WORKING PAPER 500

THE GENERATION OF INDIVIDUAL AND HOUSEHOLD INCOMES AT THE SMALL AREA LEVEL USING SYNTHESIS

MARK BIRKIN AND MARTIN CLARKE

School of Geography
University of Leeds
Leeds LS2 9JT

November 1988

THE GENERATION OF INDIVIDUAL AND HOUSEHOLD INCOMES AT
THE SMALL AREA LEVEL USING SYNTHESIS

MARK BIRKIN AND MARTIN CLARKE

SCHOOL OF GEOGRAPHY

UNIVERSITY OF LEEDS

LEEDS, LS2 9JT

1. INTRODUCTION

In this paper we describe a method for generating individual and household incomes for small areas such as postal sectors or census wards. The methodology is based upon the synthetic generation of individuals and households together with their associated attributes from aggregate data. The generation and storage of these synthetically generated households and individuals takes place in the form of lists of attributes on the computer. Although the methodology is based upon the well established techniques of iterative proportional fitting and Monte Carlo sampling, there are few applications in the spatial sciences. We have reported on the theory and application of the methods to generating micro-data for the Leeds Metropolitan District in another paper (Birkin and Clarke, 1988). In this paper we wish to focus on the specific problems faced in the generation of incomes for individuals in the micro-data samples who are differentiated in terms of age, sex, occupation, industry, and locational characteristics. It is our understanding that there have been no successful attempts to generate this type of data reported in the literature and yet there exists an enormous demand for this type of information partly from academics but mainly from the private sector.

The paper is organised as follows: firstly we describe the motivation for wanting to obtain information on income distributions at the small area level. We then describe the general principles underpinning the synthetic micro-data generation procedure before presenting, in detail, the methodology for ascribing incomes to individuals. We proceed to present a set of results for the Leeds M.D. and comment on the validity of the results. Some suggestions as to how the work can be taken further are made in the concluding section.

2. THE DEMAND FOR SMALL AREA INCOME INFORMATION

2.1 Introduction

In this section we examine the demand that exists for information concerning income distributions at the small area level and to what extent surrogates of income distributions have been used in the past, before this information was available. We shall note that most attempts at defining surrogates for income have used a combination of Census variables that are assumed to be positively or negatively associated with income. A number of generally recognised problems emerge from this type of work and are highlighted below.

2.2 Why no small area income data?

In the U.K. there exists no public information on income distributions at the small area level. The lowest spatial unit for which income data is available is the Standard Region (11 in the U.K.) as represented in such sources as the New Earnings Survey (N.E.S.) and the Family Expenditure Survey (F.E.S.). The Inland Revenue collects information at the individual level and therefore could provide data on incomes in small areas, such as postal sectors, but chooses not to on grounds of confidentiality. For the commercial sector and the academic researcher alike, important and useful information on small area incomes is not available and hence we need to resort to modelling to estimate it.

2.3 Users of small area income data

That there exists a demand for information on measures of income at the small area level is demonstrated by the number of commercial firms selling products that give measures of the characteristics of small areas to organisations marketing products and services. These organisations need to know the spatial location of different population groups so that they can target market their products, for example through direct mailing, or wish to optimise the location of stores or branches. As there is no direct measure of income available it is very difficult to estimate the amount of expenditure that is spent on particular goods or services directly. As a consequence, attempts are made to use household structure and socio-economic data to derive expenditure estimates for example, by linking this to Family Expenditure Data. With income data it would be possible to use the F.E.S. data directly.

Another major area where income data would have commercial appeal is in credit scoring. At present credit scoring agencies use geodemographic data to ascertain the credit worthiness of applicants, for example by using their postcode as a discriminator. It would be possible to refine this by adding in income data about small areas.

Academics would also find income data for small areas useful in studies which attempt to understand the relationship between activity variables (e.g. hospital utilisation) and measures of wealth/deprivation.

2.4 Surrogates for small area income data

As we have noted, the lack of small area income data contrasts strongly with the demand for information on the way in which prosperity and spending power varies across geographical space. As a consequence various methods have been used to produce measures of prosperity/deprivation and spending power. In the 1960s, much use was made of car ownership as a proxy for income. This had the advantage that it was readily available from the census and that there was an obvious correlation between levels of car ownership in a small area and average income. However, the census does not discriminate between the owner of a new Rolls Royce and the owner of a ten year old small car.

The need for more sensitive discriminators was recognised in the 1970s when greater demands were being made by private sector companies in the fields of marketing and advertising. Richard Webber (1979) devised a clustering system known as ACORN (A Classification Of Residential Neighbourhoods) based on census data relating to enumeration districts. The idea was to include a large set of census variables and identify clusters of similar areas, the general characteristics of which could be interpreted, say as middle aged, council house dwellers, mainly unskilled workers, and so on. ACORN, which was marketed in this country by CACI, proved a great success, and other companies put similar products on the markets, for example PIN, from PINPOINT and SUPER PROFILES from McIntyre. Although of undoubted value, these clustering systems do not in themselves allow for the identification of income or spending power, but rather identifies the general socio-economic profile of areas. This profile may give an indication of levels of income and expenditure but probably masks the way in which individual and household incomes are distributed within a zone.

It is therefore possible to conclude this section with the observation that there exists a demand for small area income data that is currently being met by surrogate indicators. In the rest of this paper we describe how estimates of these important items of information can be made.

3. <u>PRINCIPLES OF SYNTHETIC MICRO-DATA GENERATION</u>

3.1 Introduction

In Section 2 we have already demonstrated the need for improved small area income data and the inadequacy of existing approaches to the problem. In this section, we present an alternative methodology for estimating missing data. While some of the explicit detail of the argument is phrased in terms of the incomes problem, it will be evident that the approach has a more general application (cf Birkin and Clarke, 1988). The methodology relies upon the use of a technique known as iterative proportional fitting, which is now briefly described.

3.2 The technique of iterative proportional fitting

Let us consider initially a population with three sets of attributes. Label these attribute sets j, k, and m, and to fix ideas let us think of j as a spatial label, k the social class of an individual, and m the income of that individual. Each attribute has a series of states - here a zone of residence, type of occupation, and wage band respectively. The "occupancy matrix" for this population may be written as

$$X_i^{km}$$

the number of individuals in social class k, and income band m, residing in zone i. Alternatively, we might write this as a probability:

$$p(x_i, x_k, x_m)$$

the probability of being simultaneously in residential state i, social class k and income band m. We can combine the two representations

$$p(x_i, x_k, x_m) = X_i^{km} / \sum_{ikm} X_i^{km} \qquad (1)$$

Suppose we are now concerned to estimate the occupancy matrix (X.i.km) given the more restricted set (X.i.k) and a set of relationships between x.i, x.k, and x.m. This constitutes an algebraic representation of the problem of generating small area incomes (by social class).

