PATTERNS OF COHESION IN JAKARTA MALAY: TOWARDS A MORE OBJECTIVE METHOD OF DESCRIBING AREAL VARIATION

C.D. Grijns

1. INTRODUCTION

1.1 General background

One of the acknowledged merits of dialectology is that it demonstrated very early and very generally the complexity of the areal distribution of linguistic features in natural languages. Yet there is the paradox that dialectology so far has been unable to find adequate methods for describing the underlying order which constitutes, delimits and classifies the different varieties in a linguistic area. This fact seems to be the main reason why there is still so much uncertainty about the position of areal linguistics with regard to other fields of linguistic study, and especially about its positive contribution to the theory of language.

From this point of view one of the first tasks of areal linguistics would be the developing of better methods to describe the synchronic patterning of diatopical variation. It is obvious that especially in areas where few historical data are available, as is often the case in Austronesian studies, a reliable description of existing patterns of distribution could be an important aid to historical and comparative work. Moreover, if we could succeed in developing more valid methods for describing co-occurrence patterns in empirical linguistic data, these could also be applied to syntopical variation and thus help us to better understand the problem of linguistic 'codes' in sociolinguistics. Here again, in the Austronesian area, the tasks of linguistic description, and, where necessary, language engineering are urgent and fascinating.

Especially during the last decade rapid progress has been made in the field of data theory. At present several new techniques are available for the analysis of underlying structures in sets of empirical data. By structure we mean the pattern of relationships between the elements in a set. The type of techniques I have in view aim at a faithful description of the structure through presenting the data in a mathematical model. Via the model the relationships between the linguistic elements are measured; the measuring does *not* involve other attributes of the elements. For students of linguistic variation it is of particular interest to know that in the mathematical approach the boundaries between patterns

Amran Halim, Lois Carrington and S.A. Wurm, eds *Papers from the Third International Conference on Austronesian Linguistics*, vol.3: *Accent on variety*, 247-285. *Pacific Linguistics*, C-76, 1982.
© C.D. Grijns

are not seen as clear-cut and absolute, but as gradual and often fuzzy. Any final decision regarding classification or grouping should be based upon direct study of the empirical data.

For some years I have been exploring the possibility of using one such newly developed technique for the analysis of the complex dialect variation in the Jakarta Malay area. I would like to report here on some methodological aspects of this ongoing research. Before giving an outline of the method, the data and the model, it seems appropriate to consider how the traditional methods have been evaluated, and why they cannot be judged to be adequate; in addition we shall briefly discuss the theory and method developed by C.-J.N. Bailey.

1.2 The traditional methods

It is almost fifty years ago since Bloomfield summarised both the achievements and the potentialities of traditional dialect geography. Positively, he appreciated the contribution made to "our understanding of the extra-linguistic factors that affect the prevalence of linguistic forms" as well as to the knowledge of "a great many details concerning the history of individual forms". On account of sociolinguistic and semantic factors, however, Bloomfield saw no hope of a "scientifically usable analysis, such as would enable us to predict the course of every isogloss". On the other hand he noted that although "important social boundaries will in time attract isogloss-lines ... it is evident that the peculiarities of the several linguistic forms themselves play a part, since each is likely to show an isogloss of its own". (Bloomfield 1933:345).

Forty years after the publication of Bloomfield's *Language*, W. Winter, in his state of the art report in *Current trends in linguistics*, writes:

> ... the results hitherto achieved in the field of areal linguistics apparently do not form a coherent fabric or even a somewhat consistent pattern, but merely a patchwork quilt of colorful, but largely unrelated data and anecdotes. [One must conclude] ... that in this field nearly everything can be shown to be possible, but that not much progress has been made toward determining what is probable and to what degree, so that the time does not seem to be at hand yet for an empirically based coherent theory of areal linguistics (provided there can be such a theory for a complex field not amenable to investigation under simplified and consistent test conditions, and not just an ordered set of observations concerning events that can be shown to have taken place.) (Winter 1973:135).

Another scholar in the field of variational linguistics, C.-J. Bailey, speaks in a similar vein with regard to the results of the first hundred years of 'glottogeography':

> ... I do not believe that the present methods are ever going to bring us any nearer to the goal of defining or delimiting dialects, or that these methods are ever going to make more contributions to our understanding of the theory of language, than they already have. (Bailey 1980:234).

These judgements all point to a methodological impasse with regard to the description of dialect patterning. The traditional methods have in common that the spatial (geographical) patterning of the points of observation (the informants) is taken into account from the very beginning. The map is the main tool of the dialect geographer (cf. Goossens 1969:13). Together with the map comes a lot of other extra-linguistic information, which is highly relevant for the explanation of linguistic patterns. But these patterns themselves cannot be described on the basis of their geographical position, but only on the basis of their distribution over the points of observation, which is not the same. The two approaches should be clearly distinguished. I agree with Bailey, who advocates that the first task of dialectology is to look for language-internal patterning, the 'what-goes-with-what' approach. Bailey contrasts this line to the line taken by Trudgill (in Trudgill 1973), who concentrates on the geographical end (cf. Bailey 1980:248). (Here one may observe a parallel with the distinction between a sociolinguistics which relates linguistic patterns to social patterns, and a sociology of language which concentrates on the role of language in society.) If dialectology makes the impression of a "patchwork quilt of largely unrelated data" it is mainly because the above-mentioned distinction has not been consistently implemented. In itself, however, this cannot be the main reason for the impasse, as some of the best dialectologists have been always aware of the distinction, and particularly so the structuralists.

To begin with structural dialectology: why did Weinreich's diasystem method, the "treating of different systems together because of their partial similarity" fail to produce integrated descriptions of dialect areas? (see Weinreich 1968/ 1954). I see three reasons: (i) a full description would require a *complete* analysis of the systems which are treated together. In practice, one always has to work with (subjectively selected) subsystems. Especially in the case of semantic data the selection can only be extremely arbitrary. (For very interesting examples of semantic applications see Goossens 1969:69ff.) In brief, one has to know the position of every contrasting element under study within its total system; (ii) The similarity between the structure of different systems, even if the first condition could have been fulfilled, was not yet quantifiable; (iii) The application of structural isoglosses meets the same problems as any other isoglossic method, as Ivić very explicitly remarks ("... leaving the dialectologist in a helpless struggle with the perplexities of choice." Ivić 1962:34). It is inherent in the structural approach that heterogeneity within systems is seen as deviation from structuredness. Thus the heterogeneity of a transition area is for Kurath a case of "temporary disorganization" (H. Kurath, quoted with approval in Moulton 1968:458). This may be a good characterisation under certain circumstances. It does not offer much to go upon if one undertakes the synchronic description of a complex area.

The reason why the use of isoglosses does not lead to the description of distributional structure is that isoglosses contrast one particular feature, or a group of features, to all the other features of groups together. Since any feature may have a different relationship with any other individual feature, this means an enormous loss of information. In addition, the choice of isoglosses is almost invariably arbitrary and based on an extralinguistic criterion, since only those isoglosses are considered which join a bundle, and as long as they do so, the criterion is spatial. All the isoglossic methods have these weaknesses in common, also those which use more refined statistical techniques. If statistics are applied here, i.e. if generalisations are made on the basis of a sample, the predictions are based on geography, not on language. Ivić's suggestion to typify dialect areas according to the density, the direction, the form, etc., of the isoglosses which intersect them, has not been followed up, and highly interesting as it is, would not have yielded a description of the relationships between the linguistic features (see Ivić 1962). Guiter has succeeded in overcoming the problem of arbitrariness in the selection of isoglosses, by counting *all* the isoglosses which intersect the linking-lines between any pair of adjacent villages which are angular points of the same triangle, the total set of villages being connected in one network of triangles. This produces a valid hierarchy of boundaries and sub-boundaries in the area. It does not, however, bring out which linguistic features go together in each of the subareas, and which groups of features can be contrasted (see Guiter 1973).

Isoglosses represent dissimilarities (and for that reason they have rightly been called "heteroglosses", cf. Kurath 1972:24ff.). Another frequently used technique is based on similarity counts. For each pair of dialects or languages under investigation a similarity score is computed which is defined on a fixed set of concepts (often the one represented in the 100- or 200-word list of Swadesh). The pairs of dialects, etc., or, eventually, groups of pairs, can then be ordered according to their degree of linguistic similarity. This approach is attractive in that the ordering of the heterogeneity is not carried out on the spatial (geographical) dimension, and any fixed set of concepts can be used without leaving out any features. Thus linguistic 'nearness' between sets of variants is measured objectively. Nevertheless this is also a weak method, because only pairs of total sets can be compared, and all information is lost with regard to the specific content of the individual sets. This fact is well realised of course, and the technique is used in synchronic analysis mainly to find a preliminary grouping of dialects, etc. (For recent examples see Walker 1975, on Lampung dialects, and Anceaux 1978, on south-east Sulawesi).

Another problem which is inherent in all the methods used so far, is that there is no objective criterion to determine whether two features should be considered as compatible or not. Identical forms occurring in different dialects may have a somewhat different meaning, whereas somewhat different forms with the same, or a rather similar, meaning cannot always safely be established as compatible on the basis of known regular sound correspondences. The inevitable reduction of variants previous to their mapping or counting remains a delicate task, where the subjective opinion of the researcher plays an important part. The problem is well known as the cognation or compatibility problem. (I agree with Cadora 1979:4ff. that for synchronic purposes the latter term is more appropriate).

1.3 Bailey's theory of dialects as implicational constellations

In the meanwhile for more than a decade C.-J.N. Bailey has been developing a new theoretical approach to the problem of areal patterning. The essence of his method is that he concentrates first on language-internal patterns rather than beginning with extralinguistic distributions, as we have seen above. Within that framework his analysis is primarily time-based. Both explanation and prediction are related to the dimension of time. Explanation "is possible only when one understands how structures grow and evolve" (Bailey 1979:28). With regard to prediction, since social happenings cannot be predicted, "only the non-social side of linguistic analysis and linguistic change is fully theoretical, allowing of both explanation and prediction ... The social side is only semi-theoretical ..." (1979:36).

In order to detect this one-dimensional structure in his data, Bailey makes use of the so-called implicational scale (also known as Guttman scale, and already several times applied in sociolinguistic work: cf. DeCamp 1971, Dittmar 1973, etc.).

Table 1

							A	B	C	D
1.	A	B	C	D		1.	1	1	1	1
2.	A	B	C	-	or, numerically	2.	1	1	1	0
3.	A	B	-	-		3.	1	1	0	0
4.	A	-	-	-		4.	1	0	0	0
5.	-	-	-	-		5.	0	0	0	0
							4	3	2	1

The model is satisfied if the data show a pattern as represented in Table 1, where the rows (1., 2., ...n) represent the linguistic variables. There are as many rows as there are variants in the first row. Thus a simultaneous ordering of points of observation and of variants of one same variable becomes apparent. The theory is that observation point 1., which has all the variants, is the most original "lect" (Bailey's term), whereas variant A, which has gone through all the developments in time, is the oldest variant. Any later stage implies the next preceding stage. The variants A, B, C, D, can be perfectly ordered along the basis of the rectangle, which is interpreted as the linear dimension of time. Calculation of probabilities may in this technique determine the admissibility of violations of the model.

Bailey has demonstrated very interesting cases, where structure was found independently from geographical order (see especially Bailey 1973 and 1980). A test of the validity of his theory would include the calculation of the proportion between the amount of data which do confirm the assumption and those which do not, since one general criterion for the suitability of the model is the quantity of the data that have to be eliminated in order to satisfy the model. If too many variants have to be neglected, the model should be rejected. Bailey claims that his method can be applied on all levels of linguistic description. However the solutions which have been demonstrated so far do not include substantial sets of lexical items. Moreover, it is a precondition for the method that the linguistic history of the speech community is not disturbed by borrowings from outside or by internal discontinuity. Therefore the old factors already pointed out in Bloomfield's summary still seem to challenge the theory. Will semantic variation ever be predictable? Will it be possible to find speech communities, sufficiently homogeneous and free from unpredictable sociolinguistic variation, where the theory can be fully applied? Whatever the answer to these questions may be, Bailey's experiments are a very important effort to open up new ways in areal linguistics.

2. MULTIDIMENSIONAL SCALING OF JAKARTA MALAY DATA

2.1 General notes on the method

My own investigation also concentrates primarily on the patterning of linguistic elements and is not based on the isoglossic method. Unlike Bailey, I have not been looking so much for a new theoretical basis, but rather for a new technique which would give the perspective of a really structural description of a total linguistic area. I do believe that such a description, if successful, can contribute to new theoretical insights, and I suspect that in the case of this area especially the processes of rapid convergence and of the preserving of local identity can be studied. Since my fieldwork has been a first exploration in a completely neglected area, I have been aiming at a descriptive approach, without making any assumptions as to the expected patterning. The technique of which I am making use is a scaling technique. Scaling techniques are quantification techniques which aim at representing an empirical relational system within a formal, usually a numerical, system. Scaling techniques are based on a geometrical model and are primarily of a descriptive nature. The purpose of the procedure is to gain an insight into relations between entities in the empirical reality and to detect the 'hidden structure' in the data (cf. Kruskal and Wish 1978:7).

