COMMUNITY ECOLOGY 11(2): 187-201, 2010
1585-8553/$20.00 © Akadémiai Kiadó, Budapest
DOI: 10.1556/ComEc.11.2010.2.7

Model selection using Minimal Message Length: an example using pollen data

M. B. Dale[1,3], L. Allison[2] and P. E. R. Dale[1,4]

[1] *Griffith School of Environment, Environmental Futures Centre, Australian Rivers Institute, Griffith University, Nathan, Queensland, Australia 4111*
[2] *Dept. Computer Science and Software Engineering, Monash University, Clayton, Victoria, Australia 3800. E-mail: Lloyd.Allison@infotech.monash.edu.au*
[3] *Corresponding author. E-mail m.dale@griffith.edu.au*
[4] *E-mail: p.dale@griffith.edu.au*

Keywords: Censoring, Clustering, Complexity, Compositional data, Constrained, Gaussian, Geometric, Minimum message length, Unconstrained, User expectation, Within-cluster model.

Abstract: In this paper we examine the use of the minimum message length criterion in the process of evaluating alternative models of data when the samples are serially ordered in space and implicitly in time. Much data from vegetation studies can be arranged in a sequence and in such cases the user may elect to constrain the clustering by zones, in preference to an unconstrained clustering. We use the minimum message length principle to determine if such a choice provides an effective model of the data. Pollen data provide a suitably organised set of samples, but have other properties which make it desirable to examine several different models for the distribution of palynomorphs within the clusters. The results suggest that zonation is not a particularly preferred model since it captures only a small part of the patterns present. It represents a user expectation regarding the nature of variation in the data and results in some patterns being neglected. By using unconstrained clustering within zones, we can recover some of this overlooked pattern. We then examine other evidence for the nature of change in vegetation and finally discuss the usefulness of the minimum message length as a guiding principle in model choice and its relationship to other possible criteria.

Abbreviation: MML–Minimum Message Length.

First let him measure and order the inner world

Sándor Weöres 'Difficult Hour'

Introduction

Inductive inference is the business of trying to put order on a mass of data, i.e., coming up with an explanation for it. The data are often contradictory and contain measurement errors and other uncertainties. People do inference all of the time; it is one of the hall-marks of intelligence. Scientists do it when they form theories from experimental data. The trouble is that many theories could explain some given facts. Which is the best one? And how might we identify it? In this paper, we examine the use of the minimum message length principle (MML) for this task.

Of course we would like to choose a model close to the 'true' model underlying the data allowing for the trade-offs between model complexity and the random noise inherent in observational data. To do this we must first select the correct family of probability distributions and then adopt a consistent method of estimation of the parameters of that distribution. By embedding the possible distributions in a Riemannian space (Myung et al. 2000), we can then seek results close to the 'true' one. The minimum message length principle (Wallace 2005) provides consistent estimates for the parameters and, if there are external reasons for selecting a particular model family, permits the user to impose or suggest appropriate constraints. Given the 'wrong' model family, it will provide as close an approximation as possible to the 'true' solution. The problem of model selection has been studied by Balasubramanian (1997) in terms of the statistical mechanics of the space of probability distributions, with any finite-dimensional parametric model family forming a manifold. He concludes the Bayesian model family inference embodies Ockham's razor because for a small number of observations, simplicity and robustness will be important while for a large number of observations, the accuracy of the model family dominates. This conclusion is pertinent to the minimum message length principle employed here.

The MML approach provides a principled means of choosing between models of varying degrees of complexity. It is based on the premise that the best model of observed data, given limited observational data, is the one which provides the most compression of the data consistent with simplicity of the model. The greater the compression, the more we learn about the regularities underlying it. It provides an operational measure for Ockham's razor[1]. Ockham is credited with postulating "*if* two theories explain the facts equally well *then* the simpler theory is to be preferred". MML does exactly that. The shortest message corresponds to the maxi-

mum posterior probability, and is invariant over 1-1 parameter transforms and generally provides consistent estimators (Dowe 20078b); i.e., with increasing data the parameter estimates converge towards the true values. It is also related to Gell-Mann and Lloyd's (1996) concept of Total Information, which stresses that complexity is a subjective relationship between natural phenomena and an interested observer. Complexity relates more to the way observations are made, i.e., the model used, than in any elaborations after the model has been imposed.

Previously, Dale (2007) examined the choice of appropriate distributions for within-cluster variation, of the existence of hierarchical structure in the derived clusters (Dale and Wallace 2005) and of introducing a single factor to model within-cluster correlation (Dale et al. 2010). In Dale and Dale (2002), clustering was combined with hidden Markov models to provide a graphic model of the processes of change in salt marsh vegetation. All these analyses employed the MML principle as a basic tool for deciding which models were most appropriate. In the present paper, we propose to extend such studies and examine the use of models where spatio-temporal constraints are imposed to provide a zonation; i.e., the clusters are formed from adjacent samples. Such adjacency may be temporal or spatial, although here only one-dimensional sequences of samples are examined. Possible extensions to two or more dimensions will only be briefly mentioned.

The importance of pollen data as a record of vegetation changes cannot be disputed, although interpretation is a complex matter. The intensity of sampling within a single pollen core provides more or less detail, although some temporal averaging is always present (Joosten 2007). Furthermore, the exact source of the palynomorphs may not be precisely attributable, with some local and others regional or global (Prentice 1985, Bunting and Middleton 2009), so that spatial resolution is somewhat blurred and at multiple scales. Interpretation may also involve the use of modern analogues which poses its own difficulties (Jackson and Williams 2004) since no 'good' analogies may now exist.

The difficulties are apparent in the study of Pickett et al. (2004, see also Prentice et al. 1996, Bradshaw 1981) where interpretation was limited to recognition of biomes (see also Sugita 1993, 1994, 2007a,b), providing a rather coarse description of vegetation. However, these difficulties afflict any analysis and we shall not be concerned with them here. But they do impact, together with the monotonicity of the depth-time relationship, to make standard time series models questionable. Orlóci and He (2009, see also Orlóci 2010) examine some interesting methods based on assumptions that the record provides a sequence of 'directed' and 'chaotic' phases, whose trajectories can be cross-related between profiles, while Orlóci et al. (2006) provide a multiscale analysis and discuss various error sources. Dale et al.'s (2010) study suggests that patterns other than zonation may be more appropriate; in no case was the addition of a factor to a cluster (indicative of trend in the cluster) found desirable. If there was a 'directed' phase it was not a simple monotone pattern of change.

Many ecologists would argue for a correspondence analysis as a more suitable ordination (seriation) method rather than the component analysis used in Dale et al's. (2010) study, and Legendre and Gallagher (2001) have indicated data transformations which would allow such an analysis. The imputation of possible hierarchical structure to the relationships between zones is another possible device, and it is also easy to obtain a measure of similarity between clusters (and to calculate the similarity of new samples to clusters). These inter-cluster measures can also be used as a basis for organising the clusters. The procedure involves using the model for one cluster as a basis for calculating the likelihood of the other and *vice versa*.

Zonation is not the only constraint which might be used to incorporate the sequential information. Walker (1966), Walker and Wilson (1978), Orlóci (2010), Orlóci and He (2009) and Orlóci et al. (2006) have all proposed alternatives which have some similarities among themselves, though most of these methods require intensive sampling and dating information. It may be possible to utilise MML for these purposes as well. Dale et al. (2010) provide an evaluation of some aspects of this question.

Important properties of pollen data

This study combines constrained and unconstrained analyses with some novel within-cluster distributions, the latter chosen to reflect known idiosyncratic properties of the data. The data used for illustration are taken from palynology, where the sampling depth is usually monotonically related to time. With such data, the earliest (Von Post 1916, 1924), and still the commonest, analysis starts with zonation, a constrained clustering. The zones are usually taken to represent stable vegetation communities which change in response to environmental, usually climatic, changes. But is a zonation a good model? Or is it a product of confirmation bias (Gale and Ball 2002), restricting usage of other models? Certainly there will remain variation within each zone which is not modelled (except as a random distribution), so further compression may be possible.

Pollen data have two further characteristics of significance which might also be worth incorporating in our model. The first is that much such data is compositional in nature, i.e., the values recorded represent a proportion of some count

1 Occam's Razor has been expressed in several ways of which the commonest are probably: "Entities should not be multiplied unnecessarily" or "nothing ought to be posited without a reason given, unless it is self-evident". This does not mean simplifying syntax at the expense of semantics. The theory of relativity is syntactically complex, but semantically simple. But simplicity is not truth, only a guide to learning. Truth need not be simple.

rather than absolute numbers. Perhaps it is desirable to utilise a within-cluster distribution which reflects this? Second, as with most vegetation data, there is a preponderance of zero values, due to the absence of the taxon from the sample. Again it may be desirable to explicitly model this, rather than blithely assume that a Gaussian distribution is adequate. Pollen data are severely affected by the zero problem since identification of palynomorphs is often restricted to high level taxa which do not provide a high quality of ecological information (Dale and Clifford 1976, Riddle and Hafner 1999). Furthermore, the samples within a profile do not represent a large area, so that random absence is quite likely for many of the palynomorphs; only the commonest taxa will be regularly present.

The overall aim is to explore patterns within pollen data, including those where the user expectation of zones is not incorporated in the analysis.

Overview of methods and analyses

Table 1 identifies four analyses, one unconstrained and three constrained, with the constrained analyses differing in the assumed model for the distribution of the quantitative values.

