‘Entropy’ on Covers and Its Application on Decision Tree Construction

Zhimin Wang

***Abstract*— Decision tree is a popular classification tool. To automatically construct a good decision tree, people have introduced entropy as a heuristic for attribute selection to deal with the intractable nature of finding an optimal solution with regard to the size of a tree. To solve a special kind of decision tree construction used in biological taxonomy, we need consider polymorphic attributes, against which a single instance may hold different values. To properly evaluate polymorphic attributes during tree construction, we propose the conditional form of a novel ‘entropy’ measure called ‘disconnectivity’ as the heuristic. In parallel to the theory of generalized entropy, ‘disconnectivity’ is also generalized to a family of measures.**

***Index Terms*— cover, decision tree, entropy, polymorphic character**

I. Motivation

Our work was originated in trying to answer questions about identification key construction from taxonomy community. In the following, we will give a general background introduction of those specific problems.

A taxonomist normally collects specimens from the field and annotates them afterward. If he considers a specimen as a newly found species, he must give it a name following internationally defined nomenclature standards [8] [10] [9], and must define properties (attribute-value pairs) that differentiate it from other species. Theoretically, others can use those attributes to identify biological entities in the field, laboratory, or classroom. Besides this, to allow users to efficiently identify species, a taxonomist or field biologist often constructs taxon-by-character matrices (often simply “character matrices”), and phylogenies (not our focus in this work) and keys based on those matrices. A taxon-by-character matrix is a table with columns marked with characters and rows marked with species, whose cells are the state of the species of the row against the character on the column. Identification keys or simply “keys” in fact are the decision trees of computer science, in which every internal node is a test on some character with each branch labeled with a possible state of that character, and every leaf marked as a taxon. In table 1 we may have the following species by character matrix, which is used as a running example in [6]. For this matrix, figure 1 is a possible key for those species in those rows.

TABLE I. AN EXAMPLE OF A TAXON-CHARACTER MATRIX

Species	*Primary Color*	*Slender*	*Long Bill*
Murre	White	No	No
Gray Jay	Gray	Yes	No
Egret	White	Yes	Yes
Turkey	Gray	No	No

Although it is common for taxonomists to build keys by hand, computer-based taxonomic identification tools have been adopting well known decision tree induction algorithms from computer science since the 1970s with the goal of minimizing the average number of steps required to identify a taxon. It has been proved that the problem of the optimal decision tree construction is NP-complete [5]. To cope with this difficulty, researchers have introduced several heuristic measurements to help to approximate the minimal tree, among which information gain and Gini index are widely used [4]. The idea is that we evaluate partitions of the set of taxa, where each partition is based on a character. The partition is defined by assigning to a set all the taxa that have the same state on that character. Then we select the character, whose corresponding partition has the highest evaluation score using the heuristics mentioned above, as a new test node of the decision tree, which in turn divides the current set of taxa into several subsets, for each of which we can apply the same procedure recursively until subsets become singletons.

However, for taxonomists there is still one problem that needs to be addressed: some taxa may have multiple possible values for a character. Such a character is called “polymorphic” by taxonomists. Characters with either of those two problems divide the set of taxa under consideration, say S, not into partitions but into covers (possibly overlapping subsets whose union is S). On covers, a direct application of entropy or Gini index will make these two measures lose monotonicity with regard to the refinement relationship, which we will discuss in the next section. Consequently it makes information gain and Gini index not appropriate for evaluating those characters during the tree construction. To illustrate polymorphic characters, we can change the above matrix into the form shown in Table 2.

Manuscript received March 24, 2011. This work was supported in part by the U.S. NSF Grant 0646266

Zhimin Wang is with the Harvard University Herbaria, Cambridge, MA 02138 USA (phone: 617-495-1948; fax: 617-495-9484; e-mail: zhimin.wangzm@gmail.com).

