WORKING PAPER 494

RUIN - REALLY USEFUL INFORMATION - FOR URBAN AND REGIONAL ANALYSIS: METHODS AND EXAMPLES

MARK BIRKIN AND MARTIN CLARKE

School of Geography
University of Leeds
Leeds LS2 9JT

June 1987

SYNTHESIS: A SYNTHEtic Spatial Information System for Urban and Regional analysis with methods and examples

Mark Birkin and Martin Clarke

School of Geography
University of Leeds
Leeds LS2 9JT
UK

1. Introduction

1.1 Preliminary comments

An increasing demand is being made by academics, planners, market research firms, commerce, and industry in general for the making available of information on population, the economy and associated activities that is useful in relation to the problems being studied. In this paper we report on the development of a method that allows for the integration of data pertaining to metropolitan regions to generate a data base that we shall argue is of more potnetial use in a number of application areas than what currently exists. In this introductory sector we focus on three areas: the deficiences of existing information systems; the motivation for generating SYNTHESIS (Synthetic Spatial Information System); and an initial description of the main features of SYNTHESIS.

1.2 Deficiences of existing information systems

Considerable effort has been devoted to the development of spatial information systems in recent years. Although the situation with respect to census-based geography was much improved through the release of systems such as SASPAC there is still great difficulty in linking together the various official survey data such as FES, New Earnings Survey (NES), NOMIS, and so on. On the non-residential side the picture is worse. The demise of the Census of Production means that we are denied from obtaining interesting information on industrial and retail activity, and in many cases useful data capture is undertaken by private market research companies (e.g. Dun and Bradstreet) and only available at a price. The activity patterns of residents in cities remain even more of a mystery. Travel to work and migration data are in principle available from the census but with a considerable time delay factor. Other activity patterns such as shopping and leisure trips, educational flows, etc. have to be derived from one-off surveys or extracted from independent data sets. Almost always there are problems with the definition of consistent spatial units. As a consequence many potentially interesting pieces of integrated analysis simply cannot be undertaken unless suitable approximations and assumptions are made. The recently released BBC Domesday laser disk while containing a plethora of information on every aspect of life in Britain is neither sufficiently flexible or specified at a fine enough level of spatial resolution to be useful to the urban analyst.

The serious deficiencies of existing spatial information systems has been recognised by ESRC in its recent research initiative to establish regional research laboratories with a special function to coordinate the assembly and dissemination of information pertaining to Metropolitan areas. Precisely how the regional laboratories are to achieve this function on relatively small sums of money remains to be seen. That a demand for useful information concerning the spatial distribution of stocks and activities exist is clearly demonstrated by the demand for census based products such as CACI's ACORN package, the PIN classification developed by Pinpoint, and McIntyre's SUPERPROFILE, despite their known limitations.

A second major deficiency of many information systems is that they are out of date. Reliance on 1981 census data in 1987 will ignore major trends such as the increase in unemployment, the growth of owner occupation fuelled in part through council house sales, the continuing counterurbanisation in many areas, and so on. There is a compelling need to update information systems. Where this is not possible through on-going surveys, then model-based forecasting methods can be used, as we will describe later.

A third deficiency is that very little data is available at the micro-level. Although many countries, such as the US and Sweden, encounter few problems in making micro-data sets available from the census, in the UK barriers are erected to prevent its release. However, the ingenuity of the modeller can in part remove these barriers by synthetically generating micro-data from aggregate distributions. The method is fully described in section 2.

In summary, therefore, the lack of integration, flexibility and spatial uniformity in existing information and data systems, coupled with the lack of micro-data means that their usefulness is unreasonably restricted. Lack of contemporary information implies that we may simply be getting it wrong. These are one set of reasons for developing SYNTHESIS.

1.3 Motivation for developing SYNTHESIS

A second set of reasons for developing a synthetically generated micro-data base such as SYNTHESIS relate to a number of model based issues. First, there is a continued interest in the development of micro-simulation models in a number of different policy related areas (Clarke and Holm, 1986, Orcutt, Merz and Quinke, 1986). One of the basic advantages of microsimulation methods compared with conventional aggregate approaches relates to efficient representation. For systems that are characterised by a good deal of heterogeneity and when interest focusses on a significant number of variables a micro-level representation is often much more efficient that a corresponding aggregate counterpart (Clarke, 1985). A basic prerequisite for these models is the availability of a population (of households, banks, firms, etc.) specified at the micro-level. In a more general context the development of comprehensive models using a mixture of micro- and macro-based methods (Birkin and Clarke, 1986, Clarke and Wilson, 1986) needs a good accounting system, and for reasons we shall describe in section 4 a micro-data base provides an efficient and consistent approach to accounting.

1.4 What is SYNTHESIS

In the rest of the paper we describe the main components of SYNTHESIS. SYNTHESIS is a sample micro-data base generated from a number of different aggregate tabulations such as the census, FES, and so on. The exact list of attributes included will depend on the study being undertaken. In the application we are developing in Leeds we are concerned with stock and activity variables pertaining to individuals and households at the enumeration district level. This information can be augmented by survey data or data collected by local government, health authorities and so on. Another feature of SYNTHESIS is that it is updatable through the use of microsimulation methods (Clarke, 1986, Clarke and Holm, 1987). We describe the principle features of updating in section 4.

The synthetic sampling procedure to generate micro-data from a variety of aggregate data sources is underpinned by a method known as ITERATIVE PROPORTIONAL FITTING. The theoretical and practical considerations behind this method are well known and the way in which we have used it to general samples of households and their individual members with a relevant set of attributes is described in the next section. We also examine other ways of generating useful micro-data, such as the merging of data files. We also stress the importance of linking activity variables into the household and individual attribute list. Section 3 is devoted to the presentation of some example results from SYNTHESIS to illustrate the consistency of the approach and its potential usefulness. In section 4 we describe in more detail the variety of uses that we envisage for SYNTHESIS. Finally, in section 5 we draw some preliminary conclusions and chart some directions for future research.

2. Methodology and approach adopted

2.1 Introduction and review of micro-data generation

The purpose of this section is to describe the theory behind the generation of synthetic micro-data from aggregate distributions, to examine alternative approaches, and to present an outline of the way we have generated SYNTHESIS for the Leeds study area. Recall that we view the need for micro-data both in relation to the development of a powerful and flexible information system as well as forming the input to micro-simulation models. The worst and most common situation is that no micro-data is available (or made available). In this case the methods described in section 2.2 will have to be used. Alternatively it might be the case that some micro-data is available. However it is almost always the case that it will not contain the full list of attributes that is required. In this case it is possible to generate extra attributes using conditional probability distributions generated either directly from data or through the IPF routine described below. An interesting special case occurs when two (or more) micro-data sets are available. It may prove possible to perform a statistical matching or merging of the two data sets in the following way. Let us assume that the first data set contains a set of attributes (a,b,c,d) and the second set (a,c,d,e). Also assume that the attributes are defined in the same way and are sampled from the same population. The task is to generate a single data set containing the attribute set (a,b,c,d,e). The statistical matching is performed by assigning say an individual from the first set to the second set in such a way that the weighted difference in the overall set of common attributes (a,c,d) is minimised. Techniques for solving this problem and for incorporating additional information are emerging (e.g. Radner et al, 1980) although as Paas (1986) points out they are based on fairly ad hoc heuristic methods.