Unconstrained clustering

The clustering program used here, SNOB (Boulton and Wallace 1970, Wallace and Dowe 1994, 2000), implements MML unconstrained clustering and permits a variety of distributions within clusters for the attributes to be specified (Wallace and Dowe 2000). Multistate distributions are available for discrete variables, while for quantitative values either Poisson or Gaussian can be used. Provision has also been made for correlation between attributes in a multivariate Gaussian distribution (Agusta and Dowe 2003) and extensions for clustering sequences are also known (Molloy et al. 2006). The data may also contain missing values. In the unconstrained analysis used here only the Gaussian distribution was employed and no missing values were present.

The result is a non-hierarchical clustering, with fuzzy assignment of samples to clusters; this is necessary if the analysis is to provide consistent estimates of the cluster parameters. The number of clusters is estimated by the analysis, together with their *a priori* likelihood and cluster parameters for every attribute in every cluster are also calculated, together with the significance of any deviation from the population parameters. The parameters are also specified to an accuracy determined by the program. For the Gaussian model, the parameters are mean and variance, while for multistate attributes a frequency table is given. The assignment of samples to clusters, and their probabilities of belonging are also calculated. There is a limit on the size of the smallest clusters, which must have at least three members. However, the coefficients of belonging can be used to indicate outlying members. All this information contributes to the message length calculated.

Table 1. Main analyses used in this paper, showing message lengths for all analyses. All within-zone analyses use a Gaussian model.

Approach	Model
Clustering unconstrained	**No zonation constraint** 1. Gaussian
Zonation (constrained)	**Model within Zones**
2. Gaussian	Not applicable
3. Geometric	Not applicable
4. Censored Gaussian	Gaussian

Constrained clustering

The user must then make a decision concerning the introduction of the spatio-temporal constraint, since this has some repercussions. In these studies the only constraint employed is adjacency, which results in the formation of spatially coherent zones (models 2-4). With the introduction of the constraint, the fuzzy assignment is lost, and the assignment of samples to clusters is replaced by a statement of the bounds of the zone. This simplifies the model compared with unconstrained clustering, but loses consistency. The zonation will still have, normally, a greater lack of fit (the exception is if the optimal unconstrained clustering results in a zonation). By using a dynamic programming algorithm (Baxter and Oliver 2006, Oliver et al. 1998, Fitzgibbon et al. 2000), the algorithm provides a globally optimal result. The unconstrained clustering may not be optimal; algorithms for optimal unconstrained clustering are known (Vinod 1969, Schader 1979, Rahwan and Jennings 2008, Sombattheera and Ghose 2008) but they are NP-complete and hence extremely time-consuming. The heuristics used within the SNOB program seem to be effective but may still result in a suboptimal solution. Thus any difference in message length between unconstrained and constrained solutions is only a lower bound. The difference in message lengths provides the odds in favour of the lower-valued result, since the message length represents, approximately, -log (probability). Dale et al. (2007) have argued for using both constrained and unconstrained analyses, since they provide complementary information.

For the first data characteristic, the compositional nature of the data (model 3), we consider the use of the Geometric distribution in place of the Gaussian. The former is commonly used to represent the distribution of failures before attaining a success and this might be a useful way of examining pollen counts. It is a special case of the negative binomial distribution where $\Pr(X=k) = (1-p)^{k-1}$; p is the probability of success at any single trial. Thus, for a geometric distribution, there is only one parameter, p, to be estimated for each attribute in any cluster (Larossa 2005). But does the decrease in number of parameters result in too great a loss of fit?

For the second characteristic, the zero value problem here called censoring (model 4), Babad and Hoffer (1984) have examined the problem of absent data, although earlier Williams and Dale (1962, see also Williams 1969, 1971) proposed a partition which overcomes it. Here we recode the data in accordance with the Williams-Dale partition by cod-

ing the presence/absence information as a binary attribute, and using the numeric values separately with absence entries regarded as missing values. Such a model involves 3 parameters, probability of presence, mean and variance when present, whereas a Gaussian model requires only 2, mean and variance. The question is whether the increase in complexity results in much better fit such that the overall message length is reduced.

One problem which has been completely ignored here is that of irrelevant attributes. It is often assumed that many of the palynomorphs do not contribute much information to the processes of change, resulting in the omission of aquatic taxa or restriction to subsets such as tree taxa only. Blum et al. (1995) have addressed the question of learning in such situations. MML seeks overall simplicity and might therefore combine several models into a more complex one. For clustering this will result in extra clusters. Perhaps introducing an hierarchical organisation for the clusters would mitigate this problem (Dale and Wallace 2005).

Intra-zone analysis

It is possible to further analyse the constrained results. Within each zone, if it contains enough samples, we can apply an unconstrained analysis to identify further patterns which might otherwise remain hidden. For example, a trend may be recognisable from changes in the mean values for (some) attributes or from the pattern of (sub)clusters. We have used this analysis only for zones produced by the overall preferred model. Such a nested analysis could be regarded as an hierarchical analysis, such as that used by Dale and Wallace (2005). This would allow the message lengths for the zonation and the intra-zone clustering to be united, but we have not attempted this yet.

If several profiles are available, cross-relationships are usually established using any dating information which might be available or else by examining the pattern of changes. Formal methods of establishing cross-relationships have recently been proposed in the trajectory analysis of Orlóci (2010). This requires intensively sampled profiles with good dating. A distinction is drawn between zones showing directional trends and those showing only 'chaotic' patterns. The zones are thus predominantly determined by number and form of temporal changes, identified through changes in the constituent attributes. Matching between profiles is based on the pattern of trend and chaos since there will often be little similarity in the palynomorphs present in the several profiles.

Data and Explication

Data

The data are 42 samples taken from a core from Isla Clarence (Lat. 54.12 S: Long. 71.14 W) in Chile (Markgraf 1983). The samples range in age from about 9 Ka to the present, and cover the period following deglaciation, which commenced about 10 Ka ago. Since then, there appear to have been several fluctuations of ice-cover and the vegetation, at the earliest stages, would also change as new taxa return from refugia.

A few radiocarbon dates are available for these data, but insufficient to be able to closely examine their relationship with depth. Other studies suggest that such a relationship is often sigmoid, with a largely linear form in the centre of the data, but with compression at the ends. Here we cannot use the few dates to place the clusters. For other attributes comparison can made with the population mean to identify discriminating taxa, but any cluster or zone with a similar mean to the population may not appear to be significantly differentiated. The SNOB program can also discriminate between clusters using variances which would for example enable the distinction of a dense core cluster within a more diffuse one. Such cases seem to be uncommon.

In total, 17 pollen taxa were identified, and the abundances of pollen in these taxa were recorded for each sample. Several taxa were at family level or above, which makes interpretation somewhat coarse; for example high values of pollen from Poaceae could have several interpretations. However, Dale and Clifford (1976) have shown that even higher level taxa exhibit ecologically informative patterns. Pollen counts were used directly in the analyses, without normalising using some pollen sum.

Postglacial history

The general postglacial history of a nearby region is provided by Bennet and Porter (2001). Our expectations regarding the vegetation of this area must rely at least in part on information about the present vegetation of the area and other neighbouring areas, as well as knowledge of the (present) habitat of the various taxa. The present vegetation of this Magellanic region is of four main types (Paez et al. 2001, Bennett and Porter 2001, Fesq-Martin et al. 2004), arranged along rainfall and exposure gradients. First, in the wettest areas, is a form of heath, with cushion plants. Second, as rainfall decreases, the heath is replaced by evergreen/deciduous *Nothofagus* forest (the deciduous *Nothofagus pumilio* (Poepp. & Endl.) Krasser and the evergreen *Nothofagus betuloides* (Mirb.) Blume). Third, with further reduction in rainfall and at higher altitudes, the *Nothofagus* forest is replaced by stunted deciduous *Nothofagus antarctica* (Forster) Oerst. forest. Fourth, in the driest areas, a Patagonian steppe type is found, primarily xeric shrubland with large areas of cushion-like and dwarf shrubs, and grassy steppe. In addition to these 4 major types, there are possibly further types associated with special environments, for example protected moist areas on the heath with richer soils.

The moisture and temperature gradients provide an obvious source for our expectation of an axis within clusters, and they are closely associated so that a single factor might suffice to represent them both. It has been suggested that the latitude of the South-westerly systems that bring rain has varied considerably (Douglass et al. 2000) especially in the earliest

Table 2. Isla Clarence. Comparison of analyses of complete data.

Analysis	1-class Message Length (nits)	Optimal Number of Clusters	n-class Message Length (ML) (nits)	Difference In ML 1-class – n-class	% Capture
Unconstrained Clustering		**Cluster #**			
Gaussian	4863.8	6 (for details of sample assignment to clusters see Table 3)	3539.9	1323.4	27.2
Constrained Clustering		**Cluster # (zone boundaries)**			
Constrained Gaussian	4419.7	3 (1-24, 25-32, 33-42)	3810.9	161.1	4
Constrained Geometric	3678.9	3 (1-32, 33, 34-42)	3542.0	136.9	4.7
Constrained Censored Gaussian	2469.1	4 (1-24, 25-32, 33, 34-42)	2428.1	47.5	2

periods. This could be recognised in changes reflected in within-cluster factors.

The temporal sources are the obvious ones of immigration of taxa and a warming climate. This is less obviously represented by a single factor although a single seriation axis might be a possible alternative. In any case, there are likely to be different series at different times, and possibly multiple series at the same time, so we are better avoiding any specification of the nature of the temporal correlation.