The choice of the numerical system (the scale, or the scale model) depends on the nature of the data and the assumptions the researcher wishes to make regarding the expected structure. The analysis which I am carrying out at present is based on a non-metric multidimensional scaling technique for the analysis of categorical data, as will be described in the next two sections.

2.2 The data

In order to keep our exposition of the procedures as concrete as possible, we shall use a terminology which directly refers to the particular data under study. These data comprise the results of a linguistic survey carried out in 1970 in 470 points of observation ('villages', i.e. *desas,* or, in Jakarta, *kelurahans*) throughout the total Jakarta Malay area. This area includes the administrative territory of Jakarta (DKI-Jakarta) as well as a number of surrounding subdistricts in the districts of Bogor, Bekasi and Tanggerang (see Maps 1 and 2 pp.276, 278). The informants all belong to the Jakarta Malay or 'Betawi' speech community. One major assumption in collecting the data has been that this speech community is socially sufficiently homogeneous to justify the neglect of social differentiation. (On the exclusive social function of this vernacular as folk speech, see Grijns 1977). It was also assumed that the conditions under which the questionnaire was administered (always through interviews in Indonesian by Indonesian fieldworkers) have been sufficiently constant to keep undesired situational variation at a minimum.

The questions in the questionnaire are the variables. The more than 600 questions are divided into several sets of variables, each of which is analysed separately. The first set, which was used as a training ground, comprises 50 lexical items, many of which have been chosen from the Swadesh 'basic vocabulary' lists (as given in Samarin 1967:220-223). Other primarily lexical sets refer to kinship, agricultural tools, fishing tools, kitchen tools, flora and fauna, adjectives and intransitive 'verbs', etc., whereas some sets exclusively comprise phonological or morphological questions.

The data are organised as mutually exclusive, and exhaustive, response categories in a rectangular matrix. A concrete example may illustrate this. The variable 'new' in the context 'a new shirt' elicited the following variants: baru, baru/bagus, baru', anyar, anyar/énggal, bagus, cakep, baru/cakep, jempolan, utuh. For the first set of 50 items these variants were grouped into five lexical categories as shown in Table 2.

Table 2

1.	baru (occurring 339×), baru' (21×), frequency	360
2.	anyar (8×), anyar/bagus (1×)	9
3.	bagus (25×), baru/bagus (8×)	33
4.	cakep (5×), baru/cakep (1×)	6
5.	jempolan (1×), anyar/énggal (1×), utuh (2×), missing data (58×)	62
Total frequencies		470

For the analysis it is assumed that these categories are indeed mutually exclusive, which is another working hypothesis, of which the relative validity for the empirical reality can be seen from the arrangement above. A different grouping is possible, of course, and has indeed been applied when the same variable was included again in another set, as has been done with most of the 50 variables of this first set, for reasons of testing and comparison. In all cases the final category contains the residual forms, i.e. those forms which have a very low frequency, or which are somewhat suspect, etc. In all the other sets the first category contains the missing data.

In the matrix the cells contain the response categories. Thus each horizontal row corresponds with the profile of response categories on which a particular village scores; each column corresponds with the series of response categories observed with regard to a particular variable.

Table 3 shows a very small section of the matrix for the first four of the 470 villages, where the numbers in the cells are the category number. (Variables: 77-82.)

Table 3

	Variables					
Villages	77	78	79	80	81	82
1	4	9	9	6	6	6
2	3	3	4	3	3	3
3	3	3	9	6	6	6
4	5	3	3	3	3	6

Reading along the rows, one sees, for example, that villages 1 and 3 have the same profiles for variables 79-82. When reading along the columns, however, we cannot make the same type of comparison, since the numbers in each column represent mutually completely independent categories of different variables. In order to make the scores comparable and countable, horizontally as well as vertically, the matrix has been converted (i.e. rewritten) into a zero-one matrix as follows (see Table 4):

Table 4

	categories													
Villages	3	4	5	3	9	3	4	9	3	6	3	6	3	6
1	0	1	0	0	1	0	0	1	0	1	0	1	0	1
2	1	0	0	1	0	0	1	0	1	0	1	0	1	0
3	1	0	0	1	0	0	0	1	0	1	0	1	0	1
4	0	0	1	1	0	1	0	0	1	0	1	0	0	1
Variables	(77)			(78)		(79)			(80)		(81)		(82)	

In this latter matrix the categories take the place of the variables, and it is indeed not the variables which will be quantified, but the categories. What we are going to study is no longer, as in the isoglossic approach, the relationship between a particular linguistic feature (i.e. a category) and all the other features together, nor is it any longer the relationship between pairs of languages or dialects (i.e. villages) on the basis of their partial similarity. This matrix is the starting-position for an analysis of the relationships between all the categories simultaneously, and between all the villages simultaneously. It is important to state that only the data in this matrix will be analysed. This means that no external information, such as knowledge of the geographical position of the villages, will influence the analysis. Nor will any a priori weighting of the data take place. The aim is the best possible (i.e. isomorphic) representation ('picture') in the model of the empirical data as we have observed them.

2.3 The model

At this point we have to face the fact that the quantification itself, i.e. the attributing of numerical values to the categories and to the villages, is too technical a process to be accessible for users of the method (including this author) who have not passed through an advanced mathematical training. The basic principles can be understood, however, without knowledge of the algorithm involved. The model is a variant of the multidimensional scaling techniques and has been developed by De Leeuw and others. It is generated by the computer program HOMALS-1.

Like other multidimensional scaling techniques, HOMALS is strongly geometrically oriented. The model represents each village and each response category as a point in an Euclidean space. (One could visualise a three-dimensional model as a 'cloud' of points). The ultimate object of the procedure is to obtain a perfect one-to-one correspondence between the position of the points in the total model and the position of the villages as well as of the categories in their mutual relationship in the total set of empirical data, as organised in the matrix. In order to make the patterning of the villages and the categories optimally comparable, HOMALS represents both in a joint space.

The position of the points in the model is defined by co-ordinates on a system of axes which have the same origin, i.e. the zero co-ordinate. There is one axis for every dimension on which the analysis is carried out. A point in the model whose projection on a particular axis coincides with the origin has the score zero. If the projection lies on one side of the origin, the score is negative; on the opposite side the scores are positive. On a plane one can plot the position of a point on any pair of two dimensions, given the scores of the point. The HOMALS program provides the well-known type of scattergrams. However, as soon as the number of points is too large, too many points coincide in the plots. Moreover, an analysis on four dimensions requires six plots, five dimensions require ten plots, and one soon faces the problem of how to compare so many maps. Since I am working with relatively large sets on five dimensions I have called in the help of other techniques to reduce the information contained in the HOMALS scores, as will be discussed later on.

If the distribution of the categories were entirely random, the number of dimensions required for the perfect representation in the model of n categories would be $n - 1$. Where contrasting patterns exist, a reduction of the dimensionality is possible. Since the program aims at the lowest acceptable dimensionality, the direction of every axis is computed in such a way as to obtain an optimal dichotomisation of the data space, i.e. the dichotomy optimally corresponds with existing contrasts in the data. Accordingly, the opposition between the positively scoring categories or villages on a particular dimension and the negatively scoring items is an important clue for the interpretation.

The HOMALS technique is non-metrical, i.e. the position of the points in the model is not determined on the basis of absolute distances, but of an *order* of distances. This has enabled the designers of HOMALS to organise the model in such a way that both the homogeneity within groups of points and the heterogeneity between groups is maximised.

The particular geometric characteristics of HOMALS are summarised in Van Rijckevorsel and De Leeuw 1978, page 5, as follows (their terms 'subsets' and 'elements' having been replaced by 'categories' and 'villages'; the numbering is mine).

> [1]Categories and villages are in a joint space.
> [2]Villages that share most categories with other villages are *representative* and therefore *central* in space.
> [3]Villages that share the least categories with all other villages are *unique* and therefore *excentrical* in space.
> [4]Villages that share a unique group of categories are *homogeneous* and therefore *contiguous* in space.
> [5]Unique groups of categories are *heterogeneous* and therefore separated in space.

To the first characteristic the following note can be added: a village's score on a particular dimension is the sum of the scores of the categories which score on that village; a category's score is the mean of the scores of the villages which score on that category. Thus categories with a typical profile of scores on a series of dimensions are closely associated with villages that have a similar type of profile on the same dimensions, and conversely. This implies that if the characteristics of a group of villages can be interpreted, the clue is given to the interpretation of a group of categories, whereas villages can be typified by the qualities of the categories on which they score.

Let us now, after this general view of the procedures, turn to the new possibilities which the method offers, and exemplify these on the basis of concrete data.

3. SOME APPLICATIONS

3.1 Early studies

Very early applications of multidimensional scaling in Austronesian linguistics are to be found in two papers which were read at the Montreal Conference of May, 1973, by Paul Black and David and Gillian Sankoff (Black 1976; Sankoff and Sankoff 1976). Black successfully aimed at a spatial representation of the relationships between twelve dialect varieties of Bikol, a Philippine language. The two-dimensional configuration of the model, if superimposed on an atlas map, is strikingly congruous to the geographical position of the dialects. Sankoff and Sankoff explored the relationship between twenty-six Austronesian speech varieties in the Morobe Province (Papua New Guinea) in the same way and with very similar results. (cf. also the earliest application of this method by Henrici, in Henrici 1973, for the classification of twenty-eight Bantu languages).

Both studies were based on a set of lexical similarity percentages. They give proof that for a rough spatial representation of non-hierarchical relationships between dialects or languages both the type of data (the percentages) and the scaling technique are suited. It is important to realise, that if the positions of the points in the model (the dialects) had been considerably different from the positions on the geographical map, the result would have been equally valid. For an explanation of the differences one should then have looked for extralinguistic causes. Black has in fact done this in order to explain some minor discrepancies in the case of Bikol, by making a distinction between 'coastal' and 'mountain' dialects (cf. Black 1976:55).

3.2 Taking advantage of the joint space technique

My own investigation also began with geographical plotting of the scores of the individual varieties (i.e. the 470 villages) as a safe testing method. I made a separate map for the scores on each dimension. The variables were those of the first set mentioned in Section II, labelled HALS 1-50. Each of the five maps revealed at least one clearly patterned subarea. As an example I publish here the map for the third dimension, because it was the surprising interpretation of this map which made me decide to continue the analysis with the HOMALS program. As can be seen immediately, the positive-negative dichotomy in Map 3 (p.279) coincides almost perfectly with the administrative distinction between the area of DKI-Jakarta and the surrounding areas. Where exceptions occur the scores are zero, or in a few cases 1 or 2. We also see that the Mauk and Sepatan subdistricts are vaguely associated (i.e. with low scores), with DKI-Jakarta. Of the surrounding areas the western part is much more marked than the eastern.

The next step was to test the essentially new possibilities which HOMALS offers. Quite arbitrarily we selected for every dimension those categories which scored minimally 5 (+5 or -5). The occurrences of these categories we plotted also geographically. The evidence was clear: due to the representation of the villages and the categories in a joint space, high-scoring villages for a particular dimension showed the occurrence of a relatively high number of high-scoring categories for that same dimension. Again using a threshold value of 5, this time for the village scores, we made a combined map for the five dimensions and found several subareas which are distinguishable by a particular combination of positive or negative scores, as shown in Map 4 (p.280). This map demonstrates that (in terms of the fifty variables under study), there must be at least the following separate (sub)dialects: (a) Mauk + Sepatan, (b) Ciputat and surroundings,

(c) Gunung Sindur, (d) Cengkareng + Grogol Petamburan + Tanah Abang + Kebayoran Baru, (e) North-East Jakarta, and (f) Pasar Rebo. This also means that villages which belong to one of these (sub)dialects should be identifiable by their particular score profile on the five dimensions together; the same holds for the categories occurring in these villages. Below we shall give several examples of this identification procedure. We first deal with the procedure as it can be practised somewhat impressionistically by the researcher himself. For the second, more refined procedure, the help of further mathematical techniques is indispensable.

3.3 An example of rough grouping based on congruous score profiles

We now turn first to the list of village scores in the HOMALS output and try to find some unique patterns. The subdistricts of Mauk and Sepatan are clearly marked by the almost complete absence of the positive symbol in Map 4 (p.280). Moreover, inspection of the list of scores reveals that the all-negative score profile is uniquely found for the villages 435-460 (village 453 being eliminated because too many data are missing). As another unique feature of this group of villages we note the high or extremely high negative values for the second dimension. As we can see, these villages completely and exclusively cover the Mauk-Sepatan area. There is an abrupt transition which precisely coincides with the borderline between Sepatan and Teluk Naga. Some examples of the typical village score profiles are: -0 -27 -2 -7 -8 (village 449); -1 -14 -2 -3 +0 (village 437); a complete list is given in Table 7 and will be discussed later.

With regard to the values given here and throughout this paper and in the maps we should note that, on the basis of our general experience with the data, the distinction between negative and positive scores has been neglected for the values ranging from +1 to -1; thus village 437, which scores +0 on the fifth dimension, is not considered as deviating from the - - - - - score profile pattern. It should be noted that the figures we give for the scores are rounded figures. We write figures of HOMALS such as -0.012908, 0.058216, -0004503, etc., as integers: -1, +5, -0, etc.