2.2 The theory of iterative proportional fitting

The procedure of micro-data generation is heavily based on the methods of contingency table analysis. This is a well established technique that appears in a multitude of disguises from balancing factors in spatial interaction modelling through to the RAS method in economic accounting. Suppose we are interested in the basic task of generating a vector of individual characteristics $\underline{x} = (x_1, x_2, x_3, \ldots x_m)$. To begin with, we want to generate a joint probability distribution for this attribute vector, $p(\underline{x})$. Once this is done, we may synthetically create or extract individuals from the distribution. Of course, information is typically not available for the full joint distribution, so we have to construct it as a product of conditional and marginal probabilities. The basic idea is to build up one attribute at a time, so that the probability of certain attributes is ("conditionally") dependent on existing attributes

$$p(\underline{x}) = p(x_1)\ p(x_2/x_1)\ p(x_3/x_2,x_1) \ldots p(x_m/x_{m-1}, \ldots, x_1) \qquad (1)$$

A modelling task here is to define the order of attribute dependencies, and we tackle this in a practical context in section 2.3 below. The second problem, which we wish to deal with in this section, is how to absorb as much information as possible in constructing the individual conditional probabilities on the RHS of (1). Consider the relatively simple problem of modelling the joint probability distribution $p(x_1,x_2,x_3)$ [which may be easily converted to a conditional distribution by imposing a rigid assignment on two of the attributes, say x_1 and x_2], subject to known joint probabilities $p(x_1,x_2)$ and (x_1,x_3).

To outline a solution procedure, let $p^i(x_1,x_2,x_3)$ be the ith approximation to the three attribute joint probability vector, and let

$$p^1(x_1,x_2,x_3) = 1/N_1N_2N_3 \tag{2}$$

where N_j is the number of possible states associated with the attribute vector x_j. The vector is then adjusted in PROPORTION to known constraints

$$p^2(x_1,x_2,x_3) = p^1(x_1,x_2,x_3) * p(x_1,x_2)/\sum_{x_3} p^1(x_1,x_2,x_3)$$
$$p^3(x_1,x_2,x_3) = p^2(x_1,x_2,x_3) * p(x_1,x_3)/\sum_{x_2} p^1(x_1,x_2,x_3) \tag{3}$$

We then ITERATE through the equations (3) until a FITTED distribution is obtained when the probabilities are convergent within some acceptable limit. The method is therefore known as ITERATIVE PROPORTIONAL FITTING and generally exhibits fast and reliable convergence properties (Fienberg, 1970; Clarke, 1984).

It is quite straightforward to generalise this procedure to a larger number of attributes, although the notation is slightly difficult. Let:

$Z_k(x)$ be a subset of the set of attribute vectors, $E(\underline{x})$, for which marginal joint probabilities are known;

$W_k(\underline{x})$ be the complement of $Z_k(\underline{x})$ i.e. $W_k(\underline{x}) = E(\underline{x}) - Z_k(\underline{x})$

Then

$$p^1(\underline{x}) = 1/\sum_{i=1}^{m} n_i(x_i) \tag{4}$$

$$p^2(\underline{x}) = p^1(\underline{x}) * p(Z_1(\underline{x}))/\sum_{W_1(\underline{x})} p^1(\underline{x})$$

$$\vdots \tag{5}$$

.
.
.

$$p^{k+1}(\underline{x}) = p^k(\underline{x}) * p(Z_k(\underline{x}))/\sum_{W_k(\underline{x})} p^k(\underline{x})$$

and iterate in (5) until convergence.

Some important properties of the IPF method are discussed by Fienberg (1977). Let us return initially to the three attribute case, where the problem is to estimate (x_1,x_2,x_3). There are five different kinds of model which may be applied according to the extent of the information which is available about the constrained marginal probabilities (Fienberg, 1977, Chapter 3).

(1) The model of independence assumes that each of the variables is independent of the other two:

$$p(x_1,x_2,x_3) \text{ s.t. } p(x_1),\ p(x_2),\ p(x_3)$$

(2) A model of joint independence assumes that one of the attributes is independent of the other two, which are both related:

$$p(x_1,x_2,x_3) \text{ s.t. } p(x_1),\ p(x_2,x_3)$$

(3) The conditional independence model assumes that two of the variables are independent of one another, but both are dependent on the third:

$$p(x_1,x_2,x_3) \text{ s.t. } p(x_1,x_3),\ p(x_2,x_3)$$

(4) With an absence of second order effects situation, all the attributes are pairwise correlated:

$$p(x_1,x_2,x_3) \text{ s.t. } p(x_1,x_2),\ p(x_1,x_3),\ p(x_2,x_3)$$

(5) Second order models allow for interrelationships between all three types of attribute:

$$p(x_1,x_2,x_3) \text{ s.t. } \bar{p}(x_1,x_2,x_3)$$

The second order models are clearly something of a special case and do not concern us here, as we assume that the data we do have takes the form of actual cell entries (so any second order information would comprise a complete representation of the problem in question). The IPF routine then has the appealing feature of providing a solution to any of the other forms of model, although strictly speaking iteration should only be required for Model 4 in the three attribute case.

As a notational device, Fienberg adopts the convention of writing a constraint set simply as a list of attributes within square brackets, so $(p(x_1,x_2,x_3)$ becomes simply [123]. Under this convention, a full list of the nine possible models for the three attribute case is given in Table 1. One feature of these models which we will assume to be consistently true is that they are <u>hierarchical</u>, which means that any higher order term is consistent with all lower order terms, so we cannot include the set [123] without the six lower order terms [12], [13], [23], [1], [2], and [3]. The inclusion of any non-consistent lower order term is possible within a non-hierarchical model structure, but such models may not be handled by the IPF technique.

Extensions to problems of a higher dimensionality are relatively straightforward, although of course the variety of possible models tends to escalate rapidly. Fienberg explains that for higher order problems 'the method of iterative proportional fitting can always be used to compute the maximum likelihood estimates' (1977, p.61). The one problem that can arise is that of model selection. In general, an increase in the number of model "parameters", or the amount of information included via the constraints, will improve the fit of the model. There may be cases in which it is beneficial to try and reduce this complexity, and to apply simpler models with a similar performance, but this approach will not usually be advantageous for the fitting procedures considered here. Thus, as a general rule, it will be useful to include all known information as constraints, subject to the hierarchical data principle (i.e. no duplication of higher order information).

We may summarise the outcome of the IPF procedure as follows, therefore:

(1) All known information is retained, and may be generated anew via reaggregation;

(2) Although no new information is actually generated, maximum likelihood estimates are provided to missing cell probabilities;

(3) Any model incorporating partial information may be treated in this way. In practice, the maximum possible information should be included through the constraints.

2.3 The generation of a micro-database for Leeds MD

In this section, we attempt to demonstrate how some of the methods and concepts introduced previously may be used to synthetically generate an initial population for Leeds Metropolitan District. At this stage, there remain attributes that could be added. In particular, we are still awaiting much of the information about population activities, that will be made available to universities through MATPAC, hopefully later in 1987. We are, however, able to illustrate the methodology, and to provide a micro-level specification of an initial population to demonstrate some key features of SYNTHESIS in section 3.

To begin with, an obvious point to make about micro-data is that it can be based on two different kinds of unit - the individual or the household. In this example, we create a synthetic sample of 50,000 heads of household, but further individuals are added at a later stage. Altogether, we ascribe five characteristics to households, and another seven to individuals, as shown in Table 2. The method by which this sample is created is illustrated as a flow diagram in Figure 1, and we now discuss the steps involved briefly in turn.

A fundamental characteristic of the approach is our focus on spatial disaggregation. The first step in the modelling exercise is to assign each of the 50,000 sample households to a location - these locations being the 1565 enumeration districts (EDs) within

Leeds MD at the 1981 census. In practice, it is useful to combine this process with the ascription of the sex, age and marital status (MS) of household heads, since these features are all cross-classified within Table 26 of the 1981 census Small Area Statistics (SAS). In this case we need to generate a very large cumulative frequency distribution comprising 1565 zones x 4 age groups x 3 marital statuses x 2 sexes, and sample from it using the Monte Carlo method.

As we observed in section 2.3 above, this procedure already involves a crucial though not fully explicit modelling decision. It would have been equally possible to start off by assigning some other household attribute, such as tenure, or an individual attribute like socio-economic status. In a formal sense, we assume there is some sort of continuum of attribute dependence: some characteristics are very important in 'determining' others (at least in a statistical sense - the association need not be causative), and thus need to be assigned at an early stage; while others are more dependent, and need to be introduced after the ascription of those characteristics on which they are supposed to depend. For practical purposes, such distinctions will be largely intuitive, so what we are assuming here is that location, sex, age and MS are the most fundamental characteristics of household structure. Ultimately, of course, the proof of this pudding is in the eating, and we attempt to demonstrate the consistency of the simulated distributions (in aggregate) with the parent database as we go along.