Finally, in analysing pollen data we might expect that some indication of the temporal stages of re-vegetation and response to postglacial climate changes might be reflected in a tendency to form zones. That is, the clusters would be temporally homogeneous indicating a relatively stable vegetation for some extended time period. Green (1982) has suggested that major changes in the vegetation he studied were always associated with a large-scale disturbance from fire thus giving rather abrupt limits. Obviously, it is possible to constrain the analysis so that temporally coherent zones are identified but even unconstrained analyses might be expected to show signs of clear temporally distinct zones even if these are not completely coherent.

Results

Table 2 gives the overall summary results. All four analyses of the whole population identify clusters as present, but the two Gaussian analyses, constrained and unconstrained, have larger message lengths than either Geometric or censored Gaussian for the n-class solutions. All analyses provide some capture of structure, with the least for the constrained censored Gaussian and the most for the unconstrained Gaussian. The unconstrained (Gaussian) analysis is markedly better than the constrained Gaussian (odds $1:e^{-271}$), not significantly different from the Geometric (odds $1:e^{-2.1}$) and decidedly worse than the censored Gaussian (odds $1: e^{1111.8}$)[2]. Thus the censored Gaussian model is identified as the 'best' result, having the minimum message length. Inspection of the zones (or contiguous sample segments) produced show that the censored Gaussian identifies four zones which combine both the constrained Gaussian and geometric zones. These four zones were shown in a standard pollen diagram (Fig. 2 of Dale et al. 2010) and appear to identify acceptable boundaries. Here the zone boundaries are identified in Table 2 so that their correspondence can be noted. It should be noted that an unconstrained clustering using the censored Gaussian model cannot have a greater message length than the constrained censored Gaussian, although equality is a possibility if the unconstrained result was itself a zonation.

In terms of capturing structure, the unconstrained Gaussian is by far the best with 27.2% compression, and the other three, while still showing a significant difference from a 1-cluster solution, do not exceed 5% compression. Clearly either the noise levels in the constrained analyses are high or they do not abstract all the structure possible. Zonation is apparently not an efficient method for identifying patterns, at least in this case.

Unconstrained Gaussian analysis (Model 1)

The unconstrained Gaussian result assignment of samples to clusters is shown in Table 3. Two clusters, 4 and 5, are primarily found in the deeper (older) samples, with cluster 6 slightly later. Clusters 1, 2 and 3 are largely in the upper levels, with occasional appearances of cluster 4, but there is

2 The odds refer to the likelihood of the smaller message length differing from the larger

Table 3. Assignments of samples, unconstrained Gaussian analysis (model 1). Sample number, message length and cluster. All probabilities of belonging are close to 1.0.

Sample depth cm	Length and Cluster	Sample depth cm	Length and Cluster
15	51.0 3	235	80.3 1
35	50.7 3	245	81.7 1
45	61.8 2	255	62.1 2
55	48.8 3	265	90.6 6
65	48.8 4	275	80.8 1
75	52.5 4	285	86.9 6
85	64.3 2	295	95.2 6
95	66.8 2	305	86.6 6
105	78.7 1	315	87.4 6
115	76.6 1	325	87.1 6
125	49.9 3	335	86.7 6
135	85.1 4	345	49.9 3
145	49.9 3	355	90.1 4
155	84.7 4	365	72.6 5
165	52.3 3	375	92.4 4
175	65.0 2	385	82.9 4
185	51.7 3	395	68.0 5
195	58.4 3	405	65.9 5
205	80.5 1	415	67.1 5
215	55.6 3	425	65.8 5
225	85.7 4	435	81.6 5

no obvious tendency to form clear zones. However, a good deal of compression is achieved, much more than in any other result, as indicated by the large reduction in message length (1324.4 nits; odds $1:e^{-1324.4}$ compared to the single cluster solution).

The unconstrained result allows fuzzy assignment, but little fuzziness was apparent in the assignments made, with all probabilities of belonging close to 1.0 for at most a single cluster.

Constrained Gaussian (Model 2)

The three clusters obtained do not provide a convincing alternative to the unconstrained Gaussian result. There is certainly compression compared to the 1-cluster solution, although the overall compression of 4% is not large. The zones are visually acceptable, with sample 33 forming a zone alone. The separation between samples 24 and 25 reflects a marked change in tree pollen quantities, though the cause of such a change is not obvious. Overall, the evidence for zonation as a 'good' model is not compelling.

Constrained Geometric (Model 3)

Recognising the compositional nature of the data further improves the model effectiveness compared to the constrained Gaussian, though not the unconstrained Gaussian. The 3 clusters of the geometric result include a singleton cluster but lose the quantitative distinctions identified in the constrained Gaussian analysis with the formation of zones 1 and 2 involving samples 1-24 and 25-32. Overall it does not appear that the compositional nature of the data is of great importance. The % capture (4.7%) is not high although overall the analysis is about as effective as the unconstrained Gaussian. But again the evidence for zonation as a model is not compelling.

Constrained censored analysis (Model 4)

Recognising the preponderance of absence values, the constrained censored Gaussian analysis identifies both the singleton sample and the Sample 24-25 boundary, thus providing all the information of the other 2 constrained analyses concerning zone structure. Even the 1-cluster solution is much preferred to the unconstrained Gaussian, although the % capture is again small. It seems that, if we take account of the presence/absence effect, zonation becomes a considerably better model than the others shown here. But much information may still be omitted, with such a low capture percentage.

In Figure 1 we show the probabilities of presence for all attributes in the censored analysis. The *Nothofagus* species provide discrimination in all cases, but other species also contribute both in terms of their presence and of their abundance when present. The singleton sample is not of course well characterised but zone 4 is largely identified with herbaceous taxa while zones 1 and 2 have more arboreal taxa. *Misodendron* is present in about half the samples in all zones. The differences can be summarised as follows:

- Zone 1. Ericaceae, *Nothofagus antarctica* and *N. betuloides* ubiquitous, *Nothofagus pumila* common, and Cyperaceae sporadic. Possibly forest with patches of more open vegetation.
- Zone 2. *Nothofagus antarctica, N. betuloides* and *N. pumila* with *Gunnera abundant.* The vegetation is high forest.
- Zone 3. *Gunnera* increasing with *Nothofagus betuloides. N. antarctica* and *N. pumila.* This isolated sample is probably an intermediate stage in the forest invasion, with *Gunnera* important as a nitrogen-fixing initial coloniser.
- Zone 4. Some *Nothofagus antarctica*, while Ericaceae, Cyperaceae, Poaceae, *Gunnera* and Polypodiophyta are more abundant. The *Nothofagus* may not be local in this instance, suggesting an open shrubland with a large herbaceous component. Most taxa have a low probability of occurrence indicating overall sparsity of vegetation cover.

Overall these zones do reflect the four major vegetation types of Bennet and Porter (2001).

Cluster sub-structure of the zones in the constrained censored Gaussian analysis

The zones for the population having been clearly distinguished, we now turn to the substructure of the zones for the censored Gaussian analysis, excluding the singleton Zone 3. An unconstrained clustering was used and Table 4 shows the summary results. All three zones show substructure, zone 1 with 4 clusters and zones 2 and 4 having 2 each. Table 5 shows the assignment of samples to sub-clusters within each zone. The main distinguishing features of each cluster are described below starting with the youngest.

Zone 1 shows the most complex pattern commencing with cluster 6, then alternating between 11 and 24 and later 23, before moving towards cluster 24 entirely. Again the environmental evidence is not available to permit surmises as to the causes of this variation. A simple trend does not seem to be a useful model, though, with oscillation between types more apparent.

Zone 2 is the most surprising since the 2 sub-clusters alternate, a pattern not identified previously. Villa-Martínez and Moreno (2007) have reported major variation in *Nothofagus* vegetation around these times, but not the periodicity.

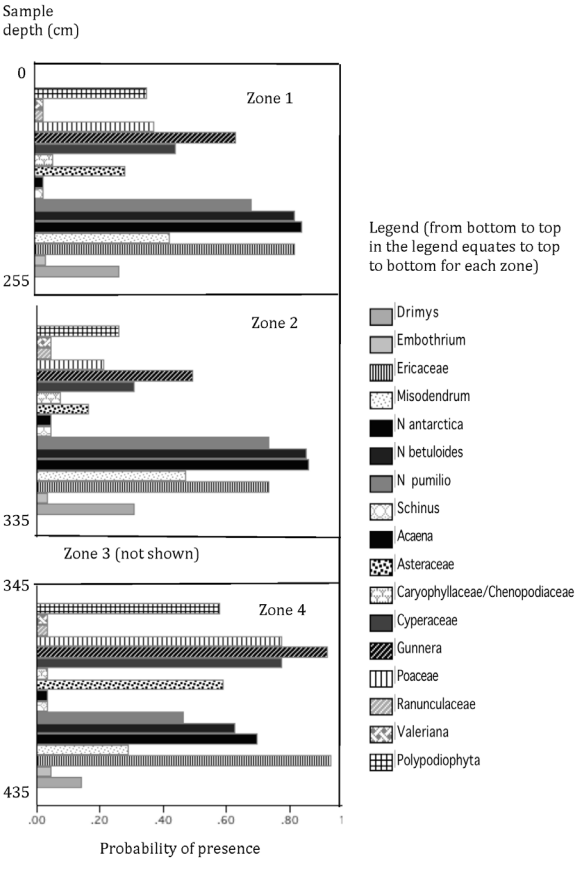

Figure 1. Attribute probabilities of presence for the four zones.

