**DIMACS Technical Report 2002-04**
**March 2002**

# The Controversy About Continuity in Clustering Algorithms

by

Melvin F. Janowitz[1] [,2] [,3]
DIMACS
Rutgers University
Piscataway, New Jersey 08854

[1]Permanent Member and Associate Director
[2]Research supported in part by NSF Grant CCR 00-87022
[3]Presented in part to the Classification Society of North America in June, 2001

---

DIMACS is a partnership of Rutgers University, Princeton University, AT&T Labs-Research, Bell Labs, NEC Research Institute and Telcordia Technologies (formerly Bellcore).

DIMACS was founded as an NSF Science and Technology Center, and also receives support from the New Jersey Commission on Science and Technology.

## ABSTRACT

It is not meaningful to talk about continuous cluster algorithms for use with data having only ordinal significance. An ordinal version of continuity for such data is introduced and studied.

# 1 Introduction

This paper deals with a controversy that arose some years ago about the role of continuity in cluster analysis. The reader should recall that cluster analysis is a branch of multivariate data analysis that tries to recapture any internal structure that a data set might possess. The input to a clustering algorithm is often a finite set of objects together with a set of attributes that the objects might posses in varying degrees, or else some measure of similarity or dissimilarity among the objects. The output is generally a classification of the objects or else some nested sequence of classifications. It is assumed that the reader has some familiarity with the subject. There are many references available. Among them is the informative book by Jain and Dubes [1]. For our purposes, we shall view clustering as the transformation of an input measure of dissimilarity to a nested sequence of suitable classifications.

Some years ago, N. Jardine and R. Sibson [8] introduced a mathematical model for cluster analysis. This model involved an assumption of continuity, and had a somewhat controversial consequence. It stated that if the output is all possible nested sequences of partitions of the underlying set of objects, then the only allowable clustering algorithm within the model is a technique called single-linkage clustering. This observation caused a storm of controversy. It is our goal to show that the axiom causing the problem was something other than continuity, and that indeed, continuity is not even an appropriate consideration within this particular model for cluster analysis. Background material is supplied in Section 1, a version of the Jardine-Sibson axioms is presented in Section 2, and a discussion of the role of continuity is given in Section 3. Section 4 explains the exact role of continuity within this framework. Finally, Section 5 launches an investigation into the nature of ordinally continuous cluster methods.

An argument will be given to indicate why continuity should be replaced by a notion of *ordinal* continuity. The idea is to note that metric properties of $d$ and $d'$ rather surprisingly imply an ordinal connection that is denoted symbolically by $d \precsim d'$ and is defined by $d(a,b) < d(x,y) \Longrightarrow d'(a,b) < d'(x,y)$. This amounts to saying that the strict weak order associated with $d$ is contained in the one associated with $d'$. Ordinal continuity for a cluster method $F$ then becomes the assertion that $d \precsim d' \Longrightarrow F(d) \precsim F(d')$. It will be argued that continuity should be viewed as a strengthening of notions involving preservation of global order equivalence as developed in [3]. All of this will be made more precise in the sequel.

# 2 Background Material

We begin by establishing some basic notation and terminology. Throughout this section we shall be working in a fixed finite nonempty set $P$ of objects to be clustered, together with some notion of dissimilarity between pairs of objects. A clustering function is supposed to somehow organize $P$ into a collection of nonempty subsets so that objects in the same subset (*cohesion*) are more similar than objects that come from different subsets (*separation*). The problem is that maximizing cohesion and maximizing separation are often conflicting goals. Our purpose is to construct a general enough mathematical model for this, so that clustering techniques other than single-linkage clustering can satisfy the model. We shall also carefully examine the assumptions of the Jardine-Sibson model, and especially try to show that continuity should not even be an issue that is raised.

Accordingly, we assume that we are given a finite nonempty set $P$ of objects having cardinality $p$. A mapping $d : P \times P \mapsto [0, \infty)$ is called a *dissimilarity coefficient* (DC) if it satisfies

**(DC0)** $d(a, a) = 0$ for all $a \in P$.

**(DC1)** $d(a, b) = d(b, a)$ for all $a, b \in P$.

The DC $d$ is called *definite* if

**(DC2)** $d(a, b) = 0 \Longrightarrow a = b$.

It is called a *semimetric* if it satisfies

**(DC3)** $d(a, b) \leq d(a, c) + d(c, b)$ for all $a, b, c \in P$.

A definite semimetric is called a *metric*. Finally, a DC $d$ is called an *ultrametric* if

**(DC4)** $d(a, b) \leq \max\{d(a, c), d(c, b)\}$ for all $a, b, c \in P$.

When convenient, we shall follow the notation in [8]; in particular, we shall denote the set of dissimilarity coefficients on $P$ by $C(P)$, and the set of ultrametrics on $P$ by $U(P)$.

**Definition 2.1** For $d_1, d_2$ both DCs, we write $d_1 \leq d_2$ to denote the fact that $d_1(a, b) \leq d_2(a, b)$ for all $a, b \in P$. We define $d = d_1 \vee d_2$ by the rule $d(a, b) = \max\{d_1(a, b), d_2(a, b)\}$ for any $a, b \in P$. More generally, if $\{d_i\}$ is a family of dissimilarity coefficients whose images are bounded above by some fixed constant, then $\bigvee_i d_i$ is defined by

$$(\textstyle\bigvee_i d_i)(a, b) = \sup_i d_i(a, b).$$

Similarly, $\bigwedge_i d_i$ is defined by

$$(\textstyle\bigwedge_i d_i)(a, b) = \inf_i d_i(a, b).$$

Note that any DC $d$ may be viewed as a vector in Euclidean space of dimension $p(p-1)/2$. This naturally defines a metric topology on $C(P)$. Any of the usual metrics can be used to define a metric on the DCs. In particular, we might define $\Delta_0$ by the rule

$$\Delta_0(d_1, d_2) = \max\{ \mid d_1(a,b) - d_2(a,b) \mid : a, b \in P\} \ ,$$

We next agree to let $\Sigma = \Sigma(P)$ denote the set of reflexive symmetric relations on $P$. For $R_1, R_2 \in \Sigma$, we write $R_1 \subseteq R_2$ to indicate that $aR_1b \Longrightarrow aR_2b$. Then $P \times P$ is the largest member of $\Sigma$, while $R_\emptyset = \{(a,a) : a \in P\}$ is its smallest. A mapping $C : [0,\infty) \mapsto \Sigma(P)$ is called a *numerically stratified clustering* (NSC) if it satisfies:

**(NSC1)** $0 \leq h \leq h' \Longrightarrow C(h) \subseteq C(h')$.

**(NSC2)** There exists an $h \in [0,\infty)$ such that $C(h) = P \times P$.

**(NSC3)** Corresponding to any $h \geq 0$ there is a $\delta > 0$ such that $C(h) = C(h+\delta)$.

The NSC $C$ is *definite* if

**(NSC4)** $C(0) = R_\emptyset$.

It is a *hierarchy* if

**(NSC5)** $C(h)$ is an equivalence relation for all $h \in [0,\infty)$.

**Definition 2.2** *For $C_1, C_2$ numerically stratified clusterings, write $C_1 \leq C_2$ to denote the fact that $C_1(h) \subseteq C_2(h)$ for all $h \geq 0$.*

**Lemma 2.3** *For any DC $d$, define $Td : [0,\infty) \mapsto \Sigma$ by the rule*

$$Td(h) = \{(a,b) : d(a,b) \leq h\}.$$

*Then $Td$ is a numerically stratified clustering.*

**Proof:** It is clear from the definition of a DC that $Td(h)$ is always reflexive and symmetric. If $h \leq h'$, then $(a,b) \in Td(h) \Longrightarrow d(a,b) \leq h \leq h' \Longrightarrow (a,b) \in Td(h')$, so $Td(h) \subseteq Td(h')$. This establishes (NSC1). To establish (NSC2), let $k = \max\{d(a,b) : a, b \in P\}$, and note that $Td(k) = P \times P$. Note now that $Td(h)$ can only change at the levels where $d$ takes values. But there are only finitely many such levels. If $Td(h) \subset P \times P$, let $h'$ denote the smallest member of the image of $d$ that lies strictly above $h$. If $0 < \delta < h' - h$, it is clear that $Td(h) = Td(h+\delta)$, thus establishing (NSC3). ∎