It is also necessary to direct our attention to the nature of the sampling problem involved in this exercise. Table 3 presents a comparison of the simulated age-sex-marital status distribution, aggregated across the whole city, with counts from the published census County Reports. The variation between the two distributions is, of course, induced by sampling error, and Table 3 gives us an essentially qualitative feel for the magnitude of this error.

Since the distributions of Table 3 are strict samples from the parent population, and no <u>modelling</u> has been introduced into the procedure so far, we can perform statistical analysis on Table 3 to verify the efficiency of the (Monte Carlo) sampling procedure itself.

In Columns 2 and 3 of Table 4 we present upper and lower 95% confidence bounds, assuming independence between each of the population groups. Column 4 shows that all the estimates of Table 3 fall within the appropriate bounds, bar one (cell 8), which is the pattern one might expect of an unbiased random sample. Even this cell is well within the appropriate 99% confidence limit (687).

A similar method of broaching this problem is to compute a goodness-of-fit statistic directly. In this case, under random sampling, the Z-statistic of Column 5 would fall within ± 1.96 (Standard Deviations) 95% of the time, and ± 2.58 99% of the time. Alternative samples with similar properties may be obtained by varying the seed values, ie by changing the random number sequences used in the simulation. Two examples are presented in Columns 6-9 of Table 4.[1]

These results provide evidence for the adequacy of the sampling procedure. In subsequent analyses we will therefore be able to concentrate on hypotheses concerning the modelling assumptions which relate the sample data to the parent population.

The age group definitions to this point are somewhat coarse (there are four groups: 16-29; 30-44; 45-59/64; and 60/65+) and the next stage is to disaggregate them. First of all, we break age down into five year groups, which can be done useing SAS Table 21, and thereby utilising a conditional dependence of age on sex, MS and location. To decompose the five year age probabilities into single year ones, however, we need to use national population estimates (OPCS, 1983a), which are based on sex only.

At step 3 we wish to determine the country of birth (COB) of the household head, and naturally we wish to make this characteristic dependent on the attributes which we already know, that is location, sex, MS and age. Unfortunately, a full, five-way disaggregation is not available here, so we need to partition the problem. Initially,

---

Note:

1. The random number sequences are generated through the NAG routine GØ5CAF, which may be initialised by a call to GØ5CBF (IPAR), where IPAR is the seed, or generator, of the random number sequence (Numerical Algorithms Group 1982). In Table 4, the generators used were 66000 (Column 4), 20000 (Column 6) and 66 (Column 8).

we neglect the locational aspect and focus on producing a joint distribution is not known, but we do have information from national census tables (OPCS, 1983b) on COB by sex by MS, and COB by sex by age (in five year groups). The full array may then be estimated using IPF as shown in Figure 3. Observe that rather biased estimates to the age by marital status distribution are produced at this point (for example, there is rather a large proportion of married and widowed/ divorced persons in the 16-19 age group!). This is because no age-MS constraints are imposed at this stage. Rather we use the results of Figure 3 to construct the probabilities of COB given age, sex and marital status. At the next stage in the exercise, we then weight the probabilities from step 3.1 according to the known ethnic composition of each ED from SAS Table 4, and determine ethnicity on the basis of known age, MS, sex and location.

The simulated distribution of household heads by country of birth is compared with published totals for the city of Leeds in Table 5. Notice here that our use of the iterative proportional fitting technique does not help us to improve the aggregate fit. Rather it is an attempt to correlate attributes in the most efficient way possible, eg. to obtain the correct rates of household formation by country of birth groups within different age bands. In the case of country of birth one also has to bear in mind the considerable difficulties associated with the original census question on this issue. Basically this involves a conflict between the notion of the perceived 'ethnic status' of the individuals and their actual birthplace (see Rees and Birkin, 1983, for a fuller discussion). In subsequent analysis here, we assume the ethnic status of the individual to be defined by the country of birth of the household head.

The procedure for generating a spouse for the household head is relatively straightforward. We simplify the process by assuming that we only need to create partners for married heads, thus side-stepping the problem of the "de facto spouse". In this case the sex and marital status of the spouse follow trivially, and for convenience we also assume that the married couple have a shared ethnic origin as noted above. The age of spouse given the age of head is estimated using another national data set, this time the 'Household and Family Composition' summaries (OPCS, 1983c). These tables give age of wife by age of husband in five year groups (Table 24) and these can be interpolated linearly to single year groups.

Once a spouse has been created (if appropriate) we can go on to generate children using a combination of national and Small Area tables. Basically, we can use national tables here to generate a distribution of family sizes according to the age of mother. Applying these probabilities to the actual wives in a particular ED allows us to produce an expected schedule of family sizes, which can be compared with the actual distribution of family sizes observed for that ED in SAS Table 18. The expected family size distributions may then be weighted appropriately before sampling takes place. Finally, once the number of children is determined, their ages may be added on the basis of family size and mother's age, again using national tables. The whole process may then be repeated for unmarried household heads as potential one parent families.

If all households were composed of simple, nuclear family units then the basic procedures for generating individuals would now be complete. Unfortunately, this is not the case, and we also need to account for the presence within households of individuals who are dependent on the head of household, but neither a spouse nor a child (eg. an elderly relative); and also the possibility of non-dependent individuals (eg. a lodger). Once again, a two-tiered strategy is adopted. From Table 18 of the Small Area Statistics one can obtain probabilities that individual households are simple (nuclear) or complex. By compounding the probabilities which have been generated at earlier steps in the exercise, we are able to infer an age-sex schedule of 'missing persons' not yet accounted for. This procedure is far from straightforward, and we therefore adopted the simplification of creating this missing persons list for the whole city. Such persons can be allocated to the complex households which were just determined.

We are now in a position to undertake a comparison between the age-sex composition of the simulated and actual populations. The first stage is a city-wide comparison, which is presented in Table 6. Next, in Table 7, we present a spatial comparison of population totals. Both of these distributions are, we would argue, most satisfactory. Thirdly, we have picked out a distribution which is relatively fine in scale both spatially and sectorally - five year age groups by ward. Information relating to the first four wards in Leeds MD is presented in Table 8.

Two kinds of reason can be identified for expecting the synthetic ward level distributions to be a little less exact than the city-wide ones. In the first place, because the age-sex characteristics were not derived directly for individuals but added by a statistical modelling procedure, it is likely that initial (sampling) errors may be compounded, and that these errors are more likely to become significant at finer levels of resolution. Secondly, as we observed above, there is a direct modelling assumption concerning non-nuclear families which was applied at an aggregate spatial scale. Some distortions may be induced if we then consider the spatial implications of such a model, as we are now doing. Since one group which might be prominent in complex households is the elderly, we can infer that this effect may account for some of the variation within these age groups which is exhibited in Table 8.

It is difficult to suggest a satisfactory goodness-of-fit test for the data of Table 8, but once again we have adopted a Z-statistic (as in Table 4), basically to standardise for variations in the sample size. Given that only 7 of the 72 statistics fall beyond the 95% confidence level (and of these, 3 are beyond the 99% level) it would be hazardous to assert that the distributions of Table 8 are significantly biased, even as a random sample from the parent distribution. Since they are in fact the products of a modelling procedure, we have every reason to be satisfied with the effectiveness of that procedure.

For summary purposes, the composition of households determined at the last step may be broken up into five categories, as in Table 3. Taken with location, these household categories provide a means for estimating tenure from SAS Table 29. Although it is possible to adopt a finer classification, we focus on only three tenure classes here: owner-occupation, council rented, and others (primarily private rented).

