# Some New Classifications of Census Enumeration Districts in Britain
## A Poor Man's ACORN

Martin Charlton, Stan Openshaw, and Colin Wymer
*University of Newcastle-upon-Tyne*

**Abstract:** *This article reports some results from a project, sponsored by the Economic and Social Research Council, that is concerned with developing a classification of the 130,000 Enumeration Districts for which 1981 Census of Population data are available. There is a brief account of the development of cluster analytic techniques that can handle 130,000 cases without having to resort to subterfuge. It is thought that the methods developed here can deal with up to 400,000 cases, thereby bringing many countries within the range of numerical taxonomy. The paper discusses the problems of obtaining stable results and investigates their sensitivity to a number of operational factors. The resulting area types are discussed briefly and an attempt is made to place the work in a broader context. It is hoped that the results described here may help to prepare the way for similar studies in other countries.*

## 1. INTRODUCTION

There seems to have little enthusiasm among regional scientists for large-scale taxonomic exercises. The initial keenness of the 1960s did not survive into the 1970s, mainly because of the swing away from empirical analysis, seemingly intractable computational problems presented by the size of the more interesting datasets, and problems in obtaining data in computer-readable form.

It is hardly surprising, then, that the first studies concerned with the areal characteristics of entire countries, using fine resolution data at a subadministrative area scale, are a fairly recent phenomenon. Such studies are as much dependent on the availability of national sized computer datasets as on the development of computationally tractable methods of spatial analysis. The best examples so far are a series of national classifications, at ward and enumeration district level, performed in the late 1970s by Richard Webber (1978, 1979). These national classifications attempt to summarize the socioeconomic characteristics of residential areas by identifying areas with similar census characteristics. The purpose is that of multivariate description and exploratory analysis. There is no real theory, and these studies are concerned only with spatial description and generalization. Yet the results have been quite remarkable for a number of reasons. First, they offered a summary of the residential characteristics of the entire United Kingdom. Second, they represented the most detailed large-scale studies of geographical differentiation ever attempted. Third, they confirmed that residential area characteristics are geographically recurrent. Finally, they proved easy to communicate to others and to use; indeed, they have since found a number of important commercial applications in marketing research (CACI, 1980, 1983). The principal constraint on academic use, apart from massive disinterest, is the need to pur-

Address correspondence to Martin Charlton, Centre for Urban and Regional Development Studies, the University, Newcastle-upon-Tyne NE1 7RU, England.

Journal of Economic and Social Measurement 13:69–96 (1985)

0747-9662/85/$03.30

chase the classification from CACI. Other problems relate to concern about the lack of adequate theoretical underpinnings, and criticisms of the taxonomic procedures used (Openshaw et al., 1980).

However, it is useful to view large-scale taxonomic exercises in a different light. In many countries, census data are the most important of all datasets that provide details of local area characteristics, and it is obviously better to study the data of the finest possible geographical scale so as to reduce the effects of aggregation problems and avoid the accompanying loss of information. Regionalization methods offer a unique opportunity to study data for complete countries, at the smallest geographical scale, by providing an efficient means of summarizing the multivariate patterns that the data contain. A single study of this sort has the potential to replace over 20 years' worth of factorial ecologies and to describe all of the characteristic types of residential area likely to be found. Of course, initially, the objective is that of data explanation and the exercise is issue driven, but the potential for both further applied and more theoretical studies is tremendous. The problem at first is very daunting because of the enormity of the task, but one imagines that regional scientists will recover and discover ways of relating the results of existing theories of urban social structure and means of generating new ones. Such matters are, however, left for the future. Attention here is focused on the methodological aspects of performing national classification exercises, and on the substantive one of interpreting the results.

Section 2 describes the development of classification procedures able to provide a comprehensive classification of national census data sets for the smallest available geographical units and their application to 1981 census data for Britain. Section 3 provides a discussion of the results that have been obtained, and Section 4 comments on more general issues of importance to this kind of exercise.

## 2. CLASSIFYING LARGE AREAL DATA SETS

### 2.1. Background: The ACORN Inheritance

The 1981 census for Britain was reported at a number of geographical scales, the most detailed being that of census enumeration districts (EDs). There are about 130,000 census EDs in Britain and they provide complete coverage of the country. For each ED, some 4,500 counts are available, providing coverage of demographic, social, housing, and economic characteristics of the residence based population. The problem is merely that of reducing the 130,000 areas into a small number of different and distinctive areal types by use of a classification or regionalization procedure.

The first serious attempt to produce a national classification of EDs was that by Richard Webber (Webber, 1979). He selected a systematic sample of about 4000 EDs, devised a set of 40 variables, and classified them into 60 clusters using a nonhierarchical iterative relocation procedure. The remaining 116,000 EDs were allocated to whichever of the clusters they were most "similar" to. The 60 clusters were then subjected to a stepwise agglomerative fusion process to yield a set of 11 families. Descriptive labels were conjured up for each of the families and the results declared to be generally useful as a description of British residential areas at the time of the 1971 census.

In the late 1970s, Webber moved to CACI, which started marketing slightly earlier ward-based classifications; wards are about eight times larger than census EDs. The ward classification known as ACORN (A Classification of Residential Neighbourhoods) attracted considerable interest for marketing research purposes. It consisted of 36 clusters and seven families, and had been obtained in a manner similar to the ED classification, although there were only about 15,000 wards in 1971. The success of the ward-based classification, and especially of the cluster profiles, resulted in a modified ED level classification in which the EDs were individually assigned to whichever ACORN cluster they were most similar to, in order to give finer precision and to allow matching with unitary postcodes.

The publication of 1981 census data presented CACI with a major dilemma. Obviously, they had to update their highly popular 1971-based ACORN classification. However, there would be considerable consumer resistance if the "NEW ACORN" were substantially different from the "OLD ACORN." The result was a new ED level classification that explained as much of the variance of the data as possible, and yet provided a good match to the existing ACORN types. The new ACORN has two additional clusters and, compared with the old one, about 50% of the EDs were now assigned to different categories.

It would appear that both the new and old ACORN suffer from a number of major faults:

1. They were not, apparently, subjected to any detailed evaluation.
2. They used raw data in the computation of similarities and, owing to the presence of correlated variables, the similarities would be heavily biased by the choice of variables.
3. The "old ACORN" was a seemingly poor classification that had only been based on a minute sample of the data (see Openshaw et al., 1980), and yet marketplace pressures forced CACI to retain much of it for quite different 1981 census data (for example, some of the variable definitions had changed, and the importance of various indicator variables had changed considerably since 1971).
4. The new ACORN was based on, seemingly, a far more comprehensive analysis, yet it was constrained to match existing ACORN types.
5. There appears to be a feeling among some commercial users that the new ACORN is not as powerful a discriminator between different areal types as perhaps it could be.

### *2.2. Alternatives to ACORN*

The taxonomic methods used for the ACORN classification were never what may be regarded as state-of-the-art or conventional. There are alternative taxonomic procedures that could be used to classify all the data. Methods based on samples are excluded because of the unknown sampling properties of the resulting classification and the risk of bias or types of area being completely missed, as was the case with the 1971 ACORN (Openshaw et al., 1980).

Despite the size of the 1981 data set, a number of different methods could be used. Sibson (1973) describes a fast single linkage technique and Defays (1977) a group average version that could probably cope *provided* something like $1.69 \times 10^{10}$ similarities could be computed, each involving an inner product computation with 40 multiplications and additions. If this problem could be overcome, then there is the remaining difficulty that such methods tend to provide very poor results.

Another possibility is to use a contiguity-constrained hierarchical agglomerative method. If the contiguity constraint is properly implemented, then this sort of regionalization technique can be highly efficient computationally (Openshaw, 1974). There are problems in obtaining contiguities for 130,000 EDs, although approximate Thiessen polygon-based contiguities could have been used. Other difficulties concern the need to utilize a two-stage contiguity-constrained/unconstrained approach (Openshaw, 1976). Viable methods exist that could be used to implement this solution, but, before this happened, a better solution became possible.

The final possibility is to develop a computationally efficient method that can handle 130,000 EDs without the need for a contiguity constraint (although one could be imposed later) and also yield what would appear to be as good a classification as is likely to be achieved by any method. The preferred method is an iterative relocation procedure that approximately minimizes the within-cluster sum of squares via a nonhierarchical approach. A number of algorithms exist that one might use as starting points; for example, KMEANS (Spath, 1980) and several in Clustan IC (Wishart, 1977).

These methods have the tremendous pragmatic advantage that, generally, they seem to produce "good" results and are fairly robust. The problems involved in their use are basically twofold: that of computational feasibility, and that of deciding how many clusters are needed. The latter problem is a matter of subjective judgment, while the former one is far more fundamental. Fortunately, the former proved fairly easy to solve by tuning a standard algorithm to minimize the number of arithmetic operations, which made it possible to handle fairly easily up to 20,000 cases (Openshaw, 1982). Additional modifications resulted in a program in which 90% of the total CPU time was spent in a five-line inner product loop. Replacement of this by an assembly-language routine, written in such a way as to utilize a pipeline capability on the local university computer, made it possible to classify, easily, data sets of 200,000 or more cases.

### *2.3. The Super-CCP Algorithm for Large-Scale Area-Based Classifications*

The method employed here is based on the CCP package described in Openshaw (1982) and uses the basic classification strategy and philosophy as outlined in Openshaw (1983). The classification process can be presented as a linear sequence of operations, although, in practice, there will often be recursion at various points.

The *first stage* is to define a set of variables that best reflect the objectives and the purpose of the study. The choice of variables is absolutely crucial, and yet it is exceptionally difficult to closely match variables [permute any $K$ ($K < 100$) from 4500] with purpose. Hence, the aim is that of a broadly based descriptive classification, so the variables have been selected to represent a wide range of residential area characteristics.

The preliminary selection of 465 variables was subjected to cluster analysis to define a series of basic variable groupings. The refinement yielded the set of 55 variables included in Appendix A. Some of the variables included here reflect experience with earlier attempts at a national classification, which suggested that they were needed to strengthen the degree of discrimination between various types of areas. This is important; indeed, the variables used here might be considered to be third-generation ones, the previous two generations having been discarded. With a classification cycle time of about three months, such to-ings and fro-ings can be both time consuming and expensive but are very much in line with the recommended philosophical viewpoint. This suggests that classifications are not one-off exercises but have feedback loops and involve a kind of data safari in search of the rare and elusive meaningful result. In other words, it is a kind of inefficient but nevertheless important fine-tuning exercise.

The *second stage* is to apply an orthonormal transformation to the 130,000 × 55 raw data matrix. A correlation matrix is calculated, from which eigenvectors and eigenvalues are extracted. A set of principal component scores are then computed, with the scores on each component weighted by the size of the associated eigenvalue. For the classification exercise described below, a total of 27 components were needed to account for over 90% of the variance of the correlation matrix. This is the method used in Clustan IC, and there are no reasons why it should not be used here. Questions about nonnormal frequency distributions biasing the correlations, and thus the principal component scores, are dismissed as irrelevant. There is so much data redundancy that it really does not matter. A final question concerned the handling of missing data. Census data for EDs with less than 25 persons or eight households may be suppressed and thus become missing. A total of 18,375 EDs contained one or more missing values for the 55 variables. The simplest solution is to ignore these records when classifying the data and then assign them to the cluster to which they are most similar, using only the variables that are present at a later stage, or assign them to a "residual" cluster. These missing EDs may be regarded as containing unreliable census data in which the effects of "Barnardization"[1] may dominate their real features. Excluding them is, therefore, very necessary since it removes most, if not all, of the suspect data from the analysis and thus overcomes one of the problems with using ED data. The UK data required 1105 s of CPU time on an IBM 370/168 computer to be transformed into principal component scores.