On the second dimension the village scores for the Mauk-Sepatan group range from -27 to -7. The next lower village score is -5, which is found in the south-east of the total area, in village 80. Although this score indicates that village 80 probably has some features in common with the Mauk-Sepatan group, it cannot belong to this group, not because of its geographical remoteness, but because its score profile is different: +0 -5 +3 +5 +5. Such common features may have been independently borrowed from a common source, such as Sundanese, or the urban dialect, or Javanese.

Let us now study the categories which typically occur in the Mauk-Sepatan area. After some trying out we select those categories which score between -25 and -5 on the second dimension and which have negative values also on the other dimensions. These are the 22 categories listed in Table 5. There are 9 more categories with an all-negative score profile. These score between -3 and -1 on the second dimension. The total number of categories in this set of 50 variables amounts to 282 after the 50 final categories (missing data, etc.) have been eliminated. Table 5 shows the HOMALS category label of every individual category, its total frequency, its form and meaning, its score profile, and the possible source language(s). We now wish to take a very close look at the actual distribution over the villages of these categories. For that purpose we combine the villages and the categories under study in one large table (Table 6).

Table 5

Category score profile - - - - -; values for second dimension between -25 and -5. The HOMALS variable and category number, the total frequency of the category, its form, its meaning, its score profile, and the possible source language(s) are listed.

cat.	freq.	form, meaning	score profile	source lg.?
41.3	5×	apa maning *so much the more*	-2 -25 -1 -6 -8	Jav.
18.2	7×	wiji *(spinach) seed*	-1 -23 -1 -6 -8	Jav., Sd.
13.2	10×	anyar *new (shirt)*	-1 -20 -0 -4 -5	Jav., Sd.
6.9	11×	(pe)pedut *mist, fog*	+0 -19 -2 -3 -5	Jav.
17.9	5×	kepagut *scratched (by a thorn)*	-1 -19 -1 -4 -6	Sd., (Jav.?)
3.3	4×	atis (no context) *cold*	-0 -17 -1 -5 -5	Jav.
26.3	23×	gerah *having fever*	-1 -16 -1 -2 -2	Jav.
29.2	8×	buruan *yard (of a house)*	-0 -16 -1 -3 -2	Sd. Banten-Jav.
29.4	14×	karang *yard (of a house)*	-2 -15 -1 -3 -2	(Banten-) Jav.
5.3	25×	empuk *fat, grease*	-1 -15 -2 -3 -2	?
15.2	8×	berek, borok, burek *rotten*	-1 -14 -1 -1 -1	Jav.
45.2	32×	kakéongan, kiong(an) etc. *ankle*	-0 -12 -1 -2 -2	Banten-Jav.
32.10	18×	kuduk *(nape of) neck*	+0 -10 -0 -2 -2	from Java.
39.9	13×	(negation word +) urungan *undoubtedly*	-0 -8 +0 +0 -0	Jav.?, Sd.?
14.6	9×	pasti *right (answer)*	-0 -7 -0 -1 -2	Sd., Jav.?
27.5	11×	luas *wide (road)*	-3 -7 -3 -0 +1	?, Mal.
40.15	15×	(various exprr, with) kaga *there's not a bit left*	-1 -7 -2 -1 +0	Jav.?
32.1	59×	kamu (sekalian/semuah) *all of you*	-1 -6 -1 -1 +0	Mal.
40.12	13×	kaga (etc.) acan *there's not a bit left*	-0 -6 -1 -0 -2	Jav., Sd.?
8.5	39×	utuh *(still) in good condition (of a bicycle tyre)*	-0 -5 +0 -2 -0	Jav.
11.1	58×	(bulan) purnama *full moon*	-0 -5 -1 -0 +0	Jav., Sd.
42.2	12×	boro *let alone ...*	-1 -5 +0 -1 +0	Sd.

Table 6

HALS 1-50, villages scoring on 2nd dimension from -27 to -7, against categories scoring on 2nd dimension from -25 to -5. 0 = missing data.

villages	449	448	450	457	452	456	442	451	447	438	445	440	443	455	437	454	439	441	446	435	436	444	460	459	458	Mauk	Sepatan	subtotal	elsewhere	sum total	category score
categories																															
41.3	+	+		+	+	+																				5	0	5	0	5	-25
18.2	+	+	+	+	+	+								+												7	0	7	0	7	-23
13.2		+	+	+	+	+	+	+	+				o	+				o					o		o	7	2	9	1	10	-20
6.9	+		+		+		+		+	+	+		+					+	+				+			4	7	11	0	11	
17.9		+	+		+			+					o					o								4	0	4	1	5	-19
3.3	+	+	+																							3	0	3	1	4	-17
26.3	+		+	+	+	+	+	+	+	+		+	+		+	+	+	+	+	+	+	+	+	+	+	10	12	22	1	23	-16
29.2	+						+	+	+				+		+		+	o					o			2	5	7	1	8	-16
29.4		+	+	+	+	+		+				+				+		o		+	+		o	+		3	8	11	3	14	-15
5.3	+	+	+	+	+		+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	o	+	10	13	23	2	25	-15
15.2	+			+		+						+	o			+		o								4	1	5	3	8	-14
45.2	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+				9	13	22	10	32	-12
43.10		+	+	+		+	+					+	+			+		+								4	5	9	9	18	-10
39.9							+			+	+		+			+		+				o				1	5	6	7?	13?	-8
14.6					+			+						+												3	0	3	6	9	-7
27.5	+		+	+									o					o						+		4	0	4	6	11	-7
40.15	+	+					+		+			+		o							+			+	o	3	4	7	8	15	-7
32.1	+		+	+	+	+	+	+	+	+	+	+	o	+	+	o	+	o	+	+	+	+		+	+	9	11	20	39	59	-6
40.2										+	+				+		+			+		+				0	6	6	7	13	-6
8.5	+		+				+		+	+	+		+		+			+	+	+				+	+	4	9	13	26	39	-5
11.1	+	+	+	+	+	+	+	+		+	+	+	o	+	+	+	+	o	+			+	+		o	10	8	18	40	58	-5
42.2						+									+						+	+		+		2	3	5	7	12	-5
total	14	11	14	12	12	11	12	10	9	9	8	9	8	7	9	8	7	7	7	7	7	7	4	7	4			220	179	399	
missing data	0	0	0	0	0	0	0	0	0	0	0	0	6	1	0	1	0	8	0	0	0	1	3	1	3	total 106	total 114				
max.poss. total	14	11	14	12	12	11	12	10	9	9	8	9	14	8	9	9	7	15	7	7	7	8	7	8	7						
area	M	M	M	M	M	M	S	M	S	S	S	S	S	M	S	M	S	S	S	S	S	S	M	M	M						
village scores	-27	-26	-25	-25	-24	-23	-22	-20	-18	-17	-16	-15	-15	-15	-14	-14	-13	-13	-13	-12	-11	-11	-9	-8	-7						

3.4 Congruous profiles of categories and villages jointly tabulated

In Table 6 the rows show the category profiles and the columns the village profiles. The actual occurrence of a category in a particular village can be determined on the basis of the data matrix. In the first three columns at the right side of the table the number of occurrences in Mauk, in Sepatan, and in the total Mauk-Sepatan area is indicated for each category. The fourth column, labelled "elsewhere", indicates the number of occurrences in villages outside the Mauk-Sepatan area. The fifth column ("sum total") gives the total frequency of the category, and the final column its score on the second dimension.

If we read the columns labelled "subtotal" and "elsewhere" from the top to the bottom, we see that up to category 45.2 the number of occurrences in the Mauk-Sepatan area always exceeds (and nearly always very considerably exceeds) the number of occurrences elsewhere. From category 43.10 on we see the reverse develop. This gives proof that the categories which score between -25 and -12 on the second dimension are particularly typical for the Mauk-Sepatan area. We also checked the distribution of the occurrences "elsewhere", and we found no noticeable patterning. The categories which also occur outside Mauk and Sepatan are in the outside area scattered over approximately 150 villages, of which only 30 villages have 2 occurrences, whereas villages with 3 occurrences or more have not been found. The nine categories which have lower scores on the second dimension and therefore have not been included in Table 6, all do occur in the Mauk-Sepatan area, only with lower relative frequency.

At this point we can conclude that in this example the profile of the village scores is indeed closely associated with the profile of the category scores. We have found that the profile of a category *predicts* its occurrence in particular villages. In cases where the scores are less marked, the predictability is accordingly lower. We have been able to identify a Mauk-Sepatan (sub)dialect *and* its area on the basis of a *simultaneous* analysis of the village scores and the category scores. The use of geographical plots of the scores has greatly facilitated the discovery, but the same result could have been attained without any consulting of the maps.

3.5 A further subdivision of the Mauk-Sepatan dialect area

At the bottom of Table 6 the total occurrences per village are indicated in the row "total". The row "missing data" shows the number of missing data per village (which is relatively low in these two areas). The next row indicates the maximally possible total per village, which occurs if all the missing data in the column for a particular village represent one of the categories under study. The row labelled "area" refers to whether the village belongs to the Mauk or to the Sepatan subdistrict (M or S, repectively). In the final row the village scores are given.

From these figures we see that generally the village scores on the second dimension in Sepatan are lower than those in Mauk. The average village score for Sepatan is -14.6, and for Mauk -18.5. In Sepatan also the average number of categories (as included in Table 6) per village is lower: for Sepatan it is 8.2 (or 9.4, if all missing data are included), and for Mauk it is 9.5 (or 10.2). Out of the 22 categories only one (40.12) does not occur in Mauk, whereas six categories do not occur in Sepatan (41.3; 18.2; 17.9; 3.3; 14.6; 27.5). If we divide the categories into two groups, those scoring between -25 and -7, and those scoring -6 or -5, we find that the six categories which exclusively occur in Mauk, all belong to the higher-scoring group, whereas the one that exclusively occurs in Sepatan belongs to the lower-scoring group. Thus, if we compare Mauk with Sepatan, it is Mauk which is particularly typified by the highest scoring group of categories.

From Map 4 (p.280) as well as from the list of village scores it can be seen that there is reason to study the contrast between Mauk and Sepatan in further detail. Not only on the second dimension is there a noticeable difference between the scores, but also on the fourth and the fifth. We therefore list all the village score profiles in Table 7, separately for Mauk and Sepatan. For drawing Map 4 the value of 5 was chosen arbitrarily for all the dimensions. From Table 7 it is apparent that for the fifth dimension this value does very well, whereas for the fourth dimension a threshold value of 3 is the most suitable one. With regard to the categories, we also retain the value 5 for the fifth dimension. For the fourth dimension we hesitate between 3 and 4. If 3 is chosen, the following categories are to be included: HALS 41.3; 18.2; 13.2; 6.9; 17.9; 3.3 (see Table 5). The villages to be included are: villages 449, 448, 450, 457, 452, 456, 451, and 455. Again we set up a combined table for these villages and categories (Table 8).

The table shows that the villages which it includes form the typical area of a subvariety (in terms of the variable set HALS 1-50) of which the most typical representatives are the lexical variants apa maning, wiji, anyar, (pe)pedut, kepagut, and atis. Geographical mapping of this result yields a spatial coherent area of contiguous villages in the western part of Mauk (see Map 5, p.281). There could be some hesitation about including category 6.9 in the group of most typical categories, since it occurs 7 times in Sepatan and one of the 4 occurrences in Mauk is in village 460, which does not belong to the typical group of western Mauk. On the other hand this group is particularly typified by the highest scores on the second dimension, and (pe)pedut scores very high indeed (-19). Thus I would not eliminate (pe)pedut from the subvariety in question. It is a Javanese word, which was found by Nothofer along the borderline between West Java and Central Java, and also in the Sumedang area (etc.), not however in Banten (see Nothofer 1980, vol.2, map 16). This is an example of how the final decision about the grouping of the linguistic features or villages ultimately lies with the researcher and not with some automatic device beyond his control. If we exclude HALS 6.9 from the group, the limit value for the fourth dimension becomes 4 instead of 3.

3.6 A comparison with the traditional method

Now one might be a bit sceptical and ask whether the same result could not have been reached by the simple use of traditional word maps. This would involve the use of 50 maps, each with information on 470 points. In fact, for this experimental set, computer-plotted geographical maps for all the 50 variables have been made. Careful studying of these maps does indeed reveal that the Mauk-Sepatan area is a particularly patterned area, and one certainly would succeed in finding most of the typical features of the subdialect. But whereas the calculated grouping points to very clearly defined borderlines of the area and a very precisely and objectively definable membership of the categories of particular groups, the traditional method would leave us with many unanswered questions.