Our next concern is to begin to build in some measures of the socio-economic activity patterns of individuals. First we concentrate on determining employment patterns, depending on location, age, sex, and marital status of the individual. It is possible, from SAS Table 9, to generate appropriate probabilities that individuals may be economically active or inactive, and if active whether they are in employment or seeking work. One procedural modification which we introduce is that individuals over a given age (60 for women, 65 for men) will fall into the category of inactive and retired. Of course, this is not exactly true, but neither is it a serious oversimplication. What it illustrates in a primitive way is the possibility within a microsimulation framework of adopting rule-based approaches to the determination of individual attributes.[2]

For individuals in employment, we would now like to determine the industry in which they work, and their socio-economic status. Once again, this data is available within the SAS, but not in quite such a compact form as in the economic activity case. We are given the following data: age by sex, by location, by industry type (SAS Table 46); socio-economic group by sex, by location (SAS Table 50); and SEG by industry-type, by location (SAS Table 44). The iterative proportional fitting algorithm can now be used to combine this information in producing the full joint probability distribution of age-sex-location-industry-SEG. We may therefore sample for industry and SEG on the basis of known age, sex and location. Some summary data at the outcome of this process is provided in Table 9.

The assignment of SEG and industry labels now provides a firm basis for the extension of economic activity patterns to embrace specific jobs for individual workers. However the modelling procedures involved in this exercise are rather more complex than those so far considered here, and we therefore reserve this step for another paper (Birkin and G.Clarke, 1987). Nevertheless, we do feel that it is important to emphasise the existence of a "supply-side" within the economy, and the (spatial) interactions between the supply-side and the residential population.

Note:

2. Hence the mainstream popularity of microsimulation models in the realms of taxation and finance, where well-defined rules determine the levels or benefits accruing to individuals.

Data is available from PINPOINT's LUPIN system at a slightly more aggregate spatial scale, with known interaction flows for non-food goods by origin postal districts by 52 named destination centres within the West Yorkshire conurbation. By assigning each ED to a 'parent' postal district, we are able to allocate each household to a primary retail destination.

3. Model outputs

3.1 The simulated distribution

We are now in a position to summarise the product which we obtain at the end of the modelling procedure described in section 2.3. There are three basic attribute sets to be discussed, relating to spatial features, to household attributes, and to individual characteristics. As we have seen previously, the smallest area at which data is typically available is the enumeration district (ED). Each ED is hierarchically related to a single ward, but they may also be allocated to postcode sectors, as we observed at the end of the previous section. Hence we have a basis for flexible spatial aggregation, which will be exploited below.

For each ED there are a given number of households in the zone. The characteristics of the individual household are then the number of people it contains, its tenure, the country of birth of the household head, and the primary shopping location of the householder for non-food goods. Individuals may then be identified according to their status within the household, eg. as a head, dependent, or non-dependent. Each individual has an exact age, a sex, marital status and economic activity status. In addition, for those who are economically active and in work we have an industry and SEG label. This information was summarised earlier in Table 2.

By way of example, Figure 2 shows the first 37 records of the simulated file, providing data for the first ten households in ED AA01. Reading from the top of the file we see that this ED is the first in the file, and falls within postcode area LS20. There are 24 simulated households in this ED.

The first household, which is owner-occupied, has 5 residents, whose nationality is British. The primary non-food shopping destination for this household is zone 41 (Leeds City Centre). The head of household is a 40 year old, married male, who is employed as a "manager" in the manufacturing sector. His wife is 36 years old, and works in a "junior, non-manual" (eg. secretarial) capacity in the distribution sector. They have three children, two girls aged 15 and 12, and a boy aged 6. It is possible to infer the characteristics of subsequent households in a similar way, eg. household 2 is a widowed (or divorced) retired lady, aged 74 years.

3.2 Data requirements and computational costs

The list of Figure 2 was generated a suite of eight FORTRAN programs representing partitioned phases of the procedured outlined in Figure 1. The amount of CPU time required in the process is in part a function of the sample size, as the Monte Carlo procedure is a relatively demanding one. However many of the basic probability distributions have to be created independently of the sample size, so this ensures that there will be 'economies of scale', ie. larger samples become relatively more cheap to produce than smaller ones.

For the purposes of experimentation, the whole simulation process was repeated eight times for four sample sizes (1000, 5000, 10000, and 50000) and a further four 5000 household samples with varying seed values. These runs are summarised in Table 10, and we observe that the costs are relatively modest. The 1000 household case takes under 20 seconds CPU to generate on the Leeds University Amdahl V7 mainframe, and even the 50000 case only of the order of 120 seconds. Note that these are essentially one-off costs given that, once generated the microdata files can be retained for future use.

The data requirements of the procedure are quite extensive, mostly because of the fine level of spatial detail adopted. We will not itemise the individual data input files here, but altogether they take up over 6 megabytes of disk storage. In comparison, even the 50000 household and individual sample takes only around 1 megabyte to store. Assuming the validity of the 'information retention' argument (see above, section 2.2), this shows one aspect of the efficiency of the micro-simulation approach, although this is a less significant thread to the argument than the efficient accounting issue (see section 4.3; and Birkin and Clarke, 1986).

3.3 Extended model outputs

A basic SYNTHESIS argument about the list presented in Figure 2 is that aggregation may be performed in a manner that is spatially and sectorally flexible, to generate types of distribution which are both potentially interesting and not usually available. Suppose, for example, we are interested in unemployment patterns. Typically, the census will provide us with adequate information about individual unemployment, but less on the issue of household employment and activity patterns. It might be interesting, therefore, to try and identify the number of workers within particular households, and to match this against the size of the household. This distribution can be obtained through SYNTHESIS by isolating variables H1 and P6, and aggregating across the remainder. Figure 3 presents this distribution aggregated to the ward scale.

Although it is not necessary to comment on this distribution in detail, we can see that the largest number of zero-worker households contain only one or two individuals. The majority of these will, of course, contain persons of pensionable age. One might argue that a particularly interesting section of this distribution concerns larger households with no income from employment. Hence the next analytical step might be to extract households with three or more residents and no direct income. With the more restricted focus, it is possible to focus more explicitly on the interrelationships between wards. Thus in Figure 4 we have ranked wards according to the proportion of households with 3 or more residents, which have no income from employment. In passing, we can observe that this index provides a particularly good deprivation indicator, and those familiar with the social geography of Leeds will recognise Halton, Cookridge and Horsforth as prosperous zones; Richmond Hill, Seacroft and University ward are all depressed. For the sake of interest, the indicator is mapped as Figure 5.

The argument about spatial flexibility is amplified if we choose to aggregate from the ED level up to postcode districts rather than wards. The appropriate distributions are provided in Figure 6. Note that some of the postal sectors overlap the MD boundary (eg. LS24, LS29, WF2, WF3, WF10, WF12) while others have low population totals because of their primarily industrial or commercial character (LS2, LS3). This flexibility between census-based and postcoded data is likely to become increasingly important as more information is made available by market researchers and other organisations (eg. health authorities) with a postal sector base.

Since the SYNTHESIS concept embraces spatial interaction, it is possible to generate characteristics of supply zones in the same way. A variety of possible indicators are used to generate profiles of the various shopping centres in Figure 7. Notice that under the rather crude definition of "primary non-food" trips, Leeds city centre takes a dominant position, while many peripheral or external centres (Batley, Featherstone, Pontefract, ...) are of more marginal interest. Once again, the usefulness of this analysis becomes more clear if the concept of targetting is introduced. To this end, the centres have been ranked by size, and by the proportion of customers in social groups 1 and 2 in Figure 8.

Of course, it would be possible to extend both these examples by the introduction of a more explicit treatment of individual and household incomes; and by the extension of economic activity attributes, such as the addition of places of work. We return to the possibilities for an extended attribute list in section 5.

4. Uses of SYNTHESIS

4.1 A flexible information system

As we have already outlined a synthetic micro-data base such as SYNTHESIS can serve a useful role as an information system. Because of the flexible approach to aggregation, both spatially and sectorally, inherent in micro-data sets, many of the drawbacks of conventional information systems can be overcome. Additionally, the storage requirements, even for large attribute sets are relatively modest vis a vis aggregate counterparts.