The relations of the form $Td(h)$ are called the *splitting relations* of $d$ or of $Td$; any splitting relation other than $R_\emptyset$ or $P \times P$ is called a *proper splitting relation.*

**Lemma 2.4** *For any numerically stratified clustering $C$, define $d_C$ by the rule*

$$d_C(a,b) = \inf\{h : (a,b) \in C(h)\}.$$

*Then $d_C$ is a dissimilarity coefficient on $E$.*

**Proof:** It is obvious from the fact that each $C(h)$ is a reflexive symmetric relation that $d_C(a,b) = d_C(b,a) \geq 0$. ∎

The next result follows directly from the definition of $Td$, and the fact that for any $a, b \in P$, and any numerically stratified clustering $C$, it is true that $\{h : (a,b) \in C(h)\}$ must have a smallest element.

**Lemma 2.5** $d(a,b) \leq h \Longleftrightarrow (a,b) \in Td(h)$.

This now allows us to establish a fundamental connection between dissimilarity coefficients and their associated numerically stratified clusterings.

**Lemma 2.6** $d_1 \leq d_2 \Longleftrightarrow Td_2 \leq Td_1$.

**Proof:** Let $d_1 \leq d_2$. If $(a,b) \in Td_2(h)$, then $d_2(a,b) \leq h$, so $d_1(a,b) \leq d_2(a,b) \leq h$, whence $(a,b) \in Td_1(h)$. Thus $Td_2(h) \subseteq Td_1(h)$ for all $h \geq 0$, so $Td_2 \leq Td_1$.

Now assume that $Td_2 \leq Td_1$. If $k = d_2(a,b)$, then $(a,b) \in Td_2(k) \subseteq Td_1(k)$ forces $d_1(a,b) \leq k$, so $d_1(a,b) \leq d_2(a,b)$. Since this holds for all $a, b \in P$, we conclude that $d_1 \leq d_2$. ∎

**Lemma 2.7** *For any DC $d$, if $C = Td$, then $d = d_C$; for any clustering $C$, if $d = d_C$, then $C = Td$. In other words, $d = d_{Td}$, and $C = Td_C$.*

**Proof:** First let $C = Td$. Recall that $d(a,b) \leq h \Longleftrightarrow (a,b) \in Td(h)$. If $k = d(a,b)$, it follows that $k = \min\{h : (a,b) \in Td(h)\}$, so $d = d_C$.

Now let $C$ be a numerically stratified clustering, with $d = d_C$. We are to show that $C = Td$. This is immediate from then fact that

$$(a,b) \in Td(h) \Longleftrightarrow d(a,b) \leq h \Longleftrightarrow (a,b) \in C(h). \blacksquare$$

**Theorem 2.8** *The association $d \mapsto Td$ sets up a one-one correspondence between dissimilarity coefficients and numerically stratified clusterings. In this correspondence, $d$ is an ultrametric if and only if $Td$ is a hierarchy.*

**Proof:** The fact that there is a one-one correspondence follows from Lemma 2.7, so we only need to work with the last assertion. Accordingly, let us assume that $d$ is an ultrametric. We are to prove that $Td(h)$ is an equivalence relation for all $h \geq 0$. Since $Td(h)$ is always reflexive and symmetric, we concentrate on showing that it is transitive. If $(a,b),(b,c) \in Td(h)$, then $d(a,b) \leq h$, and $d(b,c) \leq h$. By the ultrametric inequality,

$$d(a,c) \leq \max\{d(a,b), d(b,c)\} \leq h.$$

Now assume $Td(h)$ is an equivalence relation for all $h \geq 0$. Let $a,b,c \in P$. We wish to show that $d(a,c) \leq \max\{d(a,b), d(b,c)\}$. If $k = \max\{d(a,b), d(b,c)\}$, we know that $(a,b),(b,c) \in Td(k)$. Since $Td(k)$ is an equivalence relation, it is immediate that $(a,c) \in Td(k)$, so $d(a,c) \leq k = \max\{d(a,b), d(b,c)\}$. ∎

**Remark 2.9** Cluster analysis often takes as an input a DC $d$ on a finite set $P$, and either outputs an equivalence relation or a hierarchy on $P$. In view of Theorem 2.8, we may view cluster analysis as a transformation of a dissimilarity coefficient into an ultrametric or some generalization of an ultrametric.

We close this section by considering a pair of properties enjoyed by $U(P)$.

**Lemma 2.10** *Let $\{u_i\}$ be a bounded family of ultrametrics. Then $u = \bigvee_i u_i$ is an ultrametric.*

**Proof:** It suffices to show that $u$ satisfies the ultrametric inequality. For $a,b,c \in P$, we note that for each index $i$,

$$u_i(a,c) \leq \max\{u_i(a,b), u_i(b,c)\} \leq \max\{u(a,c), u(c,b)\}.$$

It is immediate that

$$u(a,c) = \max_i\{u_i(a,c)\} \leq \max\{u(a,c), u(c,b)\}. \blacksquare$$

**Lemma 2.11** *For any dissimilarity coefficient $d$, it is true that*

$$d = \bigwedge\{u \in U(P) : d \leq u\}.$$

**Proof**: Let $d \in C(P)$ with $k$ the largest member of the image of $d$. For any $a,b \in P$, define $u_{ab}$ by the rule $u_{ab}(x,y) = d(a,b)$ if $\{x,y\} = \{a,b\}$, and $u_{ab}(x,y) = k$ otherwise. It is immediate that $u_{ab} \in U(P)$ for all $a,b \in P$, and that $d = \bigwedge\{u_{ab} : a,b \in P\}$. The Lemma is now apparent. ∎

# 3 The Jardine-Sibson axioms.

In this section we present the basic assumptions that start at the bottom of page 415 of [10]. These same assumptions can also be found in the text [8]. Additional background material may be found in [5], [6], [7], and [11]. **Warning:** The material will be presented here in quite a different form. Jardine and Sibson work with a finite nonempty set $P$ of objects to be clustered, and let $C(P)$ denote the set of dissimilarity coefficients on $P$, so we will follow that much of their notation. But they let $D$ denote a cluster method. Since it seems rather awkward to let $D$ be a function acting on $d$, we choose to denote cluster methods by $F$ instead of $D$. Instead of defining $F$ on all of $C(P)$, Jardine and Sibson define it on a subset $A \subseteq C(P)$, and assume that $Z \subseteq C(P)$ (the target set) denotes the image of $F$. They then present pairs of conditions: one on $F$ and a second one on $Z$. We choose to make a number of changes in this scheme.

For a given set $P$, it seems rather difficult to predict in advance the exact nature of the DCs that $F$ must accept as input, so we will not define $A$, and simply assume that the domain of $F$ is all of $C(P)$. Now $F$ must really operate on $C(P)$ for *any* set $P$, regardless of its size. We shall be a little sloppy about this, and just fix a $P$. We shall not mention $Z$ specifically, and just note that $Z$ is the image of $F$. Since conditions on $F$ completely determine its image, we see no need to impose conditions on both $F$ and its image. It is worth noting though that often times, $Z$ will either represent the set of ultrametrics on $P$, or at least contain this set. We now proceed with our modified view of the Jardine-Sibson framework.

**Definition 3.1** If $F, G$ are two cluster methods on $P$, we may compose $G$ with $F$ to obtain a new cluster method $G \circ F$. Here $G \circ F$ is defined by the rule $[G \circ F](d) = G(F(d))$. In particular, $F$ is called *idempotent* in case $F = F \circ F$.