In the Mauk-Sepatan area more than 125 of the HALS categories occur. There would be no objective criterion for selecting the 22 most typical categories as shown in Table 6. This is easily understood if one realises that even of these selected features so many also occur outside the Mauk-Sepatan area (19 out of 22), whereas none covers all the 25 villages, and as many as 16 cover less or much less than half of the total number of villages in the area (see Table 6). The strict borderline found by HOMALS between Sepatan and Teluk Naga appears, as a rather fuzzy transition, only on 10 of the 50 maps. On the basis of simple counting of occurrences of individual features, as is done with the traditional method, one might attach more weight to a category such as HALS 26.3 (gerah), which occurs on 22 villages in the area and only once "elsewhere", than to apa maning (HALS 41.3) with 5 occurrences only in Mauk. From Table 6 we learn that it is apa maning which contributes most, both to the general pattern and to the

Table 7: Village scores in Mauk and Sepatan

vill.	Mauk					vill.	Sepatan				
448	-0	-26	-0	-7	-11	435	-1	-12	-1	-4	-0
449	-0	-27	-2	-7	-8	436	-3	-11	-2	-2	-0
450	-1	-25	-2	-5	-8	437	-1	-14	-2	-3	+0
451	-2	-20	-2	-3	-5	438	-1	-17	-2	-3	-1
452	-3	-24	-3	-6	-7	439	-1	-13	-2	-1	-0
453	+20	-0	-0	+1	+2*	440	-2	-15	-2	-2	+0
454	-2	-14	-1	-2	+1	441	+9	-13	-1	+0	-5*
455	+0	-15	+1	-5	-7	442	-3	-22	-2	-4	-3
456	-3	-23	-1	-3	-6	443	+9	-15	-0	-0	-7*
457	-3	-25	-1	-7	-7	444	-3	-11	-2	-1	+0
458	+1	-7	-3	-1	+3	445	-3	-16	-1	-2	+0
459	-1	-8	-0	-1	+2	446	-3	-13	-1	-2	-1
460	-0	-9	-3	+0	+0	447	-4	-18	-2	-3	-1

*Villages 441, 443 and 453 have 22, 20 and 49 missing items of data respectively. This causes the high positive scores on the first dimension.

Table 8: Villages and categories of the 'Western Mauk' subvariety

	vill.	449	448	450	457	452	456	451	455		total
categories										total	freq.
41.3		+	+	−	+	+	+	−	−	5	5
18.2		+	+	+	+	−	+	−	+	6	7
13.2		−	+	+	+	+	+	+	+	7	10
6.9		+	−	+	−	+	−	−	−	3	11
17.9		−	+	+	−	+	−	+	−	4	5
3.3		+	+	+	−	−	−	−	−	3	4
total		4	5	5	3	4	3	2	2		

distinction between the western Mauk variety and the rest of the Mauk-Sepatan area. We cannot objectively determine such differences, i.e. *measure* them, without the calculating of the astronomical number of associations which the computer program is able to do. In order to emphasise this we have chosen the Mauk-Sepatan area for this first example, since it is one of the most easily identifiable subareas of the total Jakarta Malay area.

3.7 Three further examples of the rough identification procedure

Let us now, very concisely, and primarily referring to the information contained in Tables 9-14, demonstrate how along the same lines what we shall call the 'Ciputat' dialect, the 'Gunung Sindur' dialect, and the 'Cengkareng + Grogol Petamburan + Tanah Abang + Kebayoran Baru' dialect can be identified.

Again with the help of Map 4 and on the basis of the - + + - - list of village scores, we select the profile pattern - + + - -, with values on the fourth dimension between -11 and -6 for the village scores, and between -8 and -4 for the category scores. Tables 9 and 10 contain all the information needed to identify the 'Ciputat' dialect and to draw its borderlines in Map 5 (p.281). The column labelled "catt." in Table 10 indicates for each village the number of occurrences of one or more of the categories included in Table 9. (In the same way the column "catt." in Tables 12 and 13 refer to the categories included in Tables 11 and 13, respectively).

For the Gunung Sindur dialect the profile pattern is - - + + -; the threshold values are very high: +28 and +25 on the fourth dimension for the village scores, and for the category scores +25 and +10. Tables 11 and 12, and Map 5 show the grouping.

Finally, Tables 12 and 13, and Map 5 again, indicate how the remarkable dialect zone which includes Cengkareng, parts of Grogol Petamburan and Tanah Abang, and practically the whole of Kebayoran Baru, are identified. The typical pattern is here - + - - -, with village scores between -15 and -7 on the fifth dimension, and category scores between -10 and -5.

Rather than multiplying this kind of example, we conclude this section by referring to Table 15, in which for one or two particular representative villages of each of the five dialects which we have identified in this way, the full list of variants is given which score on the village in question. The table gives also the typical score profile of the villages, but only dichotomously, i.e. positive or negative scores for each dimension, without threshold values. A plus sign after a variant indicates that the category's score profile, dichotomously, corresponds with the village score profile.

3.8 The use of advanced clustering techniques

What has been demonstrated in the above sections is in fact the application of a rough clustering technique on the score profiles. It seems possible to identify in this way the most well-marked and homogeneous score patterns, but much more refined methods are needed to detect the lesser marked patterns and their distribution. The use of rounded figures instead of the calculated real figures of the HOMALS output means the loss of much valuable information. Therefore I have been using several computer programs for the clustering of the profiles. One well-known problem with these techniques is that usually several solutions are offered which are, mathematically, equally acceptable, whereas

Table 9: The 'Ciputat' dialect

Category score profile: - + + - -; values for fourth dimension between -8 and -4. The HOMALS variable and category number, the total frequency of the category, its form, its meaning, its score profile, and the possible source language(s) are listed:

cat.	freq.	form, meaning	score profile	source lg.?
9.3	27×	belagu berani *bullying*	-1 +2 +3 -5 -3	Jav.
10.2	45×	lanang *man*	+0 +1 +6 -4 -0	Jav., Sd.
13.4	6×	cakep *new (shirt)*	-0 +2 +4 -7 -1	?
28.3	13×	sapet, sepet *wing*	+0 +2 +1 -5 -0	?
28.4	42×	sewiwi, siwi *wing*	-0 +1 +7 -5 -1	Jav.
31.2	22×	gawéan *work*	+1 +1 +4 -4 -4	Jav.
39.7	30×	udah tentu *undoubtedly*	-1 +1 +3 -4 -0	cf. Mal., Jav., Sd.
40.2	5×	ora ada dikit *there's not a bit left*	-0 +2 +6 -9 -0	ora: Jav.
41.6	9×	lebih-lebih *the more so as*	-0 +1 +4 -5 -2	cf. Mal., Jav., Sd.
43.6	46×	(ge)gitok *(nape of) neck*	-0 +2 +3 -4 -1	Jav.
50.5	17×	(k)a(n)til-(k)a(n)til(an) *uvula*	-1 +2 +7 -8 -2	cf. Jav.

Table 10: The 'Ciputat' dialect area

Village score profile: - + + - -: values for fourth dimension between -11 and -6; number of categories as included in Table 9.

vill.	score profile	catt.	vill.	score profile	catt.
23	-2 +5 +11 -10 -5	4	422	-0 +1 +7 -6 -0	3
24	-2 +4 +12 -11 -4	5	423	-0 +3 +11 -6 -4	3
25	+0 +3 +8 -6 -3	5	424	-0 +4 +10 -6 -4	3
26	-2 +4 +11 -10 -4	5	425	+0 +0 +6 -6 -0	4
27	-2 +4 +11 -10 -4	3	426	+0 +3 +7 -7 -3	3
28	-1 +2 +6 -6 -3	5	427	-0 +2 +9 -10 -4	5
408	+0 +3 +7 -9 -1	5	428	-2 +5 +5 -7 -5	4
409	-2 +3 +7 -11 -2	6	429	-1 +2 +10 -11 -5	6
415	-1 +2 +6 -7 -1	3	430	-2 +3 +6 -6 -5	5
416	-2 +2 +6 -10 -1	6	431	-2 +2 +8 -10 -5	6
417	-0 +3 +11 -6 -3	5	432	-1 +1 +8 -8 -4	5
418	-0 +3 +10 -6 -1	4	433	-1 +2 +9 -9 -4	6
420	-1 +3 +7 -10 -2	5			

Table 11: The 'Gunung Sindur' dialect

Category score profile: - - + + -; values for fourth dimension between +25 and +10. The HOMALS variable and category number, the total frequency of the category, its form, its meaning, its score profile, and the possible source language(s) are listed:

cat.	freq.	form, meaning	score profile	source lg.?
6.10	9×	mèga *mist*	-6 -2 +13 +23 -6	(Jav., Sd.: *cloud*)
31.4	11×	kejaan *work*	-1 -0 +4 +20 -2	?
35.1	10×	alukan *(rather than...) it would be better to...*	-2 -0 +8 +10 -2	Sd., Jav.
39.10	8×	ora/kaga wurungan *undoubtedly*	-6 -2 +13 +25 -6	Jav.
44.3	8×	gegusi *molar (tooth)*	-6 -2 +13 +25 -6	?
45.5	8×	(me)muncangan *ankle*	-5 -3 +13 +24 -5	Sd.
47.1	13×	pelangkakan *groin*	-3 -1 +6 +11 +0	Sd.
47.3	4×	pengpelangan *groin*	-5 -2 +9 +21 -4	Sd.

Table 12: The 'Gunung Sindur' dialect area

Village score profile: - - + + -; values for fourth dimension between +28 and +25; number of categories as included in Table 11.

vill.	score profile	catt.	vill.	score profile	catt.
2	-6 -2 +15 +26 -6	6	(6	-4 -1 +8 +13 -4	3)
3	-7 -2 +15 +25 -6	6	7	-6 -3 +13 +25 -7	5
4	-6 -2 +15 +25 -5	6	8	-7 -2 +12 +27 -7	4
5	-7 -2 +14 +28 -6	7	9	-6 -2 +13 +28 -6	6

Table 13: The 'Cengkareng-GrPetamburan-TnAbang-Kebayoran Baru' dialect

Category score profile: - + - - -; values for fifth dimension between -10 and -5. The HOMALS variable and category number, the total frequency of the category, its form, its meaning, its score profile, and the possible source language(s) are listed:

cat.	freq.	form, meaning	score profile	scource lg.?
4.6	19×	gebleg (geblek)/tolol *stupid*	-5 +4 -7 -0 -10	Jav.
14.7	24×	persis, percis *correct (answer)*	-5 +4 -7 -0 -9	Jav., Sd., Dutch
17.7	29×	barèt, barèd *scratched (by a thorn)*	-3 +3 -7 -0 -7	Sd., Balin.
36.4	6×	demènin, deminin *let (him) be; that'll do*	-4 +4 -3 -2 -7	Jav.?
38.2	13×	kaga bakal *(he is) not prepared to (go)*	-3 +3 -7 -0 -6	Jav.?
40.6	22×	kaga ada barang sedikit *there's not a bit left*	-3 +3 -6 -0 -5	(barang:) Jav.?
44.4	14×	panggal, (gigi) pangkal *molar (tooth)*	-4 +4 -7 -0 -7	Balin.

Table 14: The 'Cengkareng-GrPetamburan-TnAbang-Kebayoran Baru' dialect area

Village score profile: - + - - -; values for fifth dimension between -15 and -7; number of categories as included in Table 13.

vill.	score profile	catt.	vill.	score profile	catt.
254	-5 +4 -7 +0 -14	2	190	-7 +5 -10 -0 -13	4
255	-4 +4 -7 -0 -10	2	191	-6 +5 -9 +0 -12	4
256	-5 +4 -7 +0 -12	3	192	-4 +5 -7 -0 -7	2
257	-5 +4 -6 +0 -11	2	193	-5 +4 -9 +0 -7	2
258	-5 +5 -8 +0 -13	6	194	-4 +4 -8 +0 -8	2
259	-5 +4 -8 +0 -13	3	195	-5 +4 -7 -0 -8	3
262	-6 +5 -10 +0 -15	6	196	-6 +5 -8 -1 -11	4
263	-6 +4 -9 +0 -11	3	197	-5 +5 -8 -0 -12	4
264	-6 +6 -9 +0 -13	6	198	-4 +4 -6 -0 -9	1
265	-6 +4 -9 +0 -10	4	199	-5 +4 =6 -1 -7	2
266	-6 +5 -9 -0 -13	5			
121	-6 +5 -10 +0 -14	4			
122	-6 +5 -9 -0 -14	4			
124	-6 +5 -9 -0 -12	4			

different programs give often largely overlapping solutions. Clustering also takes relatively much computer time, and the larger data sets could not directly be handled by the smaller computers on which some of the experiments were done.

Rather than going into these yet unsolved problems, I would like to demonstrate that if it is possible to find a satisfactory solution, i.e. to identify clusters which are found to be constant through different experiments and from different initial positions, the clustering approach yields very good final results. It should be realised that a satisfactory solution does imply the (mathematically interpretable) existence of fuzziness within sets, which may lead to the calculation of a varying degree of membership of subsets (clusters), and thus also to the insight that borderlines between groups are often fuzzy. It is obvious that the mathematical studies which are going on in this field are highly relevant for any research which aims at the analysis of empirical data (see Backer 1978a, 1978b).

The only detailed clustering operation on which I can report here concerns a set of 126 category score profiles over five dimensions. The 19 variables refer to fishing tools. At the end of the clustering procedure eight clusters were identified. Their standardised distribution over the villages was calculated (the standardising at the same time solves the problem of mapping the missing data), and the result was geographically plotted. With some modifications for technical reasons, which do not affect the value of the map as an example, Map 6 (p.282) shows the geographical distribution of the eight clusters. At the end of this procedure we listed for every cluster the distribution of its individual members over the villages, and we tried to find an implicational patterning, in view of Bailey's theory mentioned in Section 1. So far we cannot report any positive result. We should, however, keep in mind that this set of variables has a very high number of missing data, up to an average of 45 per cent, which is due to the technical character of the variable set, and to some degree to ambiguity caused by the mediocrity of the pictures used in the questionnaire.