The *third stage* is to generate a random classification of the 111,831 EDs into $K$ clusters, where $K$ is selected to produce the "most meaningful" classification. In practice, this is a very difficult decision to make properly, but usually the results are not too dependent upon it. Ideally, an attempt would be made to tune the classification by choosing a value of $K$ so that the resulting classification was most meaningful in terms of a priori knowledge, most easy to interpret, and most useful for the purpose of the exercise. To be used, the results must have "face value," and maximizing this subjective concept by fiddling with the number of clusters is more an art than a science. The time required here is fairly trivial: about 16 s of CPU time.

The *fourth stage* involves using an iterative relocation procedure to modify progressively the initial random classification so as to minimize the within-cluster sum of squares. In practice, this may require a large number of iterations before no further changes are made during an iteration, although usually the process would be halted once the total change in within-cluster sum of squares during a complete iteration is less than some small part of the total value. The results may also be dependent on the original starting classification and on the number of principal component scores.

Most of the computer time is expended in relocating the starting classification. Details of actual run times are given later but ranged from under 1 to over 17 hr of CPU time, depending on the convergence criteria used.

The *fifth stage* involves computing cluster diagnostics in an attempt to aid the user in labeling the clusters. The usual way of doing this is to relate cluster means to global means, scaling the differences by the size of the standard deviation associated with the global mean and, if necessary, also the cluster mean. The cluster mean values are themselves of some interest in helping to describe the different levels of characteristic variables. This requires about 250 s of CPU time.

The *sixth stage* is concerned with evaluation of the results. Some of the key virtues emerge during stage 5, but a more detailed examination of various numerical indices is of some additional use in comparing different classifications. This takes between 600 and 800 s.

The *seventh stage* usually involves returning to stage 3 and trying a different number of clusters as more is learned about the nature of the results. It is also useful to collapse the $K$ clusters into $M$ higher-order clusters ($M < K$), as this gives the flexibility of viewing a classification at two different levels of detail. Traditionally, this has been done by stepwise fusion (Webber, 1977), but there are significant advantages to be gained from regarding it as a completely separate taxonomic exercise. That is physically to aggregate the data to the $M$ clusters and then to start the classification process off again, using the cluster centroids as the raw data. This overcomes the cumulative suboptimality that is known to afflict all agglomerative clustering strategies. It also changes the correlation structure of the data to reflect its most highly aggregated state. The result is that the higher-order clusters, based on aggregated data, can employ the same nonhierarchical taxonomic technique as previously, and experiments have shown that the clusters have a greatly improved degree of clarity.

### 2.4. Comment

It is emphasized that national classifications are not, and can never be, either a one-shot process or a one-off process. The algorithmic sequence described in Section 2.3 is not linear but involves feedback and recursion. Likewise, the results are dependent on the decisions that were made concerning various operational choices. The process may seem subjective, but in reality it is no more subjective than that characteristic of other numerical methods. Indeed, it could even be claimed that the greater degree of openness involved makes the classification procedure inherently more honest. It is objective in that the results are reproducible, and really this minimal state is all that can reasonably be asked of any quantitative technique. Moreover, it also has the fundamental advantage that the results, to be useful, have to be statistically reasonable *and* meaningful. The fact that the "meaningfulness" of any classification is an arbitrary thing is less important than it is for the researcher to demonstrate in an explicit fashion precisely why the results are meaningful and possess face value. This criterion also applies to all other statistical and numerical techniques but has seldom been used. The point here—and classifications are a good vehicle for demonstrating it—is that statistical criteria by themselves are not sufficient to prove that a technique is valid and useful.

### 3.1. How Many Iterations?

In practice, the only way that the effects of the various operational decisions described in Section 2 can be evaluated is by the adoption of a strictly empirical ("suck it and see") strategy. There is, at present, no theory of areal taxonomy. Seen in purely geometric terms, we are attempting to partition a multidimensional space such that each partition contains a small hyperspherical cluster of points. Unless we have only a low-dimensional space and a small problem, then we are forced to experiment, and to learn from past experience in this field. This suggests two alternatives: either we generate a number of classifications with various numbers of clusters to establish some general trends, or we opt for, say, 42 clusters on the basis that this number looks as good as any other. The former approach requires us to choose a "useful" classification from the alternatives, and then proceed to fine-tune it. This implies that "usefulness" is a statistical attribute, and that the results have plausible face value. The latter approach saves us the problem of having to make these decisions, but denies us the possibility of investigating the "value" of alternative classifications, together with a number of pertinent methodological questions concerned with the relocation process, the labeling strategies, and the evaluation methodology. We should also bear in mind that a classification with "too many" clusters will be difficult to appraise and interpret, and thus will diminish the potential advantages of parsimony that have, in the past, made classification methods so generally useful. The immediate problem is that of simply producing a national classification.

## 3. PRODUCING A BETTER NATIONAL CLASSIFICATION FOR BRITAIN

By far the most computer-resource-consuming stage is the relocation process. With a small problem (say up to 500 cases), convergence (that is, one pass through the data with no moves taking place) may be accomplished in under 10, and certainly under 20, iterations. With a large problem, such as this, the number of moves made will, with each successive iteration, decrease asymptotically to zero. The within-cluster sum of squares, and the number of moves made per iteration, also drop rapidly in the first few iterations and then change more slowly. Observation of the relocation process suggests that, as the sum of squares curve becomes asymptotic, a small number of cases are switching between different clusters because of rounding error and the presence of a small number of "difficult-to-classify" cases, which perturb the cluster centroids by diminishingly smaller amounts each iteration. As the number of clusters increases, so it seems that the effects of this cycling become trivial.

The solution, to prevent unnecessary usage of computer time, is to adopt suitable convergence criteria for the iterative process. Table 1 shows the CPU times needed for a broad range of national classifications; in general the times required are a linear function of the number of clusters and the number of iterations; the number of iterations obviously depends on the convergence criteria. In this case, we use a proportional change in the within-cluster sum of squares of 0.1% and 0.005% as stopping criteria. With a 0.1% cutoff, the CPU time required falls between 15 min and 4.6 hr; with a 0.005% cutoff, this range increases to 27 min and 17.2 hr (using an IBM 370/168 mainframe). The adoption of less stringent convergence criteria clearly has a marked effect on the CPU times required. It should be noted that, the less strict the convergence criteria are, the greater the chance of misclassification and the greater the probability of finding poorly classified cases.

One way of assessing the effects of different numbers of iterations is to compare the classifications that are produced. It would seem that a convergence limit of 0.005% should be adequate, although, in fact, the 50-cluster classification described below was subjected to a convergence criterion of 0.00005%.

**Table 1.** Number of Iterations, Number of Moves, and CPU Time Required for Various National Classifications[a]

| Number of | Number of iterations | | Number of moves | | Cumulative CPU time (s) | |
|---|---|---|---|---|---|---|
| clusters | 0.1% | 0.005% | 0.1% | 0.005% | 0.1% | 0.005% |
| 5 | 4 | 4 | 110,364 | 110,364 | 1,639 | 1,639 |
| 10 | 14 | 17 | 229,464 | 230,743 | 5,095 | 6,161 |
| 15 | 10 | 22 | 215,773 | 233,734 | 941 | 2,303 |
| 20 | 11 | 26 | 234,752 | 257,313 | 6,010 | 13,978 |
| 25 | 10 | 28 | 243,260 | 280,594 | 6,495 | 17,767 |
| 30 | 11 | 31 | 264,759 | 306,171 | 1,945 | 6,352 |
| 40 | 12 | 28 | 267,334 | 302,111 | 2,239 | 7,582 |
| 50 | 10 | 29 | 251,385 | 303,142 | 2,764 | 9,250 |
| 60 | 8 | 26 | 234,980 | 261,160 | 3,515 | 10,443 |
| 70 | 18 | 37 | 311,153 | 344,604 | 8,860 | 18,750 |
| 80 | 18 | 32 | 314,835 | 340,767 | 10,154 | 18,009 |
| 90 | 14 | 44 | 309,735 | 364,686 | 8,748 | 27,388 |
| 100 | 16 | 47 | 312,773 | 371,429 | 10,734 | 34,778 |
| 110 | 15 | 49 | 317,891 | 373,192 | 11,197 | 36,510 |
| 120 | 17 | 45 | 314,250 | 367,349 | 13,743 | 34,921 |
| 125 | 14 | 36 | 310,625 | 352,163 | 11,827 | 30,116 |
| 140 | 18 | 39 | 339,676 | 375,560 | 16,707 | 36,163 |
| 150 | 15 | 63 | 328,000 | 413,207 | 14,803 | 62,072 |

[a] The column headings 0.1% and 0.005% refer to the percentage change in the within-cluster sum of squares. At the chosen levels, the number of iterations carried out, the number of relocations, and the cumulative CPU time in seconds are given.

### *3.2. How Many Clusters?*

The simplest solution is to evaluate a large number of alternative classifications in terms of some summary statistic, such as the total within-cluster sum of squares, and see whether there is any relationship with the number of clusters. The traditional approach is to look for a "discontinuity" in the function as an indication of where to stop, or which classification to look at. Two other statistics are also used as a measure of the performance of classifications: the value of the outer-fence[2] distance of each case to its parent cluster, and the equivalent far-out distance distribution of cases and clusters in terms of the taxonomic space in which the classification is performed. In the current exercise, plots of these three statistics against the number of clusters indicated that simple exponential approximations could be applied with a high degree of fit ($r^2 > 0.999$). These functions can be used to interpolate between the observed points and to extrapolate beyond the 150-cluster maximum that has so far been examined (see Table 2). Additionally, inspection of the functions themselves, and their second-order derivatives, can be used to provide a broad indication of where the functions have leveled out. It would seem that, beyond about 50–70 clusters, there is little advantage to be gained, but that on the other hand, by 38 clusters (the CACI choice), the functions have not yet "bottomed out." It should be noted that there measures are merely one guide to determining an appropriate number of clusters.

### *3.3. How Sensitive Are the Results?*

A final academic question concerns the stability of the resulting classification. Openshaw and Gillard (1977) have already explored the effect of different macro-level decisions, but interest here is focused on the more localized problem of what happens if the data are changed a little. To test whether small data changes have any effect, two variables were changed and the 50-, 100-, and 150-cluster classifications subjected to further iterations using the revised data. This turned out to be a good test, since the number of component scores actually increased from 27 to 29. However, the resulting, revised classifications were almost identical to the original ones. Rand (1971) coefficients of 0.979, 0.992, and 0.993 were recorded; a Rand statistic can be used to measure the similarity between classifications. The results indicate that, if anything, a 150-cluster classification is more robust than a 50-cluster one. They also indicate that the classifications themselves appeared to be stable, at least under mild perturbations.