*Condition* 1. *Appropriateness.* It is assumed that $F$ transforms any input DC $d$ into a desired form $F(d)$, and that if the input is already in the desired form, then $F$ will leave that input fixed. Another way of saying this is to require that

$F = F \circ F$; i.e., $F$ is *idempotent.*

It is now evident that the image of $F$ is simply the set of dissimilarity coefficients left fixed by $F$.

*Condition* 2. *Label freedom.* The output should be independent of any prior labelling of the underlying objects. This just says that if the members of $P$ are relabelled, the output will be the same except with the output objects appropriately relabelled. One way of guaranteeing this is to assert that for any permutation $\rho$ of $P$, and for any DC $d$,

$$[F(d)] \circ (\rho \times \rho) = F[d \circ (\rho \times \rho)].$$

The reader should observe here that $[d \circ (\rho \times \rho)](a, b) = d(\rho a, \rho b)$.

*Condition* 3. *Scale freedom.* The output should be independent of scale factors. One would like for a change in the units of measurement to not change the output classifications. Stated more precisely, we have that for any $\alpha > 0$,

$$F(\alpha d) = \alpha F(d).$$

Scale freedom seems an extremely natural property.

*Condition* 4. *Preservation of Clusters.* This condition is somewhat controversial and is certainly not satisfied by many commonly used cluster methods. The idea is that if a pair $a, b$ of objects is grouped together at some level $h$ in the input, then it must be grouped together at that same level in the output. In other words, we require for all $h \geq 0$, that $Td(h) \subseteq [TF(d)](h)$. In view of Lemma 2.6, this can be accomplished by requiring that

$$F(d) \leq d.$$

*Condition* 5. *Optimality.* The output should in same natural sense represent the best possible fit to the input data. There are many ways of doing this. One way of accomplishing this would be to require the following two items:

(5a) $F(d) \leq d' \leq d$ with $d' = F(d') \Longrightarrow d' = F(d)$.

(5b) If $\{d_i\}$ is a bounded subset of the image of $F$, then $\bigvee_i d_i$, is in the image of $F$.

*Condition* 6. *Continuity.* Historically, the most controversial of the Jardine-Sibson conditions appears to be that of continuity.

$F$ is continuous with respect to the metric $\Delta_0$.

As a reminder to the reader, this just says that for any $d \in C(P)$, and any $\varepsilon > 0$, there corresponds a $\delta > 0$ having the property that

$$0 < \Delta_0(d, d') < \delta \Longrightarrow \Delta_0(F(d), F(d')) < \varepsilon.$$

The motivation for this assumption is simply the desire to have small errors in the input data to translate into small errors in the output.

Jardine and Sibson now impose axioms which they call SET5, SET5b, and MAP6. We shall ignore these axioms, and instead proceed to investigate the consequences of Conditions 1 through 5.

**Lemma 3.2** *Let $F$ satisfy Conditions* 1 *and* 4. *Then Condition* 5 *is equivalent to $F$ being isotone in the sense that $d_1 \leq d_2 \Longrightarrow F(d_1) \leq F(d_2)$.*

**Proof:** Assume first that $F$ satisfies Condition 5. Let $d_1 \leq d_2$. Then $F(d_1) \leq d_1 \leq d_2$. By Condition 5b, $F(d_1) \vee F(d_2)$ is in the image of $F$. Note that the full strength of 5b was not needed: only the fact that $F(d_1) \vee F(d_2)$ is in the image of $F$. Since $F(d_2) \leq F(d_1) \vee F(d_2) \leq d_2$, we have that $F(d_2) = F(d_1) \vee F(d_2)$, whence $F(d_1) \leq F(d_2)$.

Suppose conversely that $F$ is isotone. If $F(d) \leq d' \leq d$ with $d' = F(d')$, we have $d' \leq d \Longrightarrow F(d') \leq F(d)$, whence $d' = F(d)$. This is Condition 5a. If $d = \bigvee_i d_i$ with each $d_i$ in the image of $F$, then $d_i \leq d \Longrightarrow d_i = F(d_i) \leq F(d)$, so $d = \bigvee_i d_i \leq F(d)$. The reverse inequality comes from Condition 4, and $d = F(d)$ is now clear, thus establishing 5b. ∎

**Lemma 3.3** *Let $F$ be a cluster method defined on $P$. Assume that $F$ satisfies Conditions* 1, 4 *and* 5*, and let $Z$ denote the image of $F$. Then for any $d \in C(P)$, it is true that $\{d_i \in Z : d_i \leq d\}$ has a largest member $d_Z$. It must in fact be true that $F(d) = d_Z$.*

**Proof:** Let $B = \{d_i \in Z : d_i \leq d\}$, and note that by Condition 4, $F(d) \in B$. Let $d_i \in B$. By Condition 5, $d_i \leq d \Longrightarrow F(d_i) \leq F(d)$. By Condition 1, $d_i = F(d_i)$, so we deduce that $d_i \leq F(d)$. But this shows $F(d)$ is the largest member of $B$, thus completing the proof. ∎

In most applications, a hierarchy is a perfectly suitable output for a cluster method. In view of this, it seem rather innocuous to assume that the image of a clustering algorithm $F$ contains all ultrametrics. But this algorithm has some rather startling consequences.

**Lemma 3.4** *Assume that the image of $F$ contains all ultrametrics. Then $F$ isotone and idempotent together imply Condition* 4*.*

**Proof:** By Lemma 2.11, any DC $d$ has the property that $d = \bigwedge_i \{u_i \in U(P) : d \leq u_i\}$. Since $d \leq u$ with $u \in U(P)$ implies that $F(d) \leq u$, it is now obvious that $F(d) \leq d$. This establishes Condition 4. ∎

The next result follows immediately from this observation together with Lemma 3.2 and Lemma 3.3. It will therefore be stated without proof.

**Theorem 3.5** *Let $F$ be a cluster method defined on $C(P)$, and let $Z$ denote the image of $F$.*

(1) *Assume that $F$ is idempotent and isotone, and that $Z$ contains the set of all ultrametrics on $P$. Then for any DC $d$ on $P$, it is true that $F(d) = \max\{u \in Z : u \leq d\}$.*

(2) *If conversely it is true that for all* $d \in C(P)$, $F(d) = \max\{u \in Z : u \leq d\}$, *then* $F$ *is idempotent and isotone.*

**Definition 3.6** A nonempty subset $B \subseteq C(P)$ is said to have the *relative maximum property* if for any DC $d$, it is true that $\{d_i \in B : d_i \leq d\}$ has a largest member.

**Remark 3.7** There is a a natural one-one correspondence between cluster methods satisfying Conditions 1, 4, 5 and subsets of $C(P)$ having the relative maximum property. If $F$ is such a cluster method, then the image of $F$ has the relative maximum property by Lemma 3.3. If $B \subseteq C(P)$ has the relative maximum property, then $F_B$ defined by $F_B(d) = \max\{d_i \in B : d_i \leq d\}$ is a cluster method satisfying Conditions 1, 4, 5 whose image is $B$. In particular, there is only one idempotent, isotone cluster method having the ultrametrics as its image. We proceed now to describe it.

For any reflexive symmetric relation $R$ on $P$, we define a new relation $E_R$ on $P$ by the rule $xE_Ry$ if there are finitely many elements $x_0, x_1, \ldots, x_k$ of $P$ such that $x_0 = x$, $x_k = y$, and $x_{i-1}Rx_i$ for $i = 1, 2, \ldots, k$. It is easy to show that $E_R$ is an equivalence relation containing $R$, and that any equivalence relation containing $R$ must also contain $E_R$. For that reason $E_R$ is called the equivalence relation *generated by* $R$, or the *transitive closure* of $R$.

