DOCUMENT RESUME

ED 149 581 FL 005 960

AUTHOR Black, Paul
TITLE Multidimensional Scaling Applied to Linguistic Relationships.
PUB DATE 73
NOTE 78p.; Expanded version of a paper presented at Conference on Lexicostatistics (University of Montreal, Montreal, Canada, May 19-20, 1973)
JOURNAL CIT Cahiers de l'Institut de Linguistique de Louvain; v3 n5-6 Dec 1976

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS African Languages; American Indian Languages; Bikol; Comparative Analysis; Descriptive Linguistics; *Dialect Studies; Glottochronology; Indonesian Languages; Language Classification; *Language Typology; *Language Variation; Linguistics; Malayo Polynesian Languages; Morphology (Languages); *Multidimensional Scaling; Phonology; Regional Dialects; Salish; *Statistical Analysis
IDENTIFIERS Konsoid; *Lexicostatistics; Lower Niger

ABSTRACT

As the several specific applications in this paper demonstrate, multidimensional scaling provides a long-needed means for investigating and describing spatial relationships among speech varieties. It is especially applicable to the relationships among varieties of a single language (or more properly, linguistic "clines"), which, as is generally known, are poorly described by the hierarchical mode of classification commonly used in comparative linguistics. But multidimensional scaling may also frequently be used to describe spatial variation which has persisted among distinct but related languages and which cannot be adequately described by an otherwise well-motivated hierarchical classification. These conclusions are illustrated by the application of multidimensional scaling to lexicostatistical percentages within four linguistic groups, located in the Philippines, Africa, and North America. (Author)

Multidimensional Scaling Applied to Linguistic Relationships*

by

Paul Black
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

As the several specific applications in this paper demonstrate, multidimensional scaling provides a long needed means for investigating and describing spatial relationships among speech varieties. It is especially applicable to the relationships among varieties of a single language (or more properly, linguistic 'cline'), which, as is generally known, are poorly described by the hierarchical mode of classification commonly used in comparative linguistics. But multidimensional scaling may also frequently be used to describe spatial variation which has persisted among distinct but related languages and which cannot be adequately described by an otherwise well motivated hierarchical classification. These conclusions are illustrated by the application of multidimensional scaling to lexicostatistical percentages within four linguistic groups, located in the Philippines, Africa, and North America.

* This is an expanded version of a paper first presented at the Conference on Lexicostatistics held at the University of Montreal on April 19-20, 1973.

Multidimensional Scaling Applied to Linguistic Relationships

by

Paul Black
Bell Laboratories
Murray Hill, New Jersey 07974

Multidimensional scaling is a relatively new technique of data analysis which is already widely used in such diverse fields as marketing, psychology, and political science, and which promises to be an equally valuable tool in the quantitative study of a variety of linguistic problems. In the lexicostatistical study of linguistic relationships, this technique provides a long needed complement to the traditional hierarchical mode of linguistic classification. While a "family tree" diagram or some other representation of a hierarchical subgrouping is an obviously appropriate way of describing the temporal hierarchy of linguistic splits through which a group of languages may have evolved from a common ancestral protolanguage, multidimensional scaling can be used to investigate and describe the spatial variation which originates in the wave-like spread of linguistic innovations within a single language, and which may also persist within the evolutionary tree to an extent sufficient to hamper the correct inference of this tree.

This paper begins by contrasting hierarchical (or 'tree') structure with spatial (or 'cline') structure in the context of a specific lexicostatistical problem, namely the description of the relationships among a dozen varieties of

Bikol. Here, hierarchical subgrouping is easily shown to be clearly inappropriate, both in terms of the structure of the lexicostatistical data and in terms of linguistic interpretation. A discussion of earlier linguistic approaches to such situations leads up to the application of multidimensional scaling to the data to produce a well defined spatial representation of the relationships, and the striking resemblance of this spatial representation with the actual geographical distribution of the varieties provides ample evidence for the appropriateness of this approach. As linguists cannot be expected to be familiar with multidimensional scaling, the next section provides a basic orientation in the mechanics and art of applying this technique and interpreting its results, and uses applications of this technique to relationships within Konsoid and Lower Niger to further characterize the range of its usefulness. The final section applies this technique to relationships within Salish in order to illustrate how it may be used to investigate the persistence of cline structure within the evolutionary tree.

The sets of data used in this paper all consist of familiar lexicostatistical percentages similar to those first used by Swadesh (1950). While there may be various linguistic difficulties involved with the use of these percentages (for an early summary and bibliography, cf. Hymes 1960), the results based on them in this paper add to the evidence for their general usefulness. Multidimensional scaling is, of course, equally applicable to more sophisticated lexicostatistical indices (e.g. cf. Kruskal, Dyen, and Black 1971 and in press) and may also prove to produce similar results when applied to nonlexical measures of linguistic similarity or difference. In this latter vein, it may be noted that informally derived spatial representations have been used by Hockett (1958: 328) to describe relationships measured by indices of mutual intelligibility and by Kroeber (1960) to describe those defined by indices of phonological and morphological similarity.

The particular approach taken in this paper began to develop during the course of lexicostatistical research on some ninety-five contemporary varieties of Indoeuropean, undertaken originally by Isidore Dyen[1] of Yale and continued by him in collaboration with Joseph B. Kruskal[1] of Bell Laboratories and myself. While the results obtained through the application of multidimensional scaling to various parts of this Indoeuropean data will eventually be published as part of a more comprehensive study of Indoeuropean lexicostatistics, their significance led me to explore the usefulness of this technique further by applying it to similar data from nearly a dozen other linguistic groups.[2] These do not represent the only nor quite the first applications of lexicostatistics. Simultaneously and independently, Sankoff and Sankoff (in press) have applied this technique for similar, but not identical, purposes, while Kirk and Epling (1972, 1973) and Henrici (in press) have applied multidimensional scaling to lexicostatistical data for somewhat different purposes. However, the present paper explains the technique and its implications for routine lexicostatistical application considerably more carefully than the other papers cited, and presents more extensive illustrations of its application.

Tree Structure or Cline Structure?

Bikol is a Malayo-Polynesian language spoken in the Philippines on the southern peninsula of Luzon and on several smaller, adjacent islands. Like any other widely spoken language, Bikol is far from homogeneous, but rather shows considerable dialectal differentiation throughout the area in which it is spoken. McFarland (1972) has investigated variation within Bikol in terms of both structual differences and lexicostatistical percentages among twelve Bikol varieties; these constitute a small, carefully selected sample of all Bikol varieties. Here McFarland's observations on phonological and morphological variation will be laid aside in order to consider what may be learned from an analysis of his lexicostatistical percentages alone. Figure 1 displays these percentages, as rounded to the nearest whole percentage, in the commonly used form of a lower half matrix without the diagonal.

Insert figure 1 about here

About all that is obvious from Fig. 1 are those things implied by the way in which the percentages were obtained. Each percentage is simply the percentage of homosemantic cognates shared by a pair of varieties in lexical samples selected in accord with some specific list of meanings (for further details, cf. Hymes 1960). Thus higher percentages tend to indicate greater lexical (and thus, presumably, greater overall) similarity. As the Bikol percentages are based on a four hundred item list of meanings, an estimate of statistical variation suggests that pairs of percentages differing by seven percentage points or more are statistically significantly different at about the five percent level.[3] In addition, it may be noted that percentages based on standard one or two hundred item samples are generally indicative of mutual intelligibility when they are above seventy or eighty percent. While Fig. 1 thus suggests that some Bikol varieties are significantly more similar than other pairs, and that many, if not most, of the pairs are mutually intelligible, this lower half matrix is hardly a visually striking revelation of the structure of the relationships.

The familiar distinction between the "family tree" and "wave" models of linguistic change (for a recent discussion, cf. Anttila 1972: ch. 15) suggests that linguistic relationships in general may conform to either of two quite different types of structure, or perhaps to a combination of both. Here the Bikol percentages will first be analyzed in order to determine whether they conform to the constraints of 'tree' structure, and are thus appropriately described by means of a hierarchical classification. While few linguists would expect hierarchical subgrouping to prove appropriate for describing relationships within a single language such as Bikol (although many use it to approximate such relationships), there would be little need to propose a radically new mode of description should the older one prove fully adequate. As the Bikol percentages will be shown to conform poorly to tree structure, however, the thesis that they conform to 'cline' structure will be considered. For the purpose at hand, a 'cline' structure will be considered to be characterized by potentially continuous variation in some sort of meaningful space. The fact that multidimensional scaling can represent the Bikol varieties as a two dimensional configuration which correlates highly with their geographical distribution will thus be offered as evidence that the percentages do in fact conform to cline structure.

How well the Bikol percentages conform to tree structure, and are thus adequately described by a hierarchical classification of any sort, depends on the extent to which they satisfy a condition known as 'ultrametric inequality'. In simple terms, the three percentages among three varieties satisfy this constraint just in case the two lowest are equal, at least within the limits of statistical variation (for a mathematical discussion of ultrametric inequality and its relevance to hierarchical classification, cf. Johnson 1967: 245). As the Bikol percentages include both subsets which satisfy ultrametric inequality and those which do not, they provide an excellent basis for illustrating the significance of this constraint.

Figure 2 shows two sets of Bikol percentages which conform quite precisely to the constraint of ultrametric inequality and are thus well represented by the accompanying tree diagrams. In 2a, the tree shows the relatively closer relationship (85%) between Oas and Libon by means of its lower branching; the upper branching is then able to represent the remaining two relationships of either of these with Pandan just because these relationships are equal (both 64%). Note that if these two values are associated with their respective nodes in the tree, all three original data values may be recovered from the tree just because they meet the constraint of ultrametric inequality. An even simpler situation is illustrated in 2b, in which the equality of all three percentages requires only one three-way branching in the corresponding tree diagram

---

Insert figures 2 and 3 about here

Figure 3 shows two sets of Bikol percentages which deviate considerably from ultrametric inequality. In 3a, all percentages differ by at least ten percentage points and thus should certainly be regarded as significantly different. The accompanying tree is an approximation of these relationships in that it shows that Sorsogon and Masbate are more closely related to each other (at 79%) than either is to Oas. But as its highest node can properly represent only a single value, it fails to show that Oas shares a much higher percentage with Sorsogon (69%) than with Masbate (58%). In 3b, it is the two <u>higher</u> percentages which are equal. As no two of the varieties are more closely related to each other than either is to the third, their interrelationships are best approximated by a tree with a single three-way branching. This tree fails to show, however, that Gubat and Masbate are less closely related to each other (at 70%) than either is to Sorsogon (at 79%).

The best (if not the only feasible) way of determining the extent to which the percentages as a whole conform to a tree structure is to actually attempt to approximate them in terms of one. Figure 4 shows two such attempts. The tree in 4a was derived from the percentages by means of an averaging algorithm similar to that described by Lyen (1962) for use in comparative lexicostatistics (although no attempt was made here to combine nodes which might not be considered significantly different); the tree in 4b was derived from that in 4a in a manner described below. Both trees are accompanied by horizontal scales from which the 'ideal' (or 'fitted') percentage value of each node may be determined; these ideal values are simply the averages of the percentages represented by the corresponding nodes. Comparing these ideal values with the ranges of percentages which they are supposed to represent provides a means of judging how well these trees (which are among the "best possible") fit the original data and hence how well this data conforms to a tree structure.

Insert figure 4 about here

A comparison of the averaging method tree (4a) with the original percentages reveals several problem areas. While the leftmost node, for example, has an ideal percentage of 64%, it represents percentages ranging from a low of 55% (between Virac and Masbate) to a high of 73% (between Daraga and Sorsogon), this latter being in fact higher than the ideal percentages of two nodes to the right. Furthermore, this variation is systematic, with all of Masbate's relevant percentages falling below the average and nearly all of those of both Sorsogon and Gubat being above it. Other nodes involve similar, if not as extreme, variation: e.g. while Pandan and Virac share a percentage of 76%, they are connected in the tree by a node at 66%, and similarly while Legaspi and Daraga share a percentage of 83%, they are connected by a node at 72%.

An algorithmic method such as that used to produce the tree in 4a does not always produce the "best" tree for a set of data, however, particularly when the data is not especially tree-like. The problem areas mentioned above suggest a number of modifications which might be applied in order to improve this tree, or at least produce alternative trees which are not much worse. Some objective means of comparing how well different trees fit the data is needed, however; this is provided by the index of 'distance' between a tree and the data it represents proposed by Hartigan (1967:1141) in his approach to fitting trees to data. Specifically, the 'distance' between the original percentage p_{ij} and a tree with corresponding ideal percentages $\hat{p}_{ij}$ assigned to its nodes is measured by the "sum of squares" type index $\sum_{i>j} W_{ij}(p_{ij}-\hat{p}_{ij})^2$, where W_{ij} is simply a weighting factor here taken to be unity (i.e. i and j are index numbers for the varieties, and $i > j$ as the computation will involve only the lower half matrix as shown in Fig. 1). This index will be zero if the tree fits the data perfectly, otherwise it will be positive. The tree in 4a has a distance index of 0.12 (with each percentage regarded as a decimal fraction between zero and one); while its fit is far from perfect, it is much better than that of, say, a tree with a single twelve-way branching at the average (70%) of the entire set of Bikol percentages, which has a distance index of 0.42.

Three substantial modifications of the original tree (in 4a) ultimately transformed it into the tree shown in 4b; while only the fir . nese resulting in a tree with a lower distance index, the other two did not produce any especially great rise in this index (note that in each case averages of the percentages were recomputed in accordance with the new tree structure). First of all, in order to better depict the lowness of the bulk of Masbat's percentages (as noted above), its high perce tages with Sorsogon and Gubat were ignored and it was made an entirely separate branch joining the tree at 60%. This modification actually improved the fit slightly, descreasing the index of distance from 0.12 to 0.11. A second modification involved ignoring Virac's high percentages with Naga and Legaspi and grouping it with Pandan; this raised the index of distance to 0.13. A third modification, in which Daraga was detatched from Oas and regrouped with Naga-Legaspi, resulted in the final tree shown in 4b, with an index of distance of 0.14. Note that it was only the effects of recomputing the averages that caused the nodes involving Oas-Libon, Buhi, and Iriga to coalesce in 4b.

Obviously the Bikol percentages do not conform perfectly to a tree structure. While they may indeed be approximated in terms of a tree, all four trees discussed above (i.e. the two trees shown in Fig. 4 and the two intermediate stages) are well motivated by various aspects of the structure of the percentages. (In this regard, note that a poorly motivated tree, such as one which would group Naga with Masbate, cannot be produced while maintaining a monotonic relation between the nodes and their ideal (i.e. average) percentages, as has been required here). While the tree incorporating only the first modification is mathematically better (with a distance of 0.11) than the other three, it is up to the linguist to decide whether this is really significantly better for linguistic purpose than the other, highly different trees discussed, as this is not mathematically obvious from the latter's only slightly higher distances of 0.12 to 0.14. And in doing this, he must decided whether <u>any</u> of the trees actually provide a useful basis for historical (or other) interpretation.

Interpretation is in fact the key here. While tree structures may be viewed as representing the history of phylogenetic splits among languages, Bikol appears to constitute a single language yet undivided by such splits. If additional, intermediate varieties of Bikol were incorporated into the study, the adequacy of a tree approximation could be expected to grow even worse, even if such a hierarchical classification were modified to incorporate major nonhierarchical trends (as in the case of the modified tree diagrams of Southworth 1964 or the incorporation of certain relevant lexicostatistical percentages into such a classification by Dyen 1965). Here many linguists would be inclined to simply forego all but the most approximate classification in favor of describing the details of variation by means of an isogloss map. And yet, many linguists have recognized the relationship between lexicostatistical percentages and the "spatial alignment" of linguistic varieties (cf. Hymes 1960: 24-5). The hypothesis that Bikol is a single language, or more precisely a linguistic 'cline' (or language or dialect "chain", "cluster", or "continuum"), characterized by more or less continuous variation throughout a geographical area, leads to the expectation that the percentages should conform to a structure capable of being represented in two dimensional space.