## 4. A POOR MAN'S ACORN

### *4.1. Towards a General-Purpose Classification*

The objective of the present exercise is to try to provide a "good" general-purpose classification, if such a thing can be conjured up. The objective here is to develop for the U.K. academic community of poor men a direct alternative to, and replacement for, CACI's ACORN as a general description of the areal characteristics of British residential areas. The results of Section 3 indicate that, while the 38 clusters of ACORN are too few, pragmatic arguments can be used against any classification with very many more clusters. Accordingly, it seems that 50 clusters would be a reasonable compromise choice, although perhaps, for some purposes, 70 or 80 clusters might have been preferable. However, even 50 clusters is too large to handle easily. Moreover, although this classification seemed to contain a reasonable amount of spatial and taxonomic discrimination, there are groups of seemingly similar clusters. So, a number of higher-order classifications of the 50 clusters were examined to see whether any sensible simplification could be obtained. A higher-order classifica-

**Table 2.** Relationships between Number of Clusters and Various Measures of Classification Performance

| Number of clusters | Sum of squares[a] | Second derivative | Outer fence[b] | Second derivative | Far out[c] | Second derivative |
|---|---|---|---|---|---|---|
| 5 | 58.8 | 0.424 | 9.57 | 0.090 | 13.93 | 0.133 |
| 10 | 52.8 | 0.095 | 8.35 | 0.019 | 12.13 | 0.029 |
| 15 | 49.6 | 0.039 | 7.70 | 0.008 | 11.18 | 0.011 |
| 20 | 47.4 | 0.021 | 7.28 | 0.004 | 10.56 | 0.006 |
| 25 | 45.8 | 0.013 | 6.96 | 0.002 | 10.10 | 0.003 |
| 30 | 44.5 | 0.008 | 6.72 | 0.001 | 9.73 | 0.002 |
| 35 | 43.4 | 0.006 | 6.52 | 0.001 | 9.44 | 0.001 |
| 36 | 43.3 | 0.006 | 6.48 | 0.001 | 9.39 | 0.001 |
| 38 | 42.9 | 0.005 | 6.41 | 0.001 | 9.29 | 0.001 |
| 40 | 42.6 | 0.004 | 6.35 | 0.000 | 9.19 | 0.001 |
| 42 | 42.2 | 0.004 | 6.28 | 0.000 | 9.10 | 0.001 |
| 45 | 41.8 | 0.003 | 6.20 | 0.000 | 8.98 | 0.001 |
| 50 | 41.1 | 0.002 | 6.07 | 0.000 | 8.79 | 0.000 |
| 55 | 40.5 | 0.002 | 5.96 | 0.000 | 8.62 | 0.000 |
| 60 | 39.9 | 0.001 | 5.86 | 0.000 | 8.47 | 0.000 |
| 65 | 39.4 | 0.001 | 5.77 | 0.000 | 8.34 | 0.000 |
| 70 | 39.0 | 0.001 | 5.68 | 0.000 | 8.22 | 0.000 |
| 75 | 38.6 | 0.001 | 5.60 | 0.000 | 8.11 | 0.000 |
| 80 | 38.2 | 0.001 | 5.53 | 0.000 | 8.00 | 0.000 |
| 85 | 37.8 | 0.000 | 5.47 | 0.000 | 7.90 | 0.000 |
| 90 | 37.5 | 0.000 | 5.41 | 0.000 | 7.81 | 0.000 |
| 95 | 37.2 | 0.000 | 5.35 | 0.000 | 7.73 | 0.000 |
| 100 | 36.9 | 0.000 | 5.29 | 0.000 | 7.65 | 0.000 |
| 105 | 36.6 | 0.000 | 5.24 | 0.000 | 7.58 | 0.000 |
| 110 | 36.3 | 0.000 | 5.20 | 0.000 | 7.51 | 0.000 |
| 115 | 36.1 | 0.000 | 5.15 | 0.000 | 7.44 | 0.000 |
| 120 | 35.9 | 0.000 | 5.11 | 0.000 | 7.38 | 0.000 |
| 125 | 35.6 | 0.000 | 5.07 | 0.000 | 7.32 | 0.000 |
| 130 | 35.4 | 0.000 | 5.03 | 0.000 | 7.26 | 0.000 |
| 135 | 35.2 | 0.000 | 4.99 | 0.000 | 7.21 | 0.000 |
| 140 | 35.0 | 0.000 | 4.95 | 0.000 | 7.15 | 0.000 |
| 145 | 34.8 | 0.000 | 4.92 | 0.000 | 7.10 | 0.000 |
| 150 | 34.6 | 0.000 | 4.89 | 0.000 | 7.06 | 0.000 |
| 200 | 33.1 | 0.000 | 4.62 | 0.000 | 6.66 | 0.000 |
| 250 | 32.0 | 0.000 | 4.42 | 0.000 | 6.37 | 0.000 |
| 300 | 31.1 | 0.000 | 4.26 | 0.000 | 6.14 | 0.000 |
| 350 | 30.3 | 0.000 | 4.13 | 0.000 | 5.96 | 0.000 |
| 400 | 29.7 | 0.000 | 4.03 | 0.000 | 5.80 | 0.000 |
| 450 | 29.2 | 0.000 | 3.93 | 0.000 | 5.66 | 0.000 |
| 500 | 28.7 | 0.000 | 3.85 | 0.000 | 5.55 | 0.000 |

[a] Sum of squares: the percentage within-cluster sum of squares.

[b] Outer fence: the percentage of cases in the outer fence category.

[c] Far out: the percentage of cases in the far out category.

tion that recognized 13 distinctive area types was chosen that provided a convenient framework within which to label and identify the principal (but not necessarily majority) characteristics of the 50 clusters.

The labels attributed to each cluster are based on a comparison of mean, or median, cluster characteristics in relation to the data as a whole. The descriptive interpretations were then tested by looking at the results for areas that are familiar to the authors. It was also found useful to be able to examine the spatial distribution of cluster types within *x* miles of particular locations. For example, if the cluster characteristics seem to imply the presence of defense bases, then we can test for the incidence of this cluster type around known military areas. Table 3 shows the proportions of the resident population and private households, and the number of enumeration districts to be found in areas belonging to each of the 13 area types; Table 4 shows the area type labels and the characteristics of the clusters in these types; and Appendix B gives descriptions of the area types in terms of the divergence of the characterizing variables from the national means. Table 5 shows the composition of the population in terms of the 13 area types in a 5-k circle surrounding some selected British towns and cities.

**Table 3.** Breakdown of Great Britain by Area Type

| Area type | Resident population | Private households | Enumeration districts |
|---|---|---|---|
| 1 | 7.47 | 7.25 | 8,387 |
| 2 | 6.33 | 7.67 | 8,635 |
| 3 | 19.40 | 20.02 | 22,068 |
| 4 | 1.65 | 1.36 | 1,800 |
| 5 | 0.44 | 0.37 | 474 |
| 6 | 12.48 | 11.45 | 13,101 |
| 7 | 3.51 | 4.36 | 5,749 |
| 8 | 11.31 | 10.10 | 10,829 |
| 9 | 19.55 | 19.70 | 21,603 |
| 10 | 0.28 | 0.34 | 409 |
| 11 | 4.92 | 4.78 | 7,382 |
| 12 | 3.69 | 4.46 | 5,242 |
| 13 | 6.48 | 5.61 | 6,174 |
| Unallocated | 3.76 | 3.46 | 18,578 |
| Totals | 100 | 100 | 130,431 |

## 5. SPATIAL PATTERNS

Finally, to demonstrate that the results of the classification have the potential to offer both concise and new perspectives on the social geography of Britain, it is interesting to map the proportion of the population belonging to each area type by local government district. Figures 1–12 show a map for each area type. As might be expected, much of the high-status residential group is concentrated in the southeast of England in the so-called "stockbroker" belts. There are some interesting exceptions, notably around Liverpool, Newcastle, Leeds, and Hull. The contrasting picture is presented by the "older council tenancies," concentrated in Central Scotland and northern industrial areas, Norwich being an exception. The manufacturing and mining areas of Britain contain the major proportions of the "blue-collar" areas, Crawley and the Kent coal field providing south eastern outliers. The "less well-off" council housing is a phenomenon associated with older industrial areas: The Scottish preponderance is simply a reflection of the far higher proportion of public housing in Scotland compared with England and Wales. The single renters are concentrated in city cores, where large Victorian terraces are converted into flatted accommodation. The newer suburban housing appears to be concentrated in a belt between London and the Liverpool/Manchester area, the inhabitants being young, relatively affluent families in commuting areas. "White-collar" areas of Britain are more dispersed than the high-status areas; a group is evident around Norwich; the London–Brighton railway provides an axis to the South Coast, but only the more affluent high-status residents appear to be able to afford houses about the London–Southampton line. The "nonpermanent" group map is a little misleading since these areas are few in number,

**Table 4.** Area Type and Cluster Labels[a]

| | |
|---|---|
| Area type 1: High-status residential | |
| 6 | Professional and qualified workers in service employment |
| 33 | Exclusive owner occupation, with service employment |
| 36 | Fewer professional workers, journey to work by car |
| Area type 2: Older council tenancies | |
| 4 | Lone female pensioner households, fewer fertile females |
| 8 | Older adults, fewer young families |
| 12 | Male unemployment, single parent families, journey to work by bus |
| 29 | Owner occupation, manufacturing employment, skilled labor |
| Area type 3: Manufacturing areas with manual labor | |
| 9 | Poor housing, male unemployment, low car ownership |
| 23 | Poor housing, male unemployment |
| 28 | Unskilled labor, journey to work on foot |
| 32 | Older adults |
| 34 | More concentrated manufacturing employment |
| 35 | Pensioner households, low economic activity rates |
| 45 | Mining employment, skilled manual labor |
| Area type 4: Multiethnic areas | |
| 20 | Indo-Pakistani and Caribbean-born predominate |
| 43 | African-born predominate, lower marriage rate |
| Area type 5: Government and military establishments | |
| Area type 6: Less well-off council housing | |
| 5 | Caribbean-born, lower marriage rate |
| 13 | Single-parent families, low working-age-population masculinity |
| 15 | Lower male unemployment |
| 16 | Overcrowding, children and youths |
| 19 | African or Caribbean-born, young families |
| 22 | Male unemployment, young families, overcrowding |
| Area type 7: Single persons in rented property | |
| 1 | Non-Commonwealth or non-EEC-born, service employment |
| 2 | Journey to work by rail, professional employees, no children |
| 3 | Single-parent families, lone female pensioner households |
| 7 | Substandard rented housing, mobile single males |
| 18 | African or caribbean-born, single-parent, six-person households |
| Area type 8: Newer suburban housing | |
| 27 | Large houses, two-car households, qualified workers |
| 42 | Older adults, pensioners, fewer children |
| 44 | Young families, owner occupation |
| 47 | Mobile, young married couples |
| Area type 9: Service employment area with nonmanual workers | |
| 10 | Single persons in rented property |
| 14 | Journey to work by rail |

**Table 4.** (Continued)

24 African, Indo-Pakistani, or Caribbean-born
37 Young families
40 Older adults, large houses, professional workers
46 Pensioners, two-car households

Area type 10: Areas of nonpermanent housing

Area type 11: Rural areas
30 Tied housing, rented accommodation
48 Less tied housing
50 Agricultural workers, homeworking, second homes

Area type 12: Retirement areas and resorts
17 Lone-female-pensioner households, journey to work on foot
38 Second homes, *fewer economically active married females*
41 Lone-female-pensioner households, low economic activity

Area type 13: Better-off council housing
11 Mobile families with children
21 Young families with children
26 Smaller households

[a] For each cluster in each group, only the characteristics that differentiate it from the other clusters in the group are given.