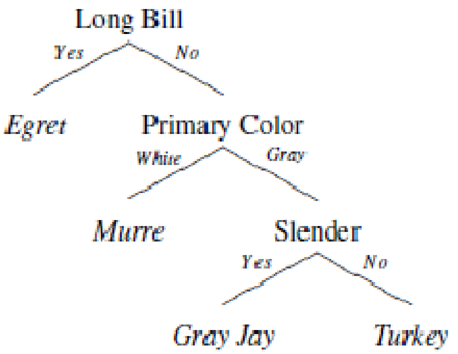

Figure 1. An example of an identification key.

TABLE II. A TAXON-CHARACTER MATRIX WITH PLOYMORPHICAL CHARACTER

Species	*Primary Color*	*Slender*	*Long Bill*
Murre	White/Gray	No	No
Gray Jay	Gray	Yes	No
Egret	White	Yes	Yes
Turkey	Gray	No	No

Now 'Primary color' will generates two subset {*Murre, Egret*} and {*Murre, Gray Jay, Turkey*} with overlap {*Murre*}.

II. Possible Solutions for Polymorphism

To discuss possible solutions for polymorphic attributes in decision tree construction, we give a real word taxon-character matrix in table 3, which shows a diagnostic table for 11 common British trees.

TABLE III. A TAXON-CHARACTER MATRIX FOR 11 COMMON BRITISH TREEFS FROM [11]

	Leaf					Bark texture	Flowers
	Pinnate, lobed, or neither	No. of pairs of leaflets of leaves or per leaf	Basic Shape	Positions on stem	Whether edges of leaves or leaflets toothed		Sexual characteristics
Ash	Pinnate			Opposite	Toothed	Smooth to rugged	Polygamous
Beech	Neither	-	Pointed oval	Alternate	Not toothed	Smooth	Monoecious
Birch	Neither		Pointed oval to triangular	Alternate	Not toothed	Smooth	Monoecious
Elder	Pinnate	1,2,3,4	-	Opposite	Toothed	Corky	Bisexual
Elm	Neither	-	Pointed oval	Alternate	Toothed	Rough	Bisexual
Lime	Neither	-	Heart-shaped	Alternate	Toothed	Scaling	Monoecious
Oak	Lobed	-	Oblong	Alternate	Not toothed	Rough	Monoecious
Plane	Lobed	-	Heart-shaped	Alternate	Toothed	Scaling	Monoecious
Rowan	Pinnate	5,6,7	-	Alternate	Toothed	Scored horizontally	Bisexual
Sweet chestnut	Neither	-	Broad lanceolate	Alternate	Toothed	Rough	Monoecious
Sycamore	Lobed	-	Heart-shaped	Opposite	Toothed	Scaling	Monoecious

In table 3, we can easily identify some polymorphic characters: number of pairs of leaflets per leaf, the texture of bark, sexual characteristics of flowers and basic shape of leaves. A coded diagnostic table is illustrated in Table 4.

There are some possible treatments of polymorphic characters. One is to generate new states by combining states that one taxon may have. For example we can combine two states 'smooth' and 'rugged' to make a new state 'smooth to rough'(the authors of Table3 named this 'smooth to rugged') for 'bark texture' and so Ash will have a distinctive state, which removes the overlap. However, in case where one taxon may have two or more states for some character as a whole, but each individual only presents one of them, the new state created by combination makes little sense for identification. Moreover, for continuous characters there is just no way to create a new state by combining states. So we do not think that combining states is a general treatment for polymorphic characters.

The other method to reduce covers generated by a polymorphic character to partitions is to divide the taxon with multi-states into sub-taxa with a weighting scheme, where we associate taxa with values representing their occurrence frequency. In fact, people in the taxonomy community have been taken the taxon abundance data into account to give abundant taxa a short identification path in the constructed key [6], although this practice is still controversial. Here if a taxon *T* has n states for a character, we will divide it into n sub-taxa (with the same taxon label) and each sub-taxon is associated with one of those *n* states and a weight, which represents the appearance frequency of that sub-taxon inside taxon *T*. The main problem with this method is that it is hard and sometime impossible to estimate those weights. Another problem with this method occurs when a taxon is polymorphic on a relatively large number, say *m*, of independent characters, for which this method will generate at least *2m* sub-taxa and very small weights, which in turn are big computational challenges. In the following sections, we will give a new method to deal with this problem by introducing some entropy-like measures that are proper to be directly applied to covers, and at the end will show some experimental results.