In his first published application of lexicostatistics, Swadesh (1950: 164) noted that 'One of the advantages of a statistical valuation of linguistic distance is that is permits a multidimensional recognition of relations', and he thus proceeded to describe relationships within Salish not only in terms of a hierarchical classification, but also in terms of a spatial representation suggestive of 'approximate geographic relations in an earlier epoch.' His spatial representation, later adopted by some other linguists, is very rough: the varieties (often distinct languages) are placed in a two dimensional arrangement of boxes, with various devices (e.g. different types of lines) being used to show differences in degrees of relationship. Other linguists, notably Kroeber (1960), attempted to represent the different degrees of relationship more directly in terms of actual physical distance. As this latter is essentially what multidimensional scaling does, it is enlightening to consider the practical difficulties involved in attempting this by hand.

To produce a spatial representation of the Bikol relationships, it is convenient to first convert the percentages, which are measures of similarity, into dissimilarity measures which can be scaled and used as actual physical distances. There are many ways in which this might be done, but suppose that each percentage is simply subtracted from one hundred percent, and that each percentage point in the difference is interpreted as a distance of one tenth of an inch. As Sorsogon and Masbate, for example, have a common percentage of 79%, points corresponding to them might be placed 2.1 inches apart; their relationship to each other would thus be represented in one dimensional space (i.e. on a straight line). Oas might then be added to the picture by placing it 3.1 inches (corresponding to 69%) from Sorsogon and 4.2 inches (corresponding to 58%) from Masbate; these relationships would then be well represented in two dimensional space (i.e. as a triangle). To add a fourth variety is not so easy, however, because each of its percentages with the first three must be made to correspond to an actual physical distance. If this proved to be a physical impossibility (and it most surely would before very many of the twelve varieties had been added to the diagram), then there are two solutions. The first is simply to resort to the next higher dimensionality when necessary (e.g. the relationships among four varieties could be represented in the shape of some sort of irregular tetrahedron). But if the dimensionality is not restricted to some very small number of dimensions, then such a representation not only might prove to be difficult to visualize and interpret (twelve varieties could require as many as eleven dimensions), but it would also be highly trivial: distance measures based on the percentages in the manner described above may generally be assumed to satisfy the constraint of triangle inequality required for such a representation.[4] The other solution would be to "adjust" the distance measures based on the percentages so that they could all be represented in a two dimensional spatial configuration, which might be expected to have some reasonable interpretation. Unless there are clear criteria for making such "adjustments" and measuring the extent of the resultant deviation from the original data, however, it will not be clear to what extent the resulting spatial representation reflects constraints in the original data and to what extent it reflects the subjective judgement of the investigator. The difficulties involved in an attempt to produce any major, nontrivial spatial representation by hand are well illustrated by the comments of

Kroeber (1960: 3) on his carefully done spatial representation of the quantified phonological and morphological relationships among nine Indoeuropean groups:

> It does not try to plot all the coefficients, but only the higher ones for each language; and as I expected, even these could not all be accomodated in a two-dimensional diagram; though it so happens that all but Armenian do accord <u>reasonably</u> <u>well</u> with their <u>nearest</u> relatives in distances measured within one plane. (emphasis mine - PB)

These are the difficulties that may now be easily handled in a well defined way by use of the highly developed technique of multidimensional scaling. It is only necessary to put the data into one of the several generally available multidimensional scaling computer programs in order to produce a spatial representation with specific properties and a measured fit with the original data. The basic details of this technique are described in the following section, but the significance of the results when it is applied to the Bikol percentages are readily apparent from Fig. 5.

Insert figure 5 about here

Figure 5a shows the two dimensional spatial representation of the Bikol relationships based on the application of multidimensional scaling; each variety is represented by a point (marked by a cross) in a plane, with the interpoint distances being based on the original percentages in a quite specific, though very complex, way. The fit of this configuration to the percentages on which it is based has been measured and may be characterized as being quite good. In 5b, this configuration has been superimposed on a map showing the geographical distribution of the varieties according to McFarland (1972), and lines have been drawn to connect the scaling locations (the crosses) with the geographical locations (the dots). Considering that the spread of linguistic innovations is affected not only by geographical distance, but also by topographical features (here, both mountains and large expanses of water) and sociopolitical factors, the scaling configuration appears to match the geographical distribution markedly well (in terms of the standard (product-moment) correlation coefficient, in fact, their distances correlate at 0.77).[5] The two greatest shifts in position, namely those involving Naga and Pandan, do not appear unreasonable. Naga and Legaspi have moved closer together because they share the highest percentage (88%) of all pairs of varieties; not only do linguists consider them both 'coastal' dialects as opposed to the neighboring 'mountain' dialects, but their surprisingly great linguistic similarity is well known to Bikol speakers as well.[6] Pandan's shift southward would appear entirely reasonable if it should be the case that its speakers are in little direct communication with the mainland, but rather communicate indirectly by way of Virac. It should also be noted, however, that the scaling placement of peripheral points is less precise than that of more central points (e.g. moving Pandan to the north by an inch in the diagram would not affect the goodness of fit as much as moving Daraga in this same manner). In any case, it seems obvious that a multidimensional scaling of the Bikol percentages shows them to conform well to the structure of a cline, and results in an especially satisfying spatial representation of these relationships.

The Mechanics of Multidimensional Scaling

Multidimensional scaling is almost always done on computers, and the details of its application and interpretation are thus best illustrated in terms of a specific computer program. The one used in this paper is the KYST program of Kruskal, Young, and Seery (1973), which incorporates two earlier approaches, that of Shepard and Kruskal (as in Kruskal's M-D-SCAL program), and that of Torgerson and Young (as in Young's TORSCA program). The SSA ('Smallest Space Analysis') program of Guttman and Lingoes represents a third distinct approach along these lines (for a detailed comparision of a variety of such programs, cf. Green and Carmone 1970). The INDSCAL ('Individual Differences Scaling') program of Carroll and Chang represents a quite different approach which is especially useful for the analysis of data comprising the subjective opinions of difference of a variety of individuals; while it is quite useful in some linguistic applications (see, for example, the results obtained by Wish and Carroll (in press)), it has not yet been used to any particular advantage in lexicostatistics. All but the most specific remarks made here in reference to KYST may be extended to refer to these other programs (for further introductory material on multidimensional scaling, cf. Kruskal 1971 and Shepard 1972).

While multidimensional scaling is a highly versatile approach whose application may potentially involve data transformations and other complexities, only a very straightforward application of the technique was generally needed in order to produce satisfactory analyses of many sets of lexicostatistical percentages. Unless otherwise noted, in fact, all the scaling discussed below involved only the most rudimentary use of nonmetric multidimensional scaling, and were produced by supplying the KYST program with input consisting of the data plus a few control cards. As these cards both provide a precise definition of the approach used and also illustrate how easy it can be to apply multidimensional scaling by means of KYST, they are listed below on the left the explained briefly on the right:

DIMMAX=4	Maximum dimensionality requested is four.
DIMMIN=1	Minimum dimensionality requested is one.

REGRESSION=DESCENDING	Apply nonmetric scaling to similiarity measures.
LOWERHALFMATRIX	The form of the data is a lower half matrix,
DIAGONAL=ABSENT	without the diagonal.
DATA	The data deck begins here.
MCFARLAND'S BIKOL	(Title card)
12 1 1	The full matrix would have twelve rows and twelve columns.
(11F3.3)	(FORTRAN format describing the data cards, which are the next eleven cards and contain the percentages essentially as shown in figure 1)
COMPUTE	Compute this application.
STOP	There are no other applications in this run; stop.

Such input produced scalings of the Bikol percentages in four dimensionalities, ranging from high of four dimensions to a low of one. Additional cards could have been added to control various aspects of the computation and printing of the results (these are otherwise controlled by default values), for transforming the data in various ways, for weighting the data values in some appropriate manner (this would have especially valuable if the quality of the percentages varied considerably, perhaps because they were based on samples of different sizes), and so on; some of these possibilities are touched upon below. In addition, other applications could have been incorporated into the same computer run by adding their cards between the COMPUTE and STOP cards.

From this input, KYST produced several pages of output for each requested dimensionality; Figs. 6 to 8 show the printout for the two dimensional scaling of Bikol only. The first half of the first page (Fig. 6) describes the 'history of computation' and illustrates the basic working methods of KYST. Starting (at iteration zero) with an initial configuration (here based on final configuration obtained in the next higher dimensionality in order to save computing time), KYST proceeds to improve this configuration iteratively until no small change can improve it further within the (here preset) limits of precision desired. Specifically, an "improvement" is simply a change in the configuration which improves its fit with the original percentages, as measured by the index of 'stress' given in the second column. While the stress of the initial configuration was 0.088 (or 8.8 percent), by the sixth iteration it has been reduced (an thus improved) to 0.069, and the improvements made in subsequent interations are so fine that they are not reflected even in the third decimal place of stress. The coordinates of the final configuration are given further down this first page, with both the letters A through L and the index numbers one through twelve corresponding to the individual Bikol varieties according to their order in the input data (i.e. the same order shown in Fig. 1).

The scatter diagram on the second page of output (Fig. 7) illustrates in what way and just how well the final configuration corresponds to the original percentages, and also provides a basis for describing the index of stress used to measure this fit. The horizontal axis of this plot represents the original percentages, which range from 55% to nearly 90%. While the vertical axis represents the distances corresponding to these percentages, there are actually two types of distances involved. Those marked by the D's in the plot represent the actual distances between points in the final configuration, while those marked by dashes are 'ideal distances', i.e. values for distance which would match the percentages precisely according to the constraint used in the scaling, but cannot actually be realized as physical distances in (in this case) two dimensions. The extent to which the actual distances deviate from the ideal distances for the various percentages (i.e. as measured vertically in the plot) provides a measure of poorness of fit between the scaling and the percentages, and the index of stress is simply a "sum of squares" measure of this deviation not unlike the index of 'distance' discussed above in connection with tree diagrams.[7] On the scatter diagram, relatively great deviation (and thus higher stress) would appear as a spread of the D's away from the dashes, while for a good fit (and low stress), most of the D's would be fairly close to the dashes (and, in the printer plot in Fig. 7, in fact, many of the D's could not be printed at all because they were too close to the dashes).

Insert figures 6, 7, and 8 about here

There are a variety of ways in which the relation between the ideal distances and the original percentages could have been constrained. Unless otherwise noted, all applications discussed in this paper involve the fairly simple constraint associated with 'nonmetric' scaling: all that is required is that the ideal distances be in a monotonic relation to the original percentages (or more precisely, an monotonic decreasing relation, since percentages are similarities and distances are dissimilarities). All this means is that higher percentages have to be represented smaller ideal distances, and thus the dashes representing these distances in Fig. 7 fall irregularly, but never rise, from left to right. The ideal distances are in fact simply values with both satisfy this constraint and at the same time are closest (as measured by stress) to the actual distances used to represent the percentages in physical space (for a theoretical treatment of nonmetric scaling, cf. Kruskal 1964a). It would, of course, be pleasing to have a stricter functional relationship between the percentages and the scaling distances; by using nonmetric scaling, however, it is not only unnecessary to postulate such a specific relationship, but the rough curve of dashes in Fig. 7 provides some basis for deciding what sort of function might prove appropriate. On the basis of this curve and a little experimentation, it was found that the formula $\underline{d} = (-\ln \underline{p})^{1.5}$ provided a good approximation of the relationship between distance $\underline{d}$ and percentage $\underline{p}$ (the latter expressed as decimal fractions between zero and one) for Bikol and several other groups.[8] To do a metric scaling involving this strict functional relationship required only the addition of a few cards to the input in order to transform the percentages appropriately and the change of the REGRESSION=DESCENDING card to REGRESSION=POLYNOMIAL=1 in order perform a linear regression on the transformed percentages; this linear relationship between ideal distance and transformed percentage may be seen in the plot of distance versus transformed percentage for Bikol in Fig. 9 (note that the dashes rise in a straight line from left to right). Note, however, that aside from the fact that this scaling had somewhat greater stress than the nonmetric one (0.103 as opposed to 0.069) due to the greater constraints, its results were otherwise virtually identical; in particular, it produced a final configuration so similar to that produced nonmetrically that it could not be distinguished by visual inspection. While the investigation of the functional relationship between distance and percentage is an interesting area, the study of cline structure by the use of nonmetric scaling does not depend on such information.

Insert figure 9 about here

The last page of printout for the two dimensional scaling of Bikol (Fig. 8) is simply a printer plot of the final configuration whose coordinates were listed at the bottom of the first page (Fig. 6). Not only are the units given along each axis somewhat arbitrary measures of distance (although chosen for mathematical convenience), but the orientation of the configuration is also arbitrary; i.e. the percentages alone provide no basis for determining what should be north, south, east, and west. A comparison with the actual geography suggests that right is approximately north and up is approximately east, and thus to get the plot shown in Fig. 5 it was necessary to both rotate that shown in Fig. 8 by ninty degrees counterclockwise and reflect it on its new vertical axis, and then to adjust its scale to fit the map.

The computation and presentation of the configurations for each of the four dimensionalities requested in the case of Bikol is similar to that discussed above (although the three and four dimensional configurations are each presented in a series of two dimensional plots showing various pairs of axes at a time). While the two dimensional scaling has been said to be the most appropriate one for Bikol, in general choosing the most appropriate dimensionality involves a number of considerations. First of all, lower dimensionalities are preferable to higher ones because they are more parsimonious and easier to use, as noted in the preceding section. As a rough guide, an n dimensional representation should involve something more than 4n varieties for adequate accuracy, at least in the lower dimensionalities under consideration here. But stress tends to increase as the number of dimensions decreases, and it is also important that stress remain relatively low. As another rough guide, stress in the 0.05 to 0.10 range tends to be indicative of reasonably good fit, with lower values indicating even better fit and higher values worse. Note also that in this search for the lowest dimensionality with reasonably good stress, it will sometimes be found that stress jumps greatly from one dimensionality to the next lower one; this is particularly good indication that the higher dimensionality is the appropriate one. The stress for the Bikol scalings, for example, rose from about 0.02 in four dimensions to 0.04 in three, to 0.07 in two, and then jumped by a factor of more than three to 0.25 in one dimension, thus suggesting that the two dimensional configuration was clearly appropriate (as was confirmed by all other indications). This jump in stress may be seen in Fig. 10, which is a plot of dimensionality versus stress for most of the scalings discussed in this paper, and which also shows the average stress curves for random data involving twelve and sixteen varieties in order to provide an idea of what really poor stress would look like (these are from Klahr 1969). But perhaps the most decisive factor in determining the most appropriate dimensionality is the existence of a reasonable interpretation for the configuration in this dimensionality. This has already been demonstrated for the two dimensional scaling of Bikol, and the linguistic basis for this interpretation suggests that in general a good fitting two dimensional configuration would be highly desirable in the investigation of cline relationships (although it is not clear that some higher dimensionality might not also prove appropriate in some cases).