**Table 5.** Characteristics of Selected British Towns in Terms of 13 Aggregated Groups (Great Britain = 100)[a]

| Town/Group | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| London | 4 | 227 | 8 | 61 | 38 | 266 | 797 | 0 | 5 | 0 | 7 | 14 | 7 |
| Newcastle | 113 | 364 | 67 | 60 | 0 | 225 | 127 | 10 | 71 | 0 | 0 | 29 | 34 |
| Edinburgh | 210 | 264 | 32 | 0 | 0 | 134 | 458 | 5 | 80 | 0 | 0 | 40 | 2 |
| Portsmouth | 33 | 91 | 163 | 0 | 320 | 64 | 216 | 15 | 150 | 77 | 21 | 136 | 45 |
| Southend | 130 | 58 | 104 | 0 | 0 | 58 | 199 | 12 | 157 | 0 | 0 | 389 | 44 |
| St. Albans | 276 | 15 | 69 | 0 | 400 | 9 | 101 | 120 | 155 | 146 | 39 | 14 | 134 |
| Motherwell | 49 | 298 | 24 | 0 | 0 | 352 | 5 | 62 | 21 | 0 | 0 | 20 | 172 |
| Northwich | 133 | 102 | 207 | 0 | 0 | 59 | 20 | 154 | 69 | 0 | 0 | 0 | 31 |
| Rugby | 75 | 61 | 108 | 90 | 0 | 33 | 0 | 169 | 162 | 0 | 58 | 35 | 128 |
| Norwich | 77 | 118 | 141 | 0 | 71 | 87 | 112 | 52 | 131 | 0 | 15 | 123 | 82 |
| Derby | 69 | 95 | 147 | 672 | 0 | 93 | 56 | 77 | 80 | 72 | 0 | 15 | 125 |
| Canterbury | 148 | 24 | 62 | 0 | 0 | 56 | 147 | 40 | 115 | 513 | 70 | 300 | 237 |
| Trowbridge | 41 | 24 | 149 | 0 | 0 | 67 | 22 | 226 | 83 | 409 | 102 | 0 | 115 |
| Wellingborough | 51 | 34 | 127 | 229 | 0 | 61 | 0 | 152 | 98 | 269 | 34 | 36 | 239 |
| Llandudno | 11 | 71 | 141 | 0 | 0 | 10 | 0 | 0 | 82 | 0 | 23 | 1101 | 71 |
| Hereford | 65 | 99 | 124 | 0 | 0 | 32 | 15 | 55 | 93 | 566 | 112 | 138 | 252 |
| Scunthorpe | 10 | 118 | 104 | 227 | 0 | 205 | 8 | 131 | 84 | 337 | 23 | 11 | 102 |
| Penzance | 64 | 39 | 92 | 0 | 0 | 45 | 106 | 0 | 66 | 0 | 200 | 1075 | 79 |
| Penrith | 0 | 0 | 150 | 0 | 0 | 0 | 0 | 49 | 117 | 0 | 262 | 560 | 111 |

[a] For each town, the composition by area type is given for a 5-k scan around the center of the town.