III. COVER AND DISCONNECTIVITY

First we will formally introduce cover and disconnectivity, which will serve as the theoretical foundation of our approach.

Let a universe U be a finite set. A cover Λ of U is defined as a set of subsets of U such that

$$U = \bigcup \{C \mid C \in \Lambda\}$$

A cover Λ_1 is a refinement of cover Λ_2 if each element of Λ_1 is a subset of some element of Λ_2. We say a cover Λ is simple if for any two different elements C_1, C_2 in Λ we don't have $C_1 \subseteq C_2$ or $C_2 \subseteq C_1$. Then we can define a partial order on simple covers of a universe by $\Lambda_1 < \Lambda_2$ iff Λ_1 is a refinement of Λ_2. We denote all simple covers on the universe U as COVER(U).

TABLE IV. TABLE 4: THE CODED TAXON-BY-CHARACTER MATRIX FROM [11]

	A	B	C	D	E	F	G
Ash	3	3,4, 5,6, 7	-	1	2	1,2	1,2
Beech	1	-	1	2	1	1	1
Brich	1	-	1,2	2	2	5	1
Elder	3	1,2, 3,4	-	1	2	3	2
Elm	1	-	1	2	2	2	2
Lime	1	-	3	2	2	1	2
Oak	2	-	4	2	1	2	1
Plane	2	-	3	2	2	5	1
Rowan	3	5,6, 7	-	2	2	4	2
Sweet Chestnut	1	-	5	2	2	2	1
Sycamore	2	-	3	1	2	5	1
Code	A	1=not pinnate or lobed, 2=lobed, 3=pinnate					
	C	1=pointed oval, 2=triangular, 3=heart-shaped, 4=oblong, 5=broad lanceolate					
	D	1=opposite,2=alternate					
	E	1=not toothed,2=toothed					
	F	1=smooth,2=rough, 3=corky,4=scored horizontally					

Definition 3.1. *Let* Λ *be a cover on U, we define* Λ^{red} = Λ − {D|$D \in \Lambda$, $D \subset C$ *for some* $C \in \Lambda$}

It is easy to see that Λ^{red} is always a simple cover on U.

Definition 3.2. *Let* Λ_1, $\Lambda_2 \in$ COVER(U)*, we define* $\Lambda_1 \wedge \Lambda_2 = \{C_1 \cap C_2 \mid C_1 \in \Lambda_1, C_2 \in \Lambda_2\}^{red}$

Theorem 3.3. *Let* Λ_1, $\Lambda_2 \in$ COVER(U)*, then* $\Lambda_1 \wedge \Lambda_2$ *is the maximal element of those in* COVER(U) *that are refinements of both* Λ_1 and Λ_2.

Definition 3.4. *Let* $\Lambda_1 \in$ COVER(S) and $\Lambda_2 \in$ COVER(T), where $S \cap T = \phi$, we define $\Lambda_1 \times \Lambda_2 = \{C_1 \times C_2 \mid C_1 \in \Lambda_1, C_2 \in \Lambda_2\}$

In the following sections 'cover' means simple cover.

A. Disconnectivity

Definition 3.5. Let the universe be U, the intolerance set of a cover Λ is defined as $\iota(\Lambda)$= {(a, b)|$\forall$C $\in \Lambda$, a $\notin$ C or b $\notin$ C}, which is composed of all the ordered pairs of elements of U that do not both belong to any single element of Λ.

Then we call |$\iota(\Lambda)$| the disconnectivity of cover Λ (denoted by Disc(Λ)), which is the total number of pairs of elements of universe U that are not both contained by any member of C. In other words, if we form an undirected graph G(V, E) such that V is the universe U and E contains all pairs of elements of U that are both contained by some member of C, the disconnectivity of cover C is the total number of 0s in the corresponding adjacency matrix of G.