Insert figure 10 about here

Two additional examples round out this elementary presentation of the basics of multidimensional scaling. The first involves the application of this technique to lexicostatistical percentages among twelve varieties of Konsoid, a Lowland East Cushitic cline spoken in southwestern Ethiopia. These percentages were calculated by Black (in press) on the basis of a nonstandard, 141 item lexicostatistical list. At first glance it is tempting to view the two dimensional scaling (Fig. 11a) as ideal, since it not only has extremely low stress (0.026), but it also agrees quite well with the geographical distribution of the Konsoid varieties (shown in 11b with the scaling superimposed).[9] A closer look at the configuration in 11a suggests, however, that the east-west variation is relatively small, and that it is along this dimension that the scaling least adequately reflects the geographical distribution (e.g. the positions of Fasha and Kolme are reversed in this dimension). This suggests that a one dimensional representation may prove nearly as adequate as a two dimensional one, and to be sure, stress is still quite good in one dimension, where it has only slightly more than doubled to become about 0.05 (cf. Fig. 10).

Finding the best one dimensional scaling of Konsoid, however, was complicated somewhat by the susceptibility of KYST to the problem of 'local minima' in one dimensional space (for this and other computational problems, cf. Kruskal 1964b). A local minimum is a value for stress which cannot be decreased by any small change in the configuration, but which may be decreased by some major change; it is as if stress is at the bottom of a "valley," but there is some yet lower "valley" located somewhere over the "hills" of stress. Figure 12a shows the first one dimensional scaling produced by KYST for Konsoid; while this has fairly low stress (0.056), it does not look quite right in comparison with the original percentages because the varieties of Bussa and Gidole are not grouped as would be expected, but are rather interspersed. Further investigation demonstrated that this was in fact a local minimum: a new one dimensional scaling (i.e. 12b) was produced by providing a new starting configuration as data input and proved to have slightly lower stress (0.055). While the difference in stress is slight, there is a clear reason, based on interpretation, for prefering this second, presumably mathematically optimal solution. This problem of local minima thus must be kept in mind, although for KYST is seldom proves to be a serious problem in dimensionalities of two or higher. A metric one dimensional scaling of Konsoid (cf. 12c) incidentally avoided this problem of local minima; note how it forced all three varieties of Gidole, which share the highest percentages in the set, to be represented by a single point (the higher stress of 0.118 is simply the result of the greater constraints involved in metric scaling, which in this case used the simple function $\underline{d} = 1-\underline{p}$).

Insert figures 11 and 12 about here if possible on the same or adjacent pages side by side

---

The application of multidimensional scaling to Lower Niger serves to illustrate the fact that a scaling can be no better than the data on which it is based. Lower Niger is spoken throughout the East Central State of Nigeria and in adjacent areas to the west and southwest; some of its varieties (especially Onitsha, Orlu, Owerri, and Ọhụhụ) constitute the core of the well known Igbo language. Williamson (1973) calculated lexicostatistical percentages among seventeen varieties of Lower Niger on the basis of a one hundred item list of meanings. The percentages among sixteen of these varieties are suggestive of a cline structure (in any case, a nonhierarchical structure), but the seventeenth (Ẹkpẹyẹ) does not appear to participate in this cline (it has fairly constant percentages of 62% to 69% with the others) and thus it has not been included in the scaling. Figure 13 shows a two dimensional scaling of the sixteen varieties (marked by crosses) overlaid on a map of their geographical distribution,[10] with lines drawn to facilitate comparison. The correlation between the scaling and the map appears to be fairly mediocre and the deviations do not appear to be open to obvious linguistic explanations: Enuanị and Ụkwụanị have switched places, and Ogbah, Echie, and Ọhụhụ also show considerable deviation. Nor is the two dimensional stress of 0.12 particularly good (the scatter diagram in Fig. 14 provides a visual impression of the extent of deviation); stress tends to be a bit high in all dimensionalities, in fact (cf. Fig. 10). These poorer results may largely be due to the fairly great amount of statistical variation (for the one hundred item list, a five percent level significant difference is around fourteen percentage points; cf. footnote 3) relative to the limited range of the percentages (which range roughly between sixty and ninety percent).[11]

Insert figures 13 and 14 about here

The Persistence of Cline Structure Within Trees

The preceding examples all involve single linguistic clines, which may well be expected to have structures of relationships well depicted by means of spatial representations. With the passage of time, of course, clines may break up into distinct and mutually unintelligible languages which, in the course of their subsequent independent evolutions, begin to manifest relationships which gradually become more and more tree-like in structure. And yet, cline-like relationships are known to persist for considerable lengths of time after the tree-like relationships become well established. The best studied case of this sort involves, of course, the problem of deriving a subgrouping of the branches of Indoeuropean : while linguists using the traditional qualitative method of subgrouping according to the criterion of shared innovations have been unable to reduce the highest node of Indoeuropean to anything less than about a ten-way split, yet they have produced evidence that the relationships among these branches are not all of the same degree, but rather appear to reflect the structure of a Proto-Indoeuropean cline which began to divide into independent languages several thousand years ago (cf. e.g. Anttila 1972:304-6). Using quite a different approach, Sankoff and Sankoff (in press) similarly provide evidence that neither a pure tree model nor a pure wave model of linguistic change is fully adequate for accounting for the lexicostatistical relationships among varieties of five clearly delimited Malayo-Polynesian groups of New Guinea.

To illustrate how cline-like relationships might be preserved within the evolutionary tree, Bloomfield (1933:317-8) provides a hypothetical example:

> ...let us suppose that among a series of adjacent dialects, which, to consider only one dimension, we shall designate as A, B, C, D, E, F, G, ... X, one dialect, say F, gains a political, commercial, or other predominance of some sort, so that its neighbors in either direction, first E and G, then D and H, and then even C and I, J, K, give up their peculiarities and come to speak only the central dialect F. When this has happened, F borders on B and L, dialects from which it differs sharply enough to produce clear-cut language boundaries; yet the resemblance between F and B will be greater than between F and A, and,

similarly, among L, M, N, ... X, the dialects nearest to F will show a greater resemblance to F, in spite of the clearly marked boundary, than will the more distant dialects.

Thus, if dialects A and B constitute a single language, F a second language, and dialects L through X a third, all dialects of a single language need not be related in equal degrees to any dialect of some other language. The same principle holds for distinct languages as well: language F, for example, might well be considerably more similar to each of the two remaining languages than either is to each other, forming a clearly nonhierarchical relationship perhaps similar to that pictured in Fig. 3b. While it is appropriate to depict the history of linguistic splits between languages by means of a tree diagram, in cases such as this there are factors which gravely interfere with correctly inferring this history. In addition, a satisfactory analysis of contemporary relationships depends on a proper sampling of the varieties involved; if, for example, dialects L and X were selected as samples of the third language and the intervening dialects were ignored, L and X might well appear as dissimilar to each other as either is to, say, F, thus making it appear as if there were <u>four</u> distinct languages (Black (in press) describes an actual occurrence of this problem). And if the intervening dialects had simply died out so that there <u>were</u> in fact four languages, there might be no way to determine when this happened and thus no way to date this actual linguistic split relative to many other splits in the tree. These problems are well illustrated by a consideration of the relationships among the Salish languages.

Salish is a group of at least twenty-six distinct American Indian languages once spoken throughout an area which now includes nearly the entire state of Washington, much of Idaho, and adjacent parts of Oregon, Montana, and British Columbia (Swadesh 1952:232). In his first published application of lexicostatistics, Swadesh (1950) calculated lexicostatistical percentages among thirty varieties of Salish on the basis of lexical samples of 165 items each; aside from those pairs of varieties which Swadesh considered to be dialects of the same language, very few pairs share percentages higher than 60%, and the bulk of the percentages lie in the ten to fifty percent range. As noted earlier, Swadesh used these percentages[12] to derive both a hierarchical classification of the varieties and an informal spatial representation of their relationships. Later, Dyen (1962) arrived at a somewhat different hierarchical classification, based on these same percentages, in the course of illustrating a procedure for lexicostatistically based classification. Even more recently, Elmendorf (1969) used an informally derived spatial configuration in his discussion about the classification of a certain subset of these varieties which he characterized as forming a "chain-relationship series." The problems involved in the analysis of the relationships within Salish thus appear to be of the sort to merit investigation by means c multi-dimensional scaling.

Insert figure 15 about here

Figure 15 shows the classification of Salish according to Swadesh (1950: 163-4); it also describes Dyen's classification where it is different and gives the abbreviations for the names of the varieties as they appear in the scaling configurations (the single letters and numbers) and as used by both Swadesh and Dyen and on the maps presented here (the two-letter combinations). It may be useful to note that different divisions (marked by Roman numerals) share percentages lower than about 20%, different branches (capital letters) share those lower than about 40%, different groups (Arabic numerals) share those lower than 60%, different languages (lower case letters) share those lower than 80%, and dialects of the same language (names connected by hyphens) share percentages above 80%. The classification of Dyen (1962: 160) differs in three ways. First, it makes the Lkungen group a separate branch, coordinate with Swadesh's other branches. Second, it divides Swadesh's Olympic Branch into two branches (the 'Satsop' and 'Lower Chehalis' branches) coordinate with these same other branches. And third, it further groups the members of Swadesh's Interior Division into two branches, a 'Lillooet Branch' containing Lillooet and the Thompson Group, and a 'Columbia Branch' containing the Okanagon Group, Columbia, and Cöeur d'Alène. Like Swadesh, Elmendorf (1960) also recognizes the unity of the Olympic Branch, but in addition further groups Lower Chehalis and Quinault together within it as a 'western' group (= Dyen's Lower Chehalis Branch) as opposed to an 'eastern' group (= Dyen's and Swadesh's Satsop Branch/Group).

A multidimensional scaling of all thirty varieties, as shown in Fig. 16, does nothing to resolve these points. The four primary divisions of Salish do indeed appear well motivated, and the two dimensional scaling does little more than to divide the varieties among these four divisions (which have been delimited and labelled by hand on this printer plot). To be sure, the varieties within Coast and Interior Salish have been arranged in a manner in accord with their classification by either Dyen or Swadesh (i.e. varieties belonging to the same subgroup are usually adjacent), but the fairly mediocre stress of 0.12 suggests that the finer relationships which are at the heart of the issue cannot be expected to be represented with sufficient precision to resolve any problems. The state of affairs becomes a bit clearer in the three dimensional configuration (not shown), which places the four main divisions in a tetrahedral arrangement; while the stress is three dimensions (0.08) may suggest fairly good fit, the fit appears to be improved most in that area of least interest, namely in the essentially hierarchical relationship between the four primary divisions of Salish. While the finer relationships become more faithfully represented in four and five dimensions, yet the representation of

---

Insert figure 16 about here

---

this primary hierarchical split remains a factor which interferes with an evaluation of the finer relationships. This illustrates an important fact about the use of multidimensional scaling: when it is used to investigate potentially cline-like relationships, hierarchical relationships clearly should be pruned from the analysis as much as possible. In this case, it seems appropriate to undertake the scaling of the Coast and Interior Divisions alone and separately.

To dispose of the simpler, and thus somewhat less interesting, case first, a two dimensional scaling of Interior Salish is shown in Fig. 17a. At first glance, the results may appear especially satisfying: not only is the stress exceptionally low (0.01), but the scaling can also be made to fit the map showing the geographical distribution of the relevant varieties extremely well (cf. 17b; this map also shows the locations of the remaining three divisions of Salish; both it and the map in Fig. 20 are based on Swadesh 1952: 234). But this is somewhat deceptive: this group contains only nine varieties, and three of these are so similar that they have been placed at a single point (labelled Sp-Ka-Pe) in the scaling. The low stress may thus be greatly attributed to the triviality of the scaling, and the good geographical fit to the fact that there are relatively few points to fit into relatively large geographical areas. Nevertheless this scaling is pleasing in that it agrees with what has already been established by other means: the Thompson and Okaganon groupings of varieties are both quite clear, and the configuration is, by the way, very similar to the informal spatial representation proposed by Swadesh (1950: 165). In addition, it also suggests that Dyen's division of Interior Salish into two groups is not particularly appropriate (it would imply that there should be more space between the Thompson (Th and Sh) and Okaganon (Ok and Sp-Ka-Pe) varieties). But even an examination of the original percentages suggests as much (e.g. the difference between the Lillooet-Thompson percentage of 50% and the Shuswap-Okaganon percentage of 57%

would appear to be significant only at about the 20% level; cf. footnote 3). Even though the scaling configuration is fairly trivial, however, it would appear to be a better motivated representation of the relationships than the hierarchical classification, which fails to make the geographical nature of the variation clear.

Insert figure 17 about here

As Coast Salish contains some nineteen varieties in seventeen distinct languages, it provides a somewhat meatier data base for the application of multidimensional scaling. That multidimensional scaling is indeed applicable here is suggested by the nature of Dyen's average percentages between his seven branches of Coast Salish; these are shown in Fig. 18 as rounded to the nearest whole percentage. If it were the case that these seven branches bore a hierarchical relationship to each other, then there should be large blocks of percentages which should be approximately the same; if Dyen's seven-way branching were in strict conformity to tree structure, then indeed <u>all</u> the percentages should be about the same. And yet the percentages range from less than 20% to more than 40% in a pattern of gradual variation which is certainly suggestive of cline structure. As all branches of Coast Salish appear to participate in this cline, it seems appropriate to apply multidimensional scaling to the entire set of nineteen varieties.

Insert figure 18 about here

A two dimensional scaling of the Coast Salish varieties proves not only to help resolve questions arising from the difference between the hierarchical classifications of Swadesh and Dyen, but also to present a picture of the cline-like relationships among the branches which is obvious from neither of these classifications. The two dimensional scaling has a reasonably low stress of 0.088, and thus appears to adequately reflect at least the larger relationships relevant here, if not the finer details of the relationships within the Coast Salish branches. The two dimensional configuration does, however, appear to suffer slightly from a certain common but extraneous effect, and in order to display this effect, hand drawn lines have been added to the raw computer printout of this configuration shown in Fig. 19. Specifically, each pair of varieties sharing a percentage of 30% or more is connected by a line. These lines demonstrate that the speech varieties essentially form a long, thin cline which for some reason insists on bending itself around into a "horseshoe." It is well known that multidimensional scaling often bends essentially linear relationships around into a horseshoe (e.g. cf. Kendall 1971), and in this case such bending could easily result from relevant differences in the percentages between more remote pairs of varieties being obscured by the effects of statistical variation.[13] Certain procedures intended to eliminate much of the "horseshoe" effect were tried, and did flatten the configuration considerably, in effect shrinking distances along the vertical axis to little more than half what they are in Fig. 19.[14] None of these procedures really eliminated the horseshoe, however, so it is possible (though doubtful) that the bending of the configuration may reflect a real aspect of the situation. In any case, the original configuration, coupled with this qualification of its nature, serves as well as a basis for discussion here as any configuration derived through more complicated techniques would.

Insert figure 19 about here

Figure 20 compares the geographical distribution (20a) of Dyen's seven Coast Salish branches with their positions according to the two dimensional scaling (20b; this differs from that shown in Fig. 19 only in scale and orientation). The largely north and south geographical alignment of all but the three southern groups does suggest that the bending observed in the scaling should be largely spurious, although it is also possible that the relationships among the northern three groups have been affected by the existence of the strait between Vancouver Island and the mainland as a potential route for contact between nonadjacent groups. With regard for the differences between the proposed classifications, it seems obvious from the scaling that Swadesh was well motivated in placing the Lkungen Group within the South Georgia Branch; in fact, the two branches proposed by Dyen appear so close together in the configuration that it was necessary to draw a dashed line between them in order to distinguish them. There is also some evidence suggesting that the Satsop and Lower Chehalis groupings are also appropriately grouped by Swadesh in his Olympic Branch (in the scaling, the two groups are visibly closer to each other than the Satsop grouping is to Twana of the Hood Canal Branch); it also seems clear, however, that Olympic should have two subgroups, as proposed by Elmendorf (1969) along the line of Dyen's distinction, rather than the three proposed by Swadesh. Note, however, that the discussion of hierarchical classification becomes somewhat academic at this point: the scaling certainly shows not only the clear major divisions at this level, and shows the less clear divisions as being less clear, but it also shows the chain-like nature of the relationships quite clearly.