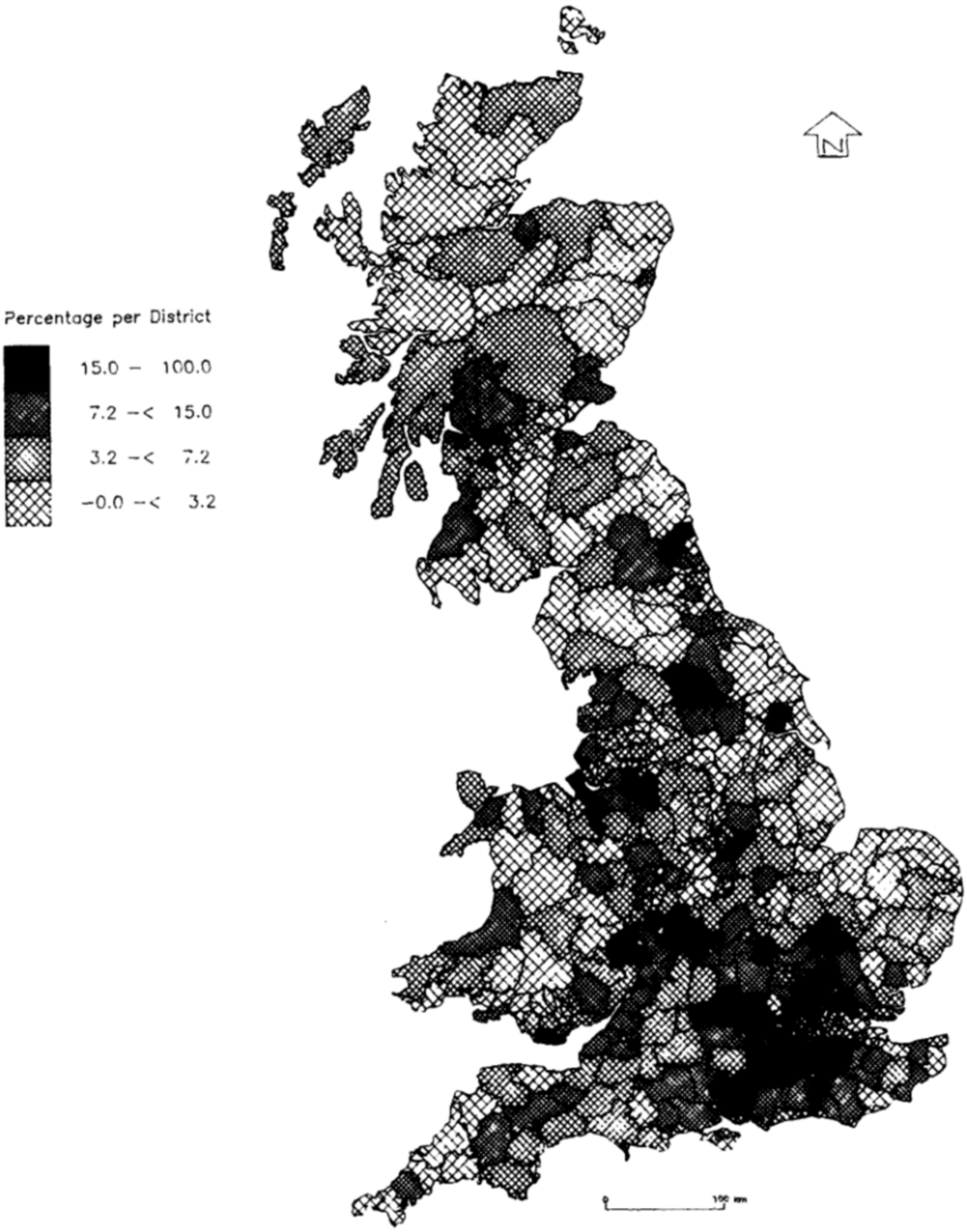

**Figure 1.** 50/13 classification using relocated centroids: High-status residential areas.

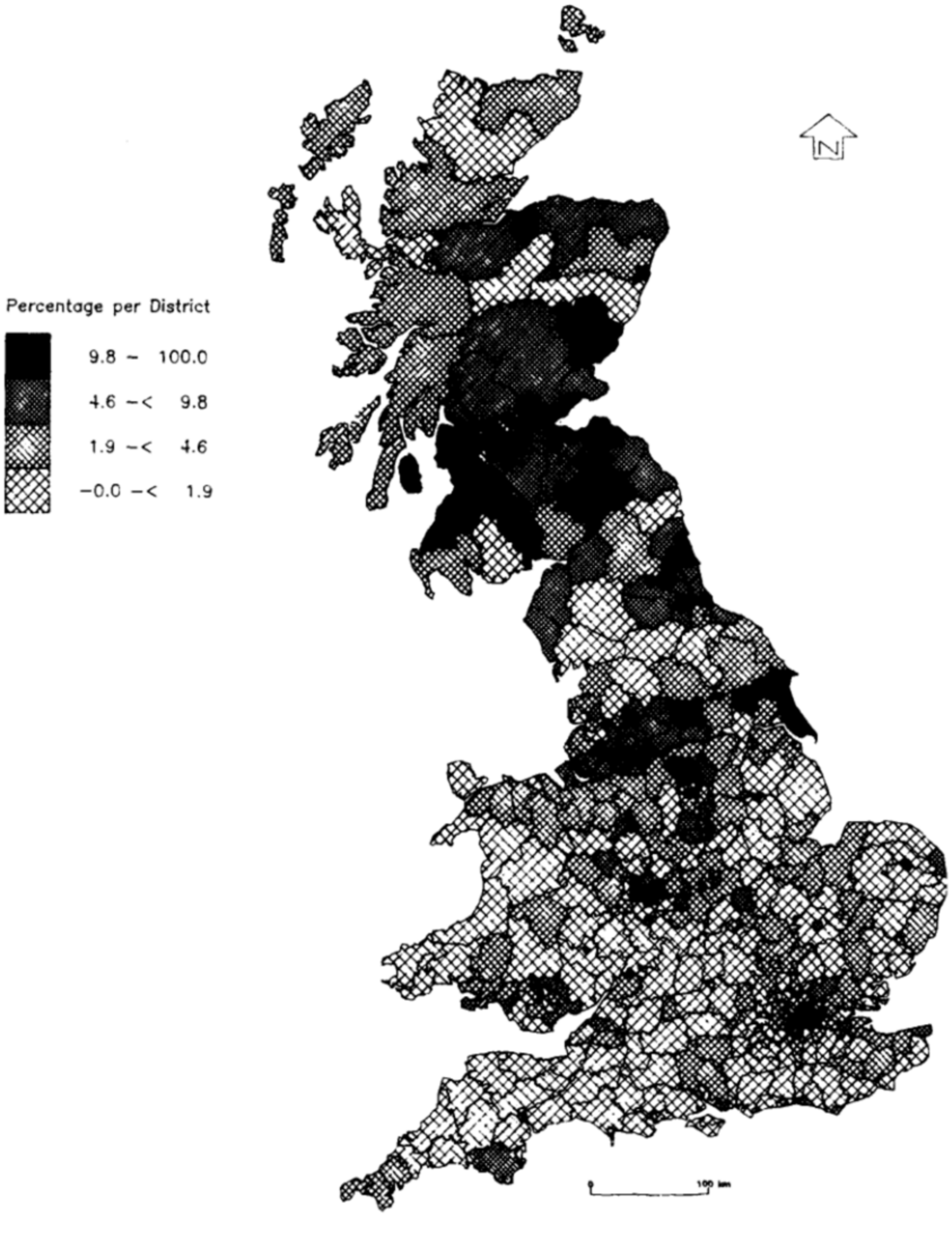

**Figure 2.** 50/13 classification using relocated centroids: Older council tenancies.

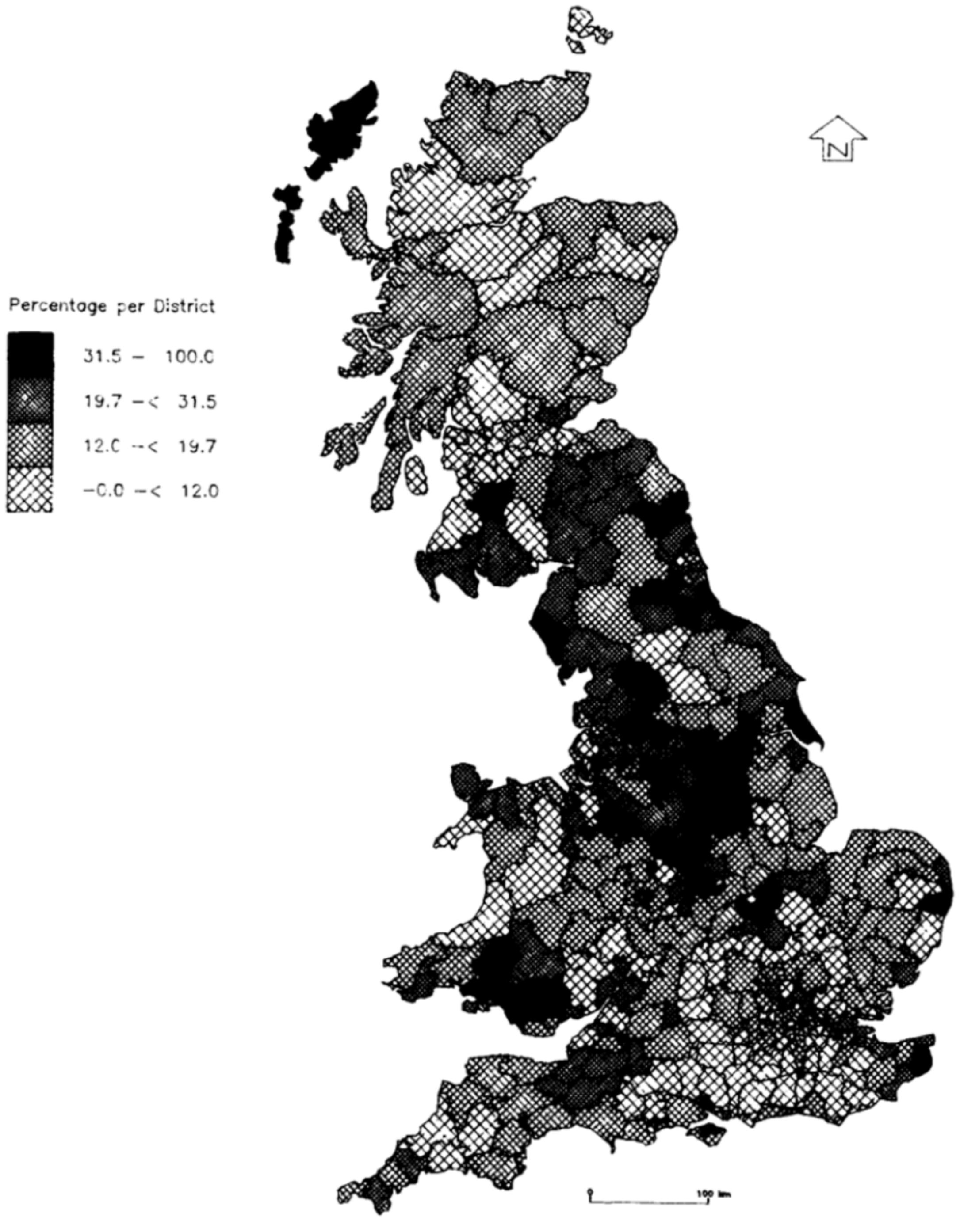

**Figure 3.** 50/13 classification using relocated centroids: Manual manufacturing employment areas.

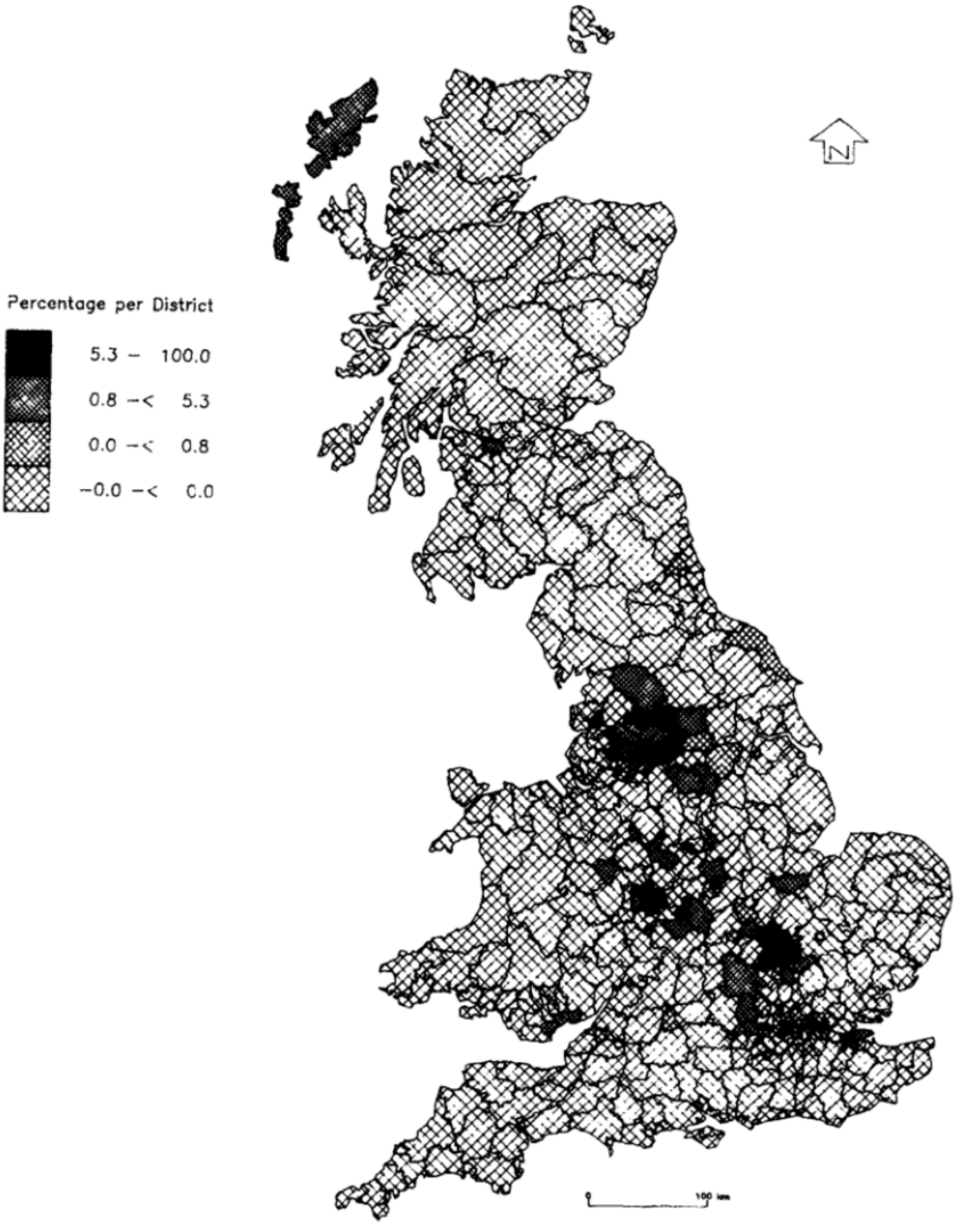

**Figure 4.** 50/13 classification using relocated centroids: Multiracial areas.

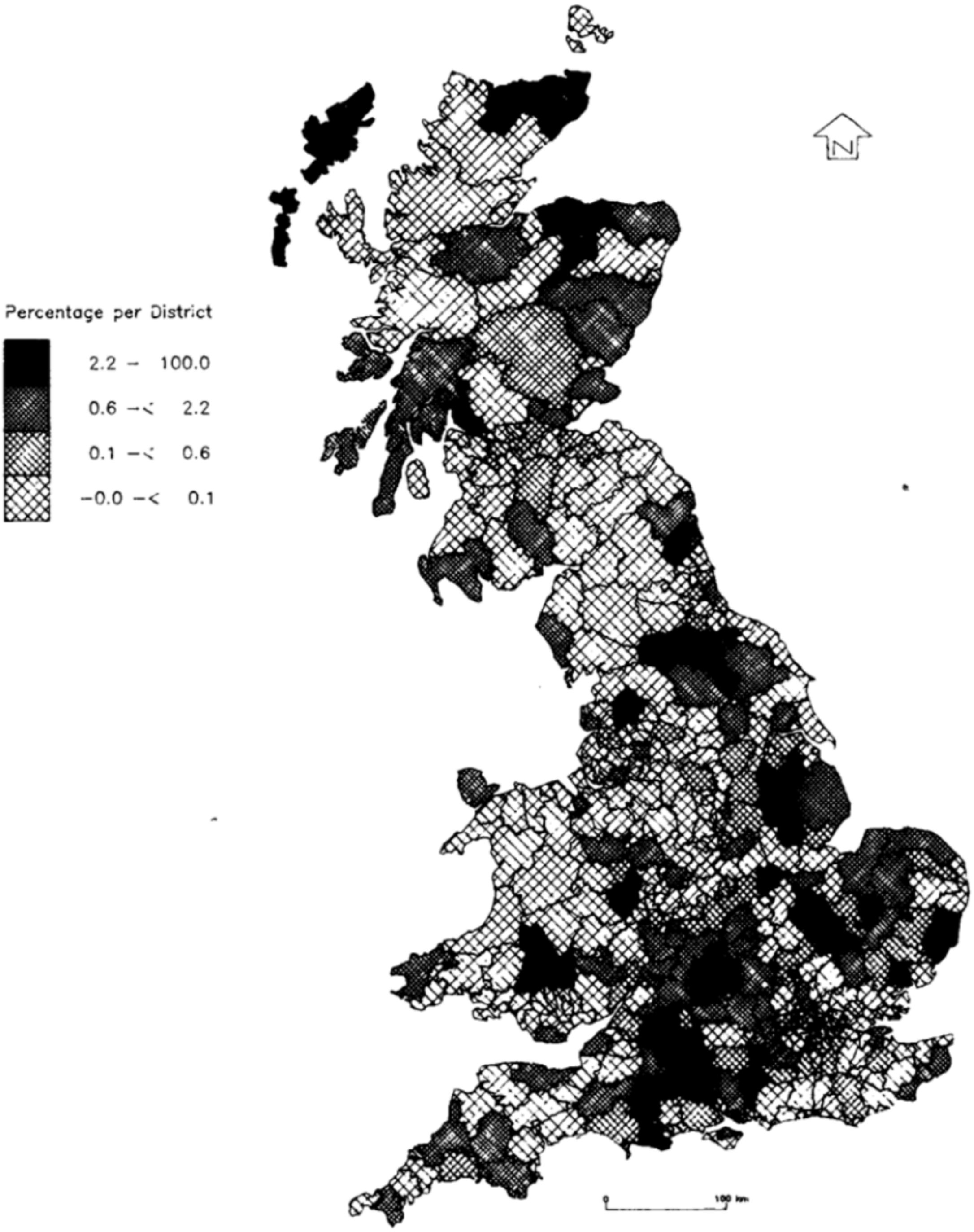

**Figure 5.** 50/13 classification using relocated centroids: Government establishments.

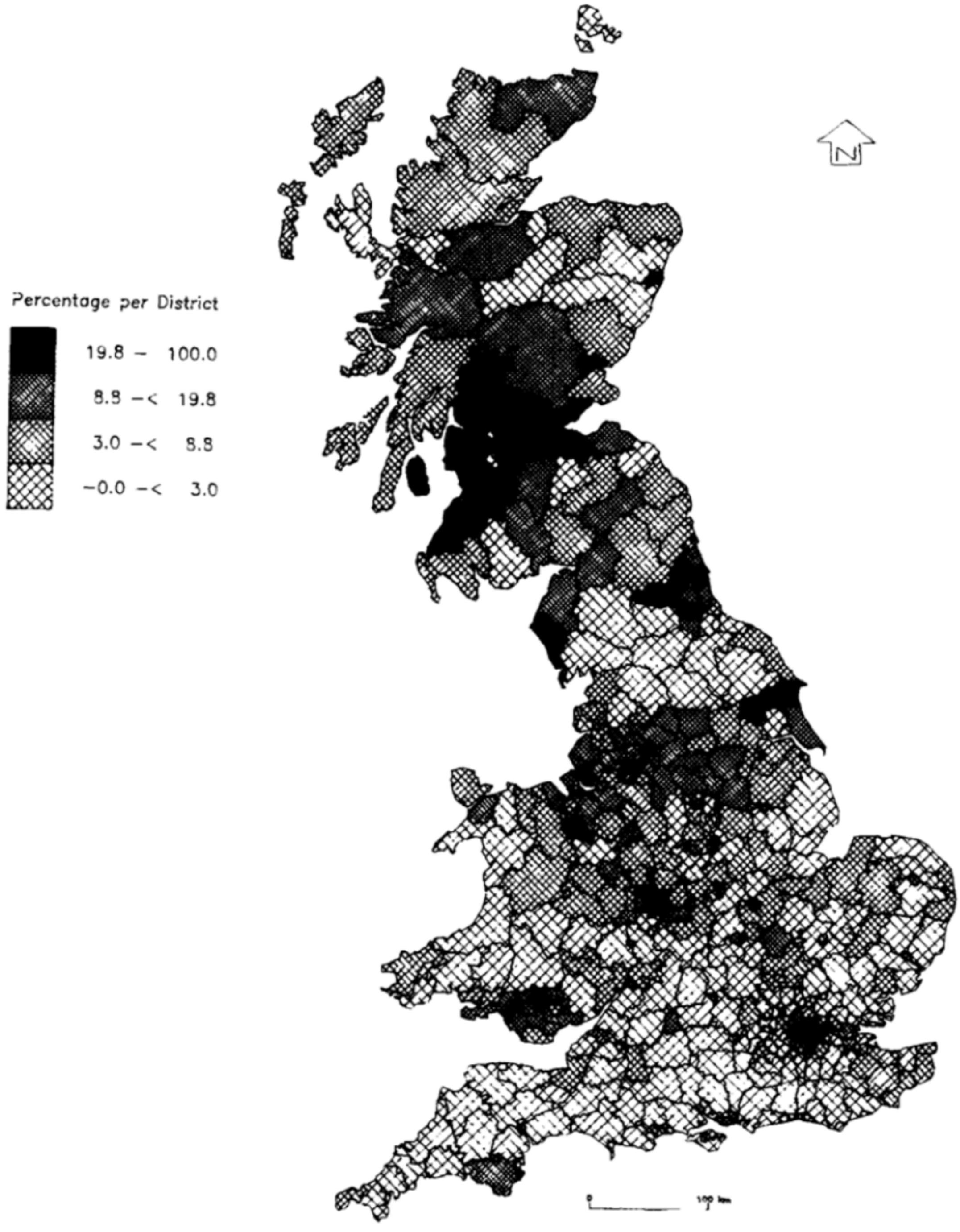

**Figure 6.** 50/13 classification using relocated centroids: Less well-off council housing.

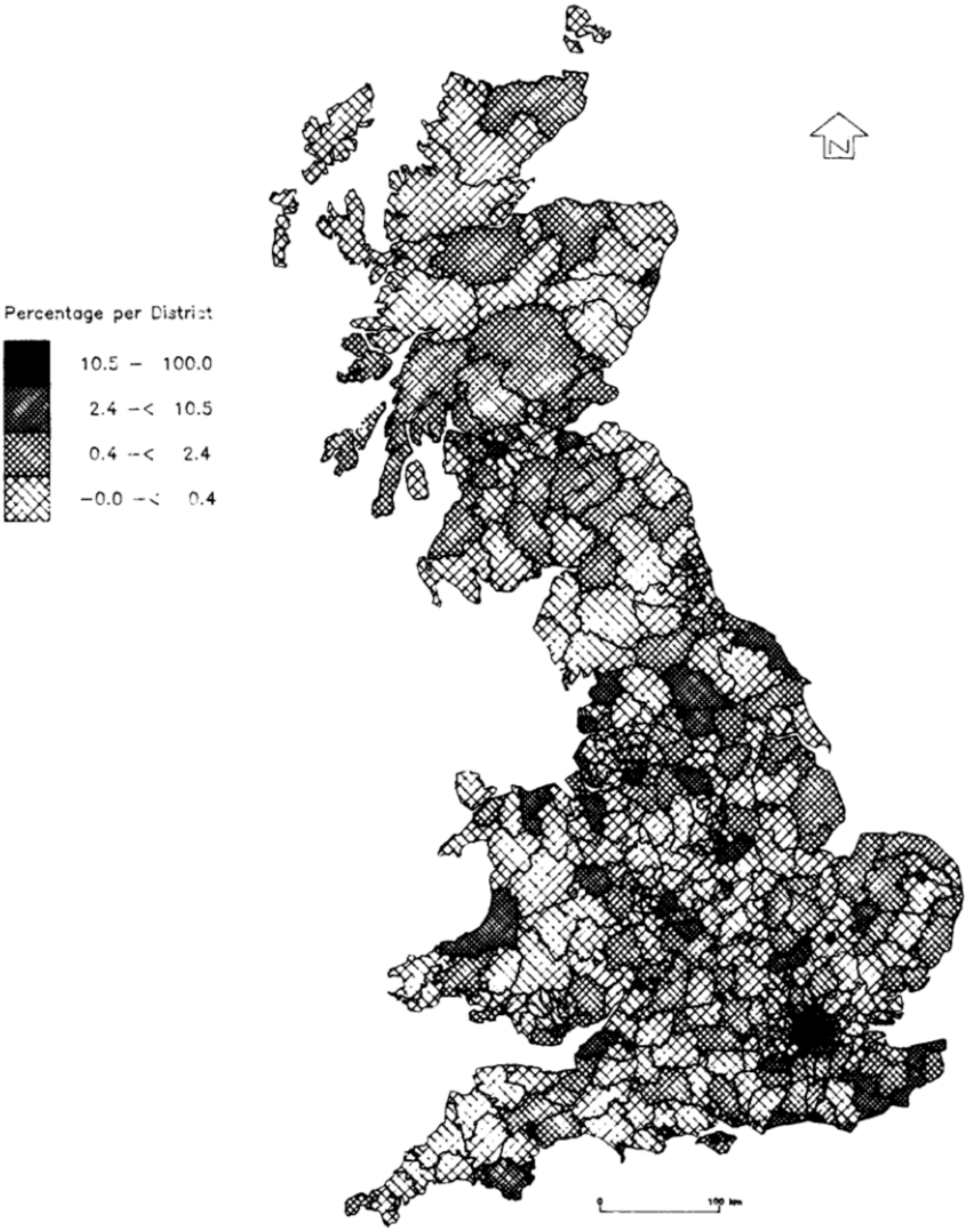

**Figure 7.** 50/13 classification using relocated centroids: Nonfamily renting households.

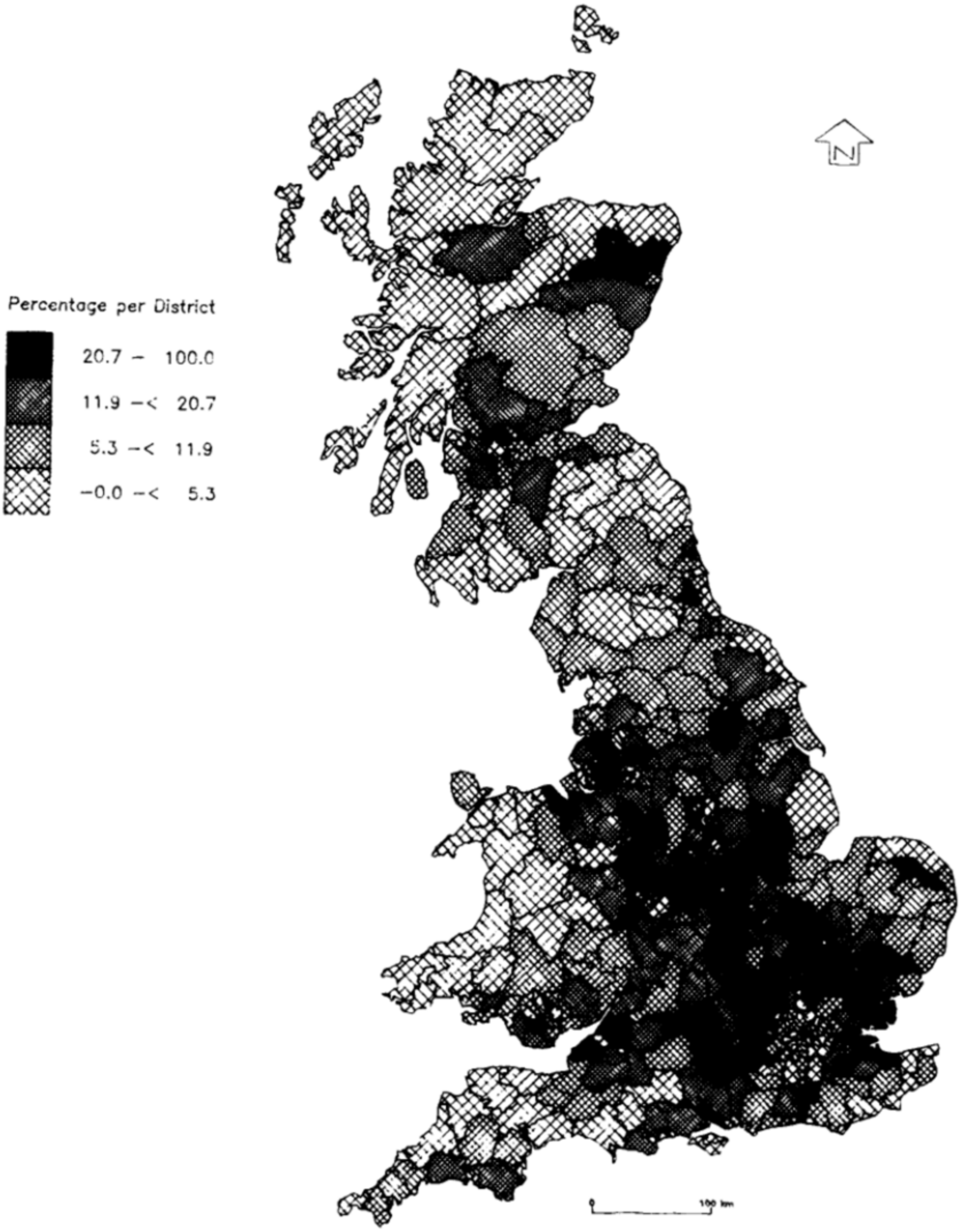

**Figure 8.** 50/13 classification using relocated centroids: Newer suburban housing estates.

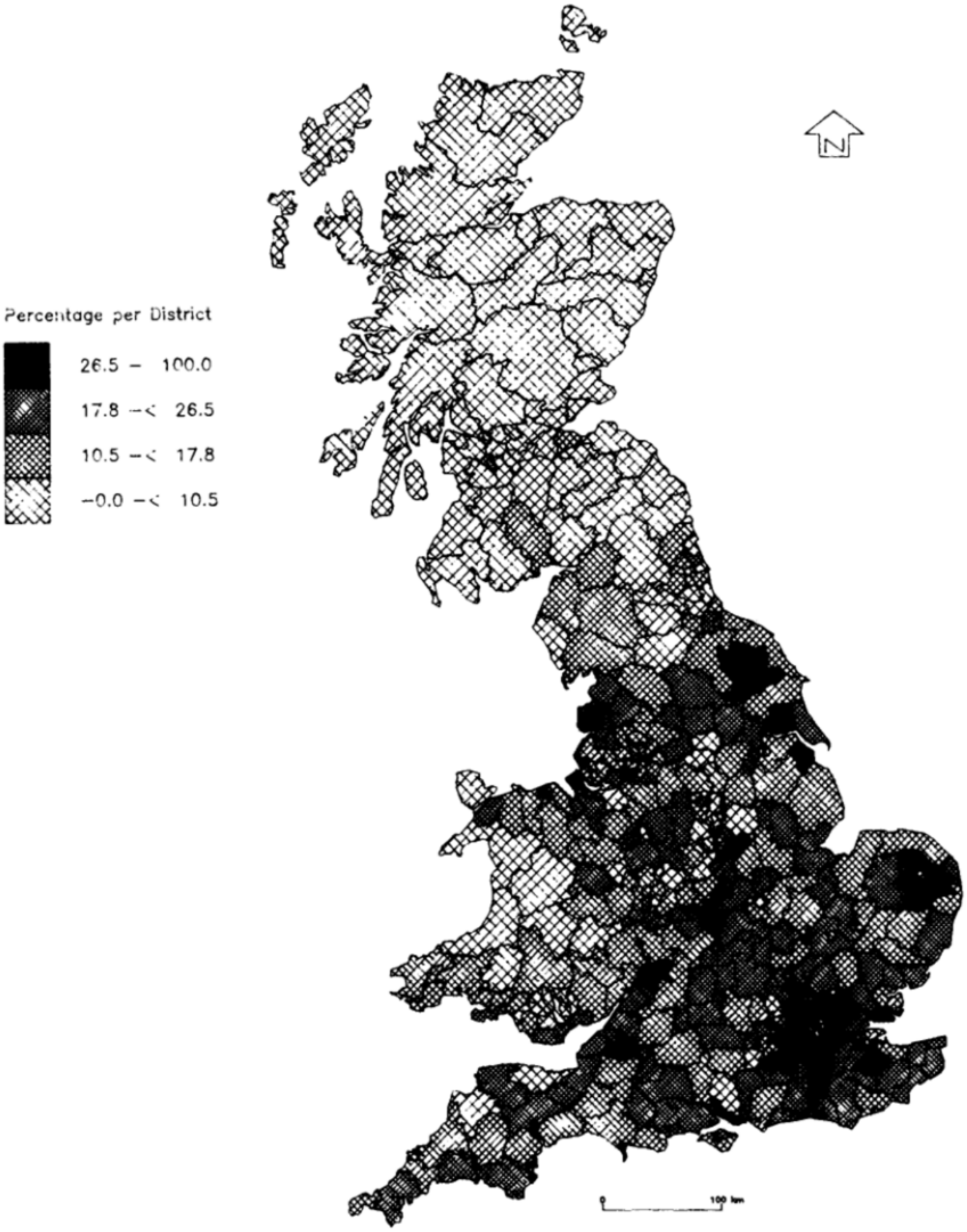

**Figure 9.** 50/13 classification using relocated centroids: Nonmanual service employment areas.

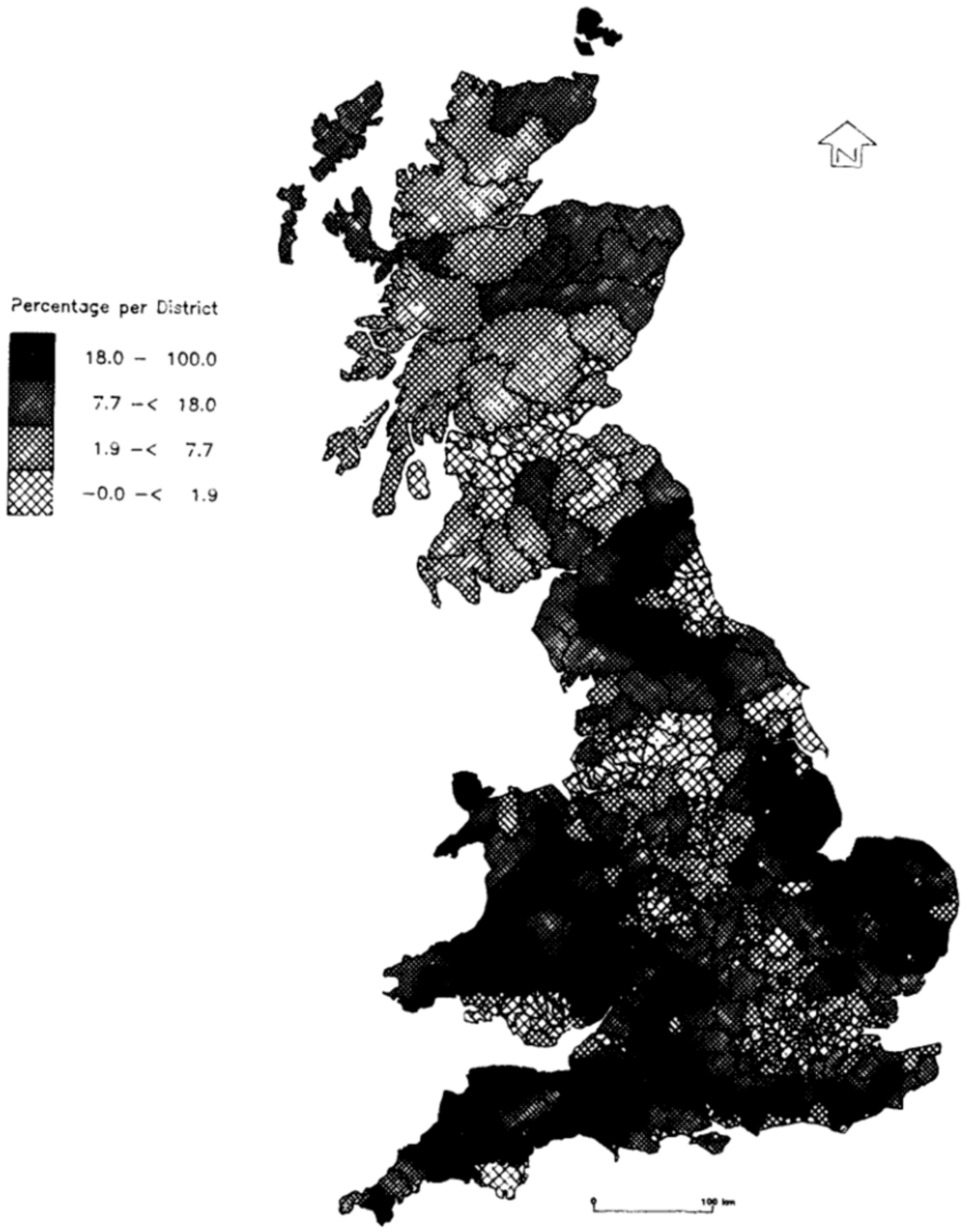

**Figure 10.** 50/13 classification using relocated centroids: Agricultural employment areas.

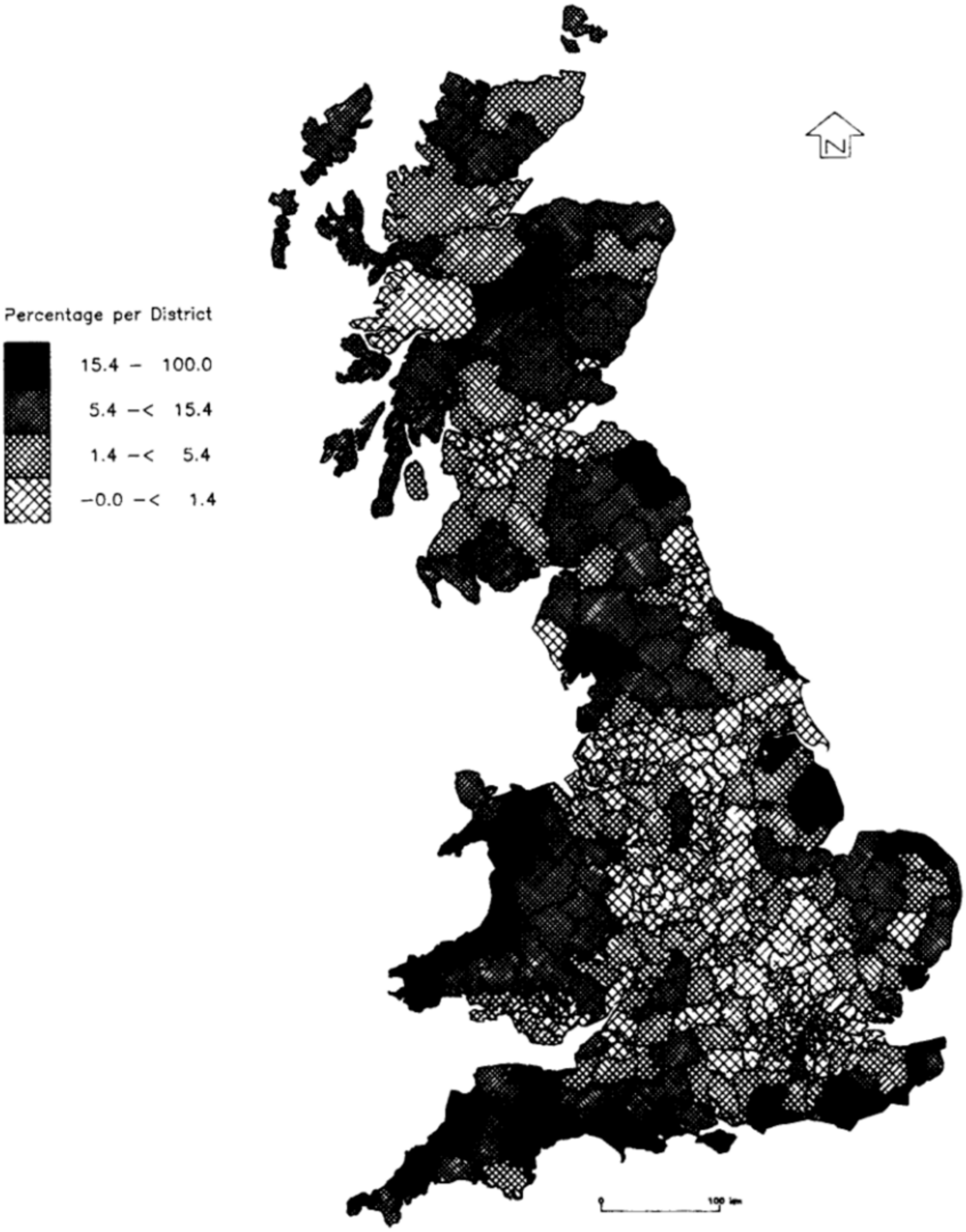

**Figure 11.** 50/13 classification using relocated centroids: Retirement migration and report areas.

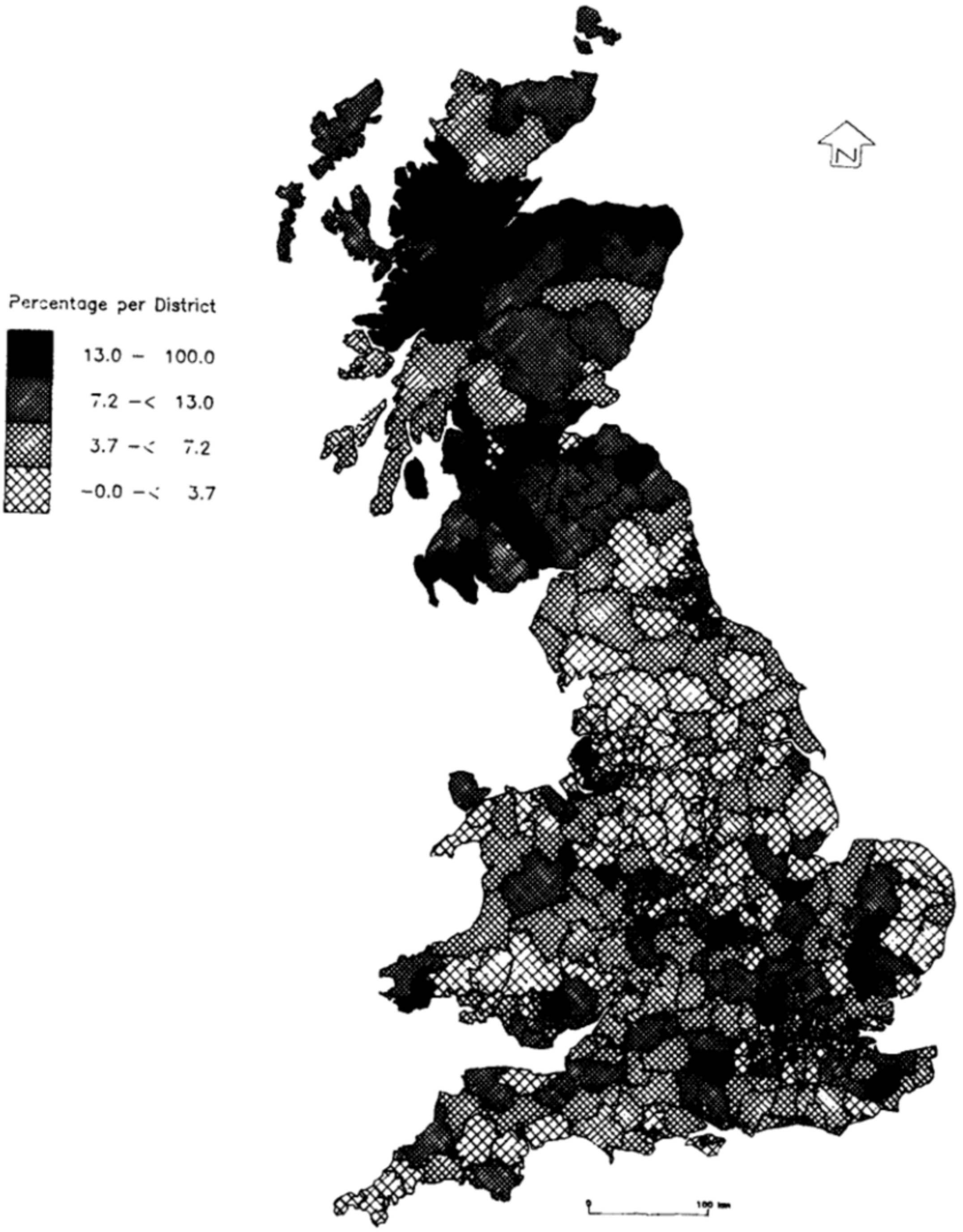

**Figure 12.** 50/13 classification using relocated centroids: Better-off council housing.

but they are concentrated in the less hilly agricultural areas of the South and East. Not surprisingly, the agricultural groups are concentrated in the traditional farming areas of Britain, in a complete contrast to the blue collar area distribution; there are still quite high proportions of this group in the suburban and "white-collar" areas of the London–Liverpool "core." The better-off council housing appears to be a largely Scottish phenomenon because of the greater incidence of public housing in Scotland; in England, however, the main locations tend to be in the less highly urbanized parts of the country.

In assessing these spatial patterns it should be remembered that local authority districts differ greatly in size, population, and internal homogeneity, and that ecological fallacies may be distorting the picture. However, the maps do provide a useful means of visualizing the local patterns of the national scale.

## 6. CONCLUSIONS

This article has described the genesis of a series of national classifications of British residential areas that have been produced for purposes of general geographical description. They are certainly not the only classifications that could be produced, but merely a sample from the set of all possible national classifications. The problem is that there is no way of measuring "bestness." Instead, the results provide an exploratory spatial description of small-scale areal census data for an entire country. They certainly have some novelty value, they appear to have face value, but whether they are useful (and which one of the classifications that is used) depends on the user being able to satisfy himself (and others) that the results are meaningful and relevant to his purpose. It is a matter for subsequent historians to debate as to whether any "national" classifications emerge.