Let $d$ be a fixed DC on $P$. For any $h \geq 0$, $Td(h)$ is then a reflexive symmetric relation. Let $E_h$ denote the transitive closure of $Td(h)$, and note that each $E_h$ is an equivalence relation. It follows that if $C : [0, \infty) \mapsto \Sigma$ is defined by $C(h) = E_h$, then $C$ is a hierarchy, so (by Theorem 2.8) $d_C$ is an ultrametric. If $u \leq d$ is any ultrametric, it follows that $C(h) \subseteq Tu(h)$ for all $h \geq 0$, so $C \leq Tu$. By Lemma 2.6, it follows that $u \leq d_C$. Thus $F(d) = d_C$ is the unique cluster method satisfying Conditions 1, 4, 5 and having its image the set of ultrametrics. This cluster method is called *single-linkage clustering*, and it coincides with the method described by Sibson [10] and by Jardine and Sibson [8]. In summary, here is what we have observed.

**Theorem 3.8** *There is one and only one cluster method* $F$ *that is idempotent and isotone, and whose image is the set of all ultrametrics on* $P$. *It is defined by the rule*

$$[TF(d)](h) = \gamma_{_E}[Td(h)] \text{ for all } h \geq 0,$$

*where* $\gamma_{_E}(R)$ *denotes the transitive closure of the reflexive symmetric relation* $R$.

It is rather curious that we did not need Conditions 2 and 3 in order to prove Theorem 3.8. Yet they both hold for the cluster method described in Theorem 3.8. Condition 2 follows from the fact that if $u$ is any ultrametric, and if $\rho$ is any permutation of $E$, then $u \circ (\rho \times \rho)$ is also an ultrametric. Condition 3 is a consequence of the fact that $u$ an ultrametric with $\alpha > 0 \Longrightarrow \alpha u$ is an ultrametric. It is equally curious that the result does not require any assumption of continuity.

Fairly strong use is made of what is called an *indicator family* in [8] and [10]. We shall approach this somewhat differently. The key item here is that single linkage clustering is defined in terms of the transitive closure mapping on $\Sigma(P)$. Using $\gamma_E$ to denote this mapping, it has has the property that

(T1) $\gamma_E = \gamma_E \circ \gamma_E$.

(T2) $R_1 \subseteq R_2 \Longrightarrow \gamma_E(R_1) \subseteq \gamma_E(R_2)$.

(T3) $R \subseteq \gamma_E(R)$ for all $R \in \Sigma$.

So we generalize this by assuming the existence of a mapping $\gamma : \Sigma \mapsto \Sigma$ satisfying (T1), (T2), and (T3). The NSC $\gamma \circ Td$ is defined by $[\gamma \circ Td](h) = \gamma(Td(h))$. We may then define a cluster method $F_\gamma$ by the rule

$$F_\gamma(d) = d_{\gamma \circ Td}. \tag{3.1}$$

If we phrase this in terms of the associated numerically stratified clusterings, we have that $[TF_\gamma(d)](h) = \gamma[Td(h)]$ for all $h \geq 0$. Let us denote the *image* of $d$ by

$$\mathrm{im}(d) = h_0 < h_1 < \cdots < h_t.$$

The dissimilarity coefficient $F_\gamma(d)$ is then given by

$$[F_\gamma(d)](a,b) = \min\{h_i : (a,b) \in \gamma[Td(h_i)]\}.$$

It is immediate that $\mathrm{im}(F_\gamma(d)) \subseteq \mathrm{im}(d)$.

**Remark 3.9** It is worth noting that if $\gamma$ satisfies (T1), (T2), and (T3), then for any $R \in \Sigma$, it must be true that

$$\gamma(R) = \min\{S \in \Sigma : S = \gamma(S),\ R \subseteq S\}.$$

**Theorem 3.10** *Let $\gamma$ satisfy Conditions* (T1), (T2) *and* (T3). *Then $F_\gamma$ satisfies Conditions* 1, 4 *and* 5.

**Proof:** Condition 1 follows from (T1), while condition 4 comes from (T3), and Condition 5 is a consequence of (T2). ∎

Before proceeding, we need to chat a bit about the way that a nested sequence of reflexive symmetric relations determines a dissimilarity coefficient. Let $d \in C(P)$ have image $0 = h_0 < h_1 < \cdots < h_t$. The *canonical representation* for $d$ is then given by

$$d \mapsto R_0 \subset R_1 \subset \cdots \subset R_t;\ h_0 < h_1 < \cdots < h_t \tag{3.2}$$

where $R_i = Td(h_i)$ for $i = 0, 1, \ldots, t$. Note that $d(a,b) = \min\{h_i : (a,b) \in R_i\}$. Indeed, a less efficient representation for $d$, called an *extended representation* may be obtained as

$$d \mapsto S_0 \subseteq S_1 \subseteq \cdots \subseteq S_t;\ k_0 \leq k_1 \leq \cdots \leq k_u \tag{3.3}$$

as long as $\mathrm{im}(d) \subseteq \{k_0, k_1, \ldots, k_u\}$, and $S_i = Td(k_i)$ for $i = 0, 1, \ldots, u$. The point is that if $k_i$ is not in the image of $d$, or if $k_i = k_{i+1}$, then $S_{k_i} = S_{k_{i+1}}$. The dissimilarity coefficient $d$ is then recaptured by the observation that $d(a,b) = \min\{k_i : (a,b) \in S_i\}$.

Conversely, if each $S_i \in \Sigma(P)$, $S_u = P \times P$, each $k_j \in [0, \infty)$, and $k_0 = 0$, the following pair of sequences determine a unique dissimilarity coefficient $d$ such that for $i = 0, 1, \ldots, u$, $S_i = Td(k_i)$:

$$S_0 \subseteq S_1 \subseteq \cdots \subseteq S_u;\ k_0 \leq k_1 \leq \cdots \leq k_u \tag{3.4}$$

To see this, we just define $d(a,b) = \min\{k_j : (a,b) \in S_j\}$, and note that

$$d(a,b) \leq k_i \Longleftrightarrow \min\{j : (a,b) \in S_j\} \leq i \Longleftrightarrow (a,b) \in S_i \tag{3.5}$$

There are two very good reasons for wanting representations of the form (3.3) as well as (3.2). First of all, if the image of $F(d)$ is necessarily contained in the image of $d$, and if $d$ is represented as in (3.2), it turns out to be useful to represent $F(d)$ in the form

$$d \mapsto S_0 \subseteq S_1 \subseteq \cdots \subseteq S_t;\ h_0 < h_1 < \cdots < h_u \tag{3.6}$$

where $S_i = [TF(d)](h_i)$. Now if $h_i$ is not in the image of $F(d)$, we have $S_i = S_{i+1}$. To illustrate this, suppose $P = \{a, b, c\}$, with $d$ given by $d(a,b) = 1$, $d(a,c) = 2$, $d(b,c) = 3$. Applying single linkage clustering produces an ultrametric $u$ whose values are $u(a,b) = 1$, and $u(a,c) = u(b,c) = 2$. The corresponding representations would be

$$d \mapsto R_0 \subset Td(1) \subset Td(2) \subset Td(3);\ 0 < 1 < 2 < 3, \tag{3.7}$$

$$u \mapsto R_0 \subseteq Tu(1) \subseteq Tu(2) \subseteq Tu(3);\ 0 < 1 < 2 < 3, \tag{3.8}$$

where $Tu(1) = Td(1)$, and $Tu(2) = Tu(3) = P \times P$.

Secondly, suppose $d$ has its canonical representation as in (3.2). If $\theta$ is a nondecreasing map on $[0, \infty)$ with $\theta(0) = 0$, then $\theta d$ is the dissimilarity coefficient defined

by $(\theta d)(a,b) = \theta(d(a,b))$. Noting that $\theta(h_i)$ might possibly equal $\theta(h_{i+1})$, an extended representation for $\theta d$ may be obtained as

$$\theta d \mapsto S_0 \subseteq S_1 \subseteq \cdots \subseteq S_u;\ \theta h_0 \leq \theta h_1 \leq \cdots \theta h_u \tag{3.9}$$

where $S_i = Td(\theta h_i)$.

It is time finally to put all of this together. We let $\gamma$ satisfy (T2), and

(T4) $\gamma(P \times P) = P \times P$.