This detailed analysis of a limited set of particularly weak data, which is very incompletely dealt with here, seems to justify the conclusion that careful clustering of the HOMALS category scores is a valid and efficient way of grouping the linguistic features according to their distribution over the villages. The approach via the category scores is more precise than if the village scores are clustered, because the missing data are integrated in the village scores, whereas for the clustering procedure they can be eliminated from the list of category scores, being a category of their own.

3.9 Dimensionality and the interpretation of the individual dimensions

Another possibility, which I have not explored so far, would be to increase the number of dimensions on which the HOMALS analysis is carried out. With the HOMALS technique it is left to the user of the program to choose the dimensionality, and in our case the number of five dimensions was quite arbitrarily chosen. HOMALS has proved to be an extremely precise technique and it is particularly devised to maximise the coherence within groups. I cannot yet estimate which technical problems would arise if one should deal with an output showing the scores for, say, ten dimensions (and how the output presentation would be organised), but further research in this direction seems needed.

Since we have not looked so far at the individual dimensions, let us return to the data set of HOMALS 1-50 and discuss each of the dimensions briefly. Map 4 reveals that every dimension marks at least one contiguous dialect area: the first

Table 15

Some sample villages. The labels refer to the 'dialect', the village number and the dichotomous score profile. Items marked by + have the same dichotomous score profile as the village.

	'Ciputat'		'Cengk.-GrPetamburan-TAbang-Keb.Baru'		'Gg.Sindur'
Variables	417 - + + - -	429 - + + - -	262 - + - - -	190 - + - - -	5 - - + + -
1. *big*	gedé	gedé	besar	gedé	gedé
2. *cloud*	asep	awan	awan	awan	mèga+
3. *cold*	dingin	adem	dingin	dingin	dingin
4. *stupid*	gebleg	tolol	tolol/bebel	geblek	bodoh
5. *fat, grease*	gajih	gajih	minyak+	minyak+	gajih
6. *mist, fog*	ampak2	ampak2	ampak2	asep	mèga+
7. *who*	siapah	—	sapah	siapê	siapa
8. *not worn out (tyre)*	bagus	utuh	bagus	bagus	bagus
9. *bullying*	belagu berani+	belagu berani+	—	lagè	belaga
10. *man*	lanang	—	lelaki	lelaki	lelaki
11. *full moon*	bulan 14-nya	bulan terang	bulan terang	bulan terang	tanggal 14
12. *narrow*	seseg	seseg	sempit	sempit	sempit
13. *new (shirt)*	baru	baru	baru	baru	baru
14. *right (answer)*	jitu	bener	percis+	persis+	jètu/cocok+
15. *rotten*	busuk	lodoh/busuk+	busuk	busuk	busuk
16. *round*	bulat	bunder/bulat	bunder/bulet	bulet/bunder	bulet
17. *scratched*	kebarèd	—	barèd+	barèd+	kebarèt
18. *(spinach)-seed*	biji	biji	biji'	biji	biji
19. *dull*	kedul	pudul	pudul+	pudul+	mintul+
20. *small*	kecit	kecil	kecil	kecil	kecil
21. *straight*	lempeng	lempeng	lempeng	lempeng	lempeng
22. *there (far off)*	di sono+	sono+	di sonoh	di sonoh	di sonoh
23. *there (near by)*	di sono	sono	di situh	di situh	di situh
24. *they*	merèka	—	diah+	diè'	dia

25. *hot (water)*	panas	panas	panas	panas	panas
26. *having fever*	panas	panas	panas	panas	panas
27. *wide (road)*	lèbar	lèbar	lèbar	lèbar	lèbar
28. *wing*	sewiwi+	siwi+	sayap	sayap	sayap
29. *yard*	—	latar+	latar	latar	pekarangan
30. *woman*	wadon	wadon	perempuan	perempuan	wadon
31. *work*	pegawéan+	gawéan	kerjaan	kerjaan	kejaan+
32. *all of you*	eluh semuanya	lu	luh semuanya	lu semuènyè	luh semuahnya
33. *how could...*	abong	(i)lokan+	masa iya...	apè iyè...+	ilokan/abong
34. *only because*	abong2+	abong2+	abong2	abong2	abong2
35. *it would be better to...*	mendingan	angguran	mendingan+	mendingan+	mendingan/angguran+
36. *let (him) be*	bagènin	bagènin	deminin	...ajè	...baé
37. *his mother*	—	—	nya'nyah+	nyaknyè'+	ibunya
38. *is not prepared to*	bader	ora bakalan+	kaga' bakal+	kagè bakal+	moal
39. *undoubtedly*	ora kudu+	—	mesti+	udè pasti	kaga wurungan+
40. *there's not a bit left*	ora pisan+	—	kaga barang dikit	abis bener	kaga pisan
41. *so much the more*	—	lebih2	apalagi+	apalagi+	komo lagi+
42. *let alone...*	—	boro2	boro2+	apè lagi+	boro lampar+
43. *neck*	tengkok	gitok+	tengkok+	tengkok+	tengkok
44. *molar*	baham	baham	baham	panggal+	gegusi+
45. *ankle*	mata kaki	mata kaki	mata kaki	mata kaki	muncangan+
46. *joint*	ugel2	pergelangan	ugel2an	ugel2	pegelangan
47. *groin*	selangkangan	pikang	pikangan+	pikangan+	pelangkakan
48. *glands in the groins*	sèkèlan	sèkèlan	kelanjeran+	klanjeran+	sèkèlan
49. *middle finger*	jari tengah	jari tengah	jeriji tengah+	jeriji tengah+	jari tengah
50. *uvula*	kantil2+	antil2an+	lak2an	lak2an	elak2an

Table 16

The highest scoring categories on the fourth dimension.
(For meanings, see Table 15.)

Var./Cat.		Positive scores	source lg.?
44.3	(ge)gusi	+25	?
39.10	ora/kaga wurungan	+25	Jav., Sd.
45.4	(me)muncangan	+24	Sd.
6.10	mèga	+23	Sd., Jav.
47.3	pengpelangan	+21	Sd.
47.1	pelangkakan	+11	Sd.
31.4	kejaan	+10	?
35.1	alukan	+10	Jav., Sd.
19.3	mintul	+8	Sd.
14.4	jitu/jètu	+7	Sd.
40.8	kaga pisan	+7	Jav., Sd.
42.10	boro lampar	+7	Sd.
29.5	pe(/pa)karangan	+6	Jav., Sd.
42.4	boro ampar (cf. 42.10)	+5	Sd.
38.10	moal	+4	Sd.

Var./Cat.		Negative scores	source lg.?
38.3	ora bahannya	-10	Jav. (ora)
40.2	ora ada dikit	-9	Jav. (ora)
33.5	lokan	-8	cf. Jav. ilok(an)
50.5	(k)antil2an, etc.	-8	Jav.,?
13.4	cakep	-7	?
18.2	wiji	-6	Jav., Sd.
41.3	apa maning	-6	Jav.
3.3	atis	-5	Jav.
9.3	belagu berani	-5	cf. Jav. gawé iagu
28.3	sapet	-5	?
28.4	sewiwi, siwi	-5	Jav.
41.6	lebih	-5	Jav., Mal., Sd.
9.6	pura, pura2	-4	Jav.? Mal.?
13.2	anyar	-4	Jav., Sd.
31.2	gawéan	-4	Jav.
39.7	udah tentu	-4	cf. Jav., Sd.? Mal.
43.6	gitok	-4	Jav.

dimension (positive) marks Pasar Rebo; the second dimension (negative) marks Mauk and Sepatan; the third dimension (negative) marks the area of DKI Jakarta; the fourth dimension (positive) marks Gunung Sindur, and finally, the fifth (positive) marks the eastern part of Jakarta proper. This list is not complete, but suffices for forming some idea of what we may expect, and particularly what we may *not* expect when trying to interpret a particular dimension. From the simple inspection of the content of some typical villages, as listed in Table 15, we learn that all these villages have a very mixed vocabulary. Now if each dimension is at least associated with an area with much mixture of Javanese, Sundanese, Malay, Balinese and other elements, it is improbable that any of our five dimensions will show a clear-cut dichotomy between, say, Javanese and Sundanese elements, Banten Javanese and Central-Java Javanese, Malay and non-Malay elements, etc. Moreover, the dimensions with HOMALS are not selected according to some underlying principle chosen by the researcher.

On the other hand, some areas are clearly more influenced by Javanese, for example, than others, and this fact could be reflected by the dimension on which these areas have particularly marked scores. This indeed is what we find if we contrast the highest scoring categories on the fourth dimension on the positive side with those on the negative side. Table 16 shows clearly more Sundanese elements on the positive side, and more Javanese elements on the negative side, and this pattern continues even if we come to the very low scores. On the basis of this general, though not absolute pattern, we can *predict* that such a *Javanese*-looking form as ora/kaga wurungan, meaning *undoubtedly*, has reached Jakarta Malay as a borrowing from *Sundanese*. The form wurung (=burung) does indeed occur in Sundanese (as a borrowing from Javanese) with the meaning *it didn't work out*. A parallel case is barèd/barèt (HALS 17.7) which has the scores -3 +3 -7 -0 -7. With the meaning *scratched (by a thorn)* it occurs both in Sundanese and in Balinese. The high negative score on the third dimension shows that it is a typical word from the urban area (it occurs indeed exclusively in Jakarta), and the zero score on the fourth dimension predicts that it has nothing to do with the contrast Sundanese-Javanese. Therefore we may assume that the source language is Balinese.

A more puzzling case is HALS 38.11, bader *(he is) not prepared to (go)*. This variant has the scores +6 +1 +2 +0 -6. The high positive score on the first dimension, combined with the high negative score on the fifth dimension, immediately points to the Pasar Rebo region (cf. Map 4). We are not surprised to find bader, which has a total occurrence of 23, as often as 16 times in Pasar Rebo. Now Pasar Rebo, in accordance with the contrast on the third dimension (see Map 3), belongs to the urban area, although some of its villages, seven in all, score +0. The low positive score of bader on the third dimension shows that it is somewhat more associated with 'rural' features than with 'urban'. The other 7 occurrences are indeed in villages which have an average score of +8 on the third dimension (villages 31, 417, 423, 424, 465, 467, 469 in Sawangan, Serpong, Ciputat and Ciledug). Within the area of Pasar Rebo bader occurs in 6 of the 7 villages which score +0. (The seventh village has no data for this variable.) (See Map 3.) Since we know, (both from a further analysis of the third dimension and from a simple geographical plotting of the frequency of occurrences of a number of Balinese features) that, generally speaking, the Balinese element in

Jakarta Malay has spread from the urban to the rural area, the positive score of bader on the third dimension does not make Balinese the most probable source language for bader. We have not been able indeed to find a parallel in Balinese. On the other hand, bader could be easily explained as a borrowing from Sundanese. In Sundanese badeur means *unmanageable, disobedient*. Javanese and Old Javanese do not seem to have any direct parallels. What is puzzling, however, is that in the case of a strong association with Sundanese we would expect a rather high positive score on the fourth dimension instead of the completely neutral +0. Should we seek then the origin of bader in Banten rather than in the Sundanese area?

In the majority of such cases as bader a straightforward historical interpretation on the basis of individual dimensions is not possible in an area where so much mixture exists and in which relatively recent migration plays such an important part as is the case in the Jakarta Malay area. One would constantly have to look for additional data from outside the area. Moreover, as is generally attested in the literature (cf., e.g., Kruskal and Wish 1978:30), the difficulty of interpreting the dimensions is inherent in the multidimensional scaling techniques. The making of few assumptions makes the interpretation load heavier for the researcher. It seems therefore methodically more fruitful to concentrate first on the most complete possible grouping of the linguistic features and the villages, before new information from outside is called in.

We conclude this section with two more general observations. The first is, that extremely high scores on one side which are not counterbalanced by (rather) high scores on the other side of a particular dimension, indicate that the high scoring features are contrasted to *all* the other features, so that not much can be expected from a detailed analysis of the low scoring side. This is the case with the second dimension in the HALS 1-50 set, where the maximum negative category score is -25, and the maximum positive score +4. Thus the Mauk-Sepatan area and dialect are set apart as a typical group over against the total remaining area and its typical features.

The second observation regards the positive side of the first dimension. In this set as well as in any other set we have analysed so far, HOMALS groups the missing data categories on one side of the first dimension. This means that villages with many missing data score very high on that side (see Table 7). This is a very convenient warning to the user that his data are unreliable from a particular point of view. It makes the geographical mapping of the village scores for the first dimension more complicated, but since the missing data form a separate category, the grouping of the categories is not affected. HOMALS also may bring out other errors in the data. In a set of phonological items, where the questions had been administered as a multiple choice, the informants had often given more than one form. I found very high scores for a long series of such double answers, and I expected to find a nicely patterned transition area. What I discovered instead was that the villages which scored so high exactly coincided with those places where one particular fieldworker had been collecting the data. He had been either too insecure or too insistent, but anyhow HOMALS had him taped.