This article has demonstrated what can be done and gives an indication of the sorts of results that can be produced. The potential of adopting a similar approach for other countries should not be overlooked. National-sized areal classifications of this sort may well prove to be one of the more useful services offered by modern geographers and regional scientists. The potential for further developments should not be overlooked. When regionalization methods are integrated with automated cartographic techniques and remote sensing imagery, then what may be termed the era of "fully automated geographies" or even "automated regional science" (ARS!) will have dawned. How we use these techniques and what we use them for is more a matter of philosophical attitudes than anything else.

ACKNOWLEDGMENTS

The authors would like to thank Katherine Bell and Karen Hunter for assistance in collating the output from several hundred computer runs, and we gratefully acknowledge funding from the Economic and Social Research Council for the research described in this paper.

## APPENDIX A: 1981 POPULATION CENSUS VARIABLES USED IN THE CLASSIFICATION

Population aged 0–4 (infants)
Population aged 5–14 (children)
Population aged 15–24 (youths)
Population aged 25–44 (younger adults)
Population aged 45–64 (older adults)
Population aged 65+ (old persons)
Masculinity of working-age persons
Married persons per 10,000 adults
Females of reproductive age
Infant : young woman ratio
African-born residents
Caribbean-born residents
India/Pakistan-born residents
Non-Commonwealth/EEC-born residents
One-year migrants
Economically active married females

Unemployed males
Students
Nonpermanent private households
Owner-occupied private households
Council/New Town rented households
"With job" private households
Unfurnished rented private households
Furnished rented private households
Second and holiday homes
Households with 1–3 rooms
Households with 7+ rooms
Households without cars
Households with 2+ cars
Households share/lack bath
Households share/lack inside w.c.
Persons per 100 rooms
Overcrowded households (>1.5 ppr)
Persons per 100 households
Six-or-more-person households
Two-adults-with-children households
Married-couple households
One E.A. adult without children
Lone-female-pensioner households
Lone-parent families
Agriculture workers
Energy and water workers
Manufacturing workers
Service, distribution workers
Travel to work by car
Travel to work by bus
Travel to work on foot
Residents working at home
Travel to work by rail
Professional workers
Nonmanual workers
Skilled manual workers
Semiskilled manual workers
Unskilled manual workers
Qualified workers

For each enumeration district, the data are expressed as a rate.

## APPENDIX B: 13-AREA-TYPE/ 50-CLUSTER CLASSIFICATION; DESCRIPTION OF AREA TYPES

### *Area Type 1*

This consists of 8387 EDs, which have high proportions of professional employees, households with seven or more rooms, employees with qualifications, households with two or more cars, and students. (This last is because students were regarded as being resident at their parents' home in the census definition.) There are low proportions of households without cars, and low proportions of skilled manual employees in these areas. This group is an amalgam of three clusters. We may conveniently label this group as "high status residential areas."

### *Area Type 2*

This higher-order cluster has 8635 EDs, and is characterized by high proportions of households with 1–3 rooms, council tenancies, lone female pensioner households, households without cars, and residents aged 65 and over. The average household size is low, as are the proportions of residents aged 25–44, fertile females, and owner-occupied households. This group contains four clusters, which are composed of areas of older council tenancies.