Table 4. Within zones unconstrained analysis of the censored Gaussian zones.

Zone	1-class Message Length (nits)	Optimal Number of Clusters, n	n-class Message Length (nits)	Difference In Message Length 1-class – n-class	% Capture
1	2041.8	4	1653.4	388.4	19.0
2	777.7	2	705.3	72.4	9.3
3	942.9	2	817.9	125.1	13.3

Table 5. Assignment of Samples within subclusters for the unconstrained analysis of the censored Gaussian zones. Columns show: sample number (depth), message length (ML) for samples (nits) and assigned cluster.

	Sample	ML (nits)	Cluster
Zone 1	15	52.2	24
	35	51.4	24
	45	67.4	23
	55	50.1	24
	65	49.8	24
	75	52.4	24
	85	65.5	6
	95	62.0	24
	105	70.5	6
	115	54.7	24
	125	51.4	24
	135	74.1	23
	145	51.4	24
	155	68.6	23
	165	51.8	24
	175	69.5	11
	185	52.2	24
	195	71.2	23
	205	70.8	11
	215	51.2	24
	225	70.8	11
	235	67.0	6
	245	66.3	6
	255	64.5	6
Zone 2	265	70.4	10
	275	88.7	7
	285	68.5	10
	295	91.5	10
	305	66.0	7
	315	86.1	7
	325	66.5	10
	335	9.19	7
Zone 3		Singleton	
Zone 4	345	54.2	14
	355	87.1	13
	365	85.2	13
	375	103.4	13
	385	87.6	13
	395	54.3	14
	405	52.9	14
	415	58.6	14
	425	53.0	14
	435	102.1	13

Again, a simple trend is not apparent.

Zone 3 is too small to analyse further.

Zone 4 has 2 major periods extending over considerable time periods but with anomalies at the beginning and end. It is likely that this pattern is related to local glacial features as well as distance to pollen sources but we do not have the evidence to confirm this.

To identify the main distinguishing vegetation characteristics between the sub-clusters, Figure 2 reports the attributes means for those attributes showing significant differences from the population value.

For Zone 1 (Figure 2a), we have the following subclusters:

- cluster 6 has moderate *Drimys,* Asteraceae, Caryophyllaceae/Chenopodiaceae and Polypodiophyta, showing links with more open earlier vegetation. It is mostly confined to the earliest samples.
- cluster 11 has some *Cyperaceae*, and high *Acaena* with lower *Nothofagus antarctica,* perhaps indicating a wetter environment.
- cluster 23 has higher *Nothofagus antarctica* with high *Drymis*, Cyperaceae, Poaceae, Asteraceae and *Gunnera* which might show a more open habitat. Ericaceae is at a relatively high level and both *Misodendrum* and Polypodiophyta attain their greatest abundance.
- cluster 24 has no high valued taxa other than *Nothofagus antarctica* and Ericaceae, but *Misodendrum* is relatively sparse, as is *Gunnera*. It is perhaps the typical forest zone being the most extensive of the sub-clusters. The samples are predominantly late in occurrence.

Zone 2 (Figure 2b) has basically a rich sub-cluster 7 and a poor sub-cluster 10. 7 has very high *Nothofagus antarctica*, with *Drimys*, *Embothrium*, Ranunculaceae and Polypodiophyta but *Misodendrum* is higher in the poorer cluster 10. This zone seems to represent the dominant forest zone but the periodic pattern is not easily explained. There have been suggestions that wind patterns shifted north and south (Villa-Martinez and Moreno, 2007) causing variations in rainfall but these have not been shown to be periodic.

Zone 4 (Figure 2c) again shows a rich and a poor subzone structure with very high Cyperaceae, Caryophyllaceae/Chenopodiaceae, Asteraceae, *Schinus, Valeriana* and *N. pumilio* in sub-cluster 13. *N. antarctica* is present at low levels in both sub-clusters, but less in sub-cluster 14 which is quite poor for all taxa. The predominance is of herbaceous taxa, and possibly there was little arboreal vegetation close to the lake at this time and the variation is related to the distance to the vegetated area. Zone 4 is, of course, composed of the earliest samples.

The within-zone analyses have been successful in identifying further patterns in these data, thus recovering some of the information lost due to the imposition of zones

Discussion

Zonation as a structuring paradigm

Comparison of the Gaussian analyses shows that zonation is not a particularly good model. Not only is the message length considerably increased but the % capture of the zonation is low. These data are much noisier, with % capture <5%, than much vegetation data which usually manage at least 20%. Other unpublished analyses do produce higher percentages in which case zonation becomes more attractive. Zonation reflects an assumption by the user of the nature of change in vegetation through time, rather than a model derived from observation. It seems that this assumption may be unwarranted in the present example, but what is important is that MML allows us to subject our assumption to testing.

The other analyses are more informative. They suggest, firstly, that employing the Geometric distribution provided no great improvement over the unconstrained Gaussian, so that incorporating the compositional aspects of the data is apparently not of major importance. Secondly, and in contrast,

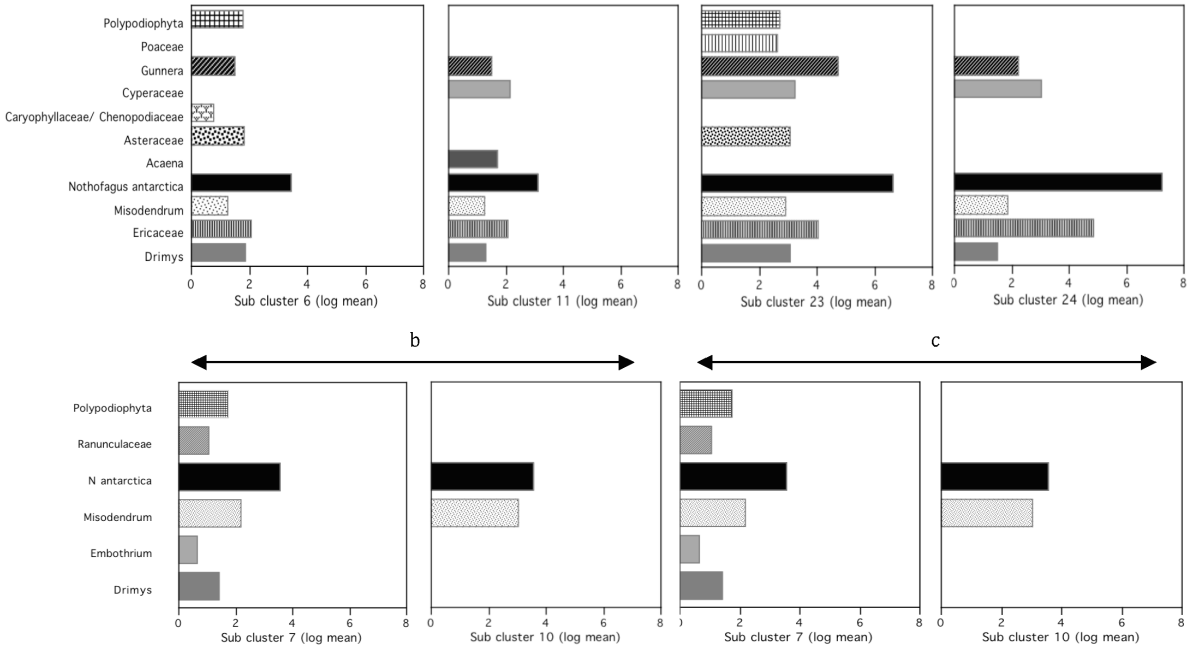

Figure 2. Attribute means for the subclusters within zones. Zone 3 is excluded.

the censored Gaussian, accounting for zero values, provided considerable improvement and provides the best result of any considered here. However, the % capture remained very low, which is not encouraging. Much variation remains which might be incorporated into the model and improve the overall capture.

There are alternative procedures to simple constrained analysis for incorporating inter-sample correlation (Visser and Dowe 2007, Visser et al. 2009) and we hope to examine these at a later date. It may also be possible to combine the within-zone with the constrained clustering instead of using a two-stage process. This would provide an overall message length for all of the pattern identified, which is not possible with the procedure adopted here. Other work on constrained clustering involves individual instance-level constraints where some items are specified as necessarily in the same, or in different, clusters (Davidson et al. 2007); that is we specify 'must-join' and 'must-not-join' rules for pairs of samples. Such constraints have been applied to segmentation (Yu and Shi 2004) for region-finding in images. However, they do not seem pertinent in the present case since there is no obvious rationale for imposing the rules.

Equally, support for the directed/chaotic patterns assumed by Orlóci (2010) does not appear to be very strong, although there is partial evidence for something similar. There seems to be more support for an ebb-and-flow ('pseudo-periodic') pattern, though more complex patterns can also be found within zones. Green (1982) has argued that most big changes in forest vegetation are associated with major disturbances such as fire, which will increase the local variability used to distinguish chaotic episodes. Thus the appearance of 'chaos' may not indicate a non-directed phase in the vegetation structure. It could also result from local recovery after disturbance. Indeed, evidence for directed phases is slim.

The evidence obtained from these analyses does not suggest that zones are well supported for these data. This does not mean that zonation should always be rejected of course and the occurrence of zones may be context dependent. User expectation of a particular structure may be based on wider evidence. However, failure to recover potential structure is not to be regarded as a trivial matter. MML permits the quantification of the losses which can be used to provide a principled decision of the usefulness of zonations.