One major virtue for the scaling representation, as opposed to a hierarchical subgrouping, of attested speech varieties in some situations is its relative insensitivity to other speech varieties which are not attested, due either to their having died out or to the inevitable imperfections in the data collection process. Suppose, for example, that a new variety of Coast Salish were discovered and found to occupy a position halfway between Twana (Tw) and the Satsop grouping. This would require only the addition of a point to the scaling, but would imply a major change in the hierarchical classification (it would suggest that all three southern groupings formed a single branch). Similarly, suppose hypothetically that a group which had until recently formed a link between the North and South Georgia groups had just recently become extinct. Then a hierarchical subgrouping of the attested varieties would suggest that the linguistic split between the two groups had occurred considerably earlier than it actually had. In the scaling, on the other hand, the gap between the two groups does not rule out the possibility of their having been "connected" by such "missing links". The Coast Salish configuration is thus a "fossilized skeleton" of a cline structure which existed sometime in the past. While the "flesh" of this structure may never be recovered, its general outline may be inferred from the shape of the configuration. (Note that the sensitivity of hierarchical subgrouping to unattested speech varieties arises because the subgrouping does not reflect the true situation. If scaling is used where it does not reflect the true situation, then it too is sensitive to unattested speech varieties.)

Insert figure 20 about here

While the Coast Salish scaling draws the larger relationships into focus, it undoubtedly leaves many of the finer relationships somewhat blurred; the stress of 0.088 is satisfactory in terms of the overall picture, but hardly low enough to suggest that the finest differences are reflected with precision. Within the South Georgia (including Lkungen) Branch, for example, the grouping of such pairs as Fraser-Nanaimo and Lkungen-Lummi is somewhat more obvious from the original percentages than from the scaling configuration. The fact that the configuration of this group does not agree closely with the geographical location is a less reliable indication, however: the fact that pairs of closely related varieties (and in the North Georgia Branch, the Comox language along) are divided between Vancouver Island

and the mainland suggests that there may have been relatively recent migrations. Note also that much of the intragroup variation may well have arisen long after the various branches became distinct languages, and may thus be largely unrelated to the larger cline relationship which dominates the scaling.

The primary purpose of this paper is to demonstrate the usefulness of multidimensional scaling in the investigation of nonhierarchical linguistic relationships. Hopefully it also suggests new avenues of research which might profitably be explored by linguists, statisticians, and perhaps also scholars of such other relevant disciplines as sociology and anthropology. The above examples demonstrate that a model of lexical change which does not take into account spatial relationships is surely a gross approximation of reality; in this regard, the 'divergence with interaction' model of Sankoff (1972) represents progress toward a more satisfactory hypothesis. From the point of view of data analysis, obviously neither hierarchical clustering nor multidimensional scaling alone is fully adequate to produce a linguistically appropriate picture of the relationships, and a more general, integrated, and yet easily applied approach would be a great boon to linguists and scholars in other disciplines faced with similar complex combinations of hierarchical and spatial variation (here some limited progress has been made by Degermann 1970). Another area of research is suggested by the possibility of correlating linguistic distance with a combination of geographical distance and topographical and sociopolitical factors. Such avenues of research may eventually lead more refined methods of inferring the course of prehistorical linguistic development and also such associated nonlinguistic phenomena as patterns of prehistorical contact and migration.

FOOTNOTES

[1]I am very grateful to both of these men for their helpful comments on portions of various drafts of this paper, as well as for their more general support and encouragement which has continued during the past several years. In fact, this paper literally could not have been written without the strong mathematical guidance of Joseph B. Kruskal, who also went to considerably trouble to advise me on how a great many details might best be presented. Needless to say, however, I must claim full responsibility for errors and omissions.

[2]I am also very grateful to the various scholars who have generously provided me with their unpublished data and various supplementary information which contributed to my pursuit of this research. These include Curt McFarland of Yale and Kay Williamson of the University of Ibadan, whose data are incorporated into this paper, as well as Patrick Bennett, Nancy Thayer, Shigeru Tsuchida, and Ralph Williams.

[3]This estimate is an approximation based on the assumption of statistical independence. It is easily shown, however, that the three percentages among three varieties are clearly not statistically independent in general. E.g. if each lexical sample contains only a single word per meaning, and if varieties A and B share eighty cognates out of a hundred and varieties A and C share forty, then varieties

B and C must share between twenty and sixty cognates because of the transitivity of the relation of cognation.

It is nevertheless convenient to make a number of simplifying assumptions, including the one of statistical independence, in order to provide a simple characterization of the extent of statistical variation inherent in the percentages. The following table permits a rapid estimate of the percentage point difference required for two percentages to be significantly different at several levels of confidence and for lexical samples of various sizes. Specifically, this percentage point difference is $\frac{C}{\sqrt{n}}$, where *n* is the number of items in the sample and C is a constant as given in the following table according to various levels of confidence:

Confidence level:	20%	10%	5%	2%	1%
C (in percentage points):	91	116	139	164	182

Thus, for a two hundred word sample, two percentages are significantly different at the five percent level (i.e. roughly five percent of the time) if they differ by more than about $\frac{139}{\sqrt{200}}$, or about ten, percentage points, which agrees with the Chi-square estimate of Dyen (1962: 153). This table provides rough, but theoretically reasonable, approximations for percentages *p* (expressed as decimal fractions of unity) and list size *n* if both $np > 5$ and $n(1-p) > 5$.

[4]Triangle inequality is simply the constraint that the linguistic distance (e.g. one hundred minus the percentage) between two varieties be no greater than the sum of their

distances to a third. If each lexical sample contains only one word for each meaning, this constraint will be automatically satisfied for reasons discussed in footnote 3. It may not be satisfied by this particular transformation of percentage into distance if lexical samples frequently contain more than one word per meaning, thus permitting situations to arise where e.g. pairs A-B and A-C could each share 95% cognates and pair B-C shares only 80%. In practice, however, such cases rarely result in any great deviation from triangle inequality, and such deviation could in any case be lessened or removed by some suitable nonlinear transformation of percentage into distance.

[5]The standard (product-moment) correlation coefficient is convenient simply because it is a common and familiar index of correlation. In theory, however, there is no reason to expect anywhere near a perfect correlation between linguistic and geographical distance. On the other hand, one might well expect the relation between the two configurations to be systematic in some way, so that one might appear to be a fairly regular distortion of the other.

It may be noted, for example, that the scaling configuration might be made to fit the map in 5b if some parts of it could be stretched while other parts were shrunk, and if it could be bent a bit as well. Thus a measure of "smoothness" of "continuity" between the two configurations, such as that suggested by She_ard and Carroll (1966), might be a more appropriate measure of their fit, although were such a measure presented here, there would be nothing to compare it to (other maps presented in this paper are far less precise with regard to geographical location). Another approach would involve treating distances over different types of terrain (e.g. land versus water) differently in terms of their correlation with linguistic distance (I am grateful to William Boyce for this and other suggestions).

[6]According to my wife, who is a Bikol speaker from Polangui, near Oas.

[7]Specifically, the index of stress is $\sqrt{\dfrac{\sum(d_{ij}-\hat{d}_{ij})^2}{\sum d_{ij}^2}}$,

where d_{ij} is the actual distance between varieties _i_ and _j_ in the configuration, and $\hat{d}_{ij}$ is the corresponding ideal distance. The same formula could have been used to measure the fit of the trees in the preceding section, and in fact the four trees proposed for Bikol would then have values for stress ranging from 0.13 to 0.15. The values for stress or other measure of fit are not directly comparable between trees and scalings, however, because the two types of structure involve different numbers of degrees of freedom.

[8]Note that a similar transformation might have been used to improve the fit of the trees discussed earlier for Bikol. The argument against tree structure thus did not primarily involve a demonstration that no tree fit the data adequately according to some measure of fit, but rather a demonstration that _too many_ trees fit the data about equally well, and no well motivated monotonic transformation is likely to change this.

[9]This map is based on aerial photographs supplied by the Imperial Ethiopian Mapping and Geographical Institute. While these permitted the centers of populations to be located reasonably well and in an uniform scale, the boundaries between populations are based largely on rough, verbal information. Note especially that many of these groups use various parts of the "uninhabited valley" to the east for farming and hunting, although there are few reasonably permanent settlements established in that area.

[10]This map is based on a more detailed one generously provided by Kay Williamson (personal communication).

[11]In addition, it appears that several of the lists incorporated more than one word for a number of meanings Note, for example, that even though Owerri and Ọh̃ụh̃ụ share 97% cognates, their percentages with Echie differ by twelve percentage points, i.e. 87% versus 75% . Possibly this also contributed to the relative poorness of the scaling results.

[12]Actually in the form of estimates of relative time interval $i = \frac{\log C}{2 \log r}$, where C is the percentage of cognates and r is a replacement rate taken to be 85% (Swadesh 1950: 158-61).

[13]In the case discussed by Kendall (1971), time was the only relevant dimension and thus a one dimensional configuration was clearly appropriate. This case involved the temporal seriation of the Münsinger-Rain grave sites

on the basis of the extent to which they shared certain attributes. After a certain length of time, two sites would no longer share any of the attributes, and of course sites more remote in time could share no less than this. The fact that temporally more remote pairs of sites appeared no more dissimilar than ones considerably less remote caused the relationships to appear as a one dimensional manifold bent to occupy a two dimensional space.

[14]In an attempt to straighten out the Coast Salish "horseshoe," nonmetric scaling was applied to three different matrices derived from the percentages. None of these succeeded, although the resultant configurations did have different stress values and in two cases quite different ratios of "length" versus "width" and so will be described in these terms. Aside from this, they were so similar to the original configuration that any of them could be used in support of the discussion in the body of the paper.

The original configuration has a "length" (maximum dimension) of about 1.3 times the "width" (minimum dimension) and a stress of 0.09 (versus 0.17 in one dimension).

The commonly used procedure of scaling the "$\hat{D}$-matrix," with cells $\underline{i}$, $\underline{j}$ computed as $\sum_h (p_{ih}-p_{jh})^2$ for the percentages $\underline{p}$, changed the ratio of the dimensions very little, but was very clearly two dimensional, with a stress of 0.05 (versus 0.20 in one dimension). (I am grateful to

J. Douglas Carrcll for unpublished information on the nature of the D^2-matrix.)

A procedure which ignored distinctions among percentages smaller that 25% produced a length to width ratio of 2.0 and had a stress of 0.04 (versus 0.07 in one dimension. (Technically, this was accomplished by replacing percentages smaller than 25% with a nominal value of 25%, and using the PRIMARY control statement to insure the "primary approach" to the ties that this created. The primary approach allows equal percentages to correspond to different ideal distances without penalty.)

The approach of Kendall (1971) produced the greatest length to width ration, namely 2.6, and had a stress of 0.04 (versus 0.13 in one dimension). This approach involved applying scaling to the "S ∘ S" matrix of cells $\underline{i}$, $\underline{j}$ computed as $\sum_h \min(p_{ih}, p_{jh})$.

BIBLIOGRAPHY

Anttila, R. 1972. An introduction to historical and comparative linguistics. New York: Macmillan.

Black, P. (in press) Konsoid: an example of extreme dialectal differentiation. To appear in the Proceedings of the Conference on African Linguistics held at Queens College on April 7-8, 1973.

Bloomfield, L. 1933. Language. New York: Holt.

Degerman, R. 1970. Multidimensional analysis of complex structure: mixtures of class and quantitative variation. Psychometrika 35.475-91.

Dyen, I. 1962. The lexicostatistically determined relationship of a language group. International Journal of American Linguistics 28.153-61.

Dyen, I. 1965. A lexicostatistical classification of the Austronesian languages. International Journal of American Linguistics, Memoir 19. Bloomington: Indiana University.

Elmendorf, W. W. 1969. Geographic ordering, subgrouping, and Olympic Salish. International Journal of American Linguistics 35.220-25.

Green, P. and F. Carmone. 1970. Multidimensional scaling. Boston: Allyn and Bacon.

Hodson, F. R., D. G. Kendall, and P. Tăutu. 1971. Mathematics in the archaeological and historical sciences. Edinburgh: University Press.

Hartigan, J. A. 1967. Representation of similarity matrices by trees. Journal of the American Statistical Association 62.1140-58.

Henrici, A. (in press) Numerical classification of Bantu languages. To appear in African Language Studies.

Hockett, C. F. 1958. A course in modern linguistics. New York: Macmillan.

Hymes, D. H. 1960. Lexicostatistics so far. Current Anthropology 1.3-44.

Johnson, S. C. 1967. Hierarchical clustering schemes. Psychometrika 32.241-54.

Kendall, D. G. 1971. Seriation from abunaance matrices. In Hodson, Kendall, and Tăutu 1971: 215-52.

Kirk, J. and P. J. Epling. 1972. The dispersal of the Polynesian peoples: Explorations in phylogenetic inference from the analysis of taxonomy. Chapel Hill: Institute for Research in Social Science.

Kirk, J. and P. J. Epling. 1973. Taxonomy of the Polynesian languages. Anthropological Linguistics 15.42-70.

Klahr, D. 1969. ›nte Carlo investigation of statistical significance _ Kruskal's nonmetric scaling procedure. Psychometrika 34.319-33.

Kroeber, A. L. 1960. Statistics, Indo-Eurpoean, and taxonomy. Language 36.1-21.

Kruskal, J. B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29.1-27.

Kruskal, J. B. 1964b. Nonmetric multidimensional scaling: A numerical method. Psychometrika 29.28-42.

Kruskal, J. B. 1971. Multidimensional scaling in archaeol gy: time is not the only dimension. In Hodson, Kendall, and Tăutu 1971: 119-32.

Kruskal, J. B., I. Dyen, and P. Black. 1971. The vocabulary method of reconstructing language trees: innovations and large-scale applications. In Hodson, Kendall, and Tăutu 1971: 361-80.

Kruskal, J. B., I. Dyen, and P. Black. (in press) Some results from the vocabulary method of reconstructing language trees. Due to appear in the Proceedings of the Conference on Genetic Lexicostatistics held at Yale on April, 1971.

Kruskal, J. B., F. Young, and J. Seery. 1973. KYST: A new multidimensional scaling program. Mimeographed manual available from Bell Laboratories.

McFarland, C. 1972. The dialects of Bikol. Paper presented at the Forty-seventh Annual Meeting of the Linguistic Society of America, held at Atlanta on December 27-29, 1972.

Sankoff, D. 1972. Reconstructing the history and geography of an evolutionary tree. American Mathematical Monthly 79.596-603.

Sankoff, D. and G. Sankoff. (in press) Wave versus Stammbaum explanations of lexical similarities. To appear in the Proceedings of the Conference on Lexicostatistics held at the University of Montreal on May 19-20, 1973.

Shepard, R. N. 1972. Introduction to volume I of Multidimensional scaling: Theory and Applications in behavioral Sciences, ed. by R. N. Shepard, A. K. Romney, and S. B. Nerlove, p. 1-19. New York: Seminar Press.