### *Area Type 3*

These areas have characteristics not markedly different from the national average, although they contain areas with higher proportions of manual workers in the manufacturing sectors and low proportions of ethnic minorities, agricultural workers, and qualified and professional workers. It has 22,068 members. There are seven clusters in this group. The general characteristics of this cluster suggest that these are areas populated by employees in so-called blue-collar jobs.

*Area Type 4*

There are 1,800 members of this area type with very high proportions of Indo-Pakistan–born residents, six-person households, African-born residents, and overcrowded households, and high proportions of residents aged 0–4, Caribbean-born residents, unemployed males, households lacking an inside W.C., and semiskilled and manufacturing workers. Characteristic of these areas are those EDs having a concentration of the ethnic minorities.

*Area Type 5*

There are 474 EDs in this group, which are characterized by high proportions of mobile, married-couples-with-children households, tied households, young children, service-sector employees' rented housing, nonmanual workers, and journey to work on foot, and low proportions of residents aged 45 and over, male unemployment, owner occupation, lone female pensioner households, and manufacturing workers. Within these areas we find military and defense establishments.

*Area Type 6*

This area type has 13,101 members, and is characterized by high proportions of council tenancies, overcrowded households, unemployed males, bus journeys to work, and households without cars, and low proportions of owner-occupied households. There are seven clusters in this group. These areas may be conveniently labeled as "less well off council areas."

*Area Type 7*

This consists of 5,749 EDs characterized by a very high proportion of single nonpensioner residents and rented accommodation, high proportions of Caribbean-born residents, one-year migrants, student residents, households with 1–3 rooms, households sharing or lacking an inside W.C., and travel to work by rail, and low proportions of residents aged 5–14, married adults, persons per household, and two-adult-with-children households. There are five clusters in this group. Webber refers to such areas as "nonfamily," as the overriding characteristic is a concentration of single-person households.

*Area Type 8*

This higher-order cluster has 10,829 members, characterized by high proportions of married couple and two-adult households, residents aged 0–4 and 25–44, married adults, fertile females, and owner-occupied housing, and low proportions of residents aged over 45, households without cars, and lone female pensioner households. There are four clusters in this group. These areas may be labeled as "young families in newer suburban housing."

*Area Type 9*

There are 21,603 EDs in this type of area. It is characterized by slightly higher proportions of professional and nonmanual workers in the service, distribution, and government sectors. There are six clusters in this group. These areas contain residents employed in so-called white-collar jobs.

*Area Type 10*

This is a single cluster of 409 members. The overriding characteristic is an extremely high proportion of nonpermanent households (for example, residential caravans and households)—11 times the national average—and a high average household size.