The maximum information one might know about these distributions would be the set of relationships p (x.i, x.k), p(x.i,x.m) and p(x.k,x.m). This problem might be written as

Estimate $p(x_i, x_k, x_m)$

given	$p(x_i, x_k)$	or	X_{i*m}
	$p(x_i, x_m)$		X_{ikm}
	$p(x_k, x_m)$		X_{*}

Now, because of the interdependencies between the known probability distributions, it is impossible to solve this problem routinely. However a range of (related) approaches are available to solve the problem involving balancing factor methods, information minimising, and mathematical programming, eg. entropy-maximising (see Wilson, 1970; Fienberg, 1977, for example). One of the most robust approaches is the method of Iterative Proportional Fitting (IPF - see Fienberg, 1970). Although the details of the method need not concern us here (see Birkin, 1987, or the original references - Fienberg, 1970, 1977), the property of robustness applies in at least two respects

- maximum likelihood estimates are always produced for the data set provided

- in principle, IPF can be used to estimate full conditional probability distributions from partial ones irrespective of the number of attributes involved. The only effective constraints are computational ones, as we shall see below.

3.3 Relationship to conventional procedures

Referring back to Section 2, we can draw a contrast between two types of procedure for the estimation of missing income data. The first is to estimate incomes directly (eg from car ownership, social class/occupation etc.). If this approach is used without the adoption of the appropriate procedures for handling interdependence within data sets, as discussed above, then it is inevitable that the models involved will be conceptually rather oversimplified ones (cf Section 4.2 below). The clustering approaches of ACORN, SUPERPROFILES, PIN et al attempt to remedy this problem by including large numbers of variables in their classifications. This may only be achieved, however, with a crucial loss of resolution, so that one is left with aggregate measures of zonal structure which may be simply inadequate in many instances.

We can see now that the IPF technique provides the potential to combine these two kinds of approach. Thus it should be possible to obtain EXACT measures of individual characteristics (eg income) based on detailed enumeration of a large variety of other characteristics (eg age, location, occupation, sex, industry, and so on). Further substance to this claim is provided in Section 4. In the meantime, there is one further problem to be considered.

3.4 The role of synthetic micro-data

We have seen above that a key feature of the IPF approach is that it allows us to quantify explicitly the interdependencies between a very large number of attributes. For practical purposes, each attribute may be associated with tens, hundreds, or even thousands of possible states, eg. a city the size of Leeds is commonly divided into over 1500 enumeration districts (EDs) or residential zones. As a simple illustration of the problems which can arise here, suppose that we are concerned with modelling the journey-to-work patterns of residents in Leeds, and suppose that we know the ED of workplace as well as that of residence. The appropriate array may be written as X.ij, and has over 2 million elements - but the working population of the city is only of the order of 300,000! Clearly the array will be a very sparse one (ie many of the elements are zeroes), and this problem can become even more crippling as we add further disaggregation by age, sex and so on (see Section 4).

One way in which this problem of proliferation of data may be tackled is to turn the problem around and, in relation to the above example, to consider the residence and workplace of each individual explicitly. The approach becomes a MICRO-DATA approach when we consider the states associated with each of the 300,000 individuals, rather than the occupancy of some large and rather sparse matrix. The approach is, furthermore, a SYNTHETIC one when we do not actually have micro-data available, but we can try to recreate it, usually using versions of the very occupancy matrices which are so replaced.

In addition to these considerations of storage of information, we can note some further advantages of the micro-level representation (and for more detail see Clarke, 1981; Clarke and Holm, 1987; Caldwell, 1987). One is that it facilitates the assignment of EXACT VALUES rather than categories in states of attributes (eg. exact incomes, ages, expenditures and so on). Another is that precise rules may be applied to individuals. So we may, for example, test persons for unemployment benefit based on exactly the same criteria which would be used if such a test were really to be performed. The same principles underly micro-level dynamics for the purposes of forecasting and updating populations (eg. Duley, Rees and Clarke, 1988).

Thirdly, we may note the possibilities for FLEXIBLE AGGREGATION. Because the data is held in relation to individual entities at very high resolution, it is usually possible to tailor outputs closely to specific needs. Examples are provided in Section 5.

Finally, it is also necessary to draw attention to one unavoidable weakness in the procedure of synthetic micro-data generation. In order to compress very large occupancy matrices into a micro-level representation (or to convert from probability distributions to imaginary individual characteristics, which amounts to the same thing) it is necessary to SAMPLE from the given distribution. The method of MONTE CARLO SAMPLING is frequently applied, but despite this frequency of application the statistical properties of the procedure are still not fully understood (eg. Clarke, 1984; Orcutt et al, 1985). Some practical efforts to quantify the levels of error introduced are presented by Birkin and Clarke (1988).

3.5 Summary

The method of Iterative Proportional Fitting (IPF) provides the framework in which it is possible to combine diverse data sets, to provide, explicitly, crucial linkages between population variables. In effect, this allows us to combine the features of the clustering and direct estimation approaches to data estimation problems. To achieve these benefits, it is necessary to adopt a switch in representation from the occupancy matrix to the imaginary or "synthetic" micro-data unit. While this change incurs a penalty in terms of sampling error, there are many other compensatory gains to be exploited.

4. A SMALL AREA INCOMES MODEL

4.1 Introduction

In this section we attempt to apply the methodology discussed in Section 3 as a means for estimating small area incomes. This process begins with an examination of the data situation, and we show how a variety of data elements may be combined using the IPF procedure (Section 4.2). It is then possible to apply the probabilities to an existing synthetic micro-data base to generate estimated individual and household income profiles. These results may be reaggregated for the purposes of evaluation (Section 4.3). This evaluation suggests that further explicit account must be taken of the spatial dimension in the analysis. A suitably revised model is presented and discussed in Section 4.4.

4.2 Data availability and sub-model structure

The principal source of information on the earnings of individuals in the United Kingdom is the New Earnings Survey (NES), published annually by the Department of Employment since 1972. The annual reports are published at impressive speed (generally within eight months of the actual survey) in six volumes, comprising

A. Report and key results
B. Analyses by agreement
C. Analyses by industry
D. Analyses by occupation
E. Analyses by region and age group
F. Hours: Earnings of part-time women workers: Holidays

We need to bear in mind now that our attempt to simulate incomes takes place as part of a larger synthetic data generation procedure within the SYNTHESIS package. SYNTHESIS focusses on linking population and activity variables from the census with data from other sources including (in addition to the NES) the FES, GHS and other more or less ad hoc surveys. It is, therefore, necessary to link our analysis of earnings to existing information within SYNTHESIS. A list of attributes generated before incomes is provided in Table 4.1. Tables within the NES which relate to these existing attributes are as follows (in 1981):

(i) Earnings by occupation by sex (Tables 122, 123)
(ii) Earnings by industry (by occupation) by sex (Tables 104, 105)
(iii) Earnings by age, sex and (standard) region (Table 128)
(iv) Sex by occupation by industry (Tables

A reasonably high level of resolution is adopted, in relation to the occupation and industry attributes, with 16 Occupation Orders and 22 Standard Industrial Categories (SICs) being considered respectively. Notice that we need to augment the information of Tables 104 and 105 with Tables 122 and 123, because the earnings by industry table are only disaggregated by the most important occupation sub-categories for each industry. In effect, therefore, we need to fill in cells which are not zero, but suppressed by the NES because of the sampling method adopted.

Note: In general, the NES will suppress any cell with less than 50 occupants. Since NES is a 1% sample, this means that any industry/occupation category may have up to 5000 workers in reality and still be suppressed.

