Journal of Epidemiology and Community Health 1993; **47**: 238–241

Do inaccuracies in small area deprivation analyses matter?

Richard Reading, Stan Openshaw

Abstract

***Objective*—To assess the accuracy of computerised matching of postcode to enumeration district (ED) and to determine whether any mismatching reduces the validity of methods to distinguish socioeconomic differences in "small area" deprivation studies.**

***Design*—Computerised and manual matching of postcodes to EDs were compared and the census based Townsend deprivation score was compared with socioeconomic data on individual families.**

***Setting*—County of Northumberland, England, 1989.**

***Subjects*—Random sample of 301 families with a child aged less than 15 months.**

***Main results*—With computerised matching only 47% of postcodes were matched to the correct ED. Eighty per cent of the deprivation scores of the computer matched EDs, however, approximated (±2) to the deprivation score of the actual ED. When EDs were divided into quintiles according to the deprivation score, accurate manual matching showed that 75% of families in the most deprived EDs were classed as deprived compared with 4% in the most affluent EDs. With the inaccuracies introduced by computer matching of postcodes, the corresponding figures were 56% and 12% respectively.**

***Conclusions*—Computerised matching of postcodes to EDs is highly inaccurate, but this has little effect on the allocation of deprivation scores. The socioeconomic inequalities shown by the deprivation score are blunted, but not eradicated, by this mismatching.**

J Epidemiol Community Health 1993; **47**: 238–241

Department of Community Paediatrics Northumberland Health Authority
R Reading
Centre for Urban and Regional Development Studies University of Newcastle upon Tyne, Newcastle upon Tyne
S Openshaw

Correspondence to:
Dr R Reading, Jenny Lind Department, Norfolk and Norwich Hospital, Brunswick Road, Norwich NR1 35R

Accepted for publication October 1992

Ecological studies of "small areas" provide a useful basis for measuring inequalities in health and for investigating the influence of material deprivation on different aspects of health.[1 2] Although wards[1 3] and postcode sectors[4] have been used for some large scale studies, the census enumeration district (EDs) is increasingly preferred as the geographical unit of analysis.[5–7] This is because these smaller areas tend to have more socially homogeneous populations and their size allows greater geographical resolution.

The principle behind these methods is that health data relating to individuals can be classified using aggregated demographic, socioeconomic, and housing characteristics obtained from census information for the ED in which they live. Cases are matched to EDs by the postcode of the address of their residence. While this is a straightforward procedure in Scotland, where EDs are defined by postcodes, this is not the case in England and Wales where postal and census boundaries do not coincide and there is not, at present, a completely accurate postcode to ED matching table available. Although such a table will be released with the 1991 census results, most studies to date that have used the 1981 census data have had to rely on a geographical method whereby the Ordnance Survey 100 metre grid reference of the postcode is matched to the nearest ED "centroid" 100 metre grid reference. This introduces potential inaccuracy as a result of the mismatching of addresses to EDs. Estimates of the extent of this mismatching range between 15% and 40%.[8 9] In addition, however, there are the more commonly acknowledged sources of inaccuracy that arise from social heterogenity within EDs,[10] the problem of the ecological fallacy in which average area characteristics may be unrepresentative of the individual cases of interest[11–13], and the possibility of social change after the census[10], let alone the controversy over whether any of the census derived indices actually measure an area's experience of material deprivation.[14 15]

As part of a project examining small area differences in child health in Northumberland, a study was carried out to assess the effect of these various sources of error, particularly mismatching, on the ability of a current computerised small area method of social classification to reflect accurately levels of material deprivation in families with young children in the community.

Methods

The data were derived from a random sample, taken from the district Birth and Immunisation Register, of 301 infants less than 15 months of age. For each of these cases three items of information were collected (see fig 1). The address was located on a census boundary map, allowing identification of the actual census ED of residence for 282 cases, the remaining 19 addresses could not be found on the map or the ED was unidentifiable. The postcode of each case was also matched to an ED using a computerised system that geographically matches postcodes to ED centroids by the nearest grid reference. (The actual system used was that which is incorporated in the Super Profiles programme).[9 16] This resulted in 278 cases being successfully matched: the system could not match the remaining 23 postcodes to an ED.