4. CONCLUDING REMARKS

4.1 The objectivity of the method

The method as described above requires a very orderly arrangement of the data, it does not eliminate data before their exact position within the total configuration has been determined and evaluated, and, generally, the more subjective and intuitive judgements come at a later, much better prepared stage as compared with the traditional methods. Above all, a very precise quantification leading to the measuring of the differences replaces the more subjective estimates made by the researcher. No incommensurate external data are called to one's aid before the analysis on the data contained in the matrix is finished.

The method even may detect inconsistencies, discrepancies and other errors or flaws in the data themselves. The precarious dependency of the researcher on probabilistic methods for solving the problem of missing data (which are, in terms of the empirical data, entirely unpredictable) is overcome, since village- and category-points are only defined by the non-missing entries of the data set (cf. van Rijckevorsel and De Leeuw 1978:7). The HOMALS technique is a highly objective tool for finding latent structures even in very weak data sets. It is sufficiently known (though not always sufficiently realised) how weak data sets are which comprise information with regard to reported or observed linguistic behaviour (cf. Moulton 1968:461 ff, disputed in Kurath 1972:16). Dialect data based on the use of questionnaires are especially weak in as far as it is generally impossible to determine the position of a given variant with regard to other, alternative, variants in the informant's repertoire.

Quite apart from the subjective elements involved in the process of data collecting, however, and also leaving out the final stage where, as always in the case of empirical data, the researcher's own judgements are decisive (the two poles are far from mutually independent, of course), I see three phases in the processing of the data where subjectivity is practically unavoidable and should be kept at a minimum. Chronologically, the second and third of these phases are the choice of dimensionality and the linking of different data sets; the first phase will be discussed last: it regards the reduction of the field data into mutually exclusive and exhaustive categories.

Although it is true that the initial choice of the number of dimensions on which the analysis is carried out is very arbitrary with HOMALS, it is always possible to increase the dimensionality as long as the results seem to justify this. There is, however, no measure for determining the best dimensionality; only relative importance of each individual dimension can be seen from its stress value (stress indicates the 'goodness of fit' of the model).

With regard to the third phase, I can only say that I have not yet explored practically, and that mathematically it is not yet known to what extent HOMALS solutions can be mutually compared. (In some other techniques, such as factor analysis, the rotation of the configurations is applied). If a regular correspondence between the solutions for different sets of variables could be found, the variables could be combined in an objective way, and the highly subjective way of combining the variables in particular sets (as done by myself) could be replaced by a structurally determined selection. I am thinking particularly of the possibility of selecting the best variables in order to study the distribution of semantic fields, or to compare distributions on different levels of description (phonological, morphological, etc.).

Finally, how can we keep the first problem under control? The reduction of the field data into a limited number of categories implies many decisions with regard to the compatibility of variants. One may choose a linguistic criterion, such as the strictly lexical, or strictly phonological, character of the data set. But there will always remain many cases of doubt: should wurungan and urungan be kept apart (see above), or itam and item *black*, (as noted by the fieldworkers), or encè' and enci'? Both mean *father's younger brother*, but have been found to come from different source languages; (see for details Grijns 1980). For the first set of 50 lexical variables I have consulted all the geographical maps. Later I found the following method giving the best results. All variants which occur more than three times are classed in a separate category. Particularly interesting variants with a total occurrence of only three of even two are sometimes kept apart. HOMALS usually gives satisfactory results for my data even with such low frequencies. After the HOMALS analysis has been carried out, it is decided on the basis of the category scores whether further reduction is warranted. This is the case if the score profile of two or more of the categories is almost identical, and, of course, if there are no linguistic considerations to keep the variants apart. In some cases, if the scores of all the categories are extremely low on every dimension, even the whole variable may be eliminated from the set, although the fact of almost random distribution contains in itself important information. Since the variables HALS 1-50 were later included also in other sets, often with a different classification of the variants, the applicability of the above sketched procedure has been amply tested. Thus in the case of the variable baru, which was originally grouped into five categories, as shown in Table 2, where the forms baru and baru' were combined in category 1, in another set of (lexical) variables baru' was set apart and scored sufficiently differently from baru to justify its retention as a distinct entity.

4.2 Possible relevance for historical linguistics and sociolinguistics

The early investigations as mentioned in the first section of Chapter 3 were all carried out in the context of historical linguistics. What a technique such as HOMALS can offer here is considerable: a check on cognation or compatibility (since very similar score profiles mean very similar associations with all the other features under study), a much fuller use of the available information (no more cognation *percentages*, but direct scores for all the individual features), and a simultaneous patterning of the varieties (dialects, languages or even larger groups) and the variants.

For sociolinguistics the same possibility exists of simultaneous grouping of the informants (without previous classification according to social groups) and of the individual variants they use. Since the program can handle large data sets, the difficulty of how to select the best variables can be overcome. One does not need to begin with some few selected variables and then build up the set, as Thelander was forced to do in his article on code-switching or code-mixing (Thelander 1976). In that case one has to move from the more obvious to the less obvious variables, and it may become increasingly difficult to enlarge the set somewhat objectively, whereas the HOMALS technique applies an objective procedure of reduction, and one is free to include in the initial data set any features which seem to be of interest for the distinguishing of the speech varieties under study.

Multidimensional scaling is not a panacea for all problems of patterning of linguistic features and informants (cf. Berdan 1978, where the application of multidimensional scaling and the related technique of principal components analysis is compared in the case of five variants of one vowel variable; it should be noted, however, that HOMALS can be seen as a non-linear form of principal components analysis, cf. Van Rijckevorsel and De Leeuw 1978:1 and 2). In such a complex variation as we find in the Jakarta Malay area, and especially if dealing with lexical variables, one would hardly even think of the possibility for an algorithm to *generate* variants. But as has been demonstrated, for our data the HOMALS procedure has unmistakably considerable *predictive* power as to linguistic patterning.

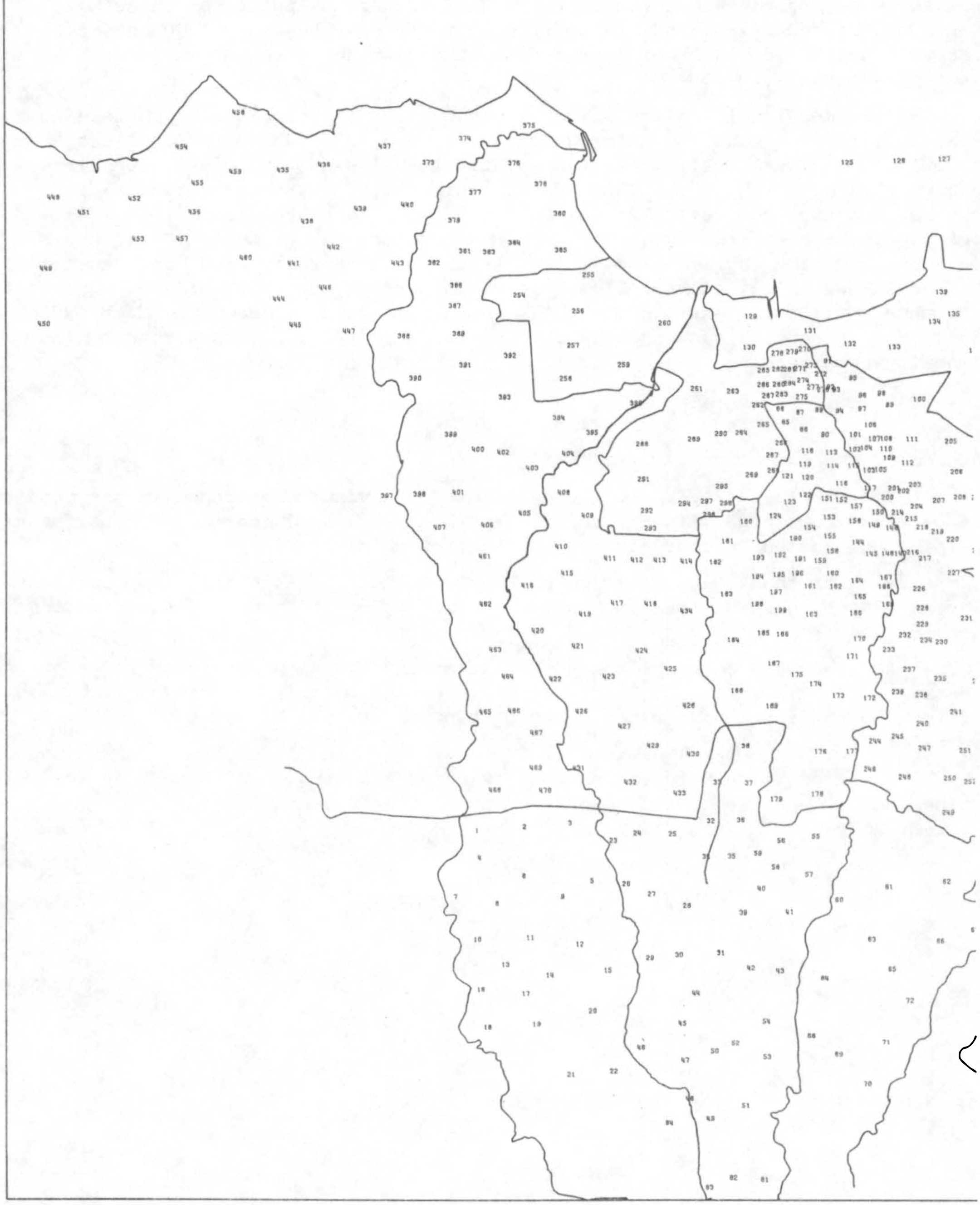
Acknowledgements

I am grateful to L.J. Th. van der Kamp for giving me guidance in the process of getting acquainted with the techniques which I am applying, as well as for his helpful suggestions with regard to the text of this paper.

I wish to thank Mike Tjok Joe of the Department of Scientific Applications in the Computing Centre, Leiden University, who has carried out all the programming work for me.

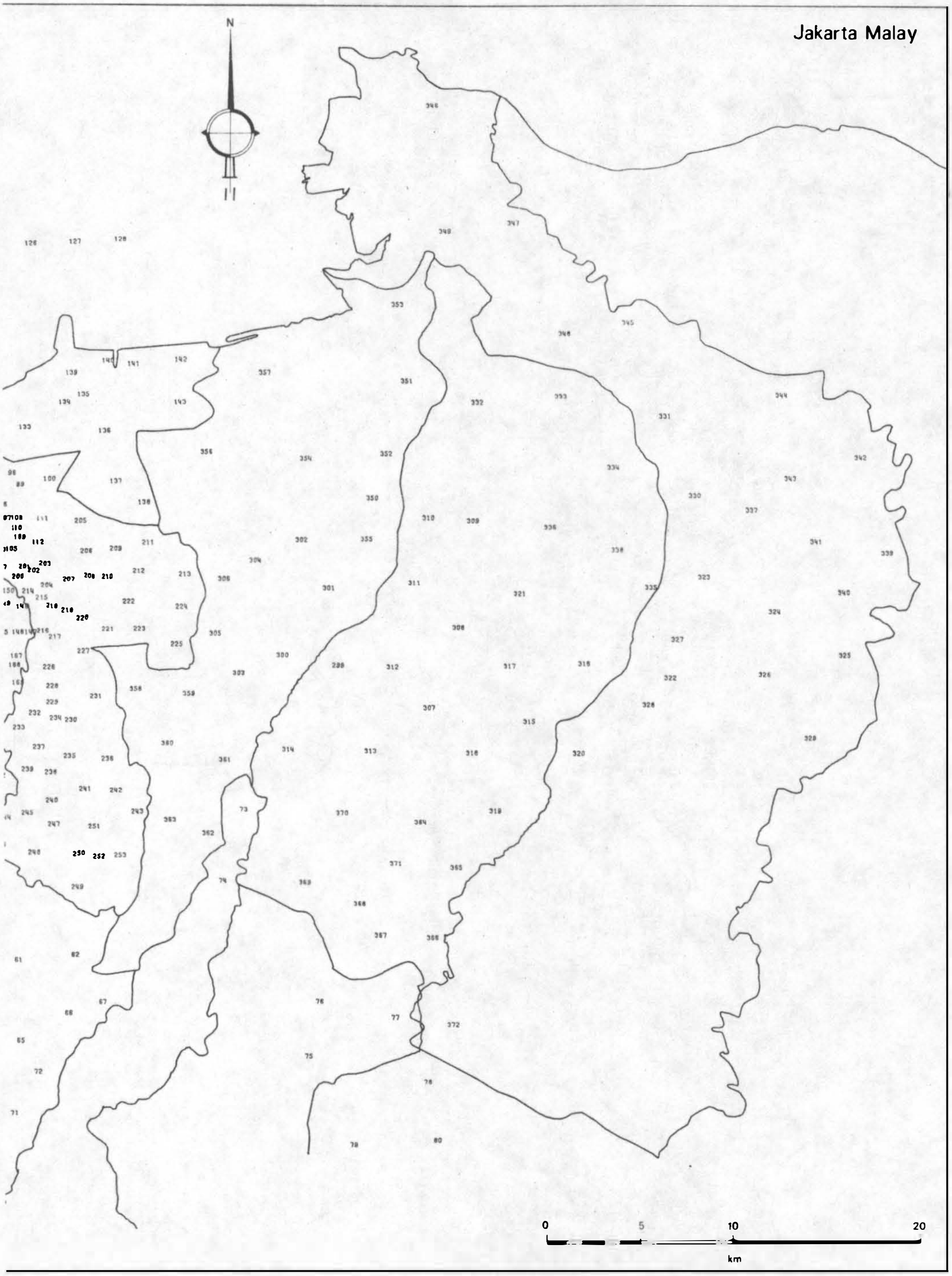
I am also very much obliged to E. Backer and H.P.A. Haas of the Information Theory Group, Department of Electrical Engineering, Delft University of Technology, for making the detailed clustering studies on the set of 'fishing tools' variables.