The application of a suitable version of the IPF methodology to this data set allows us to produce the array

$$I^{bags}_{QR}$$

where

b is occupation type (16 categories)
a is age (8 NES categories)
g is industry type (22 categories)
s is sex (2 categories)
R is a zone label (R=1: the whole country)
(R=2: Yorkshire/Humberside)
Q has four values: Q=1: mean income
Q=2: lowest decile
Q=3: median income
Q=4: highest decile

Note that the obvious missing element from this list is any small area spatial label. Here again it is because of the suppression methods adopted by NES that small area data can never be made available at any useful levels of resolution for this analysis. A further defect, from our perspective, is that the NES collects spatial data only by zone of workplace. Hence any small area tables that were available would fail to pick up the variation by ZONE OF RESIDENCE on which we are focussing here.

4.3 Model results

The array (I.QR.bags), which was discussed in Section 4.2, is a matrix with 16 X 8 X 22 X 2 X 4 X 2 = 45,026 elements, and is, therefore, too large to reproduce here. In Table 4.2 we show a small portion of the array, showing the average earnings by age and industry males in 'professional (management and administration)' jobs, for Yorkshire and Humberside. Observe that this constitutes an industry-disaggregated version of Table 128 of the NES, and is also available for individual occupation groups. Alternatively, it may be seen as an age and full occupational disaggregation of Tables 104/105, or as an age and industry disaggregated version of Tables 122/123. The disaggregation is achieved by combining this data with known information about the national age-sex-industry-occupation mix.

The information contained in (I.QR.bags) may be used to create a spatially disaggregate incomes array through application of the relationship.

$$I^{bags}_{Qi} = X^{bags}_{i} I^{bags}_{QR} \qquad (4.3)$$

For the reasons we have noted above, however, the information in X:... is held, together with other characteristics, in a micro-scale representation within the SYNTHESIS system. Hence, to generate micro-scale incomes we need to convert I.Qi.bags to an income distribution (for each b,a,g,s category), and then sample using the Monte Carlo method. One way of doing this is to assume a normal income distribution, hence

$$m^{bags}_{R} = I^{bags}_{IR}$$

$$o_R^{bags} = (I_{4R}^{bags} - I_{2R}^{bags}) / 1.281$$

Note: Normality implies that the median and mean of the distribution will be equal, and there is an equal spacing of deciles around the mean. Neither of these conditions will necessarily hold. Alternatives are the Poisson distribution or linearisation of the distribution.

Some outcomes of this model are summarised for two sets of census-based small areas in Leeds - the wards and EDs - in Table 4.3. Column 1 gives the average wages of both sexes working and resident in the zone. Columns 2 and 3 give the minimum and maximum averages for EDs in the ward. Column 4 gives the estimated wage taking into account only the occupation structure of the ward, and average wages by occupation. Column 5 is a general index of deprivation for each ward.

Of course, because this model is an attempt to synthetically generate unobtainable data, we now face a considerable problem in attempting to evaluate the outcome of this procedure! With careful use of local knowledge, however, it is possible to make further progress. For example, we can certainly say that the RANKING of zones in Table 4.3 is a plausible one (and compare Column 5 for formal 'proof' of this). However the spread of averages is much less than we would expect. In fact, the rate of variation between EDs within wards is actually rather greater than the inter-ward variation, but even here the overall variation between EDs is barely more than twofold.

A hint about the reasons for this difficulty is provided in Column 4. Here we see that allowing for variations within the occupational structure of wards produces a similar kind of pattern which lacks the requisite inter-ward variation. The obvious inference to draw is that there may be genuine spatial factors at work, which encourage high earners to exhibit different spatial choice decisions to low earners within distinct occupational or other categories. While this is, on reflection, an intuitively reasonable conclusion, it is one with far-reaching implications. Specifically, it implies that any attempt to project "income-based" measures (notably expenditure) through parallel variations like social class or age, may also be seriously compromised through neglect of these intrinsically spatial characteristics. For our own part, we now investigate a means to circumvent this difficulty in relation to the current problem.

4.4 A revised model of small area incomes

To reiterate, briefly, a difficulty arises in the existing model because spatial choice is affected by income in a more subtle way than is being picked up. Specifically, workers "trapped" in inner city zones could typically be expected to be earning rather less than average incomes for any given job. Similarly, the occupants of affluent suburbs may be expected to be doing rather better than the average.

We can begin to crystallise an understanding of the nature of the problem, as well as a means of its solution, through careful consideration of Figure 4.1. Here we present three idealised groups of individuals, who are considered to be low, medium, or high earners (eg. unskilled manual, skilled manual, and professional workers). There is always likely to be a reasonable amount of overlap between categories. For example, mineworkers, within the manual category, would typically earn more than a trainee solicitor or accountant, classified as professional.

Suppose that a zone is completely dominated by low income earners. According to our existing view of Figure 4.1, in which we would take a random sample of incomes across the distribution, this would yield a zonal average of M1. However, suppose further that the ability to reside in a particular zone is based purely on the ability to meet the costs of living in that zone (like housing,commuting, rates, etc). Under this further assumption, then if the income required to live in a particular zone is M1, then many intermediate workers, and also a few 'high income' earners may also find this a suitable zone of residence (and we reach this conclusion by extending the mean M1 VERTICALLY down Figure 4.1, rather than focussing only on one group at a time, as previously). In order that a zone be completely dominated by low income groups, therefore, we would infer that the average income must be less than L2 (and L3), the assumed minima for groups 2 and 3. Observe that L2 will, typically, be a rather lower figure than M1. A similar line of reasoning would lead us to infer that the level of zonal income necessary to support a completely professional domination would be U2, which is typically higher than M3.

We can develop a general method as follows. Suppose we were to break down the working population of Leeds into occupation groups (say), and to 'distribute' these groups by income according to known profiles (ie mean, standard deviation etc). If we initially disregard the spatial aspects, then by taking a small wage band it is possible to calculate how many individuals of each occupation are likely to be earning that amount. This distribution can be converted to a socio-economic profile, so it is possible to derive a set of 'ideal profiles' for each wage band. By analogy with Figure 4.1, these profiles will imply a high proportion of unskilled manual workers at low incomes, and a high proportion of employers and managers at high incomes. The suggestion now is that if we take a given small area, then we can derive an average income associated with the area by matching its socio-economic profile with that of the corresponding 'ideal profile'.

Finally, to allow that the new income estimates reflect only spatial trends, we need to divide gross adjusted income by a basic income which ascribes the variation due to the age-sex-industry-occupation mix of an area. A set of factors which allow for the purely spatial component to earnings ability can be derived in this way. This is shown for Leeds wards in Table 4.4.

4.5 Addition of rule-based incomes

We have now considered in some detail a method for estimating EARNED income using a variety of dependent characteristics. For completeness, however, it is also necessary to focus attention on the transfers of unearned income within our system of interest, and it is to this problem that our attention now turns.

It is possible to think of the transfer of social security benefits from the state to the individual or household in terms of a formal and often relatively straightforward set of rules of eligibility. The appropriate rules are described in HMSO Social Security Statistics (DHSS, 1981). For example, the eligibility to Family Income Supplement is jointly dependent on household structure and household income. In 1981, the basic requirement here was the family head be in full time employment, but the total family income be less than a total of 60 pounds plus 7 pounds for each dependent child. Benefit is then payable in proportion to the difference between the actual income and this minimum for eligibility.