Shepard, R. N. and J. D. Carroll. 1966. Parametric representation of nonlinear data structures. Multivariate Analysis II, ed. by P. Krishraiah, p. 561-92. New York: Academic Press.

Southworth, F. C. 1964. Family tree diagrams. Language 40.557-65.

Swadesh, M. 1950. Salish internal relationships. International Journal of American Linguistics 16.157-67.

Swadesh, M. 1952. Salish phonologic geography. Language 28.232-48.

Williamson, K. 1973. The sound system of Proto-Lower Niger. Paper presented at the Conference on African Linguistics held at Queens College on April 7-8, 1973.

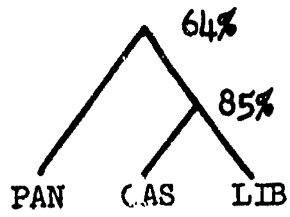
Wish, M. and J. D. Carroll. (in press) Applications of INDSCAL to studies of human perception and judgement. To appear in Handbook of Perception, vol. 2, ed. by E. C. Carterette and M. P. Friedman.

Pandan	PAN											
Virac	76	VIR										
Naga	67	81	NAG									
Legaspi	67	83	88	LEG								
Daraga	68	72	78	83	DAR							
Oas	64	67	74	75	86	OAS						
Libon	64	67	73	73	83	85	LIB					
Buhi	63	67	73	73	79	79	80	BUH				
Iriga	61	67	73	70	77	73	80	78	IRI			
Sorsogon	64	67	69	71	73	69	67	66	65	SOR		
Cubat	64	66	66	69	69	66	64	64	63	79	GUB	
Masbate	57	55	57	57	60	58	57	57	57	79	70	MAS
	PAN	VIR	NAG	LEG	DAR	OAS	LIB	BUH	IRI	SOR	GUB	

Fig. 1. McFarland's percentages among twelve varieties of Bikol. (rounded off to the nearest whole percentage)

a)

Pandan	PAN		
Oas	64	OAS	
Libon	64	85	LIB



b)

Pandan	PAN		
Libon	64	LIB	
Gubat	64	64	GUB

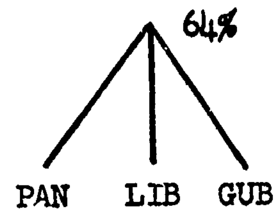
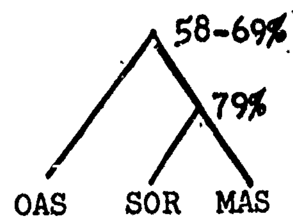


Fig. 2. Bikol relationships well represented by trees

a)

Oas	OAS		
Sorsogon	69	SOR	
Masbate	58	79	MAS



b)

Sorsogon	SOR		
Gubat	79	GUB	
Masbate	79	70	MAS

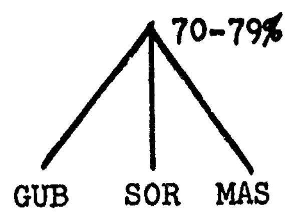


Fig. 3. Bikol relationships poorly represented by trees

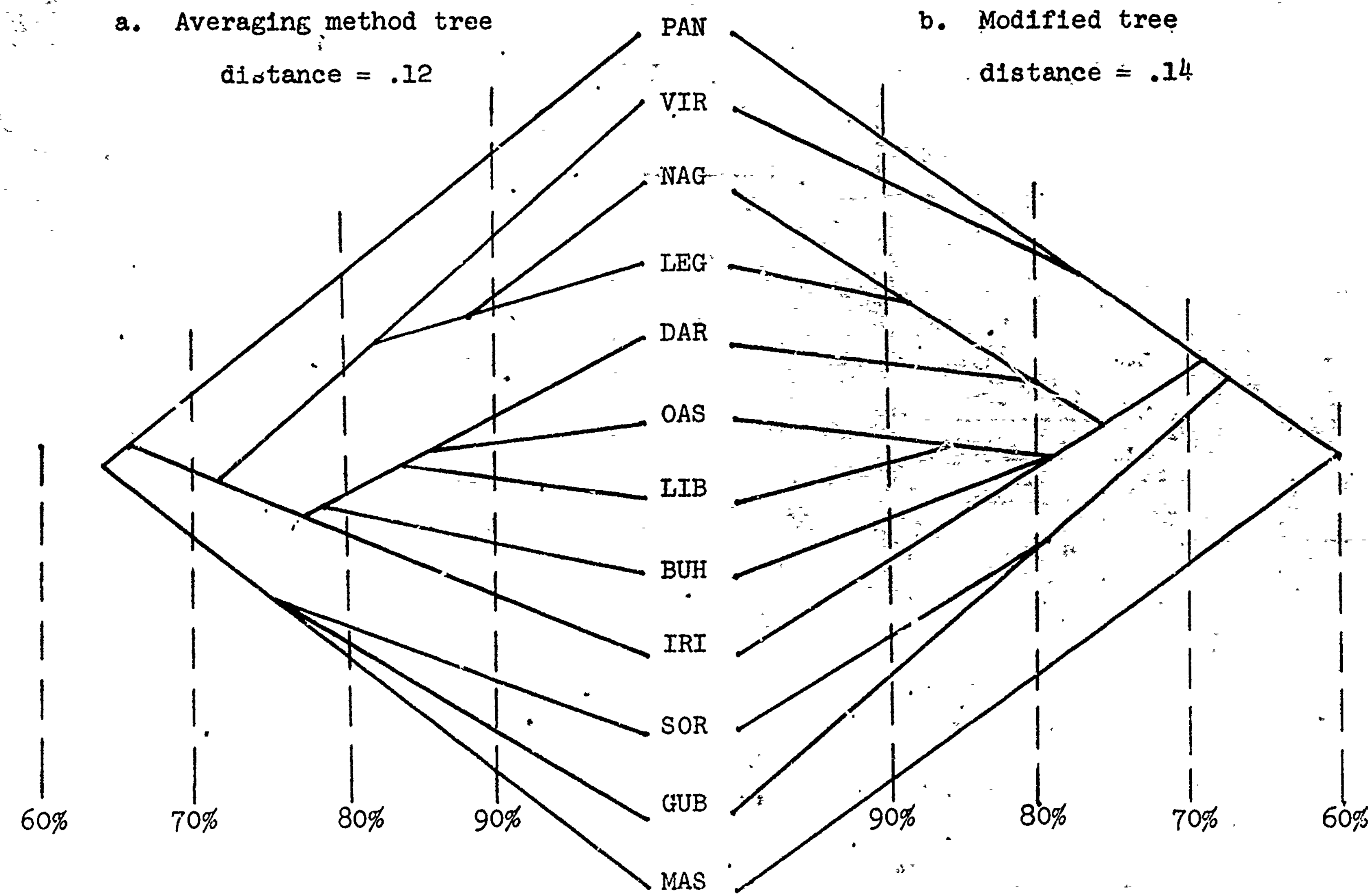


Fig. 4. Two possible trees for describing the Bikol percentages

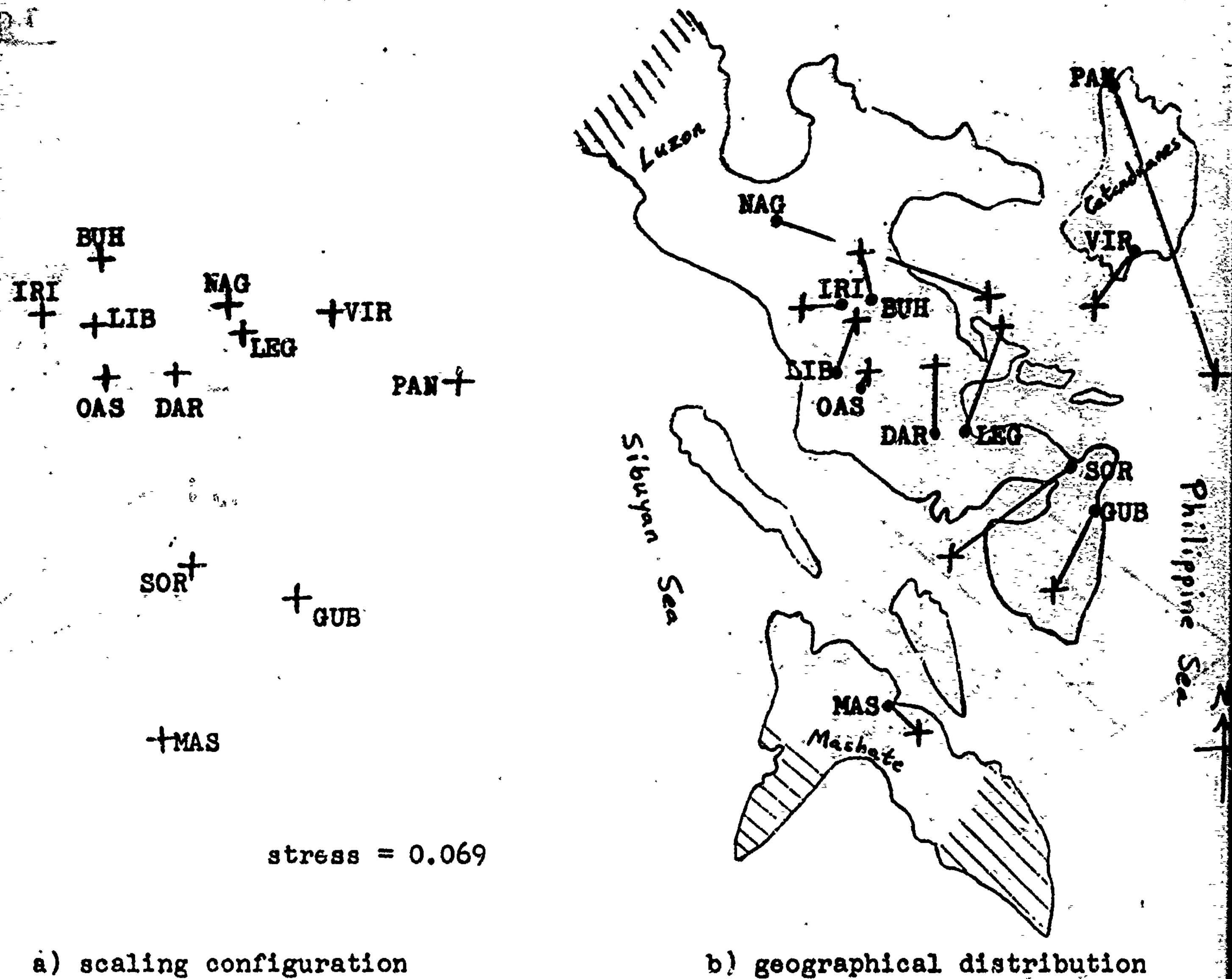


Fig. 5. Bikol: two dimensional scaling versus geographical distribution

HISTORY OF COMPUTATION. N= 12. THERE ARE 66 DATA VALUES, SPLIT INTO 1 LISTS. DIMENSION = 2

ITERATION	STRESS	SRAT	SRATAV	CAGRGL	COSAV	ACSAV	SFGR	STEP
0	0.088	0.800	0.800	0.000	0.000	0.000	0.0023	0.0050
1	0.087	0.983	0.897	-0.999	0.659	0.659	0.0021	0.0136
2	0.083	0.956	0.889	0.992	0.879	0.879	0.0018	0.0461
3	0.075	0.910	0.896	0.333	0.519	0.519	0.0009	0.0917
4	0.074	0.982	0.924	-0.717	-0.297	0.650	0.0015	0.0426
5	0.070	0.944	0.930	-0.543	-0.459	0.579	0.0006	0.0184
6	0.069	0.992	0.950	-0.569	-0.531	0.572	0.0004	0.0076
7	0.069	0.996	0.965	0.355	0.053	0.429	0.0001	0.0052
8	0.069	0.998	0.976	0.930	0.632	0.760	0.0001	0.0093
9	0.069	0.998	0.983	0.639	0.637	0.680	0.0001	0.0180
10	0.069	1.005	0.990	-0.575	-0.163	0.611	0.0005	0.0081
11	0.069	0.994	0.992	0.760	0.446	0.709	0.0001	0.0097
12	0.069	1.002	0.995	-0.890	-0.436	0.829	0.0003	0.0031
13	0.069	0.998	0.996	0.993	0.507	0.937	0.0002	0.0035
14	0.069	0.999	0.997	0.303	0.372	0.518	0.0000	0.0041
15	0.069	1.000	0.998	-0.472	-0.185	0.488	0.0001	0.0020
16	0.069	1.000	0.999	0.120	0.016	0.245	0.0000	0.0013
17	0.069	1.000	0.999	-0.133	-0.082	0.171	0.0000	0.0009

MINIMUM WAS ACHIEVED

THE FINAL CONFIGURATION HAS BEEN ROTATED TO PRINCIPAL COMPONENTS.

THE FINAL CONFIGURATION OF 12 POINTS IN 2 DIMENSIONS HAS STRESS 0.069 FORMULA 1

LABEL FOR CONFIGURATION PLOTS	FINAL CONFIGURATION	1	2
A	1	-0.125	1.416
B	2	0.394	0.820
C	3	0.553	0.258
D	4	0.373	0.304
E	5	0.217	-0.148
F	6	0.261	-0.477
G	7	0.505	-0.512
H	8	0.476	-0.391
I	9	0.708	-0.768
J	10	-1.135	0.296
K	11	-0.845	-0.219
L	12	-1.781	-0.589

DATA GROUP(S)

SERIAL	COUNT	STRESS	REGRESSION COEFFICIENTS (FROM DEGREE 0 TO MAX OF 4)
1	66	0.069	DESCENDING

Fig. 6. Bikol two dimensional scaling: page one of printout

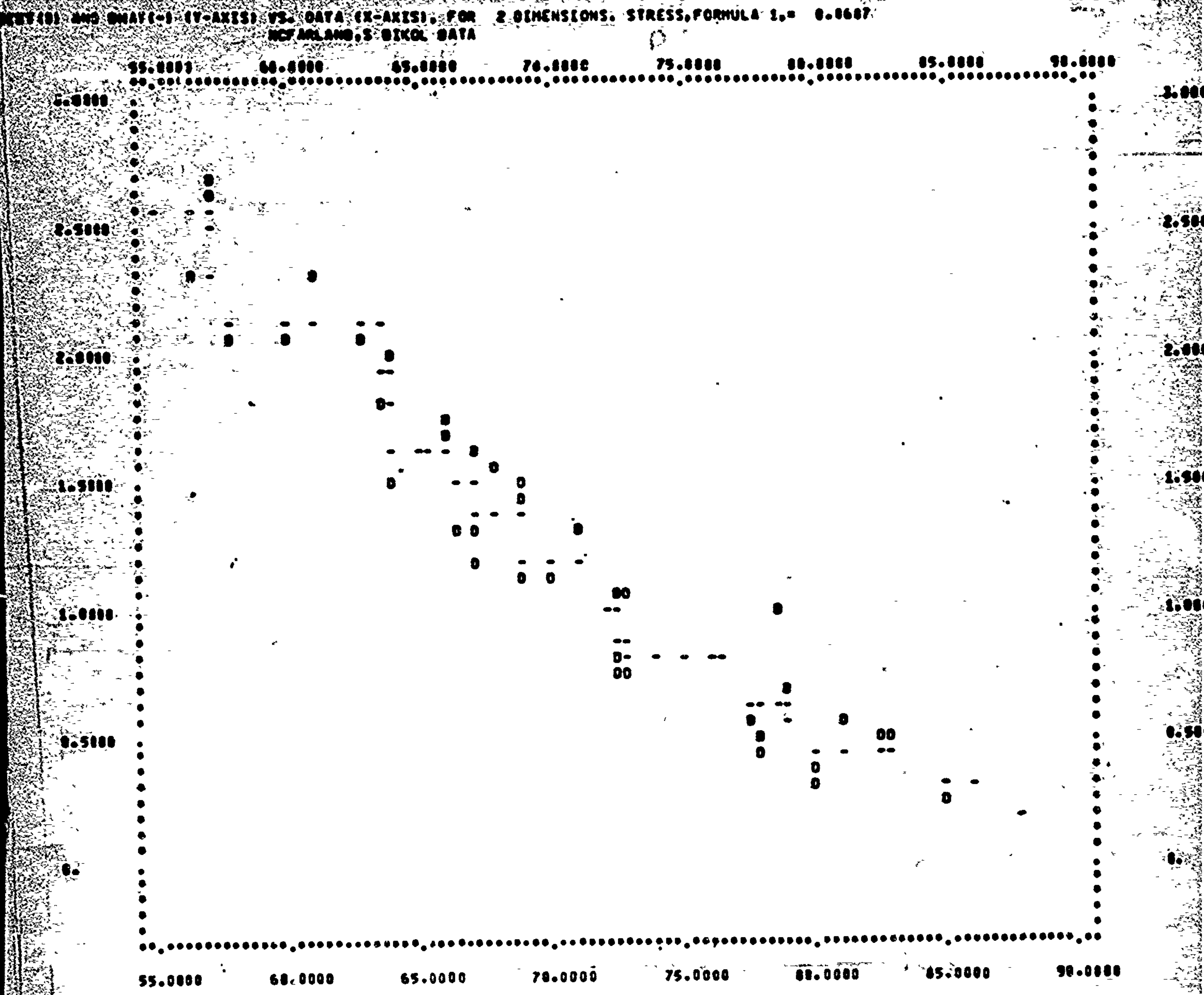


Fig. 7. Bikol two dimensional scaling: page two of printout (scatter diagram)