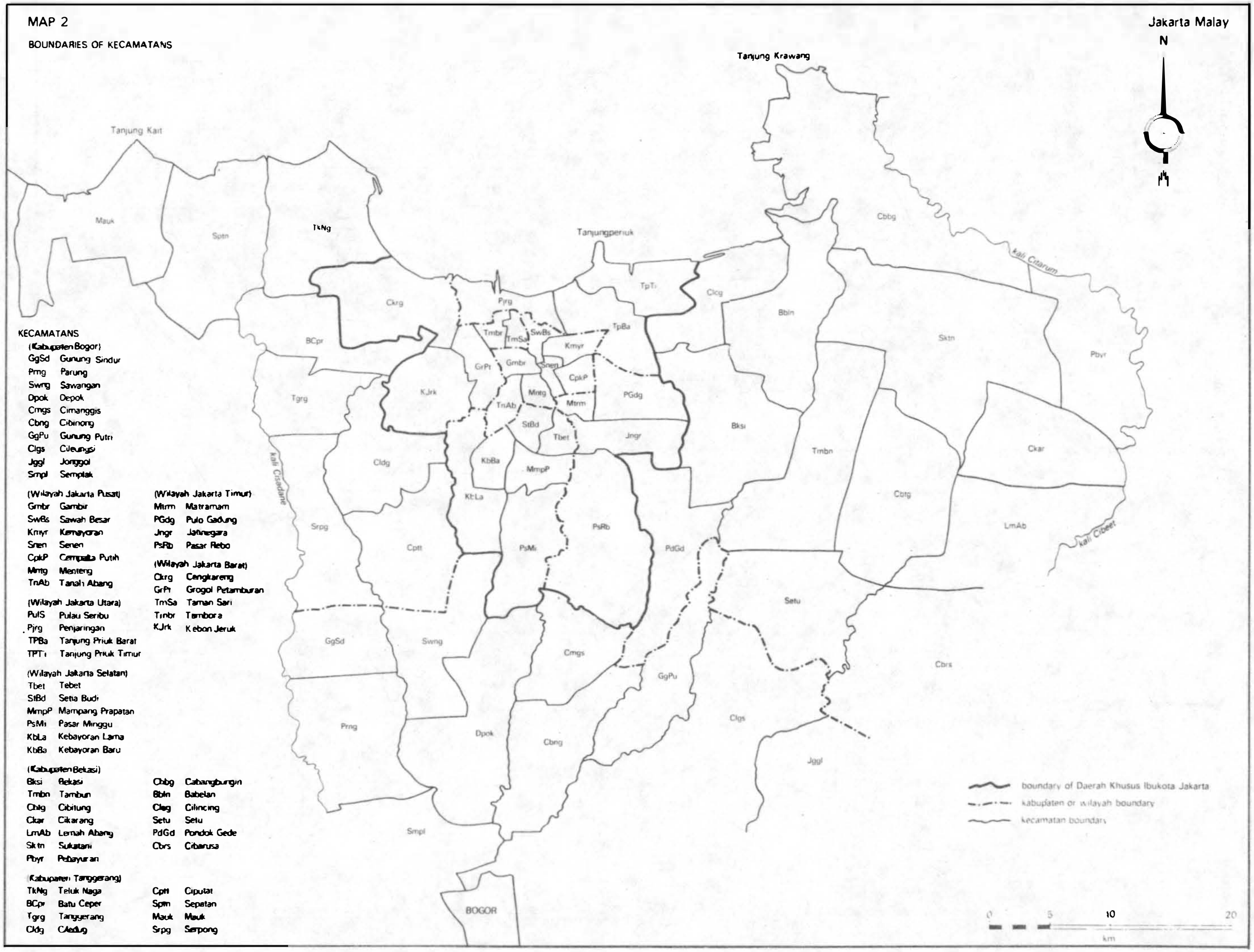
The fieldwork on which this paper is based was supported in part by the Netherlands Foundation for the Advancement of Tropical Research (WOTRO; file W 38-2).