Within zone structure

It is clear that much structure remains within zones, and that this could provide information concerning the driving forces for the fluctuations. In the earliest samples, there seems to be an alternation between more and less pollen overall, perhaps reflecting the distance to the sources, but the zone 2 variations would seem to be essentially periodic, though, without exact dating, this must remain but a reasonable hypothesis. The late stages show a more complex pattern, involving more attributes. This might perhaps be modelled by some kind of Markov process, or even a semi-Markov Lévy process (Akgiray and Lamoureux 1989) which allows modelling of the (variable) time period during which transitions are not permitted between clusters, but the number of samples within each zone is very small and would barely provide sufficient evidence. Perhaps the methods of Allison et al. (1998) would allow still more structure to be captured and avoid the problems of non-stationarity and alignment required by standard time-series approaches (Berryman 1992).

In fact, further analyses of other pollen datasets (Dale, unpublished) using the unconstrained Gaussian model do sometimes result in evidence of strong zonation, at least over parts of the profile. Data from Bega swamp (Hope et al. 2000) provide such an example but even here the zones are interspersed with other sections showing more heterogeneity.

Even when zones appear acceptable, they may have complex properties. Analysis of pollen data (Dale, unpublished) from Lynch Crater (Kershaw 1976) provide long sections assigned to the same class, with occasional interspersions either from spatially adjoining clusters, both above and below, or from members of a single cluster which appears in small numbers over half or more of the total range of samples. In both cases, the interspersed types are probably quite similar to the cluster in which they are embedded, as is shown by the fuzzy assignments. These are not common, most assignments having probability =1, but on the rare occasions when the probability of the major class is <1, the alternative, interspersed cluster provides the remainder of the probability mass; i.e., the two clusters are similar. The maximum alternative probability does not exceed $p=0.38$, with most values around $p \leq 0.04$, and the ambiguity never involves more than 2 clusters. So, with these data, a zonation would be acceptable. However, several clusters appear in sequence in multiple blocks, each at a different depth, so forming several zones of the same nature separated by various other types of zone.

Short-term studies of succession

By examining other information on succession it may be possible to explicate such overall pattern. Dale (2000) used data from Webb et al. (1967) from sub-tropical rainforest regeneration over a few years and found the earliest phases were spatially homogeneous but then dissipated into several small local areas with common temporal development. The data only cover a few years and probably the changes would not be distinguishable in a pollen profile, especially as the surrounding vegetation was pristine forest.

Over a longer time period of 14 years, Dale and Dale (2002), examining salt marsh vegetation, found that the cluster inter-relationships could be represented by a graph structure. In this there were small cliques of clusters with numerous transitions between them and these several groups were linked by rare transitions, often seemingly irreversible. The cliques themselves were arranged along two distinct axes, not necessarily orthogonal, one very short, the other much longer, diverging from a common base cluster. The cliques are possibly related to the carousel model of van der Maarel

and Sykes (1993) and the rarer changes to events of deposition and erosion affecting the water movement patterns, the equivalent of Green's (1982) significant environmental events.

There is, therefore, evidence from other research of zonation at very short times scales, possibly not detectable with most pollen data. At slightly larger time scales, there is a possible pattern of cliques of clusters, strongly interacting, with relatively brief, and occasional, transitions between the cliques. These would probably be detectable in high resolution pollen data. This evidence, which is not compelling, does not appear to support homogeneous zones nor long, directed series of changes, unless the cliques are themselves ordered. However, it is likely that different types of vegetation will show different patterns, depending on the migration of species and the diversity of the vegetation.

Spatial ordering

Using spatially ordered data, such as Gillison and Brewer's (1985) gradsects with temporal sequences replaced by environmental ones poses some additional problems; for example, time series are always asymmetric from the past to the future whereas spatial series are mostly symmetric. The two-dimensional nature of the inter-sample relationships might be addressed by using Markov random fields (Lafferty et al. 2001, Yang and Jiang 2003, Molloy et al. 2006), instead of simpler Markov processes. If time is also included, a 3-dimensional dataset results with further complexity. This would considerably complicate a study such as that of Li et al. (2001) which developed a graphical model representing the processes of change on a salt marsh. Examining the time series in the 3-dimensional is possible and would come close to meeting Murrell et al.'s (2001) desire for a merger of pattern (spatial) and process (temporal). Alternative methods, such as Wallace (1998), use estimates of the correlation existing between samples which avoids some of these complications.

Dale et al. (1988) have looked at directly clustering spatial transects using a Levenshtein similarity measure (not MML) but the results were somewhat equivocal and not particularly illuminating. In part this might have been due to the absence of 'time-warping' transformations so that the overall length of the sequences assumed too great an importance. A re-analysis of these data using MML is being considered, avoiding alignment of the sequences using the model-based method described by Powell et al. (2004).

The efficacy of MML and some possible alternative criteria

There remains the question of whether MML has proved useful in selecting a model or whether alternative selection criteria might need to be added. What properties do we require of a model selection procedure? Plausability is an obvious candidate, implying both accuracy and coverage, and MML is a measure of this. Another suggestion is coherence (Thagard 1978), defined as compliance with background knowledge.

The models we are considering are statistical models, rather than, say, the individualistic models espoused by Galitskii (1999) or Grimm (1999) or the qualitative models of Garrett et al. (2007). Probably models such as the Bayesian networks of O'Donnell et al. (2006) or Comley and Dowe (2005) provide our ultimate target, allowing simple representations derived from limited data but with richer representations given massive data. Such networks represent a mapping of the processes operating to change vegetation. Gell-Mann and Lloyd (1996) propose a related measure (total information) but stress the subjectivity of the relationship between natural phenomena and an interested observer, which relationship forms complexity.

MML combines simplicity with fit to data. To these two criteria we might add others, such as predictability, explicability, actionability, generalisability and interestingness. Does simplicity contribute to these?

1. *Predictability* (Gower 1974, Mac Nally 2000, Crutchfield and Young 1989, Shalizi and Crutchfield 2001). This is the basis of Popper's falsification approach and provides model validation; Fisher (1992) regards it as a pessimistic approach to induction. But what do we want to predict? Presumably we would like to extrapolate to novel situations. Solomonoff (2008) recognised three kinds of probabilistic induction concerned with extrapolation:

> i. Extrapolation of a series of strings and/or numbers, which is the case here.
>
> ii. Extrapolation of an unordered set of strings and/or numbers.
>
> iii. Extrapolation of an unordered set of pairs of elements which may be strings or numbers.

Presumably there is a fourth category for extrapolation using a series of strings and/or numbers. Sunehag and Hutter (2010) discussed long term prediction using compression of the previous history together with side information. MML is appropriate for selecting the compressed representation of the history as well as assessing the effects of ancillary information.

V'yugin (1999) finds series are almost always predictable and general solutions for Solomonoff's cases are known and have been shown to be consistent, that is, to converge to provide precise predictions as more data are examined. Solomonoff (2008) himself used complexity, Ockham's razor and a Bayesian approach similar to MML differing only in the choice of the *a priori* probabilities. Dowe et al. (1996) suggested using probabilistic prediction as an alternative to categorical 'yes/no' and ecologically this makes sense, providing information on risks of specific choices. MML has been used to obtain regression models (Fitzgibbon et al. 2002), autoregressive models (Fitzgibbon et al. 2004) or decision graphs (Fitzgibbon et al. 2002, Needham and Dowe 2001, Dale et al. 2007). Prediction is improved by combining

multiple models, appropriately weighted (Yin and Davidson 2004, Zhang and Lu 2010) and MML allows the use of the '*a posteriori'* probabilities of the models as weights. In sum, MML seems to have many properties which are desirable for prediction.

2. *Explicability* (Kodratoff 1986, Diday 1988, Sommer 1995). If we seek to explicate the patterns in terms of environmental and other processes, it is plausible that a simple model will be more easily explicable than a more complex one. Incorporating ancillary information on associated environmental variables is possible (Visser et al. 2009). Sober (1994) tied simplicity to "informativeness" in a specified context, combining comprehension with predictability. Fitzgibbon et al. (2003) show that the Minimum Message Length principle can be viewed as an epitome criterion that produces epitomes having appropriate properties, such as easy point estimation, human comprehension and fast approximation of posterior expectations. Predictability and explicability can be combined when using structurally isomorphic models, but predictability by itself requires only the use of behaviourally isomorphic models (Yamada and Amoroso 1971). Again it appears that MML can expect to contribute to the goal of explication.

3. *Actionability* (Adomavicius and Tuzhlin 1997). The objective here is to seek results which may be directly applied in the solution of one or more problems and this must involve a specification of the problems to be addressed. It also involves predictivity since we must be able to predict the results of some actions taken. MML is not directly involved in assessing actionability.

4. *Generalisability* (Zhu and Rohwer 1995). Results should have wide applicability, but over what domains? Simple models generalise more widely than complex ones, the latter being prone to overfitting. MML provides the means for preventing such over-exuberance! Wallace (1996) suggests MML minimises our surprise over future events, which is related to generalisability and predictability. Popper (1992) has argued that simpler models are more general and therefore more easily falsifiable, since there are more possible tests.

Generalisability might be regarded as automatic formation of hypotheses. In this context, Srinivasan et al. (1994) discuss the extension from propositional theories to first order logic, choosing the one which is least likely to have explained the data by chance, while using any relevant prior knowledge. Their proposal uses the degree to which an hypothesis allows compression of the observed data in the presence of noise. In fact, a class of concepts is only PAC (probably approximately corrrect) learnable if the hypothesis is smaller than the data (Li and Vitanyi 1989). MMLis clearly related to this approach.