*Area Type 11*

This has 7382 members, which are characterized by high proportions of agricultural workers, "with jobs" private households, homeworking, households with 7+ rooms, and households with 2+ cars. There are three clusters in this group. This group is very clearly an "agricultural/rural" one.

### *Area Type 12*

This higher-order cluster has 5242 members characterized by areas with a high proportion of second holiday homes, lone female pensioner households, and residents aged 65 and over, and a lower proportion of females of reproductive age and persons per household. There are three clusters in the group. This group is mainly found in retirement migration areas and resorts.

### *Area Type 13*

There are 6174 members. They are characterized by areas with high proportions of council households, two-adults-with-children households, population aged 5–14, and more persons per household. There are three clusters in this group. This group represents the "better-off" council housing areas.

## NOTES

1. "Barnardization" refers to a method used by the Office of Population Censuses and Surveys to preserve the anonymity of persons living in EDs with small populations. It consists of the addition of a quasirandom pattern of $-1$, 0, and $+1$ to the cell counts in the 100% sample Small Area Statistics.
2. The distance criteria used here are based on the robust measures advocated by John Tukey (Tukey, 1977), and are regarded as being more reliable than those based on standard deviation measures. The standard CCP program incorporates both types of criteria.

## REFERENCES

CACI (1980) The Classification of Residential Areas and Its Use in Marketing. London: CACI Inc.

CACI (1983) The 1981 ACORN Classification. London: CACI Inc.

Defays, D. (1977) An efficient algorithm for a complete link method. Computer Journal 20:364–366.

Openshaw, S. (1974) A regionalisation algorithm for large data sets. Computer Applications 3–4: 39–80.

Openshaw, S. (1976) A regionalisation procedure for a comparative regional taxonomy of the UK. Area 8:149–152.

Openshaw, S. (1982) A Portable Suite of FORTRAN IV Programs (CCP) for Classifying Census Data for Districts and Counties: An Introduction and User Guide. Newcastle upon Tyne: C.U.R.D.S.

Openshaw, S. (1983) Multivariate analysis of census data: The classification of areas. In D.W. Rhind (ed.). A Census Users Handbook. London: Methuen, pp. 243–264.

Openshaw, S., and Gillard, A.A. (1978) On the stability of a spatial classification of census enumeration district data. In (P.W.J. Batey, ed.). Theory and Method in Urban and Regional Analysis. London: Pion, pp. 101–119.

Openshaw, S., Cullingford, D., and Gillard, A.A. (1980) A critique of the national census classifications of OPCS-PRAG. Town Planning Review 51:421–439.

Rand, W.M. (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66:846–850.

Sibson, R. (1973) SLINK: An optimally efficient algorithm for the single linkage clustering method. Computer Journal 16:30–34.

Spath, H. (1980) Cluster Analysis Algorithms. Chichester: Ellis Horwood.

Tukey, J.W. (1977) Exploratory Data Analysis. London: Addison-Wesley.

Webber, R.J. (1978) The national classifications of residential neighbourhoods: An introduction to the classification of wards and parishes. P.R.A.G. Technical Paper TP23. London: C.E.S.

Webber, R.J. (1979) Census enumeration districts: A socio-economic classification. OPCS Occasional Paper No. 14. London: H.M.S.O.

Wishart, D. (1977) Clustan 1C User Manual. Edinburgh: P.L.U., Edinburgh University.