In addition, a questionnaire was sent to the child's health visitor asking for verification of the address and whether or not the child's family was

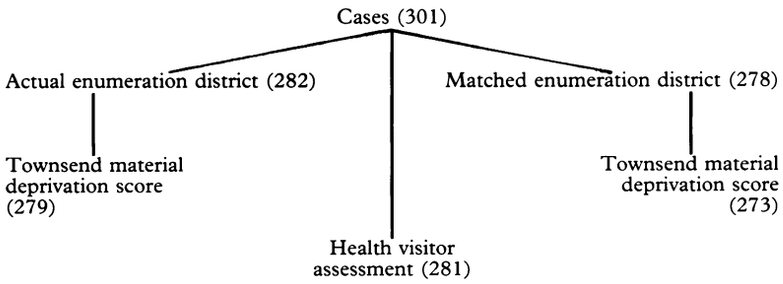

Figure 1 Information collected for each of the cases in the study (numbers in parentheses are the numbers of cases for which information was available)

considered to be deprived. Clear criteria were given for classifying a family as deprived (see table I). A total of 281 questionnaires were returned, nine children had moved, been adopted, or died, and of the remaining 272 families, 85 (31%) were classified as deprived.

Table I Deprivation criteria

(1) Neither parent in full time employment.
(2) Family do not own a car.
(3) Family live in rented, council or tied accommodation.
(4) Family live in overcrowded home (more than 1 person per room)
(5) Family on social security benefits.
(6) Single parent family.
(7) More than two children per adult in the family.
Two or more of above to be classified deprived

The Townsend material deprivation score was calculated for both the actual census ED of residence and the matched ED. This score is produced by adding the Z-scores of four census variables: the proportion of unemployed, economically active adults; the proportion of households without the use of a car; the proportion of households with more than one person per room; and the proportion of households that are not owner-occupied. The method of calculating the score was the same as that described by Townsend *et al*[1] but census variables for ED rather than electoral ward were used. A score could not be calculated for around 2% of EDs because census information was suppressed or because the matched EDs were outside the county. Cases with complete information were thus classified by the deprivation score of the ED they were matched to, the deprivation score of their actual ED of residence, and as deprived or not from the health visitor's assessment.

The study received ethical committee approval. We were concerned to ensure that the collection of this information about families without their knowledge was acceptable. It was felt that this was health service information, the information would not leave the health authority, specific details of the family were not recorded, and it was explicit in the study that the details were based on health visitor opinions and circumstantial knowledge, rather than purporting to be accurate details based on confidential data about individual families.

Results

MISMATCHING OF EDS

Table II shows how well the computer system matched postcodes to their actual ED for the 268 cases for whom information on both was available. Less than 50% of postcodes addresses are correctly matched. Of the remainder, most were matched to the correct ward but there were still 41 of 268 (15%) cases who were not. A very small number was matched to the wrong county; it is presumed that these anomalies would be excluded from any subsequent analysis. Most mismatched cases, even those mismatched by ward, however, were matched to an ED adjacent to the actual ED of residence. These errors reflect the lack of common boundaries for unit postcodes and census EDs, the limit of resolution to the nearest 100 metre grid reference, the irregular shape of the boundaries of both postcodes and census EDs, and the fact that the grid references of both the postcode and the census ED were not necessarily at the geometric centre of the respective areas.

MISCLASSIFICATION OF DEPRIVATION SCORE AS A RESULT OF ED MISMATCHING

The question arising from the previous section is whether the extent of mismatching shown has a noticeable impact on the way that cases are classified by the deprivation score of their apparent ED of residence. Figure 2 is a scattergram of the Townsend material deprivation scores of the matched EDs plotted against those of the actual ED of residence. This shows that most scores cluster about the line of equivalence. The calculated Townsend scores of the EDs in this study range from −8 (at the most affluent end of the spectrum) to +7 (at the deprived end). Just under 50% are exactly the same, corresponding to the correctly matched EDs, but 60% of the Townsend scores of the matched EDs are within ±0·5 of that of the actual ED, 70% are within ±1, and 80% are within ±2. Postcoded addresses that are matched to the wrong ED tended nevertheless to have a similar deprivation score to that of the actual ED of residence, and this presumably reflects the tendency of neighbouring EDs to share similar socioeconomic characteristics, so reducing the impact of postcode to ED mismatching.