Definition 3.6. We the denote the set {(a, b)|$\exists$C $\in \Lambda$, a∈C and b∈C} as $\tau(\Lambda)$. Then we have $\iota(\Lambda) = U \times U - \tau(\Lambda)$.

Theorem 3.7. Disconnectivity is anti-monotonic with regard the partial order < on covers

Proof:

Let $\Lambda_1 < \Lambda_2$ and (u_1, u_2) be any pair of elements in $\tau(\Lambda_1)$. So u_1 and u_2 belong to some element C in Λ_1. Because Λ_1 is a refinement of Λ_2, we have a member C^0 of Λ_2 such that $C \subseteq C^0$ and so u_1 and u_2 are also both contained by C^0, which means $(u_1, u_2) \in \tau(\Lambda_2)$. Now we can conclude that $\tau(\Lambda_1) \subseteq \tau(\Lambda_2)$ and $\iota(\Lambda_1) \supseteq \iota(\Lambda_2)$, which means that *Disc*(Λ_1) is bigger than that of *Disc*(Λ_2)

Definition 3.8. Let U be a universe, Λ a cover on U. We define the normalized disconnectivity as NDisc(Λ)= $\frac{Disc(\Lambda)}{|U|^2}$.

B. Examples

Let the universe be {1, 2, 3, 4, 5, 6} and C1 = {{1, 2, 3}, {3, 4, 5}, {6}} a cover on it, then the disconnectivity of C1 is *18*. The vertices of its graph G are {1, 2, 3, 4, 5, 6} and the edges are {(1, 2), (2, 3), (1, 3), (3, 4), (3, 5), (4, 5), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)} and the adjacency matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Similarly for cover C2 = {{1, 2}, {3, 4, 5}, {6}}, its disconnectivity is *22*. Its corresponding adjacency matrix looks like

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

C. Comparison between disconnectivity and GINI index

Applying disconnectivity to partitions can also measure their evenness. Let the universe be {1, 2, 3, 4, 5, 6} and partition C1 be {{1, 2, 3}, {4, 5}, {6}}, then the disconnectivity of C1 is *22*. Its adjacency matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

For partition C2 = {{1, 2}, {4, 5}, {3, 6}} its disconnectivity is *24*, which is bigger than C1 because of its more even distribution of elements. Its adjacency matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

Theorem 3.9 Let P be a partition of U, then normalized Disconnectivity on P is in fact the Gini-index of P.

Proof:

Let P={ P_1 , $P_2 \dots , P_n$ }. Because there is no overlap between any two elements of P, we have

$$Disc(P) = \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} |P_i| * |P_j|$$

$$= \sum_{i=1}^{n} |P_i| * \sum_{j=1}^{n} |P_j| - \sum_{i=1}^{n} |P_i|^2$$

$$= |U|^2 - \sum_{i=1}^{n} |P_i|^2$$

So the normalized disconnectivity is

$$\frac{Disc(P)}{|U|^2} = 1 - \sum_{i=1}^{n} \left(\frac{|P_i|}{|U|}\right)^2$$

which is the Gini-inde of *P*.

D. Generalized forms

Theorem 3.10 Let $\Lambda = \{B_1, B_2 \dots, B_n\}$ be a cover of S, we have $|\tau(\Lambda)|$ =

$$\sum |B_i|^2 - \sum |B_i \cap B_j|^2 \dots + (-1)^{n+1} |\bigcap_{i=1}^{n} B_i|^2$$

It follows immediately from the above theorem that NDisc(Λ) =

$$1 - \sum \left(\frac{|B_i|}{|S|}\right)^2 + \sum \left(\frac{|B_i \cap B_j|}{|S|}\right)^2 \dots (-1)^n \left(\frac{|\bigcap_{i=1}^{n} B_i|}{|S|}\right)^2$$