Under the social security rules which applied in 1981, we have identified three kinds of state benefits to which individuals were potentially eligible. These are unemployment benefit, old age pensions, and supplementary benefit. A further three types of benefit were identified as being potentially available according to the characteristics of households, and these are Family Income Supplement, Child Benefit and One Parent Benefit. The various benefits can be assigned to the members of our synthetic data base by LIST PROCESSING. Each individual or household is read from a data file and tested for eligibility to benefits. Any benefits which are payable can then be recorded, and the individual or household incomes incremented appropriately. Thus, in the case of Family Income Supplement, the first test is that the head of household should be in full-time employment. If this condition holds, then we need to count up the number of dependent children, and calculate the minimum threshold for FIS. We can then identify the actual household income, and compare this with the threshold to determine the eligibility to benefit. If all relevant conditions are met to this point, then the amount of benefit can be calculated and assigned.

Those unearned incomes which are not due to state transfers, and not governed by such simple sets of rules, are obviusly harder to deal with. Through the 1980s this sector has also been becoming gradually more important, and is likely to continue to do so, comprising such elements as share income, inherited wealth, and gains from personal pensions or other income plans. It is also necessary to think about capital assets, most obviously property, and about negative income transfers including taxation, rates, mortgage repayments, and the servicing of other types of debt.

None of these more complex types of income transfer are introduced explicitly here. The outcomes we present in Section 5 therefore relate to gross incomes, derived from both paid employment and state benefits, but not including those types of private income outlined above. This is purely for convenience, and does not mean to deny the importance of these 'miscellaneous' transfers. We reserve this area for our future research agenda, as outlined in Section 6.

5. MODEL OUTPUTS

In this section, we present a series of maps and tables which illustrate some outcomes and potential uses of our small area incomes model. First of all, note that it is possible to identify income distributions in terms of both individuals and households. This capability is a useful model outcome of much hard work in modelling household structure from an earlier stage of the SYNTHESIS procedure (Birkin and Clarke, 1988). Table 5.1 shows how the model outcomes translate into household income distributions, where the data has been aggregated into broad income bands. The same information is represented using even broader categories in Figure 5.1, where the distribution of pie sectors indicates the importance of the different household income groups within the various zones.

Similar information is shown in Table 5.2, and Figure 5.2, where the focus is on the frequency distribution of individual earnings of the two sexes. For ease of presentation, we concentrate on a comparison between the whole city, one of its most affluent wards (Wetherby) and one of the least affluent (Burmantofts). The relative utilities of household and individual data will, of course, be likely to vary for different purposes. For example, if we are considering fuel consumption or expenditure on food, then household income and composition will clearly be the key factors. In relation to expenditure on clothing or entertainments, on the other hand, individual age, social class and income may be more important criteria. Even here, however, the situation will clearly be influenced by the status of an individual within the household, and this will be even more obviously true for economic dependents. In short then, a major feature of SYNTHESIS is the ability to combine a detailed model of individual earnings with a model of household structure, as the uses to which the information may be put are greatly extended.

Another feature of SYNTHESIS which can be demonstrated by practical example is the ability to proceed to much finer levels of resolution than that used in the above figures and tables. One way in which we can do this which is at the same time both playful and insightful, is to generate particular household and income categories like 'YUPPIES' (young professional persons, living alone) or 'DINKIES' (dual income households, with no children). Distributions for these groups are shown in Table 5.3, and Figures 5.3a, 5.3b. Clearly information like this is of great potential use in both commercial applications, like targetted marketting, and in studying the changing social geography of the city from an academic standpoint.

As a final illustration, the information now assembled permits us to begin to look at economic dependency in the spatial context. Specifically, what happens if we start to identify income sources for particular zones of the city? This concept is illustrated in Figure 5.4, which shows the contribution to the total income within a zone which is generated by state transfers. The same kind of idea can easily be extended to higher levels of resolution, for example by looking at income within zones from particular industries, or even particular firms. This kind of analysis has tremendous policy implications, which we reserve for further discussion elsewhere. In Section 6, we will now try to draw together these possibilities for future development.

6 CONCLUDING COMMENTS

We have described the methodology for generating estimates of small area income distributions for cities or regions of the U.K., and presented a set of results for Leeds Metropolitan District. A number of limitations and possible extensions of the work so far can now be outlined.

The first obvious comment is that the results presented above pertain to 1981 and need to be updated for more recent years. A methodology for achieving this have been developed in Leeds and details can be found in Clarke (1986) and Duley,Rees and Clarke (1988). Secondly, we have no means of validating the results against any base data set. This is a general problem in missing data estimation procedures - the very reason for the exercise is to generate data that doesn't exist. A number of possible strategies exist for verification, from submitting the results to the Inland Revenue for scrutiny through to collaborating with, say, a financial service organisation who would have a large, but incomplete data base on incomes from , for example , mortgage applications. For the future, questions could be appended to the census on levels of household income, although this is highly unlikely.

A number of refinements could be undertaken to improve the income estimates within the model. In particular, unearned income through sources such as interest on savings and stocks and shares could be added, though data on these is poor. One potentially interesting area is the link between asset accumulation in the housing market and inheritance. The first generation of large scale owner occupiers are now dying and bequeathing their property to their offspring. With the recent surge in property values. the sums inherited are often very large and can clearly transform a household's prosperity level. By linking a model of asset accumulation in the housing market with an income model it would be possible to begin to perform an analysis of the types of effects being experienced. Some preliminary work on this has been undertaken and the results can be found in Clarke, Longley and Williams (1988).

To conclude, we feel that the methodolgy presented in this paper provides a novel way of generating estimates of income distributions for small areas. Coupled with the other advantages of a micro-simulation approach it can produce profiles of small areas that are useful to a wide variety of different users.

7 REFERENCES

Birkin, M (1987) An Iterative proportional fitting procedure (IPF) Computer Manual 26, School of Geography, University of Leeds

Birkin, M and Clarke, M (1988) SYNTHESIS: a Synthetic Spatial Information System, methods and examples. **Environment and Planning A**, forthcoming

Caldwell, S.B. (1985) Broadening Policy Models: alternative strategies, in Orcutt,G H , Merz, J and Quinke,H. (eds) **op cit**

Clarke, M (1981) A first principles approach to modellin socio-economic interdependencies using microsimulation **Computers, Environment and Urban Systems 6** 211-27

Clarke,M(1984) Integrating Dynamical Models of Urban Structure and Activities: an Application to Urban Retail Systems. unpublished PhD thesis, University of Leeds

Clarke, M (1986) Demographic processes and household dynamics: a microsimulation approach, in Woods, R and Rees, P.H. (eds) **Population Structures and Models**, Allen and Unwin, London

Clarke, M and Holm, E (1987) Microsimulation methods in Spatial Analysis and Planning, **Geografiska Annaler, 69 B** p 145-164

Clarke, M, Longley P, and Williams, H.C.W.L. (1988) Microanalysis and Simulation of Housing Careers: subsidy and accumulation in the U.K. housing market. **Papers of the Regional Science Association** forthcoming

D.H.S.S. (1981) **Social Security Statistics 1981** H.M.S.O. , London

Duley, C, Rees, P.H. and Clarke, M (1988) A microsimulation model for updating households in small areas between censuses. Paper presented to the workshop on Multistate Demography: measurement, analysis and forecasting, Kerkebosch Castle, Zeist, Netherlands, November 1988

Fienberg, S (1970) An iterative approach for estimation in contingency tables **Annals of Mathematical Statistics, 41** p907-917

Fienberg, S (1977) **The Analysis of Cross-Classified Categorical Data** M.I.T. Press, Cambridge, Mass.