4.2 Updating information systems

A question often asked is how reliable is 1981 census data in 1986? In some cases the answer to this is unquestionably: 'not very good'. Methods for updating demographic profiles of areas are well established (Rees, 1986) but if the attention is focussed on households and a wider set of attributes other than age and sex then few attempts have been made at updating. Micro-simulation models offer considerable potential in the modelling of household dynamics (Clarke, 1986, Rees, Clarke and Duley, 1987). The main principle is to take a micro-data set and update individual and household attributes using LIST PROCESSING. This involves deriving conditional probabilities for such events as birth, death, migration, marriage, divorce, and so on and to invoke Monte Carlo sampling methods to determine whether eligible individuals undergo appropriate transitions. The advantages of micro-simulation in household dynamics relate to the handling of interdepdencies - between individuals' attributes and between individual members of households. A full description of the methodology and an application to Yorkshire and Humberside can be found in Clarke (1986), whilst an outline of a corresponding approach at the small area level is presented in Rees et al (1987).

4.3 Input to comprehensive models and the development of accounting systems

We have argued the case for the development of integrated macro-micro-models on a number of occasions (Birkin and Clarke, 1986, Clarke and Wilson, 1986). Central to this argument is the need for a micro-level approach to modelling the demand side - for example the demand for public and private services, such as health, housing, education, retailing, transportation and so on. This is coupled with using the micro-data base as the accounting framework. Because the characteristics of each sample individual and household are treated explicitly, the appropriate conservation laws must be obeyed, and aggregate level transitions can always be traced back to their original components. Additionally the detailed distributional effects of policy may be assessed - taking advantage of the fact that there is no prior aggregation. This has proven particularly attractive in public policy analysis in the US and West Germany (Orcutt, Merz and Quinke, 1986). Finally, of course the integrated analysis can be coupled with dynamics through the incorporation of the household dynamics model described above.

5. Conclusions and further research objectives

We hope by now to have given the reader an indication of the main features of a synthetic micro-data base such as SYNTHESIS. Clearly there is scope for further refinement. Among the additional attributes we envisage generating for the West Yorkshire study are:

(i) Income. It has already been demonstrated (Clarke, 1984) that it is possible to construct an income generation module based on the industrial sector and occupation of the individual using New Earnings Survey data. This also accounts for wage dispersion about the mean wage for any age, occupation, industry combination, effectively giving individuals a 'wage trajectory' that can be updated over time.

(ii) Benefits and taxation. Because we have detailed information of household structure, occupation, and from (i) income, it is possible to model the flows of state benefits and taxation for each household in the sample. Ideally this needs to be coupled with information on tenure and housing finance and there is an obvious limit to the level of detail that can be applied. In the US a major application of micro-simulation methods using micro-data bases has been to assess alternative transfer income policies (see Orcutt, et al, 1976).

(iii) Housing. Using census data it is relatively straightforward to add tenure and dwelling unit size to the household attribute list. Financial information concerning house price, mortgage, rent, and so on is more difficult to obtain although we have considerable experience of using the Nationwide Building Society survey data on house purchase and the Housing Conditions Survey and aim to incorporate these into RUIN.

(iv) Journey to work. A set of associated attributes will be added as soon as the 1981 journey to work data set is released.

Further, as the micro-data is processed through a number of different models, additional attributes that pertain to the application are generated. For example in our district health care model (Clarke and Spowage, 1985) the following health related attributes were generated: morbidity condition, specialty and hospital of treatment, source of admission, type of discharge, operative procedures, length of stay and cost of treatment. The storage of these attributes is efficient and straightforward.

A number of research tasks remain outstanding. Further attention will be given to the validation of model output where possible. Analysis of sample design and related statistical characteristics of the sample will continue. It remains our objective to see SYNTHESIS as the first step in a broader programme of model design, implementation and dissemination.

Acknowledgements

The authors acknowledge the support of the ESRC through grant no. D00220001. They would also like to thank John Beaumont, formerly of Pinpoint Ltd., for supplying LUPIN data for research use.

References

Birkin, M and Clarke, G (1987) Synthetic data generation and the evaluation of urban performance : a labour market example. Paper prepared for the fifth European Colloquium on Quantitative Theoretical Geography, Bardoneccia (copies available from the first author).

Birkin, M and Clarke, M (1986) Comprehensive dynamic models: integrating macro- and micro-approaches in Griffith, D A and Haining, R P (eds.) Transformations through space and time. An analysis of non-linear structures, bifurcation points and autoregressive dependencies, Martinus Nijhoff, Dordrecht.

Blalock, H M (1972) Social statistics. Second edition. McGraw-Hill, Tokyo.

Clarke, M (1984) Integrating dynamic models of urban structure and activities: an application to urban retail systems, unpublished PhD thesis, School of Geography, University of Leeds.

Clarke, M (1986) Demographic forecasting and household dynamics: a micro-simulation approach in Woods, R and Rees, P H (eds.) Population structures and models, Allen and Unwin, London.

Clarke, M and Holm, E (1986) Micro-simulation methods in human geography and planning: a review and further extensions, Geografiska Annaler, forthcoming.

Clarke, M and Spowage, M (1985) Integrated models for public policy analysis: an example of the practical use of simulation models in health care planning, Papers and Proceedings of the Regional Science Association, 55, pp. 25-48.

Clarke, M and Wilson, A G (1986) A framework for dynamic comprehensive urban models. Accounting and micro-simulation approaches, Sistemi Urbani, forthcoming.

Fienberg, S E (1970) An iterative procedure for estimation in contingency tables, Annals of Mathematical Statistics, 41, pp 907-917.

Fienberg, S E (1971) The analysis of cross-classified categorical data, MIT Press, Cambridge, Mass.

Numerical Algorithms Group (1983) FORTRAN Library, Manual, Number 11, NAG, Oxford.

Orcutt, G H, Caldwell, S and Wertheimer, R (1976) Policy exploration through micro-simulation, The Urban Institute, Washington DC.

Orcutt, G H, Merz, J and Quinke, M (eds.) (1986) Microanalytic simulation models to support social and financial policy, North Holland, Amsterdam.

OPCS (1983a) Population estimates 1981, HMSO, London.

OPCS (1983b) Country of birth, Census 1981, HMSO, London.

OPCS (1983c) Household and family composition tables, Census 1981, HMSO, London.

Paas, G (1986) Statistical Match: evaluation of existing procedures and improvements by using additional information, in G H Orcutt, J Merz, H Quinke (eds.) Microanalytic simulation models to support social and financial policy, North Holland, Amsterdam.

Radner, D, Allen, R, Gonzalez, M E, Jabine, T B and Muller, H J (1980) Report on exact and statistical matching techniques. Statistical Working Paper No.5, U.S. Dept. of Commerce, Washington, DC, U.S. Government Printing Office.

Rees, P H (1986) Choices in the construction of regional population projections, Chapter 7 in Woods, R and Rees, P H (eds.) op cit.

Rees, P H and Birkin, M (1984) Census-based information systems for ethnic groups: a study of Leeds and Bradford, Environment and Planning, A, 16, 1551-1571.

