster prototypes for each leaf are given in Fig. 4.

The induced tree illustrates higher importance of cover and height cluster membership as compared to nominal crop units (i.e., WOR/SOR, GM/conventional) in determining the target variable—weed cover. The first partition in the tree (Fig. 4) is indeed between two groups of crop clusters: CC_1, CC_2 and CC_3 in one group and CC_4 and CC_5 in the other. CC_1, CC_2 and CC_3 all exhibit faster initial growth than CC_4 and CC_5, and weed cover was much smaller in CC_1, CC_2 and CC_3 than in CC_4 and CC_5. Within the group comprising CC1, CC2 and CC3, the next division was between SOR and WOR, beyond which further divisions were identified in terms of crop clusters. The group that comprised CC_4 and CC_5 was divided by treatment, conventional or GM, the weed cover being smaller for the GM treatment. No influence was seen of the crop height cluster membership. These finer divisions defined six weed time series clusters of which the medoids are shown in Fig. 4 in the boxes to the right (from C_1 to C_6). In summary, CC cluster membership was a stronger determinant of the weed cover time series than either crop type (WOR, SOR) or herbicide treatment (conventional, GM).

In the second predictive clustering tree (Fig. 5), agricultural and soil parameters were used as descriptive variables in addition to those used in Fig. 4. The crop clusters representing slow growing crops (i.e., not CC_1, CC_2 or CC_3) and then the crop unit again distinguished two weed clusters (C_6 and C_7) which were similar in composition to the corresponding ones (C_5, C_6) in Fig. 4. Therefore, crop cover cluster membership still had a greater influence on the initial division than any of the other variables. For the time profiles representing the fast growing crops (CC_1, CC_2, CC_3), the weed time series was influenced by several variables including field size, previous crops (probably indicative of the crop rotation at the site), soil nitrogen and soil carbon. While we do not investigate these relations further in this paper, they are all probably indicative of the intensity of crop production and field management. The outcome of this analysis is that all the environmental and agronomic factors had less influence on the weed time series than the crop cover cluster membership.

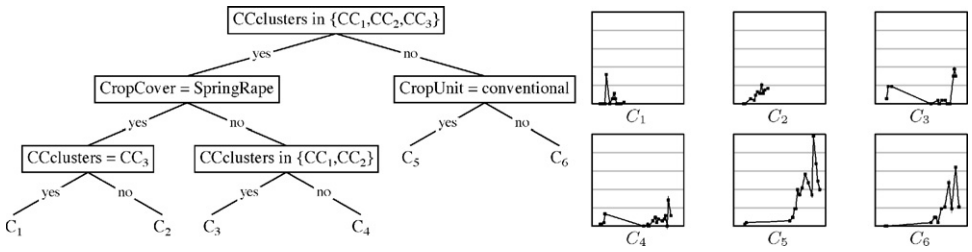

Fig. 4. A predictive clustering tree for weed cover time series, defining the clusters in terms of crop type, crop unit and crop cover/height cluster membership. The average intra-cluster variance (ICV) for this tree is 21.02. On the graphs for each of the clusters (C_1 through C_6) the *x*-axis denotes the days since sowing (ranging from 0 to 350) and the *y*-axis is percentage weed cover (ranging from 0 to 100).

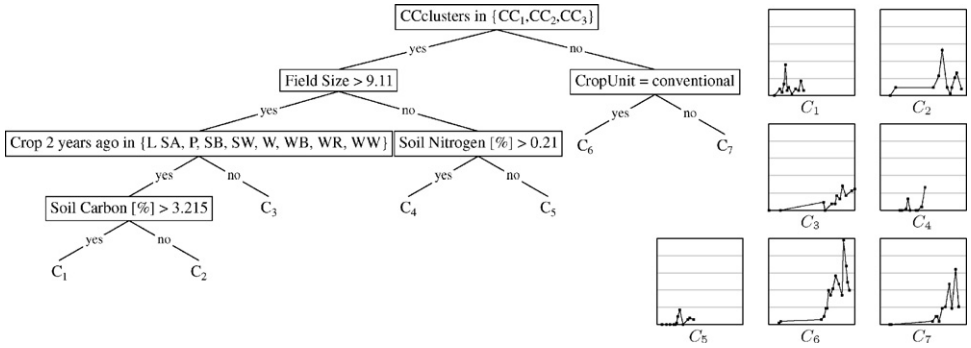

Fig. 5. A predictive clustering tree for weed cover time series, defining the clusters in terms of crop type, crop unit, crop cover/height cluster membership, and agricultural and soil parameters as descriptive variables. The average intra-cluster variance (ICV) for this tree is 19.91. On the graphs for each of the clusters (C_1 through C_7), the *x*-axis denotes the days since sowing (ranging from 0 to 350) and the *y*-axis the percentage weed cover (ranging from 0 to 100). Abbreviations for previous crops: L, linseed; SA, set aside; P, peas; SB, spring barley; SW, spring wheat; W, wheat undefined; WB, winter barley; WR, winter rape; WW, winter wheat.