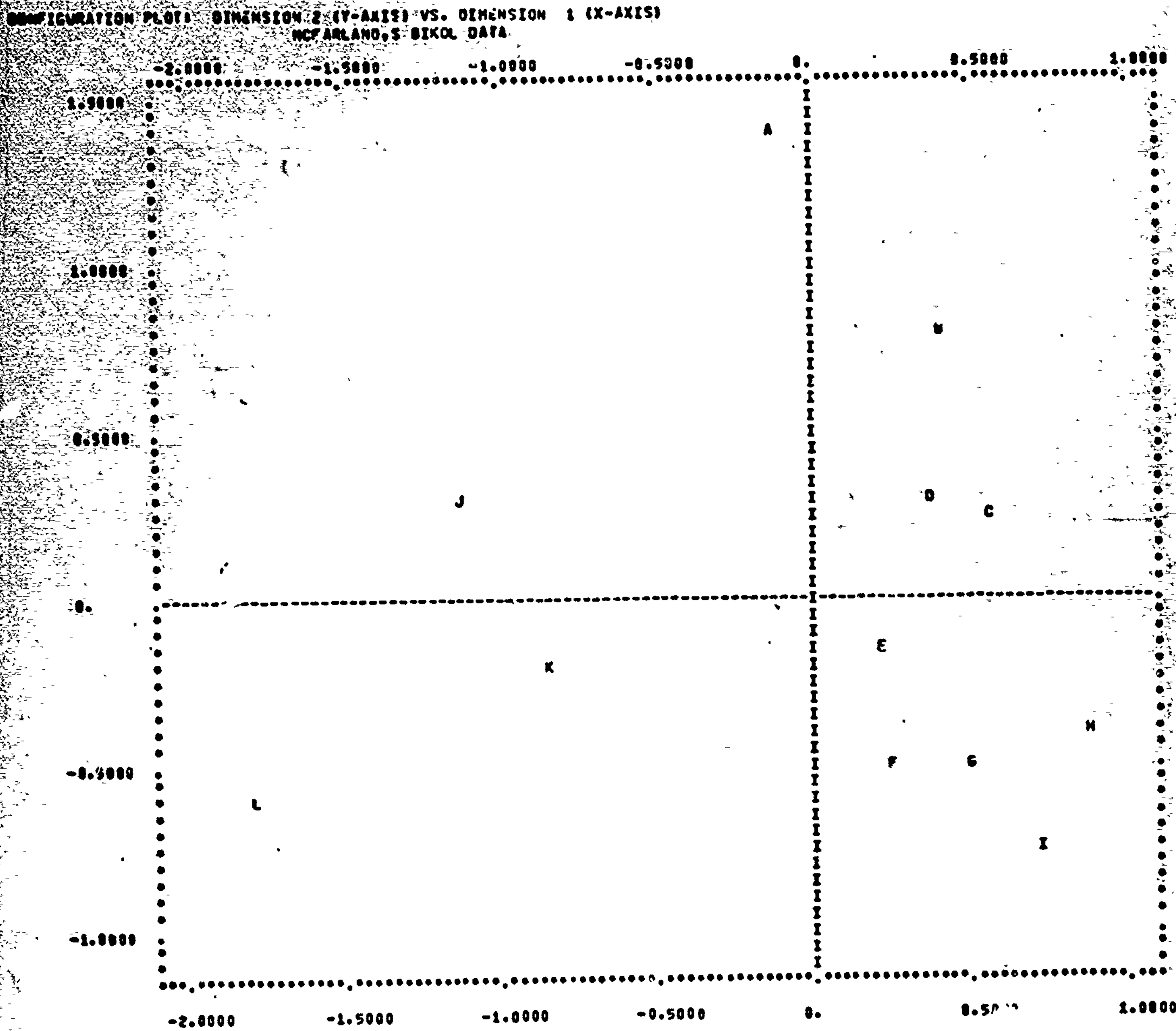


Fig. 8. Bikol two dimensional scaling: page three of printout (configuration)

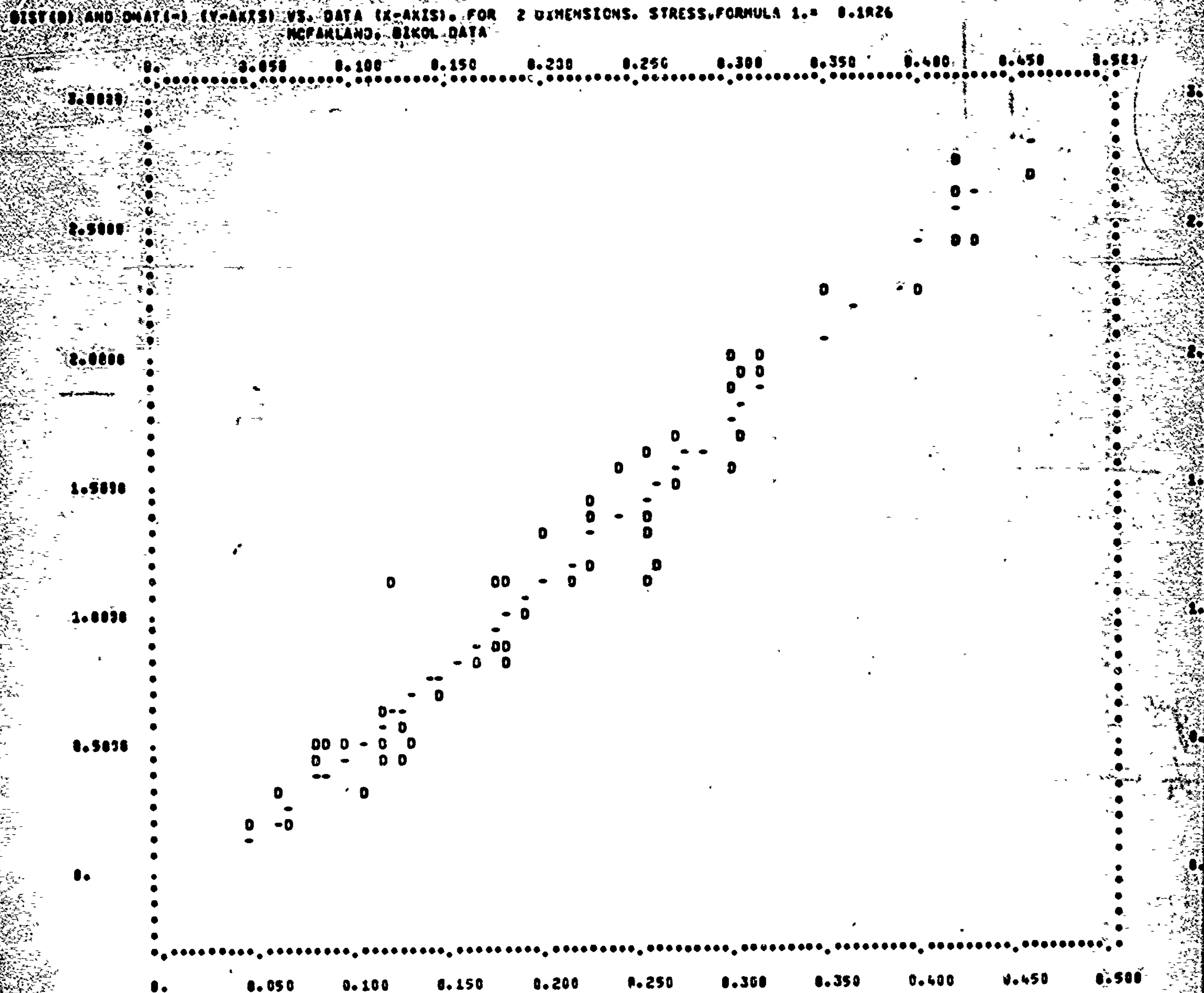


Fig. 9. Scatter diagram for metric two dimensional scaling of Bikol

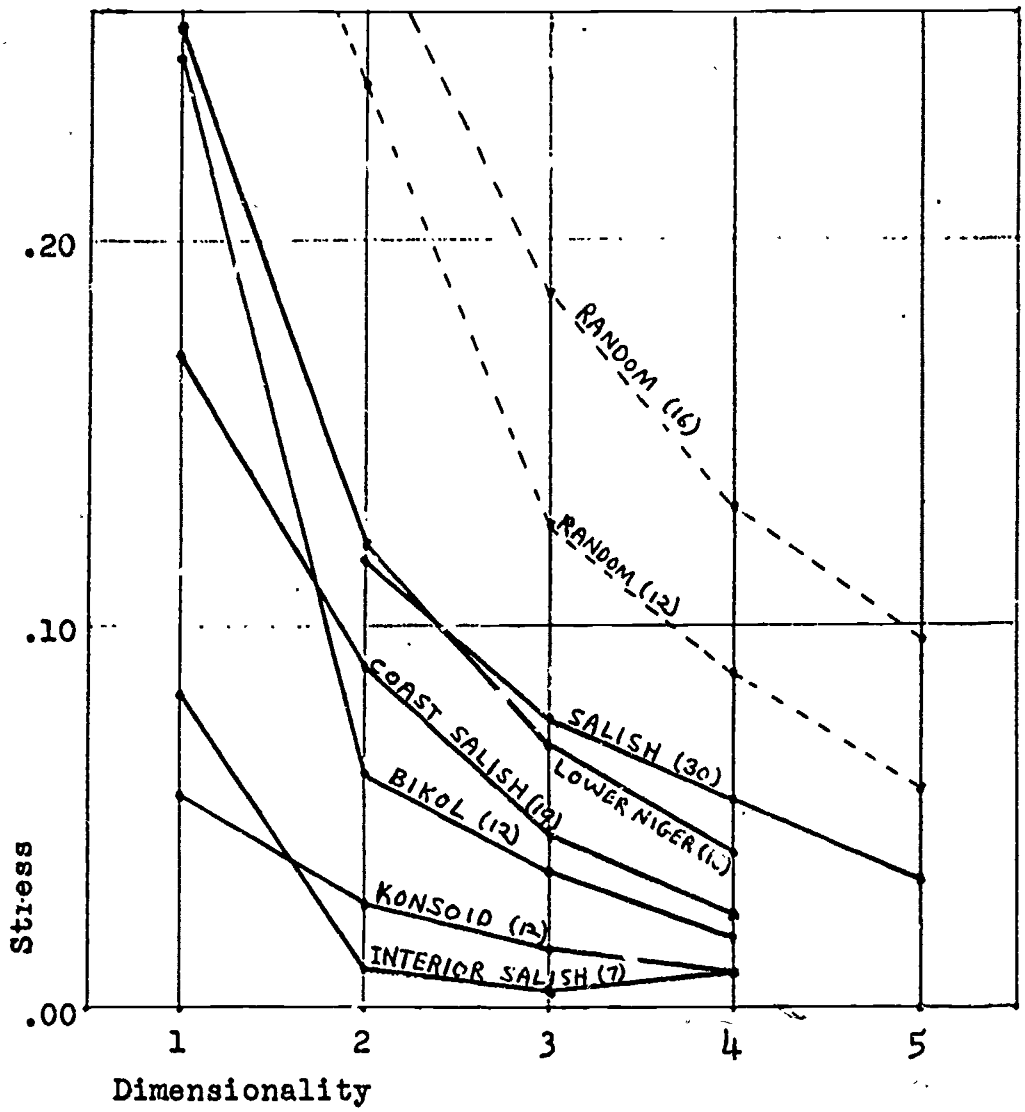
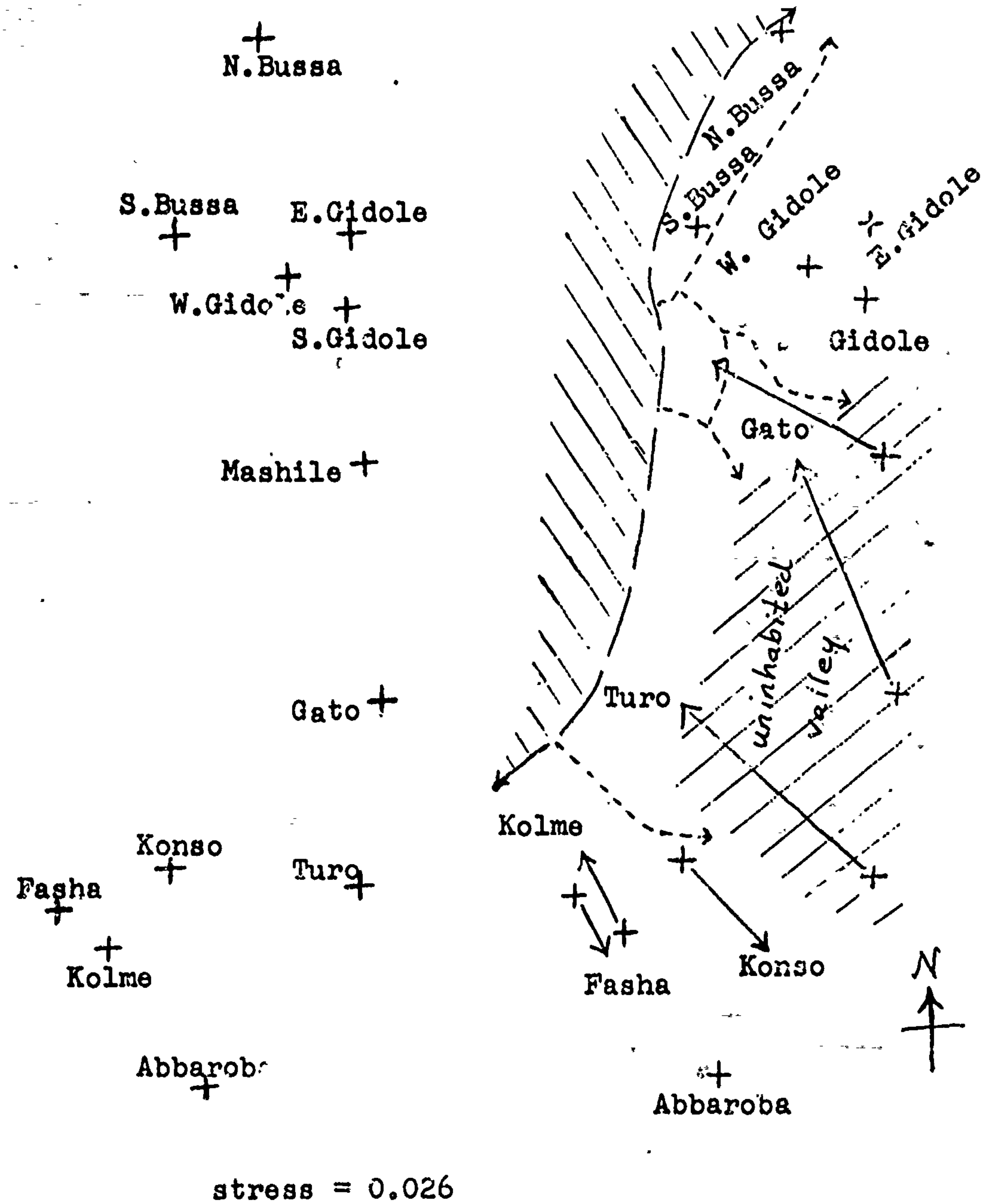
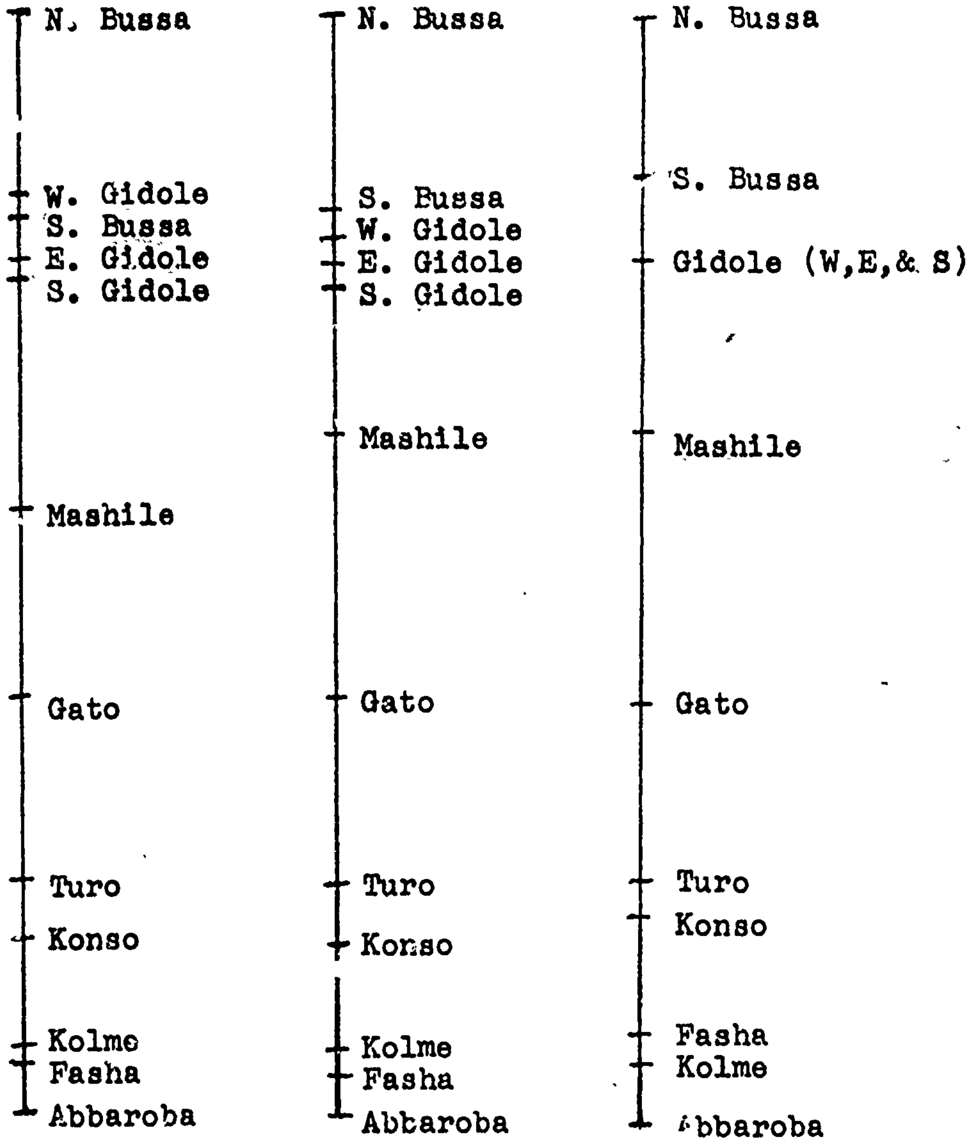