Table 1 <u>The full range of three dimensional models</u>

Model type	Constraints
1. Model of Independence	[1], [2], [3]
2. Model of Joint Independence	[1], [23]
(3 models)	[2], [13]
	[3], [12]
3. Conditional Independence Models	[12], [13]
(3 models)	[12], [23]
	[13], [23]
4. Absence of Second Order Effects	[12], [13], [23]
5. Second Order Models	[123]

Table 2a <u>Household attributes</u>

1.	LOCATION	1565	1. DAAA01 2. DAAA02 . . 1565. DABK47
2.	HOUSEHOLD STRUCTURE AND COMPOSITION	5	1. Single person, retired 2. Single person, not retired 3. Married couple, no children 4. Lone parent family 5. Married couple with children
3.	TENURE		1. Owner-occupied 2. Council rented 3. Other
4.	COUNTRY OF BIRTH OF HOUSEHOLD HEAD	7	1. Great Britain 2. Eire 3. New Commonwealth - India 4. New Commonwealth - Caribbean 5. Rest of New Commonwealth 6. Pakistan 7. Rest of the World
5.	PRIMARY RETAIL LOCATION	52	1. Hemsworth 2. Normanton . . . 52. Bradford

Table 2b <u>Individual attributes</u>

1.	STATUS WITHIN HOUSEHOLD	5	1. Head 2. Spouse of head 3. Child of head 4. Other, dependent 5. Other, not dependent
2.	EXACT AGE	86	0. 1. . . 85+
3.	SEX	2	1. Male 2. Female
4.	MARITAL STATUS	3	1. Married 2. Single 3. Widowed/Divorced
5.	ECONOMIC ACTIVITY	4	0. Inactive 1. In work 2. Retired 3. Seeking work
6.	SOCIO-ECONOMIC GROUP	7	1. Employers and managers 2. Professional 3. Intermediate/Junior non-manual 4. Skilled manual 5. Semi-skilled manual 6. Unskilled manual 7. Other/not stated
7.	INDUSTRY	7	1. Agriculture 2. Energy and water 3. Manufacturing 4. Construction 5. Distribution and catering 6. Transport 7. Other services

Table 3. Household heads: age by marital status by sex.

3A. Observed distribution

	MALES			FEMALES		
	M	S	WD	M	S	WD
16-29	19313 (3674)	5999 (1141)	619 (118)	1873 (356)	4376 (833)	1351 (257)
30-44	51337 (9767)	3274 (623)	2719 (517)	2796 (532)	1891 (360)	5031 (957)
45-59/64	63209 (12026)	3817 (726)	4660 (887)	1864 (355)	1969 (375)	7742 (1473)
59/64+	28502 (5423)	1988 (378)	6847 (1303)	1692 (322)	6499 (1236)	33438 (6362)

The upper figure is the known total (source: OPCS, 1982, Table 35)

The lower figure is the total per 50000 households.

3B. Simulated distribution

	MALES			FEMALES		
	M	S	WD	M	S	WD
16-29	3724 +1.36	1141 0.	124 +5.08	352 -1.12	855 +2.64	262 +1.95
30-44	9721 -0.47	682 +9.47	540 +4.45	526 -1.13	378 +0.80	965 +0.84
45-59/64	12023 -0.02	739 +1.79	868 -2.14	359 +1.13	371 -1.07	1461 -0.81
59/64+	5327 -1.77	384 +1.59	1289 -1.07	298 -7.45	1285 +3.96	6326 -0.57

The upper figure is the simulated cell count.

The lower figure is the percentage error, calculated as:

$$100 \times \frac{\text{Predicted value}}{\text{Actual value}} - 100.$$

Table 4. Statistical variation with alternative sample seeds.

Cell	1	2	3	4	5	6	7	8	9
1	3674.0	3559.6	3788.4	3724.0	0.8517	3670.0	-0.0686	3607.0	-1.1581
2	1141.0	1075.6	1206.4	1141.0	0.0000	1167.0	0.7701	1160.0	0.5644
3	118.0	96.7	139.3	124.0	0.5395	113.0	-0.4709	109.0	-0.8630
4	356.0	319.2	392.8	352.0	-0.2140	365.0	0.4728	352.0	-0.2140
5	833.0	776.9	889.1	855.0	0.7589	815.0	-0.6357	812.0	-0.7430
6	257.0	225.7	288.3	262.0	0.3097	291.0	1.9989*	256.0	-0.0627
7	9767.0	9593.2	9940.8	9721.0	-0.5198	9710.0	-0.6444	9746.0	-0.2371
8	623.0	574.4	671.6	682.0	2.2748*	623.0	0.0000	607.0	-0.6534
9	517.0	472.7	561.3	540.0	0.9952	516.0	-0.0443	500.0	-0.7641
10	532.0	487.0	577.0	526.0	-0.2630	561.0	1.2313	558.0	1.1069
11	360.0	322.9	397.1	378.0	0.9293	376.0	0.8283	374.0	0.7266
12	957.0	896.9	1017.1	965.0	0.2601	946.0	-0.3611	971.0	0.4537
13	12026.0	11838.7	12213.3	12023.0	-0.0314	11880.0	-1.5341	12076.0	0.5224
14	726.0	673.6	778.4	739.0	0.4818	759.0	1.2070	771.0	1.6333
15	887.0	829.1	944.9	868.0	-0.6506	909.0	0.7364	889.0	0.0677
16	355.0	318.2	391.8	359.0	0.2119	330.0	-1.3808	320.0	-1.9629*
17	375.0	337.2	412.8	371.0	-0.2084	366.0	-0.4722	391.0	0.8123
18	1473.0	1398.9	1547.1	1461.0	-0.3186	1413.0	-1.6192	1466.0	-0.1856
19	5423.0	5286.7	5559.3	5327.0	-1.3915	5513.0	1.2850	5516.0	1.3276
20	378.0	340.0	416.0	384.0	0.3074	385.0	0.3581	363.0	-0.7902
21	1303.0	1233.2	1372.8	1289.0	-0.3951	1294.0	-0.2535	1234.0	-1.9889*
22	322.0	286.9	357.1	298.0	-1.3944	315.0	-0.3957	316.0	-0.3386
23	1236.0	1167.9	1304.1	1285.0	1.3848	1240.0	0.1150	1240.0	0.1150
24	6362.0	6215.9	6508.0	6326.0	-0.4843	6443.0	1.0812	6366.0	0.0537

Column 1: Observed cell counts (from Table 3A).

Columns 2-3: Upper and lower confidence bounds to Column 1 data; calculated as

$$\hat{p} \pm 1.96 \sqrt{\hat{p}\ (1-\hat{p})\ /\ N}$$

where $\hat{p}$ is the appropriate Column 1 probability, N is the sample size (50000) (Blalock, 1972, p211).

Column 4: Estimated cell counts (from Table 3B).

Column 5: Z-values associated with Column 4, computed as:

$$Z = (p' - \hat{p})\ /\ \sqrt{\hat{p}\ (1 - \hat{p})\ /\ N}$$

where p' is the appropriate Column 4 probability (Blalock, 1972, p195).

Columns 6-9: Cell counts and Z-scores with alternative seed values (for discussion, see text).

* = Z-score beyond 95% (but within 99%) confidence bounds.

Table 5. Household heads: country of birth

Country of birth	Expected heads	Observed heads	Simulated heads	Z-score
UK	246764	46948	46986	+0.71
Ireland	3942	750	766	+0.59
New Commonwealth				
India	1949	371	348	-1.20
Caribbean	2308	439	420	-0.91
Other	1669	318	287	-1.74
Pakistan	975	185	188	+0.22
Rest of the World	5199	989	1005	+0.51
Total	262806	50000	50000	

Observed heads derived from OPCS (1982), Table 11.

Table 6 Total population by five year age groups

		MALE			FEMALE		
		Actual	Pred	Error (%)*	Actual	Pred	Error (%)*
1.	0- 4	20356	20414	0.3	19662	19406	1.3
2.	5- 9	23593	23681	0.4	22283	22042	1.1
3.	10-14	28488	28324	0.6	27477	27315	0.6
4.	15-19	29443	29468	0.1	28484	28391	0.3
5.	20-24	26243	26043	0.8	25582	25523	0.2
6.	25-29	23078	22703	1.7	22425	22246	0.8
7.	30-34	25564	25491	0.3	25383	25660	1.1
8.	35-39	20618	20350	1.3	20553	20424	0.6
9.	40-44	19522	19531	0.0	20041	19641	2.0
10.	45-49	19426	19609	0.9	19457	19720	1.4
11.	50-54	19617	19830	1.1	20501	20319	0.9
12.	55-59	20532	20686	0.8	21004	21133	0.6
13.	60-64	17133	16895	1.4	19258	19258	0.0
14.	65-69	15650	15686	0.2	19381	19294	0.5
15.	70-74	11987	11905	0.7	17623	17693	0.4
16.	75-79	7351	7341	0.1	13346	13049	2.3
17.	80-84	3348	3307	1.2	8101	7950	1.9
18.	85+	1368	1338	2.2	4486	4468	0.4