5. Discussion

The results presented above offer new insights into and knowledge about the coupling between the two types of vegetation considered – crop cover and weed cover – during their growing seasons. The clustering of the time profiles exhibited by the first variable, crop cover, had not been previously suspected and was mostly independent of whether the crop was winter or spring oilseed rape. The analysis revealed distinct shapes in the profiles (time series) of percentage crop cover, characterized combinations of morphological variables such as rate of increase, lag and maximum.

Of the five crops clusters defined by percentage crop cover (Fig. 2, Table 1) CC_2, containing only 19 time series, is the one cluster that contained mostly one type of crop, SOR. The cluster incorporates those SOR crops that grew rapidly, reaching mostly 60–80% cover after 50–60 days, after which cover did not systematically change. The three WOR crops in this cluster grew, in effect, untypically for a winter crop but in a similar way to the SOR in this cluster, nearing their maximum cover before the winter and then changing little during the next year's potential growing period.

Two of the crop cover clusters, CC_4 and CC_5, were characterised by both slower growth and low maximum cover than the others. They differ mainly in the maximum cover, 40–60% in CC_5 compared to 60–80% in CC_4. Both are indicative of crops that were not successful or 'failed' in agronomic terms, whether SOR or WOR. Highly productive, dense stands of oilseed rape were included in CC_1 and CC_3. These two clusters expanded more slowly than CC_2, but faster than CC_4 and CC_5. Both reached 80–100% crop cover, but CC_3 did so earlier in the season.

The crop cover profile membership in one of the above clusters was then a strong predictor of the shape of the weed cover profile (time series). Take, for example, the two crop clusters CC_4 and CC_5 (Fig. 2) representing crops of slow expansion and low final cover; The associated weed cover time series in Fig. 4 (i.e. C_5 and C_6) both had a similar shape characterized by a long lag with low weed cover, followed by a substantial rise in weed cover later in the growing season. A likely explanation for this profile is that herbicide management controlled the weeds initially, but because the crops did not attain high cover (not because they were suppressed by weeds but for some other reason) later germinating weed cohorts were able to exploit the space and themselves reached high cover.

The second partition in this part of the tree in Fig. 4 refers to herbicide treatment. It shows the effect of GMHT management reducing weed cover below that achieved in conventional management. This partition is due to the main effect of the GMHT treatment found in the original comparison of treatments in spring oilseed rape (Heard et al., 2003).

The other crop clusters (CC_1, CC_2 and CC_3) containing crops that would be considered successful, were associated with much lower weed cover in the left hand side partition of Fig. 4. This implies suppression of weeds by those crops. At this second partition, the season of sowing, and therefore the potential duration of weed growth, had an influence: it was associated with much shorter weed time profiles for spring than winter crops.

6. Conclusions

This paper describes a successful application of predictive clustering trees in the analysis of time series in a large, complex ecological data set. The questions addressed were (1) whether the substantial variation that existed within a large set of time series for one biological variable (crop cover) could be reduced by finding groups, or clusters, each of which comprised time series of a generally similar shape; and (2) whether the clusters were a predictor of the time series of a second biological variable (weed cover). Three methods were combined to address the question: Dynamic Time Warping to define the distance between two time series; the *k*-

medoids algorithm to partition the time series into clusters with minimal within-cluster variance; and predictive clustering trees to relate a target variable, in this case the second time series, to independent variables (input attributes), including membership in the clusters defined for the first time series. To the best of the knowledge of the authors, such an approach has never been used to analyze time series in environmental or agricultural data.

The study raises questions as to what determines the different crop time series, the associated weed time series and the conditions under which the two can coexist. The clusters were formed largely independently of whether the crop was spring or winter oilseed rape and genetically modified or conventional. The findings divert attention from crop type and weed management treatment as the main determinants of crop yield and the ecological role of weeds to more pervasive factors, such as the local combination of site, weather and field management. If these factors were understood, it may be possible to manipulate them to satisfy multiple objectives of (achieve multiple outputs from) agroecosystems or to counter any negative effects of a new technology (e.g., May et al., 2005).

This successful application of machine learning provides encouragement for the further development, and especially the application, of the methodology used here to ecological modelling in general. The results have confirmed that the applied method is robust with respect to the complexity of the data, which had missing values, time series with measurements at different time intervals, time series with different length and both numeric and discrete attributes in some of the associated agronomic and environmental variables. In addition, the sampling sites were widely spread across the territory of the UK and different personnel were involved in data collection for the different regions. Many environmental time series datasets are similar to the one analyzed here and our methodology may be used to analyze them.

Acknowledgements

The measurements of percentage cover were obtained during the Farm Scale Evaluations of GM Herbicide Tolerant Crops in the UK by staff from the Centre for Ecology and Hydrology, Rothamsted Research and SCRI.