We may still define $F_\gamma$ as in (3.1). Recall then that $[TF_\gamma(d)](h) = \gamma(Td(h))$. If $\theta : [0,\infty) \mapsto [0,\infty)$ is a nondecreasing mapping such that $\theta(0) = 0$, we shall prove that for any dissimilarity coefficient $d$, $F(\theta d) = \theta F(d)$, thereby strengthening Condition 3. Accordingly, let $d$ have the canonical representation specified in (3.2). Then

$$F_\gamma(d) \mapsto \gamma Td(h_0) \subseteq \gamma Td(h_1) \subseteq \cdots \subseteq \gamma Td(h_t);\ h_0 < h_1 \cdots < h_t. \tag{3.10}$$

Applying $\theta$ to $F_\gamma(d)$ produces

$$\theta(F_\gamma(d)) \mapsto S_0 \subseteq S_1 \subseteq \cdots \subseteq S_t;\ \theta(h_0) \leq \theta(h_1) \leq \cdots \leq \theta(h_t), \tag{3.11}$$

where $S_i = \gamma Td(\theta h_i)$. If $k_0 < k_1 < \cdots < k_u$ denotes the distinct elements of $\theta(h_0) \leq \theta(h_1) \leq \cdots \leq \theta(h_t)$, the representation becomes

$$\theta(F_\gamma(d)) \mapsto \gamma Td(k_0) \subseteq \gamma Td(k_1) \subseteq \cdots \subseteq \gamma Td(k_u);\ k_0 < k_1 < \cdots < k_u \tag{3.12}$$

On the other hand, applying $\theta$ directly to $d$ tells us that

$$\theta d \mapsto Td(h_0) \subseteq Td(h_1) \subseteq \cdots \subseteq Td(h_t);\ \theta(h_0) \leq \theta(h_1) \leq \cdots \leq \theta(h_t). \tag{3.13}$$

We also have a more efficient representation

$$\theta d \mapsto Td(k_0) \subseteq Td(k_1) \subseteq \cdots \subseteq Td(k_u);\ k_0 < k_1 < \cdots < k_u. \tag{3.14}$$

If we apply $F_\gamma$ to $\theta d$, this leads to the representation

$$F_\gamma(\theta d) \mapsto \gamma Td(k_0) \subseteq \gamma Td(k_1) \subseteq \cdots \subseteq \gamma Td(k_u);\ k_0 < k_1 < \cdots < k_u. \tag{3.15}$$

A comparison of equations (3.12) and (3.15) now establishes that $F_\gamma(\theta d) = \theta F_\gamma(d)$.

Thus Condition 3 has a rather strong generalization. Any cluster method of the form $F = F_\gamma$ satisfies Condition 3b below. This type of cluster method is called *flat* in [2]

*Condition* 3b. *Compatibility Condition.* For any nondecreasing mapping $\theta$ on $[0,\infty)$ such that $\theta(0) = 0$, and any DC $d$, $F(\theta d) = \theta F(d)$.

There is an interesting converse to this result. We shall just sketch its proof.

**Theorem 3.11** *Let $F$ be a cluster method that satisfies Condition* 3b*. There then exists a mapping $\gamma$ on $\Sigma(P)$ satisfying* (T2) *and* (T4) *such that $F = F_\gamma$.*

**Definition 3.12** Let $R \in \Sigma(P)$. We let $d_R$ be the DC defined by $d_R(a,b) = 0$ if $aRb$, with $d_R(a,b) = 1$ if $h > k$. Note that if $R \neq P \times P$, then the image of $d_R = \{0,1\}$. Note that $d_R$ may actually be viewed as a characteristic function for $R$. Indeed, if $\mathrm{im}(d) \subseteq \{0,1\}$, then $d = d_{Td(0)}$.

It will be convenient to introduce a family of nondecreasing mappings on $[0,\infty)$ as follows: Let $k \geq 0$. Define $\tau_k : [0,\infty) \mapsto [0,\infty)$ by the rule

$$\tau_k(h) = 0 \text{ if } h \leq k \text{ and } 1 \text{ if } h > k.$$

**Lemma 3.13** *If $F$ satisfies Condition* 3b*, then $F(d_R) = d_S$ for some $S \in \Sigma(P)$.*

**Proof:** Since the image of $d_R \subseteq \{0,1\}$, it is clear that $d_R = \tau_{1/2}(d_R)$. Hence $F(d_R) = F(\tau_{1/2}(d_R)) = \tau_{1/2}(F(d_R))$. But the image of $\tau_{1/2}(F(d_R)) \subseteq \{0,1\}$ shows that it is of the form $d_S$ for some $S \in \Sigma(P)$. ∎

**Lemma 3.14** *Let $d$ be a DC, and $R = Td(h)$. Then $d_R = \tau_h(d)$.*

**Proof:** We only need note that $[\tau_h(d)](a,b) = 0 \Longleftrightarrow d(a,b) \leq h$ and $[\tau_h(d)](a,b) = 1$ otherwise. ∎

**Lemma 3.15** *Let $F$ satisfy Condition* 3b*. Let $d$ be a dissimilarity coefficient with $d' = F(d)$. Then if $R = Td(h)$, it follows that $Td'(h) = S$, where $d_S = F(d_R)$.*

**Proof:** We begin by observing that $[\tau_h d](a,b) = 0$ if $d(a,b) \leq h$, and equals 1 otherwise. This just says that $\tau_h d = d_R$. Note that

$$d_S = F(d_R) = F(\tau_h(d)) = \tau_h(F(d)) = \tau_h(d').$$

The proof is completed by noting that

$$d_S(a,b) = 0 \Longleftrightarrow (\tau_h d')(a,b) = 0 \Longleftrightarrow d'(a,b) \leq h. \text{ ∎}$$

Let $F$ satisfy Condition 3b. We may define a mapping $\gamma$ on $\Sigma(P)$ by taking $\gamma(R) = S$, where $F(d_R) = d_S$. Once we do this, it is immediate that $F = F_\gamma$. It follows that $\gamma$ satisfies (T2) and (T4). This proves Theorem 3.11. Note that we did

not need the full strength of Condition 3b. We only needed it for mappings of the form $\tau_h$.

It is also worth noting that every cluster method of the form $F_\gamma$ is necessarily isotone.

# 4 Continuity Considerations

This section is devoted to a careful study of the role of continuity. We begin with some needed terminology. A mapping $\theta$ on $[0, \infty)$ is called an *order automorphism* if it is a bijection having the property that $h \leq k \iff \theta(h) \leq \theta(k)$. Recall that the metric $\Delta_0$ on $C(P)$ is defined by

$$\Delta_0(d_1, d_2) = \max\{ \mid d_1(a, b) - d_2(a, b) \mid : a, b \in P\},$$

and that to say that a cluster method $F$ is *continuous* at $d$ with respect to this metric means that for any $\varepsilon > 0$, there corresponds a $\delta > 0$ such that

$$0 < \Delta_0(d_1, d) < \delta \Longrightarrow \Delta_0(Fd_1, Fd) < \varepsilon.$$

In the Jardine-Sibson model, the following two conditions play important, though historically controversial, roles.

**Condition 3a.** (*monotone equivariance*) For any order automorphism $\theta$ of $[0, \infty)$, $F(\theta d) = \theta F(d)$ for every $d \in C(P)$.

**Condition 6.** $F$ is continuous with respect to the $\Delta_0$ metric at every $d \in C(P)$.

Condition 3a is used to establish Condition 6. A careful treatment of this may be found in Janowitz [4]. We indicate here how continuity follows without assuming Condition 3a. In fact, we shall just assume Condition 3 and isotonicity, and use this to establish continuity at all definite DCs. Accordingly, we assume that $d \in C(P)$ is definite, recalling that this means that if $a \neq b$, then $d(a, b) > 0$. We then let $M(d)$ denote the largest member, and $k(d)$ the smallest nonzero member of the image of $d$. Choose $\varepsilon$ so that $0 < \varepsilon < k(d)$. Then for $a \neq b$, $k(d) \leq d(a, b)$, so $1 \leq \frac{d(a,b)}{k(d)}$. Hence

(1) $\varepsilon \leq \frac{\varepsilon}{k(d)} d(a, b)$.