* Error : ((Predicted value/Actual value) - 1) x 100

Table 7. Population distribution by ward

				POPULATION	
		WARD	TOTAL	EXPECTED	SIMULATED
1.	AA	Aireborough	24041	4574	4654
2.	AB	Armley	22317	4246	4199
3.	AC	Barwick	21755	4139	4051
4.	AD	Beeston	17508	3331	3346
5.	AE	Bramley	21540	4098	4093
6.	AF	Burmantofts	21750	4138	4065
7.	AG	Chapel All	23180	4410	4332
8.	AH	City and Holb	20457	3892	3959
9.	AJ	Cookridge	20819	3961	3994
10.	AK	Garforth	23905	4548	4520
11.	AL	Halton	19526	3715	3601
12.	AM	Harehills	22712	4321	4252
13.	AN	Headingley	15458	2941	3145
14.	AP	Horsforth	21461	4083	4120
15.	AQ	Hunslet	15931	3031	3183
16.	AR	Kirkstall	19526	3715	3744
17.	AS	Middleton	19710	3750	3736
18.	AT	Moortown	19258	3664	3622
19.	AU	Morley N.	21413	4074	4063
20.	AW	Morley S.	22580	4296	4337
21.	AX	North	20110	3826	3711
22.	AY	Otley	23027	4321	4308
23.	AZ	Pudsey N.	22890	4355	4333
24.	BA	Pudsey S.	22155	4215	4275
25.	BB	Richmond Hill	22659	4301	4205
26.	BC	Rothwell	21093	4013	3915
27.	BD	Roundhay	20099	3824	3736
28.	BE	Seacroft	20919	3980	3965
29.	BF	University	17298	3291	3470
30.	BG	Weetwood	18118	3427	3406
31.	BH	Wetherby	22980	4372	4252
32.	BJ	Whinmoor	19174	3648	3678
33.	BK	Wortley	23190	4412	4403
		TOTAL	688561	131002	130673

Populations extracted from Table 21 of the Small Area Statistics.

Table 8. A comparison of synthetic with actual age distributions by ward

AGE	1. AIREBOROUGH			2. ARMLEY			3. BARWICK			4. BEESTON		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
0-4	283	273	0.596	243	262	-1.222	232	251	-1.250	217	212	0.340
5-9	330	314	0.884	269	272	-0.183	287	305	-1.066	195	194	0.072
10-14	413	385	1.384	290	338	-2.827**	379	359	1.031	238	227	0.715
15-19	363	350	0.685	391	379	0.609	281	304	-1.376	269	255	0.856
20-24	288	278	0.591	393	385	0.405	292	278	0.822	267	267	0.000
25-29	293	317	-1.406	319	296	1.292	250	277	-1.712	246	224	1.406
30-34	411	418	-0.347	290	293	-0.177	301	351	-2.891**	236	219	1.109
35-39	331	327	0.221	192	215	-1.663	396	325	3.582**	151	164	-1.060
40-44	279	276	0.180	209	226	-1.178	275	251	1.451	168	171	-0.232
45-49	272	254	1.094	248	247	0.064	258	231	1.685	187	185	0.147
50-54	261	244	1.055	250	249	0.063	236	242	-0.391	194	192	0.144
55-59	228	249	-1.394	248	232	1.019	245	237	0.512	194	198	-0.288
60-64	252	223	1.831	216	210	0.409	185	204	-1.399	194	195	-0.072
65-69	221	225	-0.270	216	217	-0.068	163	188	-1.961*	176	208	-2.416*
70-74	189	186	0.219	163	177	-1.098	168	153	1.159	180	181	-0.075
75-79	133	140	-0.608	125	129	-0.358	98	96	0.202	137	126	0.941
80-84	66	74	-0.985	95	73	2.259*	41	57	-2.500*	60	74	-1.808
85+	41	34	1.094	42	38	0.617	24	27	-0.613	37	32	0.822

(1) Synthetic count

(2) Actual value

(3) Z-score

* = value outside 95% confidence bound (Z=1.96)

** = value outside 99% confidence bound (Z=2.80)

Table 9. SEG by industry for persons in employment: 50000 sample for Leeds MD

	Employers +managers	Profes-sional	Junior non-man	Skilled manual	Semi-sk manual	Unsk manual	Other
Agriculture	77	1	13	67	155	1	0
	72	0	12	64	176	1	0
Energy	123	140	524	829	301	71	4
	112	142	512	843	335	68	5
Manufacturing	1512	410	2644	6031	3623	698	37
	1538	414	2713	6142	3622	742	37
Construction	470	79	386	2061	312	374	8
	491	83	401	2158	362	369	4
Distribution	2402	93	3808	2302	1906	743	24
	2367	91	3940	2323	1966	714	28
Transport	310	42	799	1392	542	245	6
	303	38	806	1413	564	260	11
Other services	1647	1321	9368	1137	2935	1195	125
	1557	1261	9213	1186	2934	1190	127

The upper figure represents actual distributions (SAS Table 44).

The lower figure represents simualted distributions.

Table 10 A suite of SYNTHESIS simulations

Run name	No. of households	No. of individuals	Seed value	CPU time
RU1000	1000	3129	0	19.09
RU5000	5000	15632	0	42.79
RU10000	10000	31344	0	64.82
RU50000	50000	156544	0	124.14
RU5001	5000	15672	1039	42.86
RU5002	5000	15771	2039	42.68
RU5003	5000	15723	3039	42.72
RU5004	5000	15724	4039	42.65