Orcutt, G. H., Merz, J and Quinke, H (1986) **Microanalytical Simulation Models to Support Social and Financial Policy** North Holland, Amsterdam, Holland

Webber, R (1979) **A Classification Of Residential Neighborhoods**, CACI, London

Wilson, A.G. (1970) **Entropy in Urban and Regional Modelling** Pion, London

TABLE 4.1 ATTRIBUTES GENERATED PRIOR TO INCOMES

a) Household attributes

- location
- household structure and composition
- tenure
- primary retail location
- car ownership

b) Individual characteristics

- status within household
- exact age
- sex
- marital status
- economic activity
- educational attainment
- occupation
- industry of employment

TABLE 4.2 PROFESSIONAL WORKERS (MANAGEMENT AND ADMINISTRATION):
MALE EARNINGS BY AGE AND INDUSTRY.

	< 18	18-20	21-24	25-29	30-39	40-49	50-59	60+
1. Agriculture, forestry and fishing (*)	64.51	98.29	100.72	138.06	168.86	176.70	172.69	162.98
2. Mining and quarrying (*)	64.51	98.29	100.72	138.06	168.86	176.70	172.69	162.98
3. Food, drink and tobacco	69.53	105.94	108.56	148.80	182.01	190.46	186.13	175.67
4. Chemicals	75.68	115.31	118.16	161.96	198.10	207.30	202.59	191.20
5. Metal manufacture (*)	64.51	98.29	100.72	138.06	168.86	176.70	172.69	162.98
6. Mechanical engineering	61.25	93.32	95.63	131.08	160.32	167.77	163.96	154.74
7. Instrument engineering (*)	64.51	98.29	100.72	138.06	168.86	176.70	172.69	162.98
8. Electrical engineering	63.91	97.38	99.78	136.77	167.29	175.06	171.08	161.46
9. Vehicles	60.31	91.90	94.16	129.07	157.87	165.20	161.45	152.37
10. Textiles	53.53	81.56	83.57	114.55	140.11	146.62	143.29	135.23
11. Clothing and footwear (*)	64.51	98.29	100.72	138.06	168.86	176.70	172.69	162.98
12. Timber and furniture (*)	64.51	98.29	100.72	138.06	168.86	176.70	172.69	162.98
13. Paper and printing	62.15	94.69	97.03	133.00	162.68	170.23	166.36	157.01
14. Other manufacturing (*)	64.51	98.29	100.72	138.06	168.86	176.70	172.69	162.98
15. Construction	54.35	82.81	84.86	116.32	142.27	148.88	145.49	137.31
16. Gas, electricity and water	65.60	99.95	102.42	140.38	171.70	179.68	175.60	165.72
17. Transport and communications	62.34	94.98	97.32	133.40	163.17	170.74	166.87	157.49

18. Distributive trades							
64.21	97.83	100.25	137.41	168.07	175.88	171.88	162.22
19. Insurance, banking and finance							
74.41	113.37	116.17	159.23	194.76	203.81	199.18	187.98
20. Professional and scientific services							
57.95	88.30	90.48	124.02	151.69	158.73	155.13	146.41
21. Public administration							
64.85	98.81	101.24	138.78	169.74	177.62	173.59	163.83
22. Miscellaneous services							
61.59	93.84	96.15	131.80	161.21	168.69	164.86	155.59

(*) For sectors 1, 2, 5, 7, 11, 12 and 14, industry-wide averages were applied due to suppression of data in the parent data base.

TABLE 4.3 Income Characteristics by Ward from the Small Area Model

ZONE (BY RANK)	AVERAGE	MINIMUM	MAXIMUM	OCC MODEL	CROWDING (*)
1. Wetherby	123.20	109.3	138.0	126.8	4.3
2. North	120.32	99.4	131.7	129.1	1.5
3. Cookridge	119.28	108.2	140.4	126.8	2.7
4. Roundhay	117.62	104.9	131.9	129.9	9.3
5. Garforth	117.05	106.4	131.7	116.5	11.4
6. Otley	116.99	104.6	138.1	121.8	5.5
7. Barwick	116.94	106.7	140.2	115.4	10.4
8. Horsforth	116.38	109.2	130.6	122.8	4.6
9. Moortown	115.70	98.3	135.6	124.7	4.9
10. Rothwell	115.41	100.2	131.6	112.3	5.0
11. Halton	115.04	105.2	124.9	117.5	9.0
12. Aireborough	112.53	99.7	131.9	119.6	5.3
13. Morley North	112.53	103.4	123.9	114.6	10.6
14. Pudsey North	112.52	99.2	121.9	117.9	10.3
15. Weetwood	112.14	94.6	120.6	124.5	14.7
16. Morley South	111.21	99.9	128.1	112.9	21.8
17. Whinmoor	110.19	95.6	120.8	112.5	14.1
18. Beeston	109.47	91.1	125.0	109.5	16.2
19. Wortley	109.45	95.4	118.2	111.5	10.7
20. Headingley	109.44	88.5	130.8	119.8	64.1
21. Pudsey South	109.44	99.0	129.1	112.1	10.0
22. Armley	107.84	100.0	132.9	119.3	29.5
23. Kirkstall	107.44	88.1	119.1	110.9	20.7
24. Chapel Allerton	107.03	89.9	135.1	110.4	60.1
25. Bramley	106.98	94.5	118.6	110.1	22.6
26. Middleton	106.65	94.5	122.6	108.4	24.1
27. Hunslet	106.40	95.8	121.0	107.0	26.7
28. City	105.58	87.5	116.9	107.8	24.9
29. Seacroft	105.23	92.1	113.5	106.1	33.5
30. Richmond Hill	105.09	92.9	119.9	105.2	43.1
31. Harehills	105.06	93.1	117.7	108.3	62.8
32. University	105.02	87.3	118.2	112.4	52.2
33. Burmantofts	104.58	94.6	118.3	107.0	25.0

(*) 'Crowding' is defined here as the number of households, per thousand, with greater than 1.5 people per room.