ACCURACY WITH WHICH THE DEPRIVATION SCORE DISCRIMINATES AREAS

Cases may be classified by the deprivation score of their true ED of residence or by the deprivation score of the computer matched ED. The former represents the discrimination of cases by small area based score which would occur with perfect postcode to ED matching; the latter represents the discrimination that occurs in practice. The information on deprivation in individual families from the health visitor's questionnaire can be used to verify how well the deprivation score represents areas with different levels of deprivation in the group of cases we are interested in; families with young children.

Table III shows the proportion of families that were classified as deprived in each of the quintiles defined by deprivation score, both of the actual

Table II Accuracy of computer matching of enumeration district (ED).

Exact match	127 (47%)
ED mismatched but ward matched	96 (36%)
ED and ward mismatched but local authority district matched	36 (13%)
Only counties matched	5
Mismatched counties	4
Total cases for which information available on both matched and actual EDs	268

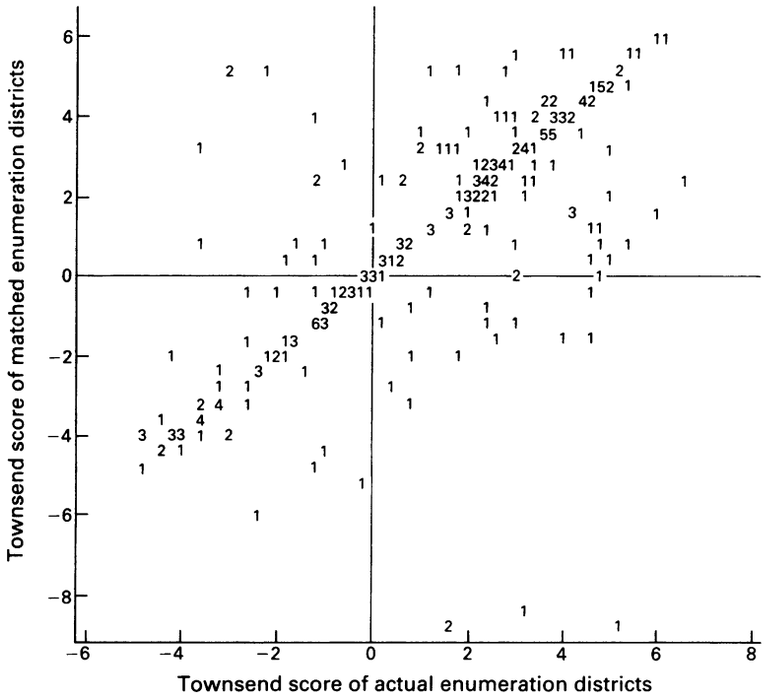

Figure 2 Scattergram plotting Townsend material deprivation score of matched enumeration district against Townsend material deprivation score of actual enumeration district of residence.

ED of residence and the matched ED. When cases are grouped by the score of the actual ED of residence, in the quintile of EDs with the highest deprivation scores, 75% of the families are classed as deprived. In the areas with the lowest deprivation scores, however, only 4% of families are classed as deprived. This represents the discriminatory power of the deprivation score with perfect postcode to ED matching.

When the cases are classified by the score of the matched ED the equivalent results show the extent by which the discriminatory power of the score is reduced by the mismatching: in the areas with the highest deprivation scores 56% of families are classed as deprived (compared with 75%), while in the EDs with the lowest deprivation scores 12% of families are so classed (compared with 4%).

Discussion

This study was undertaken to measure the extent to which mismatching of postcodes to EDs affects the results of studies from England and Wales which use census based small area methods to distinguish areas with different levels of material deprivation. In these studies health data on individuals are related to socioeconomic data for the small area, these being used as a proxy to measure the effects of material deprivation on the health of individuals. EDs are becoming the geographical units of choice in these studies because they are the smallest units for which census information is available. This assumes that the use of smaller, more homogeneous, areas will improve the accuracy of the study. We have investigated whether this reasonable assumption is justified given that any postcode based data has first to be matched to the census ED using methods that can involve considerable geographical approximation.