Fig. 10. Dimensionality versus stress for various nonmetric scalings, including average stress for two sets of random data (dashed lines); numbers in parentheses are numbers of varieties involved in the scalings.



a) scaling configuration b) geographical map

Fig. 11. Konsoid: two dimensional scaling versus map



a) nonmetric: stress = 0.056

b) nonmetric: stress = 0.055

c) metric: stress = 0.118

Fig. 12. Konsoid one dimensional scalings (see text)

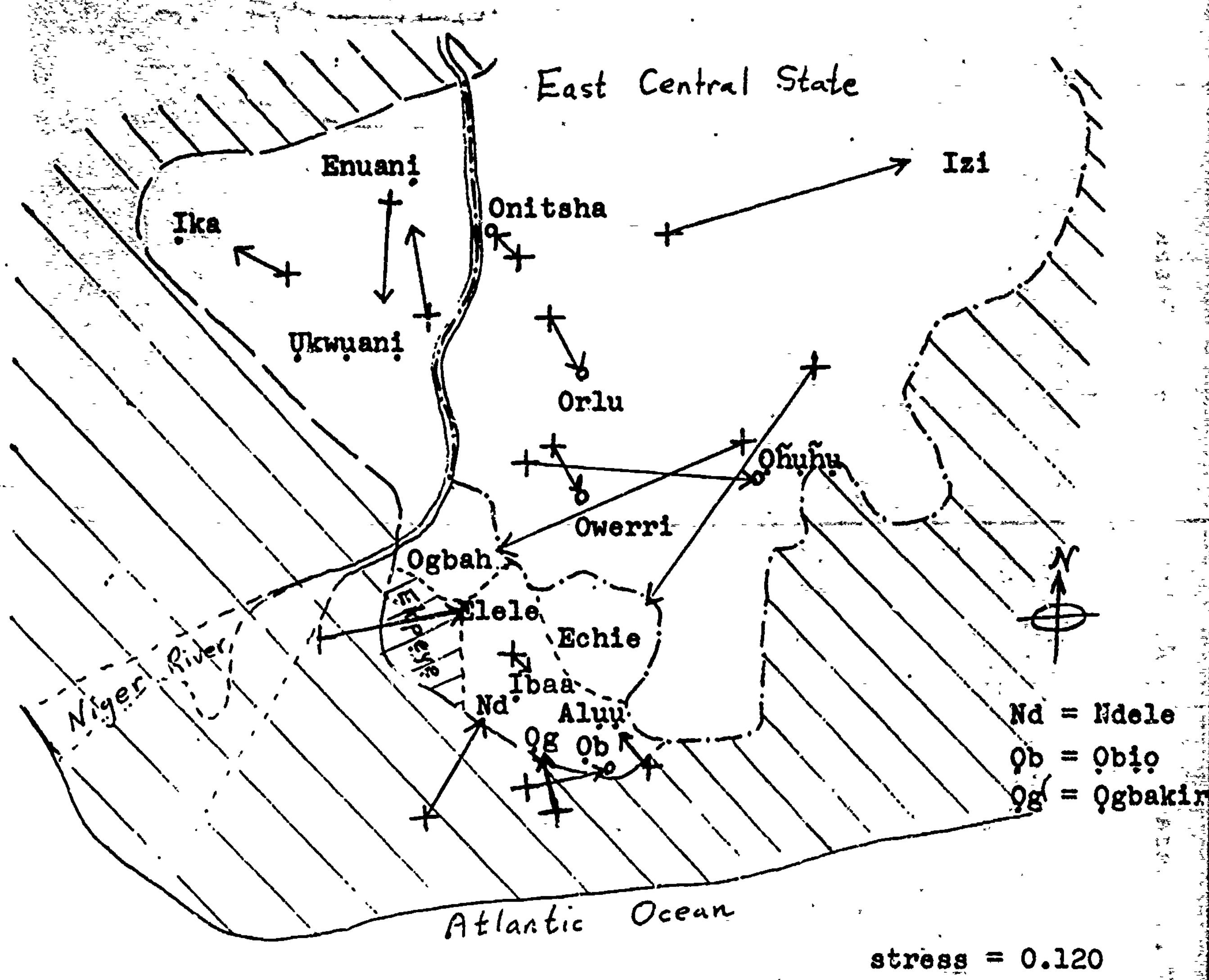


Fig. 13. Lower Niger: two dimensional scaling superimposed on geographical map

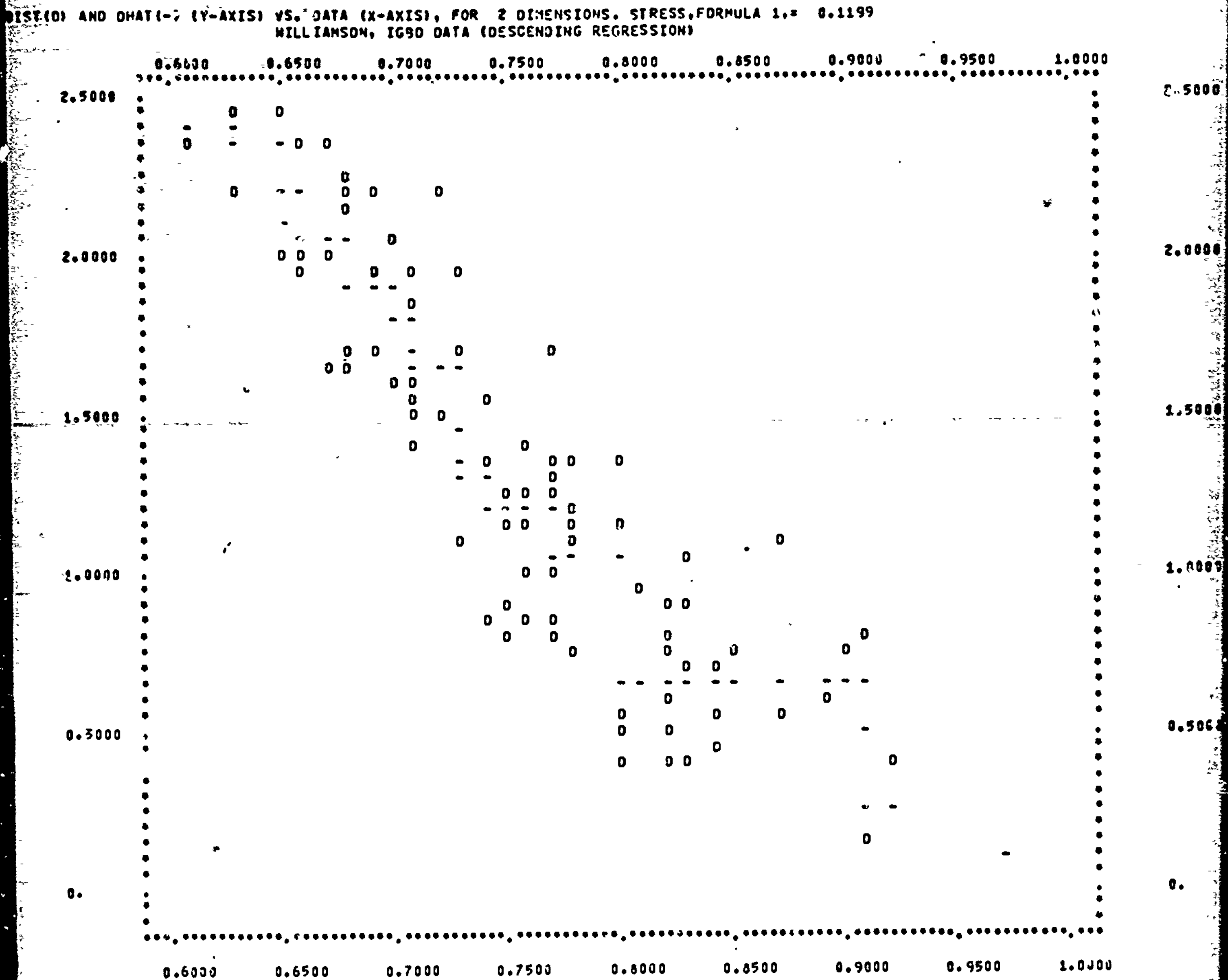


Fig. 14. Lower Niger: percentage versus distance for two dimensional scaling

Fig. 15. Salish classification according to Swadesh

Abbreviations		Swadesh's classification	Notes on Dyen's classification
		I. Coast Division	
		A. North Georgia Branch	
A	Cx	1. Comox	
B	St	2. Seshalt	
C	Pt	3. Pentlatch	
		B. South Georgia Branch	Replaced by:
		1. Nanaimo Group	
D	Fr	a. Fraser	
E	Nn	b. Nanaimo	South Georgia Branch
F	Sq	2. Sqamish	
G	Nt	3. Nootsak	and
		4. Lkungen Group	
H	Lk	a. Lkungen	Lkungen Branch
I	Lm	b. Lummi	
J	Cl	c. Clallam	
		C. Puget Sound Branch	
K-L	Sk-Sn	a. Skagit-Snohomish	
M	Ni	b. Nisqualli	
N	Tw	D. Twana (Hood Canal Branch)	
		E. Olympic Branch	Replaced by:
		1. Satsop Group	
O	Cw	a. Cowlitz	Satsop Branch
P-Q	Ch-Sa	b. Chehalis-Satsop	—and
R	Lo	2. Lower Chehalis	Lower Chehalis Branch
S	Qu	3. Quinault	
T	Ti	II. Tillamook (Oregon Division)	
		III. Interior Division	Subgrouped further into
U	Li	1. Lillooet	
		2. Thompson Group	Lillooet Branch
V	Th	a. Thompson	
W	Sh	b. Shuswap	
		3. Okaganon Group	and
X	Ok	a. Okaganon	
Y-Z- 1	Sp-Ka- Pe	b. Spokane-Kalispel- Pend d'Oreille	Columbia Branch
2	Cm	4. Columbia	
3	Cr	5. Coeur d'Alène	
4	Be	IV. Bella Coola	