TABLE 4.4 RESULTS FROM VERTICAL INCOMES MODEL

	MALES				FEMALES			
WARD	ADJUSTED INCOME	BASIC INCOME	FACTOR	FACTOR RANK	ADJUSTED INCOME	BASIC INCOME	FACTOR	FACTOR RANK
1	153.788	136.114	1.130	9	81.482	79.068	1.031	17
2	107.866	128.272	0.841	26	59.811	78.619	0.761	28
3	142.041	133.578	1.063	14	95.352	79.495	1.199	8
4	114.273	129.730	0.881	23	45.942	78.483	0.585	33
5	114.273	128.651	0.888	22	58.078	78.632	0.739	29
6	97.186	127.240	0.764	33	66.746	77.813	0.858	23
7	122.817	127.687	0.962	20	65.879	78.438	0.840	25
8	98.254	126.773	0.775	29	66.746	77.464	0.862	20
9	164.468	138.376	1.189	4	102.286	81.678	1.252	5
10	147.381	134.819	1.093	10	95.352	80.390	1.186	12
11	146.313	134.362	1.089	11	98.819	80.392	1.229	6
12	97.186	126.448	0.769	31	62.412	78.140	0.799	26
13	136.701	130.605	1.047	15	96.218	81.724	1.177	15
14	160.196	137.622	1.164	7	106.620	82.246	1.296	3
15	99.322	127.178	0.781	28	66.746	77.486	0.861	21
16	112.137	128.130	0.875	24	62.412	78.557	0.794	27
17	104.662	128.038	0.817	27	66.746	78.043	0.855	24
18	161.264	137.885	1.170	5	102.286	80.901	1.264	4
19	142.041	133.471	1.064	13	94.485	79.655	1.186	11
20	137.769	132.752	1.038	16	47.676	79.655	0.599	32
21	177.284	142.871	1.241	1	109.221	83.209	1.313	1
22	160.196	137.419	1.166	6	98.819	80.984	1.220	7
23	145.245	134.614	1.079	12	94.485	79.834	1.184	13
24	132.429	131.564	1.007	18	48.543	79.889	0.608	31
25	97.186	126.174	0.770	30	66.746	77.633	0.860	22
26	136.701	132.609	1.031	17	95.352	80.325	1.187	10
27	170.876	140.341	1.218	3	108.354	83.130	1.303	2
28	97.186	127.148	0.764	32	68.480	78.176	0.876	19
29	106.798	126.388	0.845	25	50.276	78.527	0.640	30
30	155.924	135.821	1.148	8	95.352	80.295	1.188	9
31	173.012	141.100	1.226	2	95.352	80.618	1.183	14
32	131.361	131.058	1.002	19	83.216	80.352	1.036	16
33	120.681	129.991	0.928	21	72.814	79.255	0.919	18

TABLE 5.1 Simulated Household Income Distributions

Zone	Household income <50	-100	-200	-300	-500	500+	Yuppy	Dinky
1	110	148	609	297	99	0	72	71
2	184	311	533	113	16	0	15	11
3	117	147	536	286	79	2	39	91
4	117	226	433	90	8	0	18	10
5	171	289	488	122	10	0	27	14
6	186	337	438	89	4	0	6	8
7	181	285	504	160	35	0	26	21
8	153	335	435	100	8	0	7	7
9	95	122	439	292	135	6	83	81
10	102	163	578	322	97	1	85	72
11	85	94	457	248	94	1	59	71
12	174	364	459	82	6	0	14	3
13	170	170	328	194	92	1	52	50
14	94	101	471	332	122	2	81	69
15	136	258	349	87	2	0	10	5
16	171	250	455	109	23	0	11	11
17	117	261	445	119	8	0	15	7
18	101	107	392	219	93	5	55	55
19	107	164	508	271	74	1	43	72
20	156	193	583	179	36	0	57	25
21	113	111	418	274	133	7	82	62
22	103	110	500	311	125	5	66	76
23	114	145	547	307	84	3	69	71
24	142	192	608	173	26	0	24	24
25	189	356	440	78	9	0	9	5
26	92	138	512	257	88	2	52	69
27	103	120	407	292	136	7	81	92
28	188	327	431	70	3	0	7	5
29	231	292	398	74	9	0	14	7
30	105	131	359	224	82	2	34	43
31	86	90	467	343	159	7	93	81
32	110	195	465	190	40	1	36	33
33	123	230	609	161	29	0	27	22
CITY	4426	6762	15601	6465	1964	53	1369	1344

TABLE 5.2 Personal Incomes for Two Contrasting Wards

Earned Income	Males Burm	Males Weth	Males City	Females Burm	Females Weth	Females City	Both sexes Burm	Both sexes Weth	Both sexes City
0-40	12	0	197	7	21	571	19	21	768
-50	28	3	409	43	11	1142	71	14	1551
-60	42	8	665	107	10	1581	149	18	2246
-70	68	12	975	76	37	1545	144	49	2520
-80	140	11	1698	37	53	1542	177	64	3240
-90	131	21	2330	32	57	1466	163	78	3796
-100	135	28	2823	14	51	1003	149	79	3826
-110	98	52	3136	5	32	736	103	84	3872
-120	82	51	3007	3	29	548	85	80	3555
-130	58	81	2828	0	27	414	58	108	3242
-140	25	92	2402	0	11	270	25	103	2672
-150	10	85	1971	0	8	193	10	93	2164
-160	10	85	1698	0	5	104	10	90	1802
-170	4	93	1462	0	2	52	4	95	1514
-180	0	77	1155	0	4	38	0	81	1193
-190	0	50	892	0	1	29	0	51	921
-200	1	62	677	0	3	36	1	65	713
-210	0	52	505	0	1	23	0	53	528
-220	1	40	432	0	1	30	1	41	462
-230	1	32	319	0	0	18	1	32	337
-240	1	26	232	0	0	4	1	26	236
250+	1	94	812	0	0	8	1	94	820