Definition 3.11. With the above observation and the form of generalized entropy for partitions [1], we propose a generalized formula for cover:

NDisc $_\beta(\Lambda)$ =

$$\frac{1}{1-2^{1-\beta}} \left(1 - \sum_{1,n} \left(\frac{|B_i|}{|S|}\right)^\beta + \sum_{i>j} \left(\frac{|B_i \cap B_j|}{|S|}\right)^\beta \dots (-1)^n \left(\frac{|\bigcap_{i=1,n} B_i|}{|S|}\right)^\beta \right)$$

When β = 1, we have $1 - 2^{1-\beta} = 0$ and

$$1 - \sum_{1,n} \frac{|B_i|}{|S|} + \sum_{i>j} \frac{|B_i \cap B_j|}{|S|} \dots \quad (-1)^n \left(\frac{|\bigcap_{i=1,n} B_i|}{|S|}\right) = 0.$$

Calculating the limit when $\beta \to 1$, we get NDisc $_1(\Lambda)$ =

$$-\sum_{i=1}^{n} \frac{|B_i|}{|S|} log \frac{|B_i|}{|S|} + \sum_{i>j} \frac{|B_i \cap B_j|}{|S|} log \frac{|B_i \cap B_j|}{|S|} \dots$$

$$+(-1)^n \frac{|\bigcap_{i=1}^{n} B_i|}{|S|} log \frac{|\bigcap_{i=1}^{n} B_i|}{|S|}$$

To apply disconnectivity to decision tree induction we need its conditional form:

Definition 3.12. Let Λ_1, Λ_2 *be two covers in COVER(S), we* define the conditional disconnectivity as NDisc$_\beta$($\Lambda_1 | \Lambda_2$) = NDisc$_\beta$($\Lambda_1 \wedge \Lambda_2$) - NDisc$_\beta$(Λ_2)

It should be noticed that NDisc$_2$ and NDisc$_1$ become Gini-index and Shannon entropy respectively when they are applied on partitions.

IV. Experiment

With normalized disconnectivity and its conditional form, we are ready to substitute Shannon entropy and conditional entropy with them respectively in evaluating (polymorphic) characters during decision tree construction.

To compare those two methods, entropy-division (entropy-based key construction after dividing taxa on polymorphic characters in a taxon-by-character matrix) and $NDsic_{\beta}$ method (disconnectivity-based key construction on the original matrix), we conducted experiments on three taxon-by-character matrices. Two of them are from research cited at www.hydrophiloidea.org. We designate these as Dataset1 and Dataset2, arising, respectively from [3] and [2]. The third matrix, which we denote Dataset3 is from a study on www.nature.com, and is taken from a supplementary document of [1]. Table **5** shows some basic properties of those datasets

For the entropy-division method, we first transformed those three matrices into matrices without polymorphism, which we take as input to Weka [7] to generate the key with J48 (the Java variant of the C45 algorithm). The comparison is focused on sizes of constructed keys, which is shown in Table 4.Here the tree size is measured by the sum of depths of its leaves and the average depth of those leaves.

From Table **6** we can see that we got smaller keys for all three datasets by using Disconnectivity-based methods than by using entropy- division. An interesting observation is that the number of leaf nodes in keys generated by the entropy-division method is the same as the original number of taxa in all three cases. In contrast, for Dataset1 and Dataset3 NDisc-based methods generate keys with more leaves than the corresponding original number of taxa. This means that N Disc-based methods use some polymorphic characters in those key constructions and result in smaller trees, but the entropy-division method doesn't. This illustrates that N Disc-based methods can better identify good characters among a mix of polymorphic and non-polymorphic characters. To assess the evaluation ability of entropy division method among polymorphic characters, we removed 40 characters in Dataset1 to force the entropy-division method to use some polymorphic characters. The result is shown in Table **7** where NDisc-based methods still have clear advantage as measured by tree depth, but often at little or no cost of extra leaves and even though entropy-division is forced to use polymorphic characters.