MAP 1

NUMBERS OF VILLAGES INCLUDED IN THE SURVEY

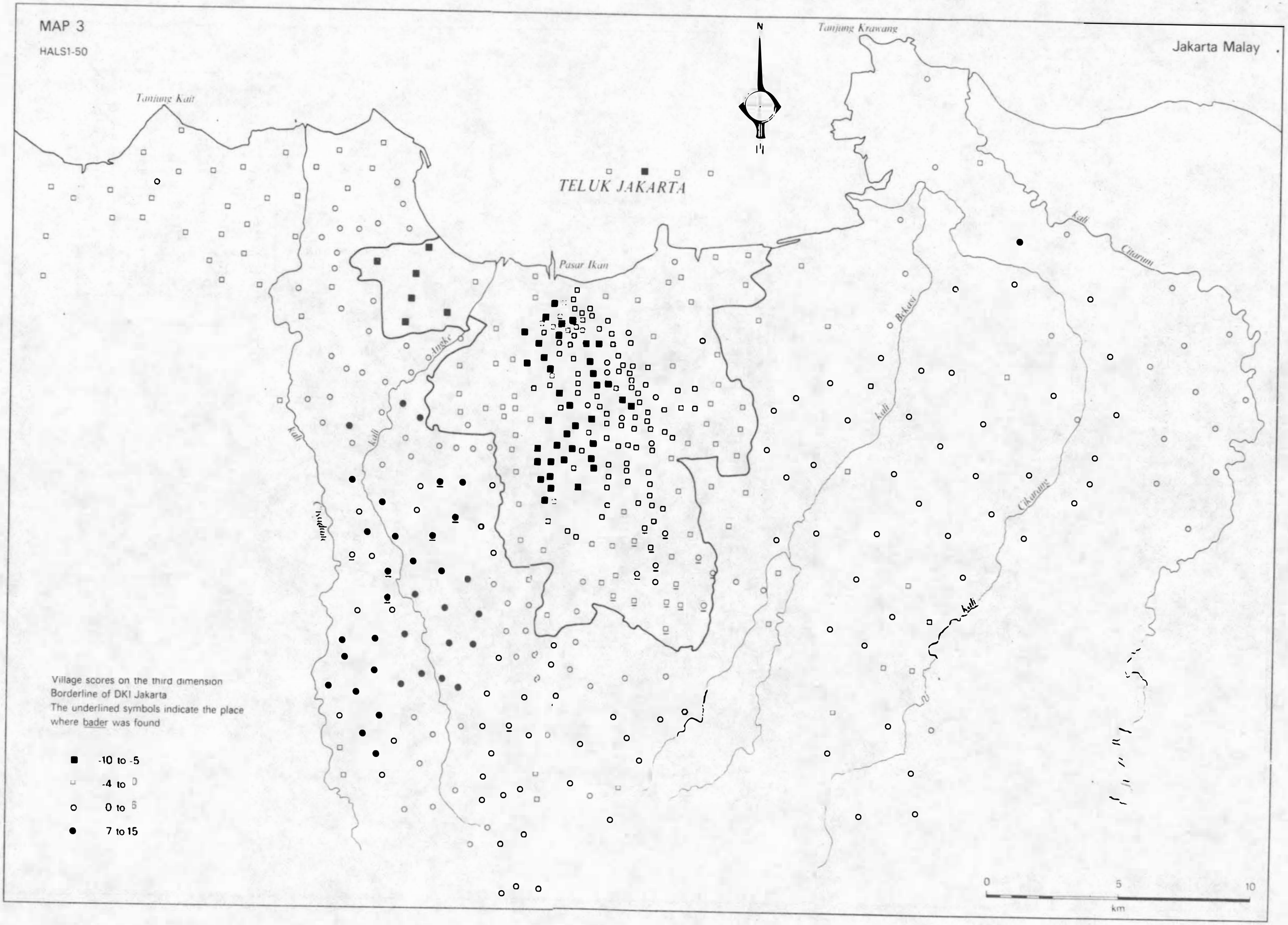

Jakarta Malay
N
0
5
10
20
km



MAP 2

BOUNDARIES OF KECAMATANS

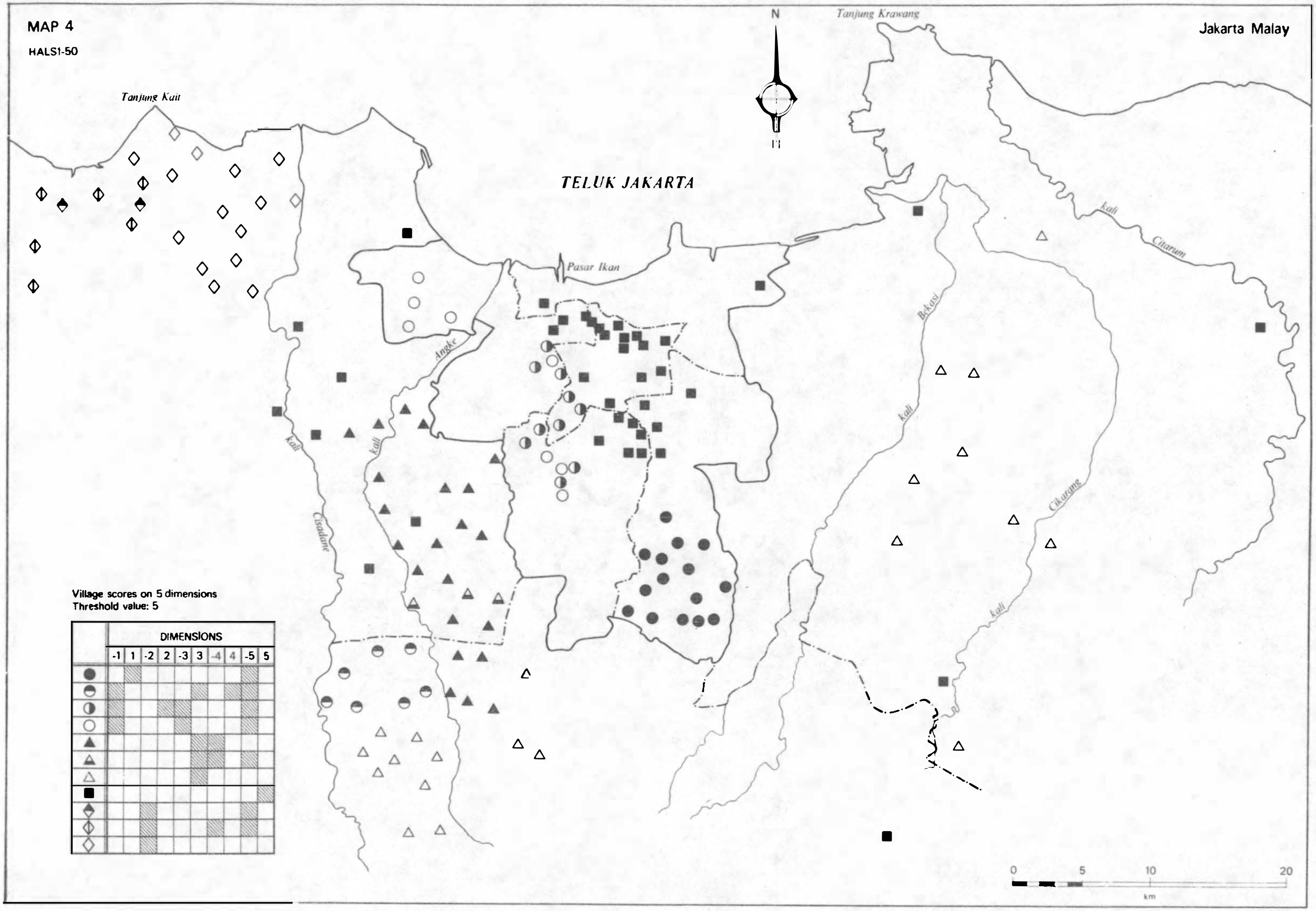
Jakarta Malay

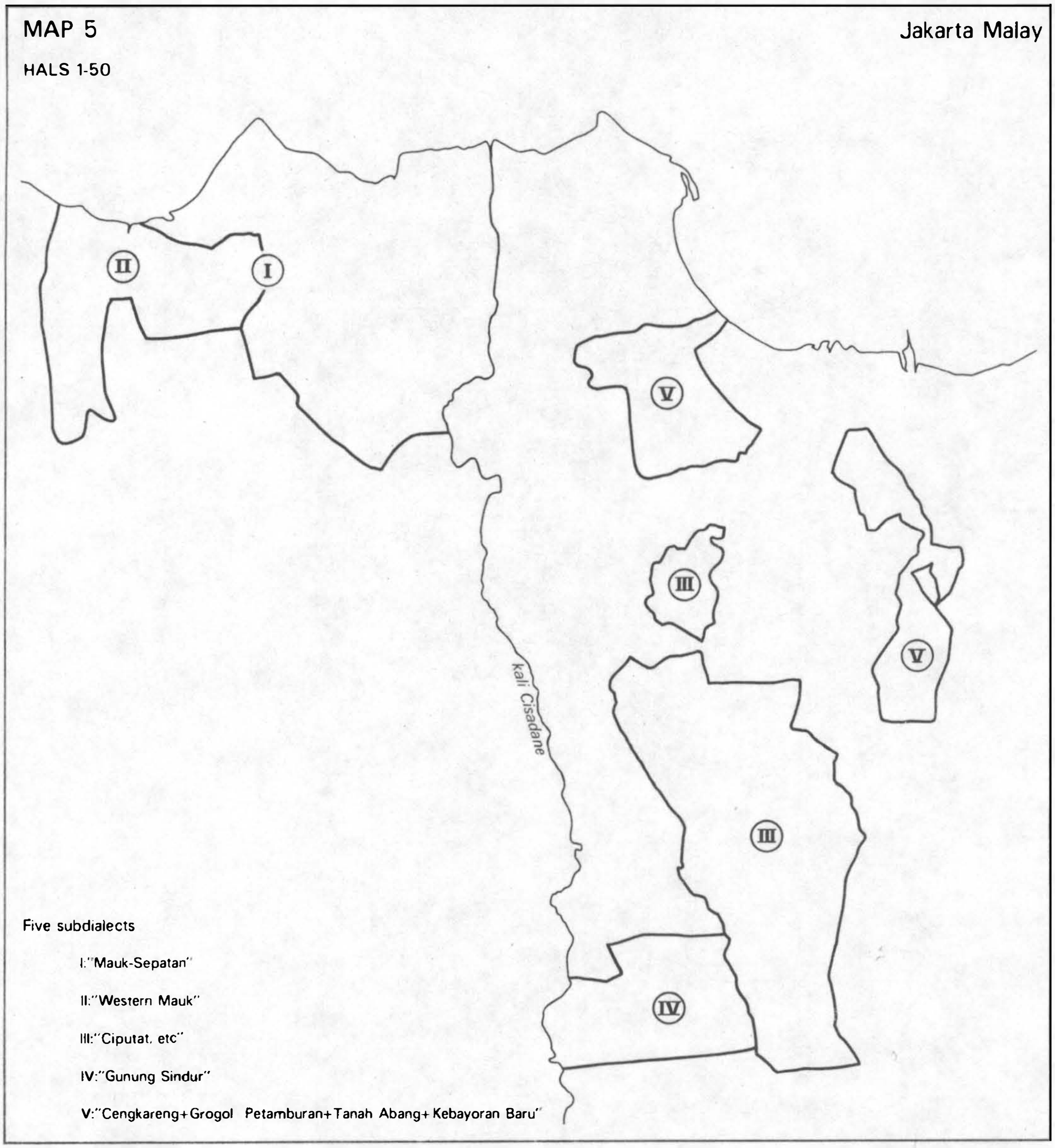


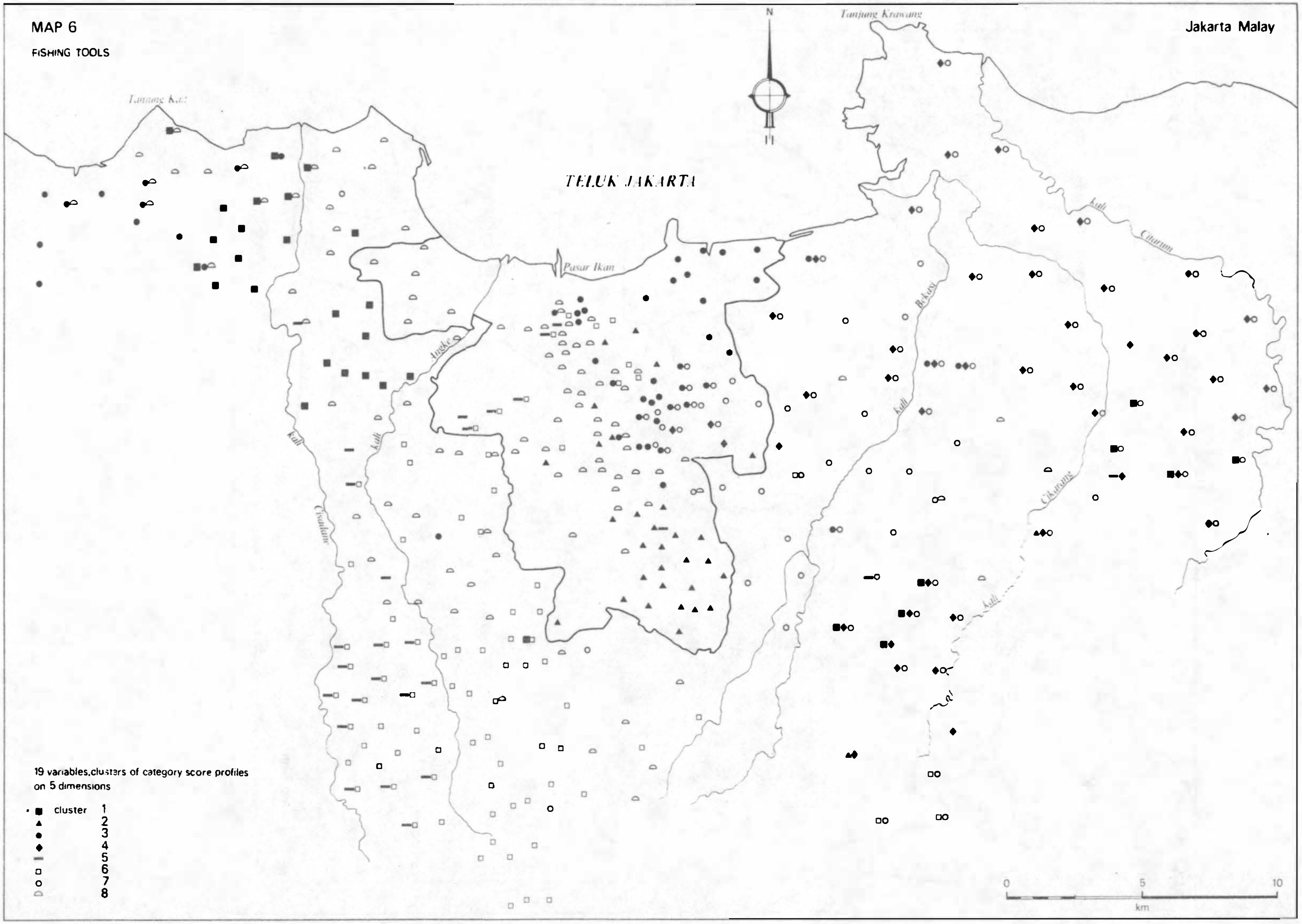
MAP 3

HALS1-50

Jakarta Malay


MAP 4
HALS1-50
Jakarta Malay
N
Tanjung Krawang
Tanjung Kait
TELUK JAKARTA
Pasar Ikan
Angke
kali
Cisadane
kali Bekasi
kali Citarum
kali Cikarang
kali
Village scores on 5 dimensions
Threshold value: 5
DIMENSIONS
-1 1 -2 2 -3 3 -4 4 -5 5
0 5 10 20
km


MAP 5
Jakarta Malay
HALS 1-50
II
I
V
III
V
kali Cisadane
III
IV
Five subdialects
I:"Mauk-Sepatan"
II:"Western Mauk"
III:"Ciputat, etc"
IV:"Gunung Sindur"
V:"Cengkareng+Grogol Petamburan+Tanah Abang+Kebayoran Baru"



MAP 6

FISHING TOOLS

BIBLIOGRAPHY

ANCEAUX, J.C.

1978 The linguistic position of south-east Sulawesi: a preliminary outline. In Wurm and Carrington, eds 1978:275-283.

BACKER, E.

1978a *Cluster analysis by optimal decomposition of induced fuzzy sets.* Delft: University Press.

1978b *Cluster analysis formalized as a process of fuzzy identification based on fuzzy relations.* Report IT-78-15. Delft: Information Theory Group, Department of Electrical Engineering, Delft University of Technology.

BAILEY, C.-J.N.

1973 *Variations and linguistic theory.* Arlington, Va.: Center for Applied Linguistics.

1979 The role of language development in a theory of language. *Arbeitspapiere zur Linguistik* 5:28-50. Institut für Kommunikationswissenschaften der Technischen Universität Berlin.

1980 Conceptualizing 'dialects' as implicational constellations rather than entities bounded by isoglossic bundles. In Göschel, Ivić, and Kehr, eds 1980:234-272.

BERDAN, Robert

1978 Multidimensional analysis of vowel variation. In Sankoff, ed. 1978:149-160.

BLACK, Paul

1976 Multidimensional scaling applied to linguistic relationships. In Dyen and Jucquois, eds 1976:43-92.

BLOOMFIELD, Leonard

1933 *Language.* New York: Henry Holt.

CADORA, F.J.

1979 *Interdialectal lexical compatability in Arabic.* Leiden: E.J. Brill.

C.N.R.S.

1973 *Les dialectes romans de France à la lumière des atlas régionaux.* Colloques nationaux du C.N.R.S. No. 930. Strasbourg 24-28 mai 1971.

DeCAMP, David

1971 Toward a generative analysis of a post-creole speech continuum. In Hymes, ed. 1971:349-370.

DITTMAR, Norbert

1973 *Soziolinguistik.* Frankfurt am Main: Athenäum.

DYEN, Isidore and Guy JUCQUOIS, eds

1976 *Lexicostatistics in genetic linguistics* II. Proceedings of the Montreal Conference ... May 19-20, 1973. *Cahiers de l'Institut de linguistique de Louvain* 3:5-6 (1975-1976).

FISHMAN, Joshua A., ed.

1968 *Readings in the sociology of language*. The Hague-Paris: Mouton.

GOOSSENS, Jan

1969 *Strukturelle Sprachgeographie*. Heidelberg: Winter.

GOSCHEL, Joachim, Pavle IVIĆ, and Kurt KEHR, eds

1980 *Dialekt und Dialektologie*. Ergebnisse des internationalen Symposions 'Zur Theorie des Dialekts', Marburg/Lahn, 5-10 September 1977. Wiesbaden: Franz Steiner.

GRIJNS, C.D.

1977 A la recherche du 'Melayu Betawi' ou parler malais de Batavia. *Archipel* 17:135-156.

1980 Some notes on Jakarta Malay kinship terms: the predictability of complexity. *Archipel* 20:178-212.

GUITER, Henri

1973 Atlas et frontières linguistiques. In *Les dialectes romans de France*...:61-109.

HENRICI, Alick

1973 Numerical classification of Bantu languages. *African Language Studies* 14:82-104.

HYMES, Dell, ed.

1971 *Pidginization and creolization of languages*. Cambridge: University Press.

IVIĆ, Pavle

1962 On the structure of dialect differentiation. *Word* 18:33-53.

KRUSKAL, Joseph B. and Myron WISH

1978 *Multidimensional scaling*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Beverly Hills and London.

KURATH, Hans

1972 *Studies in area linguistics*. Bloomington: Indiana University Press.

MOULTON, William G.

1968 Structural dialectology. *Language* 44:451-466.

NOTHOFER, Bernd

1980 *Dialektgeografische Untersuchungen in West-Java und im westlichen Zentral-Java*. Teil 1: *Text*; Teil 2: *Karten*. Wiesbaden: Otto Harrassowitz.

RIJCKEVORSEL, Jan van, and Jan de LEEUW

1978 *An outline to HOMALS-1*. Department of Datatheory/Faculty of Social Sciences, University of Leyden.

SAMARIN, William J.

1967 *Field linguistics*. New York: Holt, Rinehart and Winston.

SANKOFF, David, ed.

1978 *Linguistic variation: models and methods*. New York: Academic Press.

SANKOFF, David and Gillian SANKOFF

1976 Wave versus *Stammbaum* explanations of lexical similarities. In Dyen and Jucquois, eds 1976:29-41. Also in G. Sankoff *The social life of language*, University of Pennsylvania Press, 1980:143-151.

THELANDER, Mats

1976 Code-switching or code mixing? *International Journal of the Sociology of Language* 10:103-124.

TRUDGILL, Peter

1974 Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Language in Society* 3:214-247.

WALKER, Dale F.

1975 A lexical study of Lampung dialects. *NUSA: Linguistic Studies in Indonesian and Languages in Indonesia* 1:11-22.

WEINREICH, Uriel

1968 Is a structural dialectology possible? In Fishman, ed. 1968:305-319. Also in *Word* 14 (1954):388-400.

WINTER, Werner

1973 Areal linguistics: some general considerations. In T.A. Sebeok, ed. *Current trends in linguistics* 11:135-147. The Hague: Mouton.

WURM, S.A. and Lois CARRINGTON, eds

1978 *Second International Conference on Austronesian Linguistics: proceedings*. *PL*, C-61.