Figure 4.1 Income distributions by Class

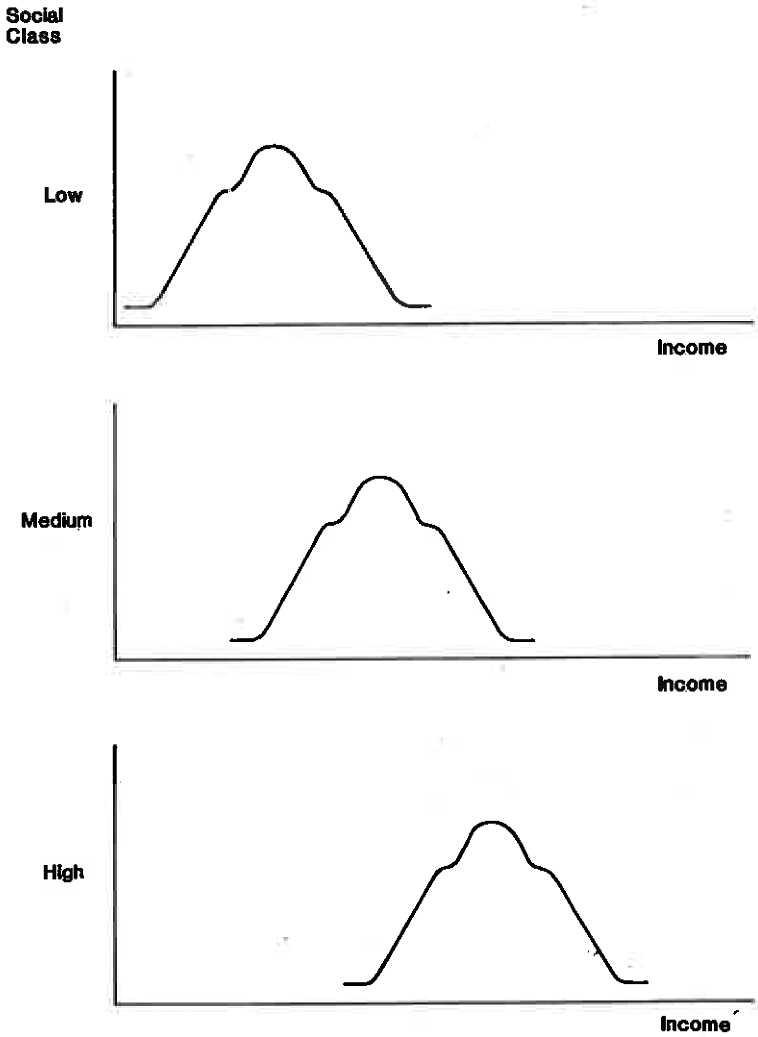

FIGURE 5.1 Household Income Thresholds

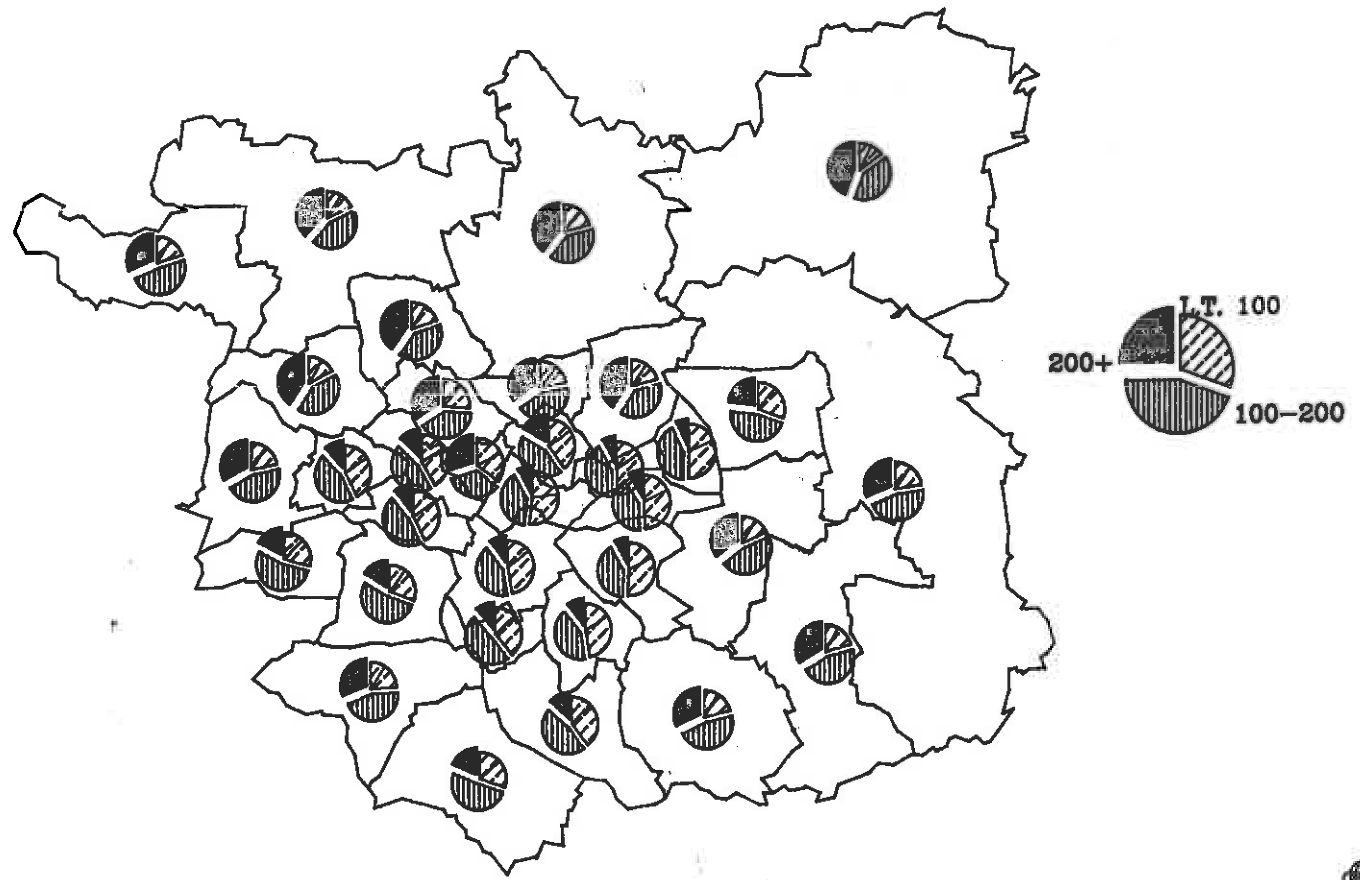

Figure 5.2 Income Frequency Distributions for Two Contrasting Wards in Leeds M.D.