V. Conclusion

In this paper, we have presented an entropy-like measurement, Disconnectivity, for covers and developed its theory in parallel to the theory of generalized entropy on partitions. Our advance, a generalized measure to covers, includes the representation as one parameter family of functions, and generalized conditional form. In the end, we have demonstrated the application of this measure on decision tree construction on polymorphic characters. We believe that for problems, where cover is the natural model and refined statistical model is not available, disconnectivity and its derived form are valuable heuristics.

TABLE V. PROPERTIES OF THE EXPERIMENTAL DATASETS

	Dataset1	Dataset2	Dataset3
Original Number of taxa	30	21	53
Number of Characters	121	157	218
Number of Polymorphic Characters	36	4	65
Number of taxa after the division	1041	158	17153

TABLE VI. COMPARISON OF THE DIVISION-BASED METHOD AND THE DISCONNECTIVITY-BASED METHOD

Measurement	Dataset	Tree Construction Method		
		Entropy-Division	$NDisc_1$	$NDisc_2$
Sum of Depth	1	156	117	123
	2	71	56	54
	3	289	207	228
Average Depth	1	5.03	3.55	3.62
	2	3.38	2.67	2.57
	3	5.56	3.76	3.35
Number of Leaves	1	30	33	34
	2	21	21	21
	3	53	55	68

TABLE VII. COMPARISON OF THE DIVISION BASED METHOD AND THE DISCONNECTIVITY-BASED METHODS ON DATASET1 WITH REDUCED CHARACTERS

Measurement	Tree Construction Method		
	Entropy-Division	$NDisc_1$	$NDisc_2$
Sum of Depth	163	129	127
Average Depth	5.26	3.79	3.74
Number of Leaves	31	34	34

References

[1] Anderson, J. S. A stem batrachian from the Early Permian of Texas and the origin of frogs and salamanders. Nature (2008).

[2] Beutel, R. G., and Komarek, A. Comparative study of thoracic structures of adults of Hydrophiloidea and Histeroidea with phylogenetic implications (Coleoptera, Polyphaga). Organisms Diversity and Evolution 4, 1-2 (2004), 1–34.

[3] Beutel, R. G., and Leschen, R. A. B. Phylogenetic analysis of Staphyliniformia (Coleoptera) based on characters of larvae and adults. Systematic Entomology 30, 4 (2005), 510–548.

[4] Calvo-Flores, M. D., Contreras, W. F., Galindo, E. G., and Pe´rez-Pe´rez, R. Xkey: A tool for the generation of identification keys. Expert Systems with Applications 30, 2 (2006), 337 – 351.

[5] Hyafil, L. Constructing Optimal Binary Decision Trees Is NP-Complete. Information Processing Letters (1976).

[6] Manoharan, A., Stamberger, J., Yu, Y., and Paepcke, A Optimizations for the ecopod field identification tool. BMC Bioinformatics 9, 1 (2008), 150.

[7] Markov, Z., and Russell, I. An introduction to the WEKA data mining system. SIGCSE Bull. 38, 3 (2006), 367–368.

[8] MCNEILL, J., BARRIE, F. R., BURDET, H. M., DEMOULIN, V., HAWKSWORTH, D. L., MARHOLD, K., NICOLSON, D. H., PRADO, J., SILVA, P. C., SKOG, J. E., and WIERSEMA, J. H., Eds. International Code of Botanical Nomenclature (VIENNA CODE). Regnum Vegetabile 146. A.R.G. Gantner Verlag KG., 2006.

[9] Sneath, P. International Code of Nomenclature of Bacteria: Bacteriological Code. ASM for the International Union of icrobiological Societies (IUMS), the Bacteriological, 1990.

[10] the International Commission on Zoological Nomenclature. INTERNATIONAL CODE OF ZOOLOGICAL NOMENCLATURE. The International Trust for Zoological Nomenclature, 1999.

[11] Payne, R., and Preece, D. A. Identification keys and diagnostic tables: a review. J. R. Statist. Soc. A (1980)