Multiplying through by $-1$, and adding $d(a, b)$ to the resulting inequality produces the fact that $d(a, b) - \frac{\varepsilon}{k(d)} d(a, b) \leq d(a, b) - \varepsilon$. In other words,

(2) $(1 - \frac{\varepsilon}{k(d)})d(a,b) \leq d(a,b) - \varepsilon$.

A similar argument produces

(3) $d(a,b) + \varepsilon \leq (1 + \frac{\varepsilon}{k(d)})d(a,b)$.

Since $0 < \varepsilon < k(d)$, we know that $0 < \frac{\varepsilon}{k(d)} < 1$, so $1 - \frac{\varepsilon}{k(d)} > 0$. Now let $0 < \Delta_0(d,d') < \varepsilon$, and note that for all $a, b \in P$, $0 < \mid d(a,b) - d'(a,b) \mid < \varepsilon$. Combining (2) and (3) produces

$$(1 - \frac{\varepsilon}{k(d)})d(a,b) \leq d(a,b) - \varepsilon < d'(a,b) < d(a,b) + \varepsilon \leq (1 + \frac{\varepsilon}{k(d)})d(a,b).$$

Thus

(4) $(1 - \frac{\varepsilon}{k(d)})d(a,b) < d'(a,b) < (1 + \frac{\varepsilon}{k(d)})d(a,b)$.

Since this is true for all distinct $a, b \in P$, we have

(5) $(1 - \frac{\varepsilon}{k(d)})d < d' < (1 + \frac{\varepsilon}{k(d)})d$

Since $F$ preserves the ordering of dissimilarity coefficients, it follows that

(6) $F[(1 - \frac{\varepsilon}{k(d)})d] \leq F(d') \leq F[(1 + \frac{\varepsilon}{k(d)})d]$ .

By Condition 3, we have

(7) $(1 - \frac{\varepsilon}{k(d)})F(d) \leq F(d') \leq (1 + \frac{\varepsilon}{k(d)})F(d)$ .

This implies that

$$F(d) - \frac{\varepsilon}{k(d)}F(d) \leq F(d') \leq F(d) + \frac{\varepsilon}{k(d)}F(d).$$

Hence $\mid [Fd](a,b) - [Fd'](a,b) \mid \leq \frac{\varepsilon}{k(d)}[Fd](a,b) \leq \frac{\varepsilon M(d)}{k(d)}$. Since this is true for all $a, b$ it follows that $\Delta_0(F(d), F(d')) \leq \frac{\varepsilon M(d)}{k(d)}$, thus establishing continuity at $d$. In summary, we have established that

$$\Delta_0(d,d') < \varepsilon < k(d) \Longrightarrow \Delta_0(F(d), F(d')) \leq \frac{\varepsilon M(d)}{k(d)}. \tag{4.16}$$

Here is a formal statement of the situation.

**Theorem 4.1** *Every isotone cluster method that satisfies Condition* 3 *is continuous at all definite dissimilarity coefficients.*

We turn next to the situation where $d$ is not definite. We may still establish a weakened form of continuity. We say that $F$ is 0-*continuous* at $d$ if for any $\varepsilon > 0$, there corresponds a $\delta > 0$ such that

$$0 < \Delta_0(d_1, d) < \delta \text{ with } d(a,b) = 0 \text{ implying that } d_1(a,b) = 0 \Longrightarrow \Delta_0(Fd_1, Fd) < \varepsilon.$$

Similarly, to say that $F$ is *left-continuous* at $d$ is to say that for any $\varepsilon > 0$, there corresponds a $\delta > 0$ such that

$$0 < \Delta_0(d_1, d) < \delta \text{ with } d_1 \leq d \Longrightarrow \Delta_0(Fd_1, Fd) < \varepsilon.$$

The reader should note that left continuity is a special case of 0-continuity. We leave to the reader the routine verification that as long as $d(a,b) = 0 \Longrightarrow d_1(a,b) = 0$, equation (4.16) holds. To show that this is the strongest possible result, we now present an example.

**Example 4.2** Define $F$ to be single-linkage clustering on all definite DCs, and $F(d) = d$ otherwise. Then $F$ is isotone, idempotent, and monotone equivariant. Furthermore, $F$ is continuous at every definite DC, and 0-continuous elsewhere.

In order to establish continuity at dissimilarity coefficients that are not necessarily definite, the situation is a bit more complicated. One condition that will guarantee this is a somewhat stronger version of Condition 3.

**Condition 3c.** $F(\psi D) = \psi F(d)$ for all mappings of the form $\psi(h) = \alpha \cdot h$ for some positive $\alpha$, or of the form $\psi(h) = \max\{0,\ h - k\}$ for some $k > 0$.

**Theorem 4.3** *Every cluster method $F$ that is idempotent, isotone, and satisfies Condition* 3c *is in fact continuous.*

**Remark 4.4** There is an interesting partial converse to Theorem 4.3. By [4], every continuous monotone equivariant cluster method $F$ is in fact of the form $F_\gamma$, and therefore isotone. If it is also isotone, and its image is the set of all ultrametrics, then it is in fact single-linkage clustering.

What really comes out of this is that cluster methods of the form $F_\gamma$ (see equation (3.1)) satisfy a condition stronger than continuity. This condition may be thought of as a type of Lipschitz condition.

**Condition 6a.** For any dissimilarity coefficient $d$, there correspond positive constants $k, M$ such that

$$0 < \Delta_0(d, d') < k \Longrightarrow \Delta_0(F(d), F(d')) < M\Delta_0(d, d').$$

It is of interest that the constants $M, k$ in Condition 6a are independent of the choice of the cluster method $F$; one only needs $F$ to be of the form $F_\gamma$.

# 5 Continuity versus Monotone Equivariance

It is easy to make a case for using continuous cluster methods. Any time we measure anything, there are inevitable (hopefully small) errors in the measurement process. Ideally, one would like that small errors in the input to a clustering algorithm should result in small errors in the output. But this is precisely what continuity guarantees. However, in the Jardine-Sibson model for cluster analysis, continuity seems intimately related to notions of monotone equivariance. But continuity is related to what happens when dissimilarity coefficients are close together, while monotone equivariance is the antithesis of this in that it tells us that the actual values of $d(a, b)$ and $d(x, y)$ are not important – only whether $d(a, b) < d(x, y)$. In this penultimate section we present work in progress that makes these observations more precise, and establishes a connection between these seemingly incompatible views.

**Definition 5.1** Let $d \in C(P)$ have image $0 = h_0 < h_1 < \cdots < h_t$. The *mesh width* of $d$ is defined by

$$\mu(d) = \frac{1}{2} \min\{h_i - h_{i-1} : 1 \leq i \leq t\}.$$

The notion of mesh width is important if one is to understand what is happening in connection with continuity. It turns out that when dissimilarities are sufficiently close to each other, they are also related ordinally.