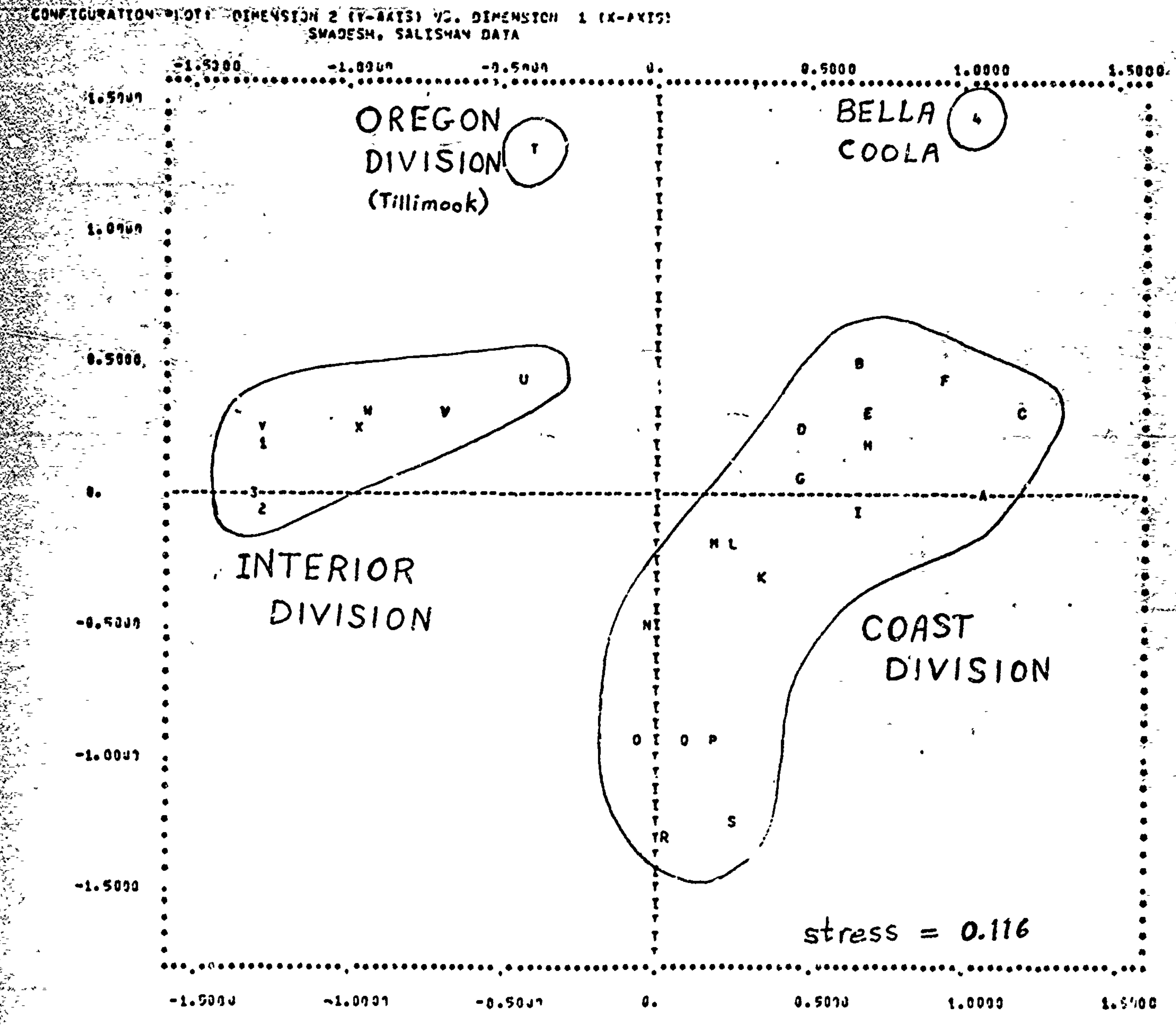
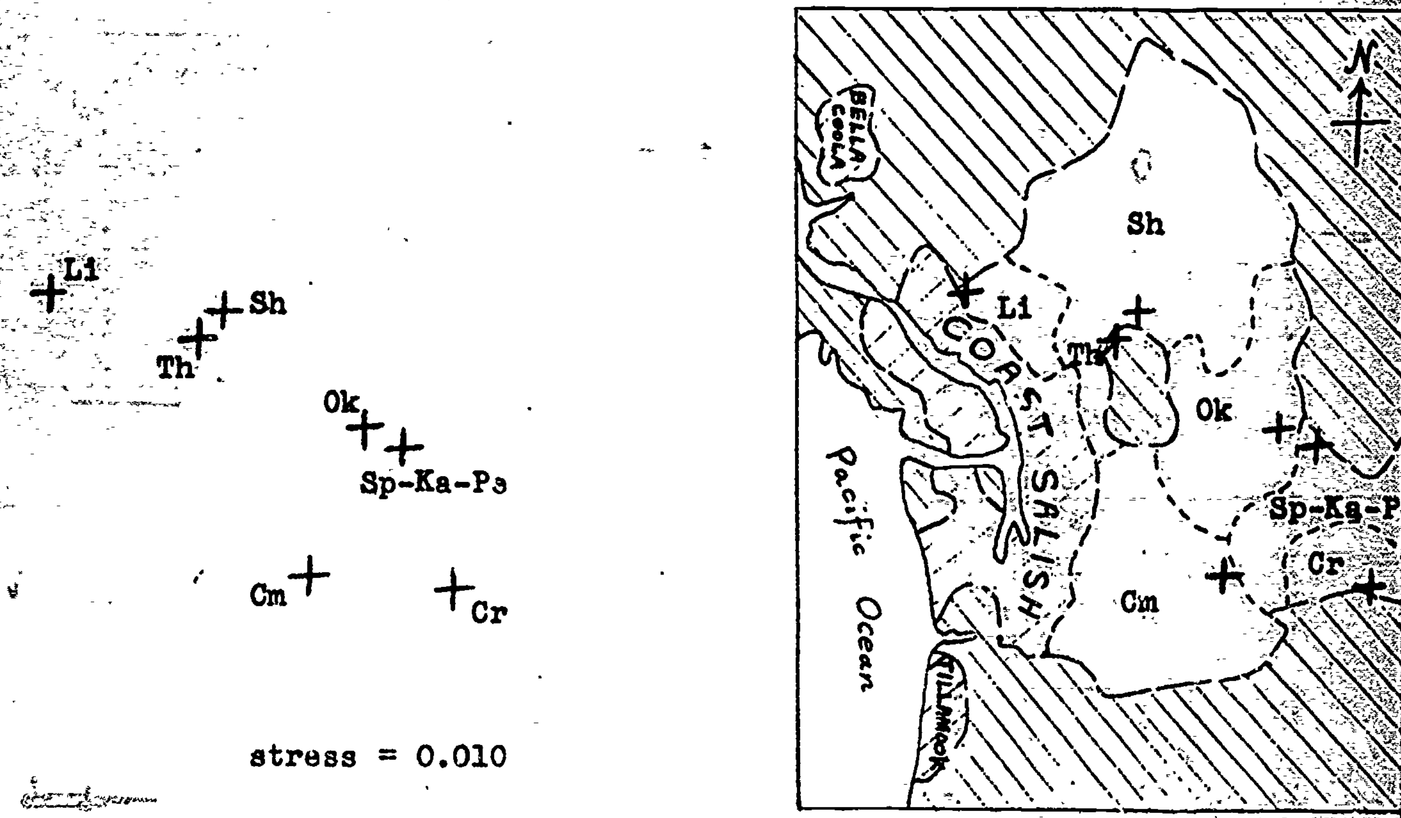


Fig. 16. Salish: two dimensional configuration showing four main divisions



a) scaling configuration

b) geographical distribution

Fig. 17. Interior Salish: two dimensional scaling versus geographical map

Swadesh's	Dyen's							
North Georgia	North Georgia	NG						
South Georgia	South Georgia	32	SG					
	Lkungen	30	39	L				
Puget Sound	Puget Sound	24	32	36	PS			
Hood Canal	Hood Canal	19	29	25	37	HC		
Olympic	Satsop	21	20	20	29	38	S	
	Lower Chehalis	16	18	20	22	25	43	LC
		NG	SG	L	PS	HC	S	

Fig. 18. Average percentages between Coast Salish branches according to Dyen (1962: 160); Swadesh's branches given for comparison.

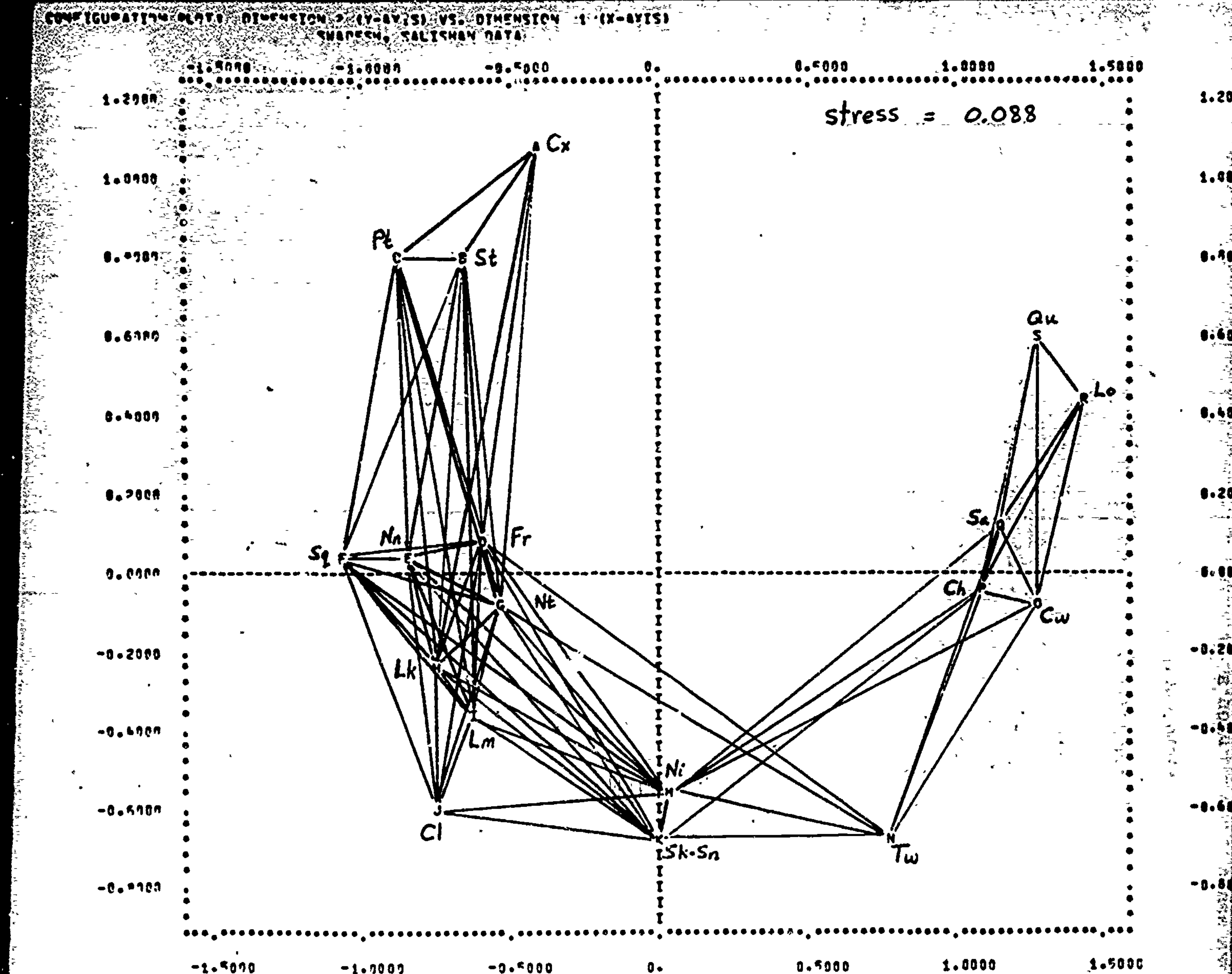
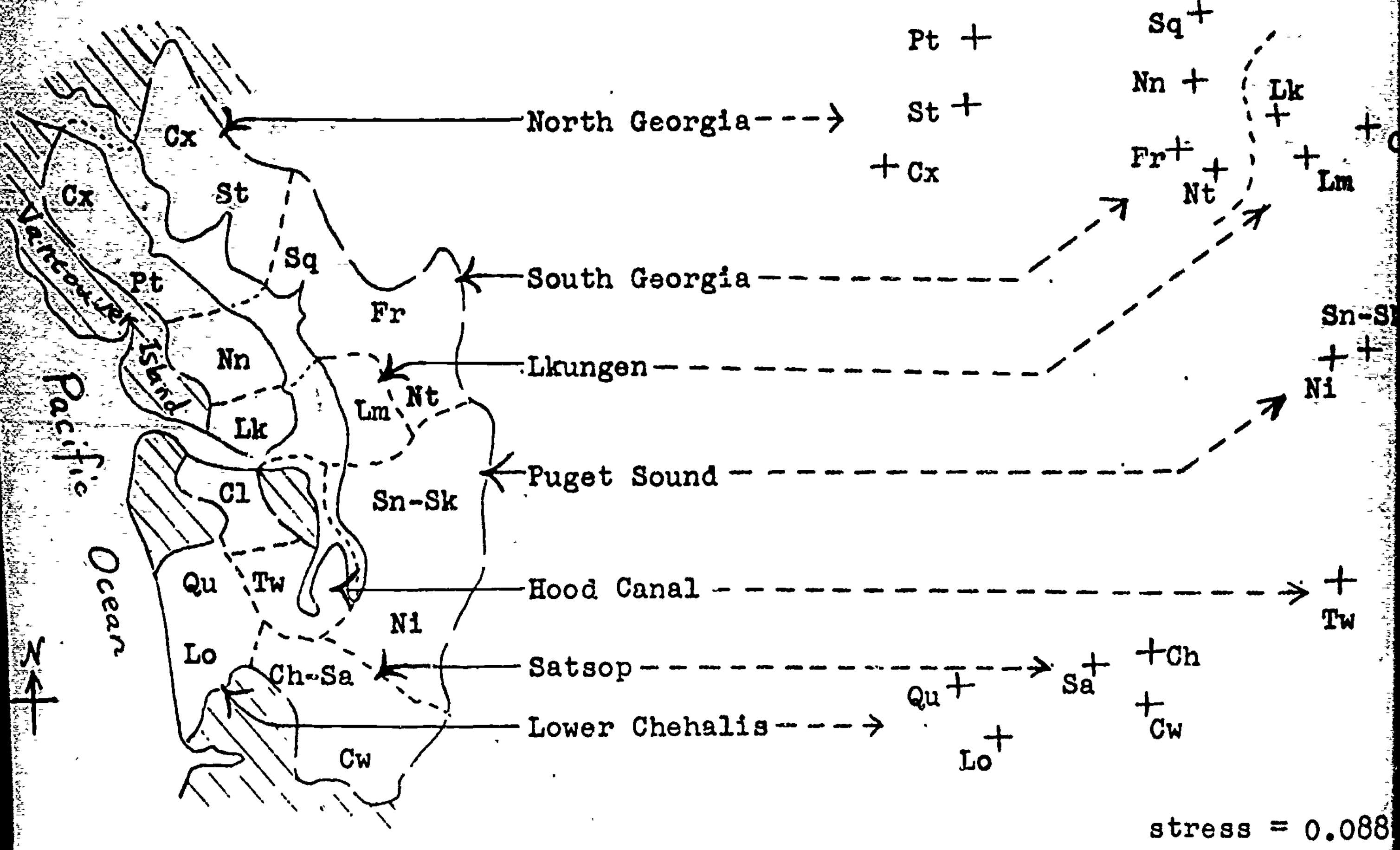


Fig. 19. Coast Salish two dimensional configuration (lines show "horseshoe" effec



a) geographical distribution

b) scaling configuration

Fig. 20. Coast Salish: map versus scaling, showing Dyen's seven branches