5. *Interestingness* (Schmidhuber 1997). Interesting results are clearly to be desired, but interest will depend on who is doing the evaluation and on the context of the evaluation. What is interesting today may not be so tomorrow. Interest can be both objectively and subjectively assessed, Hilderman and Hamilton (1999) suggesting 8 properties associated with interest. Five of these were objective measures in that they did not require application or domain knowledge, and MML could be useful in unifying and quantifying some of these. The remaining three required some expert or subjective evaluation, though this might be provided by ancillary information on dates and environmental variables. But unless our theories have explicit associated semantic descriptions, automated evaluation of interest will remain at best only partially obtainable. For example, unexpectedness has been regarded as an indicator of interesting results (Silberschatz and Tuzhilin 1996), but this involves the subjective expectations of the user, who must decide what is surprising, i.e.unlikely. MML is inherently probabilistic and could perhaps provide an objective measure of surprise. But, as noted above, Wallace (1996) argues the MML minimises surprise!

In summary, MML can provide useful information on a variety of model evaluation criteria. Simple, but adequate, models should prove more easily testable, generalisable, explicable and predictive, and may result in models which are more actionable.

User expectation

In this paper we introduced user expectation into our model by choosing constrained clustering as a suitable model and we can determine if such a postulate was desirable by assessing the effects on the message lengths. Properties such as causality (Hanson 1990, Wallace 1996, Dai et al. 1996, Neil and Korb 1998, Arnold et al. 2007) might also be considered. Perhaps by investigating at a detailed time scale we might provide more information on the operative processes and hence on their possible causes. Causation also involved relationships with other non-vegetative variables, leading back to predictiveness, but causality is a difficult concept (see e.g., footnote 169, p. 541 in Dowe 2008a). MML has been use to derive Bayesian networks (Comley and Dowe 2005). which have been proposed as a suitable formalism for representing and learning causal maps as coherent representations of causal relations among events (Gopnik and Glymour 2002).

Another possible criterion is consilience (Whewell 1847) which requires a model to be predictive, explanatory and unifying the evidence. This seems to be a combination of other criteria and therefore needs no further examination.

Overall, MML, in balancing simplicity and goodness of fit, provides a useful criterion for model selection, as does the related minimum description length criterion (Rissanen 1995, Barron et al. 1998, Lanterman 2007). It permits not only the successful comparison of models of different complexity but also the assessment of user beliefs embodied in them.

Acknowledgements: We acknowledge the provision of the data used here by V. Markgraf and the Latin American Pollen Database for the Clarence Data. We also acknowledge Hope, Singh, Geissler, Glover and O'Dea (2000) and the Indo-Pa-

cific Pollen Database for the Bega data, and Kershaw (1976) and the same database for the Lynch crater data.

References

Adomavicius, G. and Tuzhilin, A. 1997. Discovery of actionable patterns in databases: the action hierarchy approach. In: D. Heckerman, H. Mannila, D. Pregibon and R. Uthurusamy (eds.), *Proceedings 3rd International Conference on Knowledge Discovery and Data Mining.* AAAI. pp. 111-114.

Agusta, Y. and Dowe, D. L. 2003. Unsupervised learning of correlated multivariate Gaussian mixture models. *Lecture Notes in Artificial Intelligence* 2903, Springer-Verlag, Berlin. pp. 477-489.

Aitchison, S. and Kay, J. W. 2003. Possible solutions of some essential zero problems in compositional data analysis. CODAWORK'03 Girona: La Universitat. 6 pps. http://hdl.handle.net/10256/652

Akgiray, V. and Lamoureux, C. G. 1989. Estimation of stable-law parameters: A comparative study. *J. Business Econ. Stat.* 7:85-93.

Allison, L., Edgoose, T. and Dix, T. I. 1998. Compression of strings with approximate repeats. In: J. I. Glasgow, T. G. Littlejohn, F. Major, R. H. Lathrop, D. Sankoff and C. Sensen (eds.) *Proceedings 6th International Conference on Intelligent Systems in Molecular Biology (ISMB'98)*, Montreal. pp. 8-16.

Arnold, A., Liu, Y. and Abe, N. 2007. Temporal causal modelling with graphical granger methods. In: Berkhin, P., Caruana, R., Wu, X. and Gaffney, S. (eds.) *Proceedings 13th ACM SIGKDD International Conference Knowledge Discovery and Data Mining.* Association for Computing Machines, New York, pp. 66-75.

Babad, Y. M. and Hoffer, J. A. 1984. Even no data has value. *Communications of the Association of Computing Machines* 27: 748-756.

Balasubramanian, V. 1997. Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation* 9:349-368.

Barron, A. R., Rissanen, J. and Yu, B. 1998. The minimum description length principle in coding and modeling. *IEEE Trans. Information Theory* 44:2743-2760.

Baxter, R. A. and Oliver, J. J. 2006. The kindest cut: minimum message length segmentation. *Lecture Notes in Computer Science*, Springer, Berlin. 1180:83-90.

Bennett, K. D. and Porter, C. 2001. Late quaternary dynamics of Western Tierra del Fuego. Uppsala Universitet: http://www.geo.uu.se/ Institutionen för geovetenskaper: Paleobiologi: forskning.

Berryman, A. A. 1992. On choosing models for describing and analyzing ecological time series. *Ecology* 73: 694-698.

Blum, A., Hellerstein, L. and Littlestone, N. 1995. Learning in the presence of finitely or infinitely many irrelevant attributes. *J. Comput. Syst. Sci.* 50:32-40.

Boulton, D. M. and Wallace, C. S. 1970. A program for numerical classification. *Computer J.* 13: 63-69.

Bradshaw, R. H. W. 1981. Quantitative reconstruction of local woodland vegetation using pollen analysis from a small basin in Norfolk, England. *J. Ecol.* 69:941-955.

Bunting, M. J. and Middleton, R. 2009. Equifinality and uncertainty in the interpretation of pollen data: the Multiple Scenario Approach to reconstruction of past vegetation mosaics. *The Holocene* 19:799-803.

Comley, J. W. and Dowe, D. L. 2005. Minimum message length and generalized Bayesian net with asymmetric languages. In: P. Grunwald. I. J. Myung and M. A. Pitt (eds.) *Advances in Minimum Description Length:Theory and Applications* Chapter 11. MIT Press, Cambridge. pp. 265-294.

Crutchfield, J. P. and Young, K. 1989. Inferring statistical complexity. *Phys. Rev. Lett.* 63: 105-108.

Dai, H., Korb, K. B., Wallace, C. S. and Wu, X. 1996. A study of casual discovery with weak links and small samples. *Proceedings 15th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc, San Francisco USA. pp. 1304-1309.

Dale, M. B. 2000. Mt Glorious revisited: secondary succession in subtropical rainforest. *Community Ecol.* 1:181-193.

Dale, M. B. 2007. Changes in the model of within-cluster distribution of attributes and their effects on cluster analysis of vegetation data. *Community Ecol.* 8: 9-14.

Dale, M. B., Allison, L. and Dale, P. E. R.. 2007. Segmentation and clustering as complementary sources of information. *Acta Œcol.* :1-10. **VOL??**

Dale, M. B., Allison. L. and Dale, P. E. R.. 2010. A model for correlation within clusters and its use in pollen analysis. *Community Ecol.* 11:51-58.

Dale, M. B. and Clifford, H. T. 1976. The effectiveness of higher taxonomic ranks for vegetation analysis. *Austr. J. Ecol.* 1: 37-62.

Dale, M. B., Coutts, R. and Dale, P. E. R. 1988. Landscape classification by sequences: a study of Toohey Forest. *Vegetatio* 29: 113-129.

Dale, M. B., Dale, P. E. R. and Tan, P. J. 2007. Supervised clustering using decision trees and decision graphs: a ecological comparison. *Ecol. Model.* 204:70-78.

Dale, M. B. and Wallace, C. S. 2005. Hierarchical clusters of vegetation types. *Community Ecol.* 6:57-74.

Dale, P. E. R. and Dale, M. B. 2002. Optimal classification to describe environmental change: pictures from an exposition. *Community Ecol.* 3:19-30.

Davidson, I., Eter, M. and Ravi, S. S. 2007. Efficient incremental constrained clustering. In: Berkhin, P., Caruana, R., Wu, X. and Gaffney, S. (eds.) *Proceedings 13th ACM SIGKDD International Conference Knowledge Discovery and Data Mining.* Association for Computing Machines, New York, pp. 240-249.

Diday, E. 1988. The symbolic approach in clustering and related methods of data analysis: the basic choices. In: H. H. Bock (ed.) *Classification and Related Methods of Data Analysis*, North Holland, Amsterdam. pp. 673-683.

Douglass, D. C., Singer, B. S., Kaplan, M. R., Ackert, R. P., Mickelson, D. M. and Caffee, M. W. 2000. Evidence of early Holocene glacial advances in southern South America from cosmogenic surface-exposure dating. *Geology* 33:237-240.

Dowe, D. L. 2008a. Foreword re C. S. Wallace. *Computer J.* 51: 523-560.

Dowe, D. L. 2008b. Minimum Message Length and statistically consistent invariant (objective?) Bayesian probabilistic inference - from (medical) ''evidence'' *Social Epistemology* 22:433-460

Dowe, D. L., Farr, G. E., Hurst, A. J. and Lentin, K. L. 1996. Information-theoretic football tipping. In: N. de Mestre (ed.), *3rd Conference on Mathematics and Computing in Sport*. Bond University. pp. 233-241.