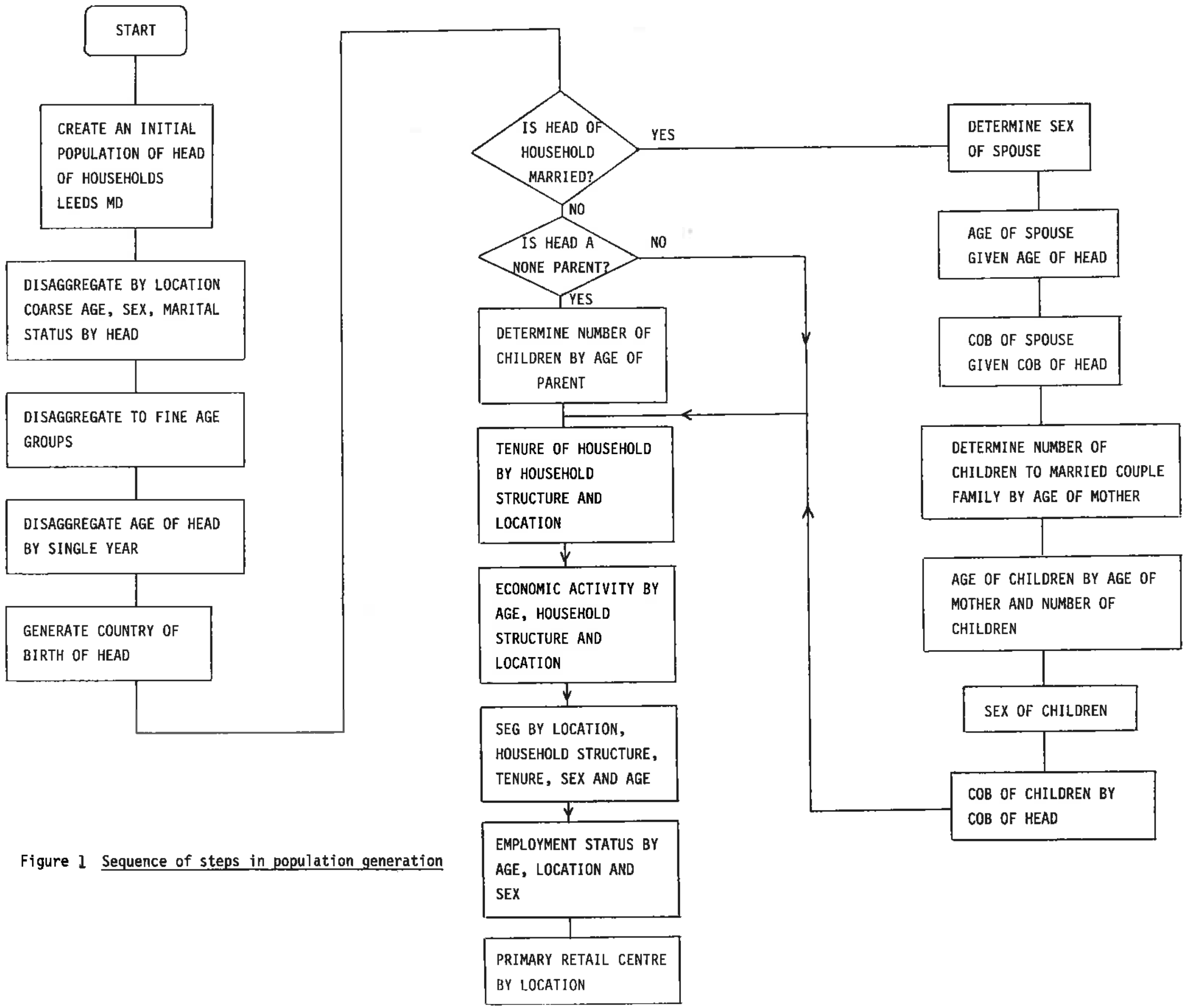

Figure 1 Sequence of steps in population generation

Figure 2. Sample output from the simulated file

```
1 AA01 LS20  24
                                  1  5  1  1  41
                                                    1   1  1  40  1  1  1  3  1
                                                    1   2  2  36  2  1  1  5  3
                                                    1   3  3  15  2  2  0  0  0
                                                    1   4  3  12  2  2  0  0  0
                                                    1   5  3  06  1  2  0  0  0
                                  2  1  1  1  41
                                                    2   1  1  74  2  3  2  0  0
                                  3  2  1  1  41
                                                    3   1  1  37  1  2  1  3  1
                                                    3   2  5  38  1  1  1  6  4
                                  4  1  1  1  25
                                                    4   1  1  66  2  1  2  0  0
                                  5  4  1  1  41
                                                    5   1  1  40  1  1  1  7  4
                                                    5   2  2  39  2  1  0  0  0
                                                    5   3  3  01  1  2  0  0  0
                                                    5   4  3  15  2  2  0  0  0
                                  6  2  1  1  41
                                                    6   1  1  64  1  1  1  3  6
                                                    6   2  2  62  2  1  2  0  0
                                  7  5  1  1  41
                                                    7   1  1  45  1  1  1  6  1
                                                    7   2  2  41  2  1  1  3  5
                                                    7   3  3  14  1  2  0  0  0
                                                    7   4  3  09  1  2  0  0  0
                                                    7   5  3  20  2  2  0  0  0
                                  8  4  3  1  41
                                                    8   1  1  55  1  1  1  3  1
                                                    8   2  2  53  2  1  1  7  3
                                                    8   3  3  12  2  2  0  0  0
                                                    8   4  3  16  1  2  0  0  0
                                  9  1  1  1  41
                                                    9   1  1  81  2  3  2  0  0
                                 10  1  2  1  41
                                                   10   1  1  25  1  2  1  7  4
```

Figure 3. Household size by income: Leeds wards

General layout

Number of incomes	Size of household: 1	2	3	4	5	6+
0						
1						
2+						

		1	2	3	4	5	6+
1	AA	276	224	41	8	2	4
		61	204	99	66	30	30
		0	186	130	106	60	39
2	AB	328	213	51	15	1	9
		128	184	78	42	31	26
		0	138	95	64	39	53
3	AC	213	189	41	5	1	1
		62	213	87	56	35	20
		0	178	111	87	42	35
4	AD	280	199	30	6	5	12
		91	117	87	44	32	15
		0	127	78	39	33	52
5	AE	305	214	40	17	11	12
		102	156	70	67	42	25
		0	107	91	56	36	49
6	AF	415	220	58	11	14	8
		95	149	73	46	42	24
		0	103	93	56	37	58
7	AG	341	185	53	18	13	17
		126	162	88	34	43	43
		0	113	101	49	29	92
8	AH	417	271	46	13	8	9
		123	164	64	42	25	27
		0	105	102	44	32	44
9	AJ	218	185	23	10	5	1
		84	189	73	60	30	28
		0	164	114	82	57	52
10	AK	176	217	29	11	5	2
		59	206	105	89	44	23
		0	163	136	114	58	58
11	AL	202	196	22	8	1	1
		56	177	70	51	17	18
		0	166	102	60	40	39
12	AM	344	198	61	16	11	16
		120	140	78	58	40	46
		0	87	81	48	48	73
13	AN	280	143	48	13	4	4
		130	104	67	44	15	19
		0	98	112	34	21	48
14	AP	234	196	24	8	1	2
		95	159	71	61	36	19
		0	171	120	100	60	62
15	AQ	253	166	42	13	9	6
		83	101	68	41	32	24
		0	70	75	36	18	45
16	AR	341	229	44	10	6	6
		118	153	71	51	17	29
		0	122	104	41	31	29
17	AS	242	170	38	12	7	7
		76	163	72	45	32	25
		0	100	88	52	45	44
18	AT	332	249	50	13	3	2
		64	167	62	39	30	17
		0	140	88	54	30	24
19	AU	228	192	38	10	3	3
		65	200	77	74	20	22
		0	168	113	97	38	37
20	AW	303	221	49	11	9	5
		89	160	89	60	35	20
		0	163	108	70	41	44
21	AX	250	219	29	11	2	0
		77	233	75	56	20	19
		0	145	107	65	42	33
22	AY	302	224	34	6	3	1
		94	200	70	55	39	28
		0	170	124	88	44	40
23	AZ	264	202	39	8	4	2
		87	188	99	64	24	12
		0	191	129	99	57	40
24	BA	279	192	38	10	5	5
		69	149	68	60	26	33
		0	190	97	85	52	61
25	BB	342	248	51	22	12	18
		114	159	76	49	44	32
		0	97	88	39	48	65
26	BC	228	173	32	11	5	1
		72	172	60	59	32	19
		0	173	108	83	46	42
27	BD	205	181	27	9	5	2
		115	160	84	47	38	16
		0	178	92	73	41	27
28	BE	295	234	45	18	13	11
		92	151	81	43	31	41
		0	80	89	39	40	49
29	BF	373	186	56	18	7	6
		134	129	75	43	22	26
		0	76	96	35	28	30
30	BG	284	215	43	6	7	2
		101	143	74	32	19	11
		0	113	87	51	39	29
31	BH	220	188	25	9	4	1
		60	210	82	64	44	28
		0	181	120	83	59	30
32	BJ	204	161	22	14	5	2
		76	162	57	41	45	31
		0	120	81	63	45	57
33	BK	296	241	36	16	8	5
		110	163	94	52	36	21
		0	174	124	77	53	32

Figure 4. Ratio of households, with greater than three members, lacking an income from employment.

Ward		Score	Rank	Ward		Score	Rank	Ward		Score	Rank
1	AA	0.084	25	2	AB	0.121	11	3	AC	0.080	28
4	AD	0.086	24	5	AE	0.131	8	6	AF	0.145	7
7	AG	0.163	6	8	AH	0.120	12	9	AJ	0.063	31
10	AK	0.078	29	11	AL	0.053	33	12	AM	0.170	5
13	AN	0.111	15	14	AP	0.058	32	15	AQ	0.117	13
16	AR	0.111	15	17	AS	0.110	18	18	AT	0.117	13
19	AU	0.094	21	20	AW	0.122	10	21	AX	0.071	30
22	AY	0.082	27	23	AZ	0.100	20	24	BA	0.111	15
25	BB	0.193	1	26	BC	0.092	22	27	BD	0.083	26
28	BE	0.186	3	29	BF	0.176	4	30	BG	0.123	9
31	BH	0.088	23	32	BJ	0.103	19	33	BK	0.189	2

Figure 5. The distribution of large, zero income households in Leeds (per thousand)