The results showed a high proportion of addresses matched to the wrong census ED. Parker and Craft (personal communication) have shown a similar degree of postcode to ED mismatching in data from the Northern Regional Cancer Register. Most mismatching in our study was to neighbouring EDs. ED boundaries are irregular, the population distribution, and hence the location of the centroid, is often eccentric, and the geographical resolution is limited to the nearest 100 metre grid reference, so it is hardly surprising that considerable mismatching exists. A further source of inaccuracy occurs because the grid reference of a postcode is given as the south west corner of the 100 metre Ordnance Survey grid in which the first house in the postcode lies. The effect of this has been investigated by Gatrell *et al*,[9] who showed that adding 50 metres north and west to the stated grid reference could reduce the mismatching of postcodes to EDs from 40% to 25%. We did not use this 50 metre correction in the postcode matching system. Although this might have improved the accuracy of matching, it is not in widespread use elsewhere. Our results, those of Gatrell *et al*, and those of Parker and Craft all suggest that in comparable small area studies, between 40% and 55% of EDs are mismatched.

Nevertheless, because neighbouring EDs often share common social characteristics, the extent of misclassification in the Townsend score was considerably less. Some 80% of cases were classified to within ±2 of the Townsend score of their actual ED of residence. Although this may seem a wide margin of error, the Townsend score is not intended to be a highly discriminating method of social classification;[1] rather it is a means whereby areas may be ranked approximately by the socioeconomic characteristics of their inhabitants.

The results shown in table III go as far as the limits of this study allow in answering the questions posed in the introduction. These were: how reliable is the Townsend deprivation score in identifying areas with high proportions of deprived families, in view of the potential inaccuracy introduced by the ecological fallacy, social heterogeneity, and post census change; and how much additional inaccuracy is introduced by mismatching of addresses to census EDs? In the absence of mismatching, the Townsend deprivation score discriminates between different types of area so that in the quintile of EDs classed as most deprived by the score, 75% of families with young children can be classed as deprived using our criteria. In the EDs classed as least deprived by the score, however, only 4% of families are classed as deprived. Hence, even in a rural county such as Northumberland where social heterogeneity within small areas is likely to be greater than in more urban settings, and using survey data for 1989 (almost 10 years after the census from which the Townsend score was derived), the score is still a powerful means of classifying areas according to the socioeconomic characteristics of the population. The EDs ranked as most deprived contain

Table III Number of deprived families at different levels of Townsend score (cases in five groups)

Townsend score	Matched EDs			Actual EDs		
	No	*(%)*	*95% CI*	*No*	*(%)*	*95% CI*
>4	20/35	57	(39, 74)	30/40	75	(59, 87)
2–4	27/59	46	(33, 59)	28/64	44	(31, 57)
0–2	21/50	42	(28, 57)	13/48	27	(15, 42)
−2–0	10/49	20	(10, 34)	8/51	16	(7, 29)
<−2	6/51	12	(4, 24)	2/50	4	(0.4, 14)

the highest proportion of deprived families, the EDs ranked as least deprived contain the fewest deprived families, and there is a steady gradient between these extremes.

Of course, false ecological inferences can still be drawn,[17] and there is a suggestion of how these might arise in table III. Even in the most deprived areas, 25% of families were not classed as deprived. This is not a failure of the area based score, it is a result of social heterogeneity and of census boundaries not being constructed to identify socially discrete areas.[12 13] The inevitability of this finding was described by Townsend,[18] who commented that deprived areas did not contain exclusively deprived families, and correspondingly, most deprived families lived outside these so called deprived areas. Table III shows that this type of effect is less when assuming that families that are not deprived live in areas that are not deprived because, although it is unlikely that everyone in an area with a high deprivation score is actually deprived, the converse is quite possible. Perhaps EDs are still too large a geographical unit.

The flattening of the gradient that occurs when the Townsend scores of the matched EDs are used to classify cases represents the added effect of mismatching on the discriminatory power of the score. As might be expected, the greatest blunting occurs at the extreme values of the score, which tend to be the two types of area most of interest, in studies of health inequality. The blunting which occurs, however, is not sufficient to eradicate observed socioeconomic differences between the EDs with different levels of deprivation score. By inference, the effect this is likely to have on the measurement of health inequalities is that their width will be reduced but they will still probably be evident.

Comments are required on three aspects of the design of this study. Firstly, the criteria for classifying families as deprived were chosen arbitrarily and consist of a combination of indicators of material deprivation and two criteria intended to identify families more likely to suffer material deprivation as a result of low income—that is single parent families and large families. To avoid double counting or including single parent or large families who are not deprived, two or more of the criteria had to be fulfilled. The reason for choosing these criteria rather than a more validated questionnaire[12] is that these were simple items which the health visitors would be likely to know about.