Fesq-Martin, M., Friedman, A., Peters, M., Behrman, J. and Kilian, R. 2004. Late-glacial and Holocene vegetation history of the Magellanic rain forest in Southwestern Patagonia, Chile. *Vegetation History and Archaeobotany* 13:249-255.

Fisher, D. 1992. Pessimistic and optimistic induction. Tech. Rep. CS-92-12, Dept. Computer Sci., Vanderbilt Univ., Nashville.

Fitzgibbon, L. J., Allison, L. and Dowe, D. L. 2000. Minimum message length grouping of ordered data. *Lecture Notes in Computer Science* 1968: Proceedings 11th International Conference on Algorithmic Learning Theory. Springer-Verlag, London. pp. 56-70.

Fitzgibbon, L. J., Dowe, D. L. and Allison, L. 2002. Univariate polynomial inference by Monte Carlo message length approximation. In: C. Sammut and A. G. Hoffman (eds.) *Proceedings 1*9th International Conference on Machine Learning (ICML'2002), Sydney, Australia, Morgan Kaufmann, San Francisco. pp. 147-154.

Fitzgibbon, L. J, Dowe, D. L. and Allison, L. 2003. Bayesian posterior comprehension via message from Monte Carlo. *Proceedings 2nd Hawaii International Conference on Statistics and Related Fields.* http://www.csse.monash.edu.au/~leighf/papers/Fitzgibbon03b.pdf

Fitzgibbon, L. J., Dowe, D. L. and Vahid, F. 2004. Minimum message length autoregressive model order selection. *Proceedings of the International Conference on Intelligent Sensing and Information Processing (ICISIP 2004)*, Chennai, India, 4-7 January 2004, IEEE Operations Center, Piscataway, NJ, USA, ISBN: 0-7803-8243-9, pp. 439-444.

Gale, M. and Ball, L. J. 2002. Does Positivity Bias Explain Patterns of Performance on Wason's 2-4-6 task? In: W. D. Gray and C. D. Schunn (eds.) *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, Routledge, p. 340-344.

Galitskii, V. V. 1999. Modelling of the plant community: An individual-oriented approach. 1. A model of the community. *Biology Bulletin* 26(2). (Translated from *Izvestia Akademii Nauk, Seria Biologicheskaya*, 2000, No. 2, pp.178-185.

Garrett, S. M. Coghill, G. M., Srinivasar, A. and King, R. D. 2007. Learning Qualitative Models of physical and biological systems. In: S. Dˇeroski, P. Langley and L. Todorovski (eds.), *Computational Discovery of Scientific Knowledge. Lecture Notes in Artificial Intelligence* 4660:248-272.

Gell-Mann, M. and Lloyd, S. 1996. Information measures, effective complexity and total information *Complexity* 2:44-52.

Gillison, A. N. and Brewer, K. R. W. 1985. The use of gradient directed transects or gradsects in natural resource surveys. *J. Ecol. Manage.* 20:103-127.

Gopnik, A. and Glymour, C. 2002. Causal maps and Bayes nets. A cognitive and computational account of causal learning and theory formation. In: P. Carruthers, S. Stich and M. Siegel (eds.) *The Cognitive Basis of Science.* Cambridge University Press, Cambridge. pp. 117-132.

Gower, J. C. 1974. Maximal predictive classification. *Biometrics* 30:643-654.

Green, D. G. 1982. Fire and stability in the postglacial forests of southwest Nova Scotia. *J. Biogeogr.* 9: 29-40.

Grimm V. 1999. Ten years of individual-based modelling in ecology: what have we learned, and what could we learn in the future? *Ecol. Model.* 115:129-148.

Hanson, S. J. 1990. Conceptual clustering and categorization: Bridging the gap between induction and causal models. In: R. S. Michalski and Y. Kodratoff (eds.), *Machine Learning: An Artificial Intelligence Approach* III, Morgan Kaufmann, San Mateo, CA. pp. 235-268.

Hilderman, R. J. and Hamilton, H. J. 1999. Heuristic measures of interestingness. In: Z. M Zytow and J. Rauch (eds.), *Proceedings 3rd European Confeence on the Principles of Data mining and Knowledge Discovery (PKDD).* Lecture Notes in Computer Science 1704, Springer, Berlin. pp. 232-241.

Hope, G., Singh, G., Geissler, E., Glover, L. and O'Dea, D. A 2000. Detailed Pleistocene-Holocene vegetation record from Bega Swamp, southern New SouthWales. In: J. Magee and C. Craven (eds.) *Quaternary Studies Meeting, Regional Analysis of Australian Quaternary Studies: strengths, gaps and future directions*, Department of Geology, Australian National University, Canberra ACT. pp. 48-50.

Jackson, S. T. and Williams, J. W. 2004. Modern analogs in quaternary palæoecology: here today, gone yesterday, gone tomorrow? A*nnu. Revi. Earth Planetary Sci.* 32:495-537.

Joosten, H. 2007. In search of finiteness: the limits of fine resolution palynology of *Sphagnum* peat. *The Holocene* 17:1023-1031.

Kershaw, A. P. 1976. A Late Pleistocene and Holocene pollen diagram from Lynch's Crater, northeastern Queensland, Australia. *New Phytol.* 77:469-498.

Kodratoff, Y. 1986. *Leçons d'apprentissage symbolique*, Cepaduesed., Toulouse

Lafferty, J., McCallum, J. A. and Pereira, F. 2001. Conditional Random Fields: probabilistic models for segmenting and labelling sequence data. *International Conference on Machine Learning* (ICML'01). pp. 282-289.

Lanterman, A. D. 2007 Schwarz, Wallace and Rissanen: intertwining themes in theories of model selection. *Internat. Stat. Rev.* 69:185-212.

Larossa, J. M. C. 2005. Compositional time series: past and present. *EconWPA Econometrics* 0510002. http://129.3.20.41/eps/em/papers/0510/0510002.pdf

Legendre, P. and Gallagher, E.. 2001. Ecologically meaningful transformations for ordination of species data. *Ecology* 270: 271-280.

Li, C., Biswas, G., Dale, M. B. and Dale, P. E. R. 2001. Building Models of Ecological Dynamics using HMM-based Temporal Data Clustering. In: Advances in Intelligent Data Analysis, 4th International Conference on Intelligent Data Analysis*, Lecture Notes in Computer Science* 2189, Springer, pp. 53-62.

Li, M. and Vitanyi, P. 1989. Inductive reasoning and Kolomogorov complexity. In: *Proceedings 4th Annual IEEE Structure in Complexity Conference*, Eugene. IEEE Computer Society Press. pp. 165-185.

Mac Nally, R. 2000. Regression and model-building in conservation biology, biogeography and ecology: the distinction between-and reconciliation of-'predictive' and 'explanatory' models. *Biodivers. Conserv. 9: 655-671.*

Markgraf, V. 1983. Late and Postglacial vegetational and palæoclimatic changes in subantarctic, temperate, and arid environments in Argentina. *Palynology* 7: 43-70.

Molloy, S., Albrecht, D. W., Dowe, D. L. and Ting, K. M. 2006. Model-Based clustering of sequential data. *Proceedings 5th Annual Hawaii International. Conference on Statistics, Mathematics and Related Fields*, 16th-18th January, 2006, Hawaii, U.S.A. 22 pages.

Murrell, D. J., Purves, D. W. and Law, R. 2001. Uniting pattern and process in plant ecology. *Trends Ecol. Evol.* 16:529-530.

Myung, J., Balasubramanian, V. and Pitt, M. A. 2000 Counting probability distributions: differential geometry and model selection. *PNAS* 97:11170-11175

Needham, S. L. and Dowe, D. L. 2001. Message length as an effective Ockham's razor in decision tree induction. In: *Proceedings 8th International Workshop of Artificial Intelligence and Statistics (AIS- TATS 2001)*, Key West, FL. pp. 253-260.

Neil, J. R. and Korb, K. B.. 1998. The MML evolution of causal models Tech. Rep. 98/17 Dept Comput. Sci., Monash University, Melbourne.

O'Donnell, R. T., Allison, L. and Korb, K. B. 2006. Learning hybrid Bayesian networks by MML. *Lecture Notes in Computer Science* 4304 :192-203. Springer, Berlin.

Oliver, J. J., Baxter, R. A. and Wallace, C. S. 1998. Minimum message length segmentation. In: X. Wu, R. Kotagiri and K. B. Korb (eds.) *Lecture Notes in Artificial Intelligence* 1394: 222-233. Research and Development in Knowledge Discovery and Data Mining, Second Pacific-Asia Conference, PAKDD-98 Melbourne Australia, 15-17 April 1998, Springer-Verlag, Berlin.

Orlóci, L. 2010. Multi-scale trajectory analysis: powerful conceptual tool for understanding ecological change. *Front. Biol. China* 4:158-179.

Orlóci, L. and He, K. S. 2009. On governance in the long-term vegetation process: How do we discover the rules? *Front. Biol. China* 4:557-568.

Orlóci, L., Pillar, V. D. and Anand, M. 2006. Multiscale analysis of palynological records: new possibilities. *Community Ecol.* 7:53-67.

Paez M. M., Schäbitz, F. and Stutz, S.. 2001. Modern pollen-vegetation and isopoll maps in southern Argentina. *J. Biogeogr.* 28:997-1021.