References

Blockeel, H., De Raedt, L., Ramon, J., 1998. Top-down induction of clustering trees. In: Proc. Fifteenth International Conference on Machine Learning, Morgan Kaufmann, San Mateo, CA, pp. 55–63.

Bohan, D.A., Boffey, C.W.H., Brooks, D.R., Clark, S.J., Dewar, A.M., Firbank, L.G., Haughton, A.J., Hawes, C., Heard, M.S., May, M.J., Osborne, J.L., Perry, J.N., Rothery, P., Roy, D.B., Scott, R.J., Squire, G.R., Woiwod, I.P., Champion, G.T., 2005. Effects on weed and invertebrate abundance and diversity of herbicide management in genetically modified herbicide-tolerant winter-sown oilseed rape. Proceedings of the Royal Society Series B 272, 463–474.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth.

Champion, G.T., May, M.J., Bennett, S., Brooks, D.R., Clark, S.J., Daniels, R.E., Firbank, L.G., Haughton, A.J., Hawes, C., Heard, M.S., Perry, J.N., Randle, Z., Rossall, M., Rothery, P., Skellern, M.P., Scott, R.J., Squire, G.R., Thomas, M.R., 2003. Crop management and agronomic context of the Farm Scale Evaluations of genetically modified herbicide-tolerant crops. Philosophical Transactions of the Royal Society of London B 358 (1439), 1801–1818.

Debeljak, M., Squire, G.R., Demsar, D., Young, M.W., Džeroski, S., 2008. Relations between the oilseed rape volunteer seedbank, and soil factors, weed functional groups and geographical location in the UK. Ecological Modelling 212, 138–146.

Hawes, C., Haughton, A.J., Bohan, D.A., Squire, G.R., 2009. Functional approaches for assessing plant and invertebrate abundance patterns in arable systems. Basic and Applied Ecology 10, 34–47.

Hawes, C., Squire, G.R., Hallett, P.D., Watson, C.A., Young, M., 2010. Arable plant communities as indicators of farming practice. Agriculture, Ecosystems and Environment 138, 17–26.

Heard, M.S., Hawes, C., Champion, G.T., Clark, S.J., Firbank, L.G., Haughton, A.J., Parish, A.M., Perry, J.N., Rothery, P., Scott, R.J., Skellern, M.P., Squire, G.R., Hill, M.O., 2003. Non-crop plants in fields with contrasting conventional and genetically modified herbicide-tolerant crops. 1. Main effects of treatments. Philosophical Transactions of the Royal Society of London B 358, 1819–1832.

Li, C., Biswas, G., Dale, M., Dale, P., 2001. Building models of ecological dynamics using HMM based temporal data clustering-a preliminary study. In: Hoffmann, F., et al. (Eds.), Advances in Intelligent Data Analysis, IDA 2001, LNCS, 2189. Springer, Berlin/Heidelberg, pp. 53–62.

Liao, T.W., 2005. Clustering of time series data—a survey. Pattern Recognition 38, 1857–1874.

Mari, J.F., Le Ber, F., 2006. Temporal and spatial data mining with second-order hidden markov models. Soft Computing 10, 406–414.

Marshall, E.J.P., Brown, V.K., Boatman, N.D., Lutman, P.J.W., Squire, G.R., Ward, L.K., 2003. The role of weeds in supporting biological diversity within crop fields. Weed Research 43, 77–89.

May, M.J., Champion, G.T., Dewar, A.M., Aiming Qi, Pidgeon, J.D., 2005. Management of genetically modified herbicide-tolerant sugar beet for spring and autumn environmental benefit. In: Proceedings of the Royal Society B, 272, pp. 111–119.

Perry, J.N., Rothery, P., Clark, S.J., Heard, M.S., Hawes, C., Design, 2003. analysis and power of the farm-scale evaluations of genetically modified herbicide-tolerant crops. Journal of Applied Ecology 40, 17–31.

Potter, C., Genovese, V., Gross, P., Boriah, S., Steinbach, M., Kumar, V., 2007. Revealing land cover change in California with satellite data. EOS Transactions American Geophysical Union 88, 269–276.

Quinlan, J.R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann series in Machine Learning. Morgan Kaufmann.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken-word recognition. IEEE Transaction on Acoustics, Speech, and Signal Processing 26, 43–49.

Shumway, R., Stoffer, D., 2006. Time Series Analysis and Its Applications, 2nd ed. Springer, New York.

Squire, G.R., Hawes, C., Begg, G.S., Young, M.W., 2009. Cumulative impact of GM herbicide-tolerant cropping on arable plants assessed through species-based and functional taxonomies. Environmental Science and Pollution Research 16, 85–94.

Viovy, N., 2000. Automatic classification of time series (ACTS): a new clustering method for remote sensing time series. International Journal of Remote Sensing 21, 1537–1560.

Zhou, X., Persaud, N., Wang, H., 2006. Scale invariance of daily runoff time series in agricultural watersheds. Hydrology and Earth System Sciences 10, 79–91.