**Lemma 5.2** *Let $d \in C(P)$. If $\Delta_0(d, d') < \mu(d)$, then*

$$d(a, b) < d(x, y) \Longrightarrow \max\{d(a, b), d'(a, b)\} < \min\{d(x, y), d'(x, y)\}.$$

**Proof:** We need only observe that

$$\max\{d(a, b), d'(a, b)\} < d(a, b) + \mu(d) \leq d(x, y) - \mu(d) < \min\{d(x, y), d'(x, y)\}. \blacksquare$$

Lemma 5.2 suggests that in the presence of monotone equivariance, closeness is not the important concept. What is important is the ordinal concept expressed by

$$d(a,b) < d(x,y) \Longrightarrow d'(a,b) < d'(x,y). \tag{5.17}$$

This is clearly equivalent to the assertion that

$$d'(a,b) \leq d'(x,y) \Longrightarrow d(a,b) \leq d(x,y). \tag{5.18}$$

**Definition 5.3** It will be convenient to write $d \precsim d'$ to denote the fact that

$$d(a,b) < d(x,y) \Longrightarrow d'(a,b) < d'(x,y),$$

and $d \precapprox d'$ to denote the fact that

$$d(a,b) < d(x,y) \Longrightarrow \max\{d(a,b), d'(a,b)\} < \min\{d(x,y), d'(x,y)\}.$$

**Remark 5.4** There is a strict weak order $W_d$ associated with any dissimilarity coefficient $d$. It is defined on the set of two element subsets of $P$ by the rule $\{a,b\}W_d\{x,y\} \iff d(a,b) < d(x,y)$. To say that $d \precsim d'$ is then equivalent to $W_d \subseteq W_{d'}$. Evidently $\precsim$ is a quasiorder on $C(P)$. As an immediate consequence of this definition and equation (5.18), we see that if $d \precsim d'$, then

$$d'(x,y) \leq d'(a,b) \leq d'(s,t) \text{ with } d'(x,y) = d'(s,t) \Longrightarrow d(a,b) = d(x,y).$$

There is a second rather intuitive way of viewing the notion of $d \precsim d'$. We recall from equation (3.2) that the canonical representation for the dissimilarity coefficient $d$ is given by

$$d \mapsto R_0 \subset R_1 \subset \cdots \subset R_t;\ h_0 < h_1 < \cdots h_t,$$

where $R_i = Td(h_i)$ for all $i$. If the values of $d$ only have ordinal significance, we are interested in $d(a,b)$ only when $a \neq b$. Thus the relation $R_0 = Td(0)$ is only of interest when there is a pair $a, b$ with $a \neq b$ such that $d(a,b) = 0$; i.e., when $R_0 \neq R_\emptyset$. Since the actual values of the $h_i$ are irrelevant, we say that a relation $R$ is a *proper splitting relation* of $d$ if $R = Td(h)$ for some $h \geq 0$, and $R \neq R_\emptyset$. This leads us to define the *ordinal representation* for $d$ by the rule

$$d \propto R_0 \subset R_1 \subset \cdots \subset R_t, \tag{5.19}$$

where $\{R_i : 0 \leq i \leq t\}$ is the set of proper splitting relations of $d$. As we shall now see, there is a close connection between $d \precsim d'$ and their associated ordinal representations.

**Lemma 5.5** *Let $d, d' \in C(P)$ with ordinal representations $d \propto R_0 \subset R_1 \subset \cdots \subset R_t$, and $d' \propto R'_0 \subset R'_1 \subset \cdots \subset R'_u$ . Then $d \precsim d'$ is equivalent to each $R_i$ being of the form $R'_{i'}$.*

**Proof:** Assume first $d \precsim d'$. We must show that corresponding to each $h \geq 0$, there is a $k \geq 0$ such that $Td(h) = Td'(k)$. We may clearly assume that $h$ is in the image of $d$. If $R = Td(h)$, let $k = \max\{d'(a,b) : aRb\}$, and note that if $d'(a,b) = k$, then $d(a,b) = h$. Then $sRt \Longrightarrow d'(s,t) \leq k$. We claim that $d'(s,t) \leq k \Longrightarrow d(s,t) \leq h$. If this were not true, we would have $d(s,t) > h$. But then $d(a,b) < d(s,t)$ would force $k = d'(a,b) < d'(s,t)$, a contradiction.

Now assume that each $R_i$ is of the form $R'_{i'}$. Suppose $d(a,b) < d(x,y)$. Let $i$ be the least index such that $aR_ib$, and $j$ the least index such that $xR_jy$. Obviously, $aR'_{i'}b$ and not $xR'_{i'}y$. Since $R'_{i'} = Td'(k)$, it is immediate that $d'(a,b) \leq k < d'(x,y)$, thus completing the proof. ∎

A monotone equivariant cluster method is really operating on equivalence classes of dissimilarity coefficients. Recalling that an *order automorphism* of $[0,\infty)$ is a mapping $\theta$ of $[0,\infty)$ onto $[0,\infty)$ having the property that $h \leq k \Longleftrightarrow \theta(h) \leq \theta(k)$, the required equivalence relation is defined by

$$d_1 \sim_g d_2 \Longleftrightarrow d_2 = \theta d_1 \text{ where } \theta \text{ is an order automorphism of } [0,\infty),$$

is called *global order equivalence* and was introduced by Sibson [11]. If the values of the input dissimilarity coefficients have only ordinal significance, Sibson argues that it is natural to use cluster methods $F$ that preserve global order equivalence in the sense that

$$d_1 \sim_g d_2 \Longrightarrow F(d_1) \sim_g F(d_2).$$

But Janowitz in [3] points out that this is not quite what is needed. One really wants to work with *weak order equivalence* which is defined by the rule

$$d_1 \sim_w d_2 \Longleftrightarrow d_1(a,b) < d_1(x,y) \text{ is equivalent to } d_2(a,b) < d_2(x,y).$$

Evidently, this just says that $d_1$ and $d_2$ have identical ordinal representations. Note that $d_1 \sim_w d_2$ is equivalent to the assertion that $d_1 \precsim d_2$ and $d_2 \precsim d_1$. It now seems natural to use cluster methods that preserve weak order equivalence in the sense that $d_1 \sim_w d_2 \Longrightarrow F(d_1) \sim_w F(d_2)$. The two notions of equivalence are connected by the fact that $d_1 \sim_g d_2 \Longleftrightarrow d_1 \sim_w d_2$ and $Td_1(0) = Td_2(0)$ Thus $d_1 \sim_g d_2 \Longrightarrow d_1 \sim_w d_2$, but not conversely. This may be concretely illustrated by taking $P = \{a,b,c\}$, and defining $d_1(a,b) = 0$, $d_1(a,c) = 2$, and $d_1(b,c) = 3$; $d_2(a,b) = 1$, and $d_2(x,y) = d_1(x,y)$ otherwise. We then have $d_1 \sim_w d_2$ but not $d_1 \sim_g d_2$. Note that $d_1$ and $d_2$ have the same ordinal representations, but different canonical representations. Indeed,

$$d_1 \mapsto R_0 \subset R_1 \subset P \times P;\ 0 < 2 < 3,$$

while

$$d_2 \mapsto R_\emptyset \subset R_0 \subset R_1 \subset P \times P;\ 0 < 1 < 2 < 3,$$

where $R_0 = Td_1(0) = Td_2(1)$, and $R_1 = Td_1(2) = Td_2(2)$. This naturally raises the question of what it means for $d' \in C(P)$ to have the property that $d' \sim_g d''$ for some $d'' \in C(P)$ such that $\Delta_0(d, d'') < \mu(d)$. It will be convenient to let $[d]_g$ denote the equivalence class containing $d$ with respect to global order equivalence, and $[d]_w$ the corresponding class for weak order equivalence.