Pickett, E. J., Harrison, S. P., Hope, G., Harle, K., Dodson, J. R., Kershaw, A. P., I. Prentice, I. C., Backhouse, J., Colhoun, E. A., D'Costa, D., Flenley, J., Grindrod, J., Haberle, S., Hassell, C, Kenyon, C., Macphail, M., Martin, H., Martin, A. H., McKenzie, M., Newsome, J. C., Penny, D., Powell, J., Raine, J. I., Southern, W., Stevenson, J., Sutra, J-P., Thomas, I., van der Kaars, S. and Ward, J. 2004. Pollen-based reconstructions of biome distributions for Australia, Southeast Asia and the Pacific (SEAPAC region) at 0, 6000 and 18,000 14C yr BP. *J. Biogeogr.* 31: 1381-1444.

Popper, K. 1992. *The Logic of Scientific Discovery* Chapter 7. Simplicity. Routledge, London. pp. 121-132.

Powell, D. R., Allison, L. and Dix, T. I. 2004. Modelling-alignment for non-random series. In: *Lecture Notes in Artificial Intelligence* 3339, Springer, Berlin. pp. 203-214.

Prentice I. C. 1985. Pollen representation, source area and basin size: towards a unified theory of pollen analysis. *Quat. Res.* 23:76-86.

Prentice, I. C., Guiot, J., Huntley, B., Jolly, D. and Cheddadi, R. 1996. Reconstructing biomes from palæoecological data: a general method and its application to European pollen data at 0 and 6 ka. *Climate Dynamics* 12: 185-194.

Rahwan, T. and Jennings. N. R. 2008. An improved dynamic programming algorithm for coalition structure generation. In: L. Padgham, D. C. Parkes, J. Mueller and S. Parsons (eds.) *Proceedings 7thInternational Conference on Autonomous Agents and Multiagent systems (AAMAS)*, Estoril, Portugal. pp. 1417-1420.

Riddle, R. R. and Hafner, D. J. 1999. Species as unit of analysis in ecology and biogeography: time to take the blinkers off. *Global Ecol. Biogeogr.* 8: 433-441.

Rissanen, J. 1995. Stochastic complexity in learning. In: P. Vitányi (ed.) *Computational Learning Theory*. Lecture Notes in Computer Science 904. pp. 196-210.

Salzberg, S. 1986. Pinpointing good hypotheses with heuristics. In: W. A. Gale (ed.) *Artificial Intelligence and Statistics*. Addison-Wesley, Sydney. pp. 133-158.

Schader, M. 1979. Branch and Bound Clustering with a generalised scatter criterion. *Oper. Res. Verfahren* 30: 154-162.

Schmidhuber, J. 1997. What's interesting? Tech. Rep. IDSIA-35-97, IDSIA, Lugano, Switzerland.

Shalizi, C. R. and Crutchfield, J. P. 2001. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *J. Stat. Phys.* 104:819-881.

Silberschatz, A. and Tuzhilin, A. 1996. What makes patterns interesting. *I. E. E. E. Trans. Knowledge Data Engineering* 8: 275-281.

Sober, E. Let's Razor Occam's Razor 1994. In: D. Knowles (ed.) *Explanation and Its Limits* Cambridge University Press Cambridge. pp. 73-93.

Solomonoff, R. J. 2008. Three kinds of probabilistic induction: universal distributions and convergence theorems. *Computer J.* 51:566-570.

Sombattheera, C. and Ghose, A. 2008. A best-first anytime algorithm for computing optimal coalition structures. In: L. Padgham, D. C. Parkes, J. Mueller and S. Parsons (eds.), *Proceedings 7thInternational Conference on Autonomous Agents and Multiagent systems (AAMAS)*, Estoril, Portugal. pp. 1425-1427.

Sommer, E. 1995. An approach to quantifying the quality of induced theories. In: C. Nedellec (ed.), *Proceedings of the International Joint Conference on Artificial Intelligence Workshop on Macine Learning and Comprehensibility*. pp. 356-359.

Srinivasan, A., Muggleton, S. and Bain, M. 1994 The justification of logical theories based on data compression. *Machine Intelligence* 13:87-121.

Sugita, S. 1993. A model of pollen source area for an entire lake surface. *Quat. Res.* 39:239-244.

Sugita, S. 1994. Pollen representation of vegetation in Quaternary sediments: theory and method in patchy vegetation. *J. Ecol.* 82:881-897.

Sugita, S. 2007a. Theory of quantitative reconstruction of vegetation I: pollen from large sites REVEALS regional vegetation composition. *The Holocene* 17: 229 - 241.

Sugita, S. 2007b. Theory of quantitative reconstruction of vegetation II: all you need is LOVE. *The Holocene* 17: 243 - 257.

Sunnehag, P. and Hutter, M. 2010 Consistency of feature Markov processes. arXiv:1007.2075v1

Van der Maarel, E. and Sykes, M. T. 1993. Small-scale plant species turnover in a limestone grassland: the carousel model and some comments on the niche concept. *J. Veg. Sci.* 4: 179-188.

Thagard, P. 1978. The best explanation: criteria for theory choice. *J. Philos.* 75:76-92.

Villa-Martínez, R. and Moreno, P. I. 2007. Pollen evidence for variations in the southern margin of the westerly winds in SW Patagonia over the last 12,600 years. *Quat. Res.* 68: 400-409.

Vinod, H. D. 1969. Integer programming and the theory of grouping. *American Stat. Assoc. J.* 64: 506-519.

Visser, G. and Dowe, D. L. 2007. Minimum message length clustering of spatially-correlated data with varying inter-class penalties. *6th IEEE International Conference on Computer and*

Information Science (ICIS 2007), Melbourne, Australia, pp. 17-22.

Visser, G., Dowe, D. L. and Uotila, J. P. 2009. Enhancing MML Clustering using Context Data with Climate Applications. In: A. Nicholson and X. Li (Eds.) Proceedings 22nd Australian Joint Conf. on Artificial Intelligence (AI'09), Melbourne, Australia), *Lecture Notes in Artificial Intelligence* (LNAI) 5866 Springer Berlin. pp. 350-359.

Von Post, L. 1916. Skogsträdspollen i sydsvenska torvmosselagerföljder. *Geol. Fören. Förhandl.* 38:384-394.

Von Post, L. 1924. Ur de sydsvenska skogarnas regionala historia under postarktisk tid. *Geol. Fören. Förhandl.* 46:83-128.

V'yugin, V. V. 1999. Most sequences are predictable. Tech. Report CLRC-TR-99-01, Computer Learning Research Centre, Royal Hollaway University of London, Egham Surrey UK.

Walker, D. 1966. The late Quaternary history of the Cumberland lowlands. *Philos. Trans. Roy. Soc.* 251:1-210.

Walker, D and Wilson, S. R. 1978 A statistical alternative to the zoning of pollen diagrams. *J. Biogeogr.* 5: 1-21.

Wallace, C. S. 1996. MML inference of predictive trees, graphs and nets. In: A. Gammerman (ed.) *Computational Learning and Probabilistic Reasoning,* John Wiley. pp 43-66.

Wallace, C. S. 1998. Intrinsic classification of spatially correlated data. *Computer J.* 41: 602-611.

Wallace, C. S. 2005. *Statistical and Inductive Inference by Minimum Message Length.* Springer, Berlin.

Wallace, C. S. and Dowe, D. L. 1994. Intrinsic classification by MML - the Snob program. Proceedings 7th Australian Joint *Conference on Artificial Intelligence*, University of New England, Armidale, Australia, pp. 37-44.

Wallace, C. S. and Dowe, D. L. 2000. MML clustering with multistate, Poisson, Von Mises circular and Gaussian distributions. Statistics and Computing 10: 73-83.

Webb, L. J., Tracey, J. G., Williams, W. T. and Lance, G. N. 1967. Studies in the numerical analysis of complex rain-forest communities I. a comparison of methods applicable to site/species data. *J. Ecol.* 55: 171-191.

Whewell, W. 1847. *The Philosophy of the Inductive Sciences* Johnson Reprint Co., New York.

Williams, W. T. 1969. The problem of attribute weighting in numerical classification *Taxon* 18: 369-374.

Williams, W. T. 1971. Principles of clustering. *Annu. Rev. Ecol. Syst.* 2: 303-326.

Williams, W. T. and Dale, M. B. 1962. Partition correlation matrices for heterogeneous quantitative data. *Nature* 196: 602.

Yamada, H. and Amaroso, S. 1971. Structural and behavioural equivalences of tessellation automata. *Information and Control* 18:1-31.

Yang, F. and Jiang, T. 2003. Pixon-based image segmentation with Markov random fields. *IEEE Transactions on Image Processing* 12:1552-1559.

Yin, K. and Davidson, I. 2004. An information Theoretic Optimal Classifier for Semi-supervised Learning. *Lecture Notes in Computer Science* 3177, Springer Berlin. pp. 740-745.

Yu, S. X. and Shi,, J. 2004. Segmentation Given Partial Grouping Constraints, *IEEE Transactions Pattern Analysis and Machine Intelligence* PAMI 26:173-183.

Zhang, H-X. and Lu, J. 2010. Creating ensembles of classifiers via fuzzy clustering and deflection *Fuzzy sets and Systems* 161: 1790-1802.

Zhu, H-Y. and Rohwer, R. 1995. Bayesian invariant measurements of generalisation for continuous distributions. Technical Report NCRG/4352, Department Computer Science, University of Aston.

Received January 15, 2010
Revised May 10, 2010
Accepted September 9, 2010