Secondly, the reason for not approaching families directly was to ensure a high and unbiased response rate. The data were being used for no other purpose than to test the reliability of a geographical classification.

Thirdly, census indicators derived from all households were used to measure deprivation, yet the validation of the area deprivation score used only families with young children. It could be argued that we should have used census indicators restricted to households containing young children, or should have sampled all households when collecting the validation data. Our reasons were that the purpose of the validation was in the context of a study on child health inequalities, therefore we were specifically interested in the distribution of deprivation among families with young children. We used the Townsend material deprivation index, which is derived from census indicators for all households, because it is convenient and commonly used in studies of health and deprivation. It is the practical problems that arise in studies of this kind that we were attempting to investigate rather than measuring the exact extent of false ecological inferences.

In conclusion, mismatching of postcodes to EDs causes blunting of apparent area based socioeconomic inequalities, but not to the extent that might have been predicted by the amount of mismatching and not to the extent that inequalities are obliterated. We believe our results may apply to any study which uses similar geographical methods to classify cases into enumeration districts for the purpose of investigating socioeconomic factors influencing health. We are not commenting here on the many studies of geographical patterns of disease incidence, where mismatching results in spatial inaccuracy. Neither do these problems affect studies from Scotland where postcodes automatically aggregate into units for which census data are available. Improvements in geographical information systems technology, which will allow more accurate geographical matching when digitised ED boundaries become available, and the release of the 1991 census small area statistics, which will include a postcode to ED table, will reduce the problems associated with mismatching. The increased accuracy is to be welcomed but if studies of health differences before and after these advances are compared it may seem that health differentials have widened simply because the blunting effect of mismatching has been reduced.

1 Townsend P, Phillimore P, Beattie A. *Health and deprivation: inequality and the north*. London: Croom Helm, 1988.
2 Carstairs V. Small area analysis and health service research. *Community Medicine* 1981; **3**:131–9.
3 Carstairs V. Multiple deprivation and health state. *Community Medicine*. 1981; **3**: 4–13.
4 Carstairs V, Morris R. *Deprivation and health in Scotland*. Aberdeen: Aberdeen University Press, 1991.
5 Curtis SE. Use of survey data and small area statistics to assess the link between individual morbidity and neighbourhood deprivation. *J Epidemiol Community Health* 1990; **44**: 62–8.
6 Reading RF, Openshaw S, Jarvis SN. Measuring child health inequalities using aggregations of enumeration districts. *J Public Health Med* 1990; **12**: 160–7.
7 Crow YJ, Alberti KGGM, Parkin JM. Insulin dependent diabetes in childhood and material deprivation in northern England, 1977–86. *BMJ* 1991; **303**: 158–60.
8 Openshaw S. Making geodemographics more sophisticated. *Journal of the Market Research Society* 1989; **31**: 111–31.
9 Gatrell AC, Dunn CE, Boyle PJ. The relative utility of the central postcode directory and pinpoint address code in applications of geographical information systems. *Environment and Planning*, A. 1991; **23**: 1447–58.
10 Carr-Hill R, Sheldon T. Designing a deprivation payment for general practitioners: the UPA(8) wonderland. *BMJ* 1991; **302**: 393–6.
11 Robinson WS. Ecological correlations and the behaviour of individuals. *American Sociological Review* 1950; **15**: 351–7.
12 Openshaw S. *The modifiable areal unit problem*. Norwich: Geo Abstracts, 1984.
13 Openshaw S. Ecological fallacies and the analysis of areal census data. *Environment and Planning* (A), 1984; **16**: 17–31.
14 Townsend P. Deprivation. *Journal of Social Policy* 1987; **16**: 125–46.
15 Morris R, Carstairs V. Which deprivation? A comparison of selected deprivation indexes. *J Public Health Med* 1991; **13**: 318–26.
16 Charlton ME, Openshaw S, Wymer C. Some classifications of census enumeration districts in Britain: a poor man's ACORN. *Journal of Economic and Social Measurement* 1985; **13**: 69–96.
17 Morgenstern H. Uses of ecologic analysis in epidemiologic research. *Am J Public Health* 1982; **72**: 1336–44.
18 Townsend P. *Poverty in the United Kingdom*. Harmondsworth: Penguin, 1979.