**Lemma 5.6** *Let $d, d' \in C(P)$. The validity of equation* (5.17) *is then equivalent to the existence of $d'' \in C(P)$ such that $d' \sim_g d''$ and $\Delta_0(d, d'') < \mu(d)$.*

**Proof:** If $d' \sim_g d''$, and $\Delta_0(d, d'') < \mu(d)$, then (5.17) follows from Lemma 5.2. Suppose conversely that $d, d'$ satisfy equation (5.17). Denote the image of $d$ by $h_0 < h_1 < \cdots h_t$, and for each index $i$, let $m_i = \min\{d'(a, b) : d(a, b) = h_i\}$, and $M_i = \max\{d'(a, b) : d(a, b) = h_i\}$. For $i = 1, 2, \ldots, t$, we then have $m_{i-1} \leq M_{i-1} < m_i \leq M_i$. Evidently $m_i \leq d'(x, y) \leq M_i \Longrightarrow d(x, y) = h_i$. So for any $\varepsilon > 0$, we may construct an order automorphism $\theta$ of $[0, \infty)$ so that $h_i < m_i \leq M_i < h_i + \varepsilon$. If $d'' = \theta d'$, we then have $d' \sim_g d''$, and $\Delta_)(d, d') < \varepsilon$. The proof is completed by taking $\varepsilon < \mu(d)$. ∎

**Remark 5.7** An immediate consequence of Lemma 5.6 is the fact that equation (5.17) is equivalent to the fact that $d$ is arbitrarily close to members of $[d']_g$ in the sense that corresponding to any $\delta > 0$, there is a $d'' \in C(P)$ such that $d' \sim_g d''$ and $\Delta_0(d, d'') < \delta$. Thus equation (5.17) is equivalent to the existence of a sequence $\{d_n\}$ of members of $[d']_g$ which converges to $d$ in the $\Delta_0$-topology. It is now clear that $d \precapprox d' \Longrightarrow d \precsim d'$ and $d \precsim d' \Longrightarrow$ the existence of $d''$ such that $d' \sim_g d''$ and $d \precapprox d''$

At the moment, we must just be content to introduce a working definition for what we shall call *order continuity*. The properties of this notion must still be investigated, and suitable algorithms must be established. If $F$ is continuous, there must clearly exist some $\delta$ such that $0 < \delta < \mu(d)$ and

$$0 < \Delta_0(d, d') < \delta \Longrightarrow \Delta_0(Fd, Fd') < \mu(Fd).$$

In other words, there is a $\delta > 0$ such that

$$0 < \Delta_0(d, d') < \delta \Longrightarrow d \precsim d' \text{ and } F(d) \precsim F(d'). \tag{5.20}$$

If $F$ preserves global order equivalence, this is clearly equivalent to the assertion that

$$d \precsim d' \Longrightarrow F(d) \precsim F(d').$$

This leads to the following tentative definition for order continuity.

**Definition 5.8** A cluster method $F$ is called *order continuous* if for all $d, d' \in C(P)$, it is true that $d \precsim d' \Longrightarrow F(d) \precsim F(d')$.

Evidently, any order continuous cluster method $F$ must also preserve weak order equivalence, so what we have here is a strengthening of the idea of a cluster method preserving weak order equivalence. Cluster methods that preserve weak order equivalence seem naturally appropriate for use with ordinal data, since the ordinal representation for the output DC is determined solely by the ordinal representation of the input DC $d$. The connection between continuity and ordinal continuity is summarized in the next Lemma. Its proof is contained in the preceding discussion.

**Lemma 5.9** *Every continuous cluster method that preserves global order equivalence is order continuous.*

We now make contact with monotone equivariance. Cluster methods are often viewed as mechanisms to summarize information. It therefore seems reasonable to assume that $F$ has the property that the image of $F(d)$ cannot have more levels than the image of $d$. Such a cluster method is said to *compress information* (see [3], p. 82). Two cluster methods $F, G$ are said to be *globally order similar* if for every $d \in C(P)$ it is true that $F(d) \sim_g G(d)$; they are *weakly order similar* if for each $d \in C(P)$, $F(d) \sim_w G(d)$. We now restate [3], Theorem 4, p. 83, and Lemma 5, p. 84:

**Theorem 5.10** (1) *A cluster method $F$ compresses information and preserves global order equivalence if and only if it is globally order similar to a monotone equivariant cluster method.*

(2) *Every cluster method $F$ that compresses information and preserves weak order equivalence is weakly order similar to a monotone equivariant cluster method.*

Examples are given in [3] to show that monotone equivariant cluster methods need not preserve weak order equivalence. In view of Theorem 5.10, it seems natural to ask when a cluster method $F$ is weakly order similar to a flat cluster method $G$. This question will be addressed in later research.

**Lemma 5.11** *Let $d \in C(P)$, and $R \in \Sigma(P)$. Then $d_R \precsim d \Longleftrightarrow R = Td(h)$ for some $h$.*

**Proof:** Suppose first that $R = Td(h)$ for some $h \geq 0$. Then $aRb \Longleftrightarrow d(a,b) \leq h$, so $d_R(a,b) < d_R(x,y) \Longrightarrow aRb$ with $xRy$ false. It follows that $d(a,b) \leq h$ and $d(x,y) > h$, so $d(a,b) < d(x,y)$, whence $d_R \precsim d$. The converse follows from Lemma 5.5. ∎

**Example 5.12** Here is an example that shows that monotone equivariant ordinally continuous cluster methods need not be continuous. If the canonical representation for $d$ is

$$d \mapsto R_0 \subset R_1 \subset \cdots \subset R_t;\ h_0 < h_1 < \cdots h_t,$$

take as the canonical representation for $F_c(d)$

$$F_c(d) \mapsto S_0 \subset S_1 \subset \cdots \subset S_t;\ k_0 < k_1 < \cdots k_u,$$

where $\{S_i : 0 \leq i \leq u\} = \{R_j : 0 \leq j \leq t \text{ and } R_j \text{ is an equivalence relation}\}$. The values of $k_j$ are chosen by the rule that if $S_j = R_i$, then $k_j$ is taken to be $h_i$. We leave to the reader the routine verification that $F_c$ is monotone equivariant and ordinally continuous. We show with an example that it is not continuous.

Take $P = \{a, b, c\}$, and define $d$ and $d'$ as follows:

$$d(a,b) = d(a,c) = 1;\ d(b,c) = 2.$$

$$d'(a,b) = 1,\ d'(a,c) = 1 + \varepsilon,\ d'(b,c) = 2,$$

where $0 < \varepsilon < 1$. We then have

$$F_c d(a,b) = F_c d(a,c) = F_c d(b,c) = 2$$

$$F_c d'(a,b) = 1,\ F_c d'(a,c) = F_c d'(b,c) = 2.$$

To show $F_c$ is not continuous, just notice that $\Delta_0(d, d') = \varepsilon$, while $\Delta_0(F_c d, F_c d') = 1$.

# 6 Conclusion

A framework was introduced in which properties of various clustering algorithms could be compared. Working within this framework, a connection was established between ordinal and metric properties of cluster methods. Theorems 4.1 and 4.3 show that continuity follows from both ordinal and algebraic considerations. Theorem 3.8 characterized single-linkage clustering (a continuous cluster method) in terms of it being idempotent and isotone. An ordinal generalization of continuity was introduced, and results of Jardine and Sibson [8] were extended to investigate the nature of this notion of ordinal continuity and its connection with continuity and monotone equivariance. There is a need to further study both continuity and ordinal continuity both for monotone equivariant and more general cluster algorithms.

# References

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data,* Prentice-Hall, Englewood Cliffs, 1988.

[2] M. F. Janowitz, *An order theoretic model for cluster analysis*, SIAM J. Applied Math. **34** (1978), 47-60.

[3] M. F. Janowitz, *Preservation of Global Order Equivalence*, Journal of Math. Psych. **20**, (1979), 78-88.

[4] M. F. Janowitz, *Continuous L-Cluster Methods*, Discrete Applied Mathematics **3** (1981), 107-112.

[5] N. Jardine and R. Sibson, *The Construction of hierarchic and non-hierarchic classifications,* Computer Journal **11** (1968), 177-184.

[6] N. Jardine and R. Sibson, *A model for taxonomy,* Math. Biosci. **2** (1968), 465-482.

[7] N. Jardine and R. Sibson, *Choice of methods for automatic classification,* Computer Journal **14** (1971), 404-406.

[8] N. Jardine and R. Sibson, *Mathematical Taxonomy,* Wiley, New York, 1971.

[9] C. Jardine, N. Jardine and R. Sibson, *The structure and construction of taxonomic hierarchies,* Math. Biosci. **1** (1967), 173-179.

[10] R. Sibson, *A model for taxonomy II,* Math. Biosci. **6** (1970), 405-430.

[11] R. Sibson, *Order invariant methods for data analysis*, Journal of the Royal Statistical Society Series B **34** (1972), 311-349.