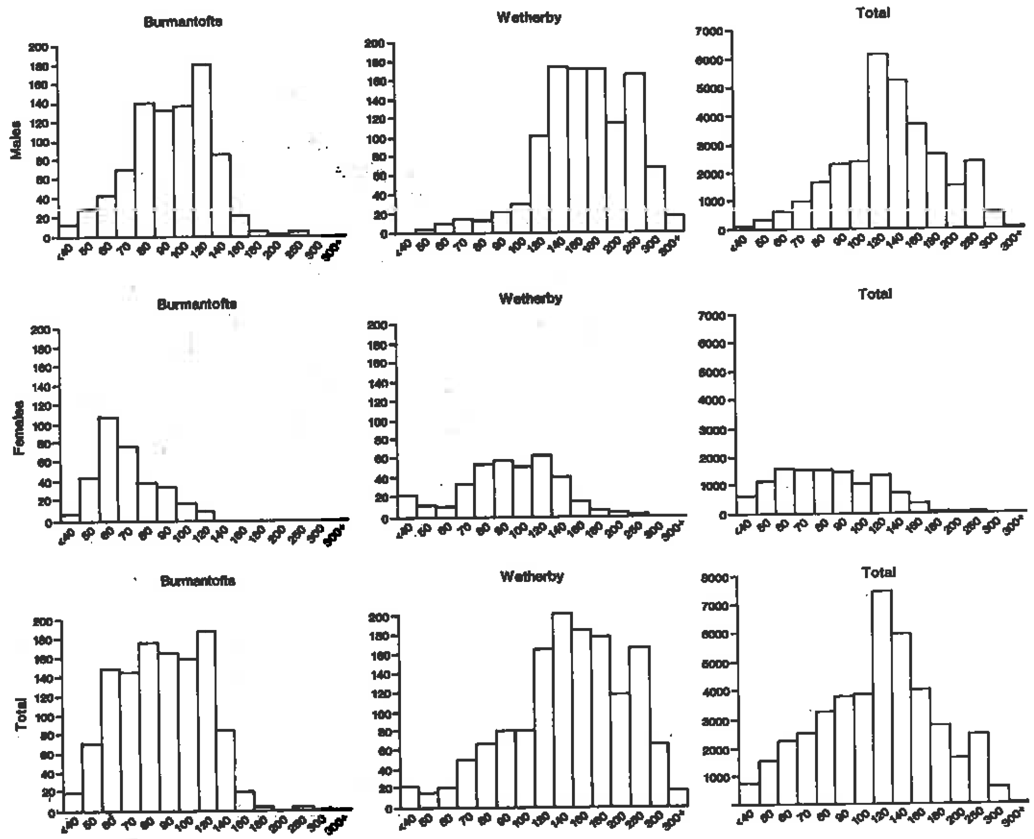

FIGURE 5.3a Young Professional Persons

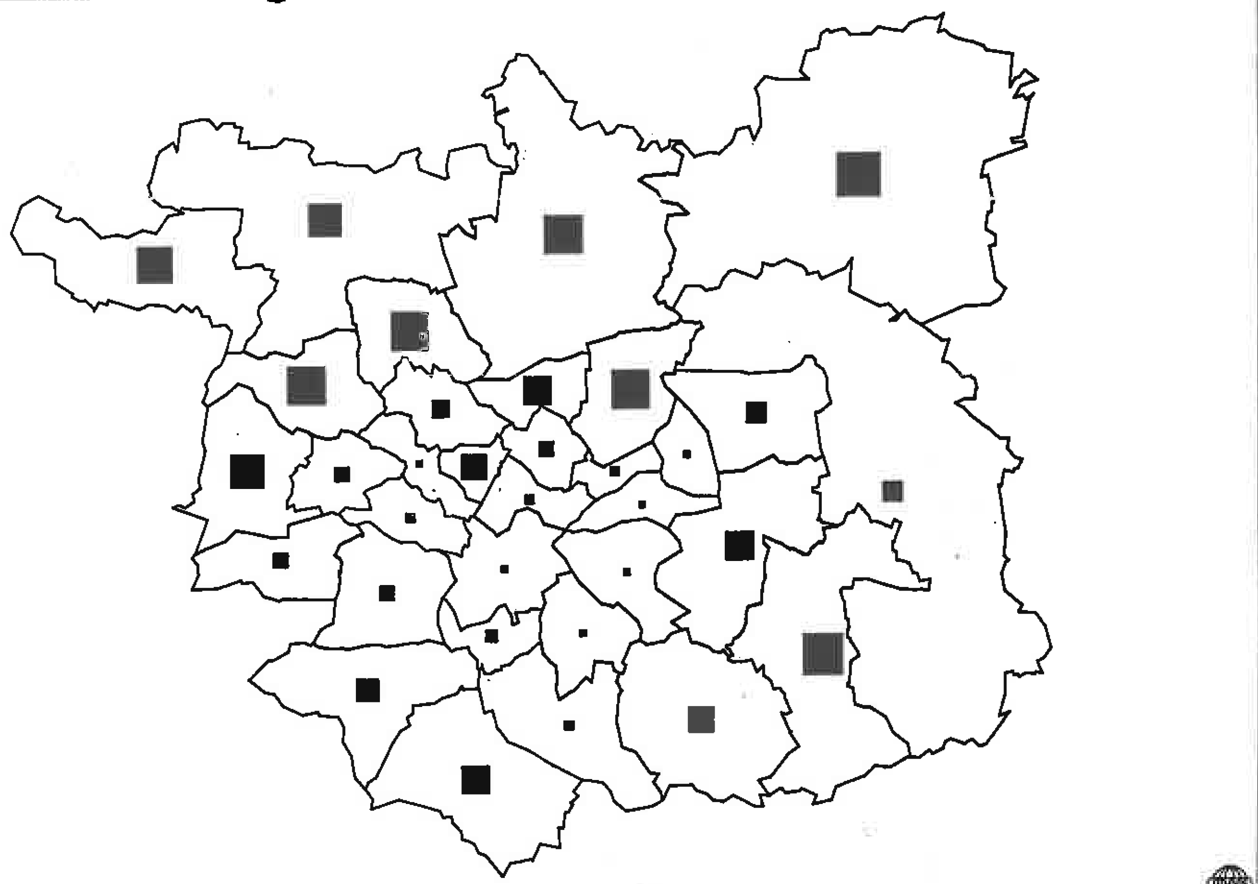

FIGURE 5.3b Dual Income Households Without Dependents

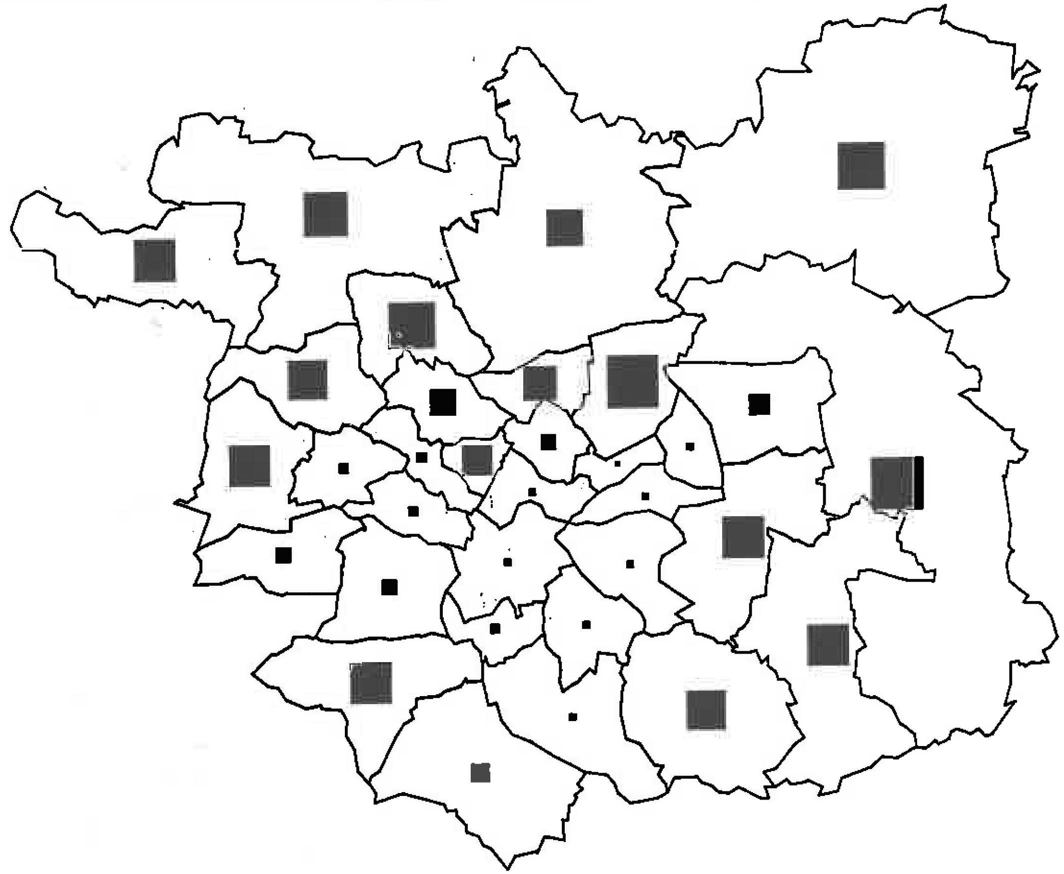

FIGURE 5.4 Economic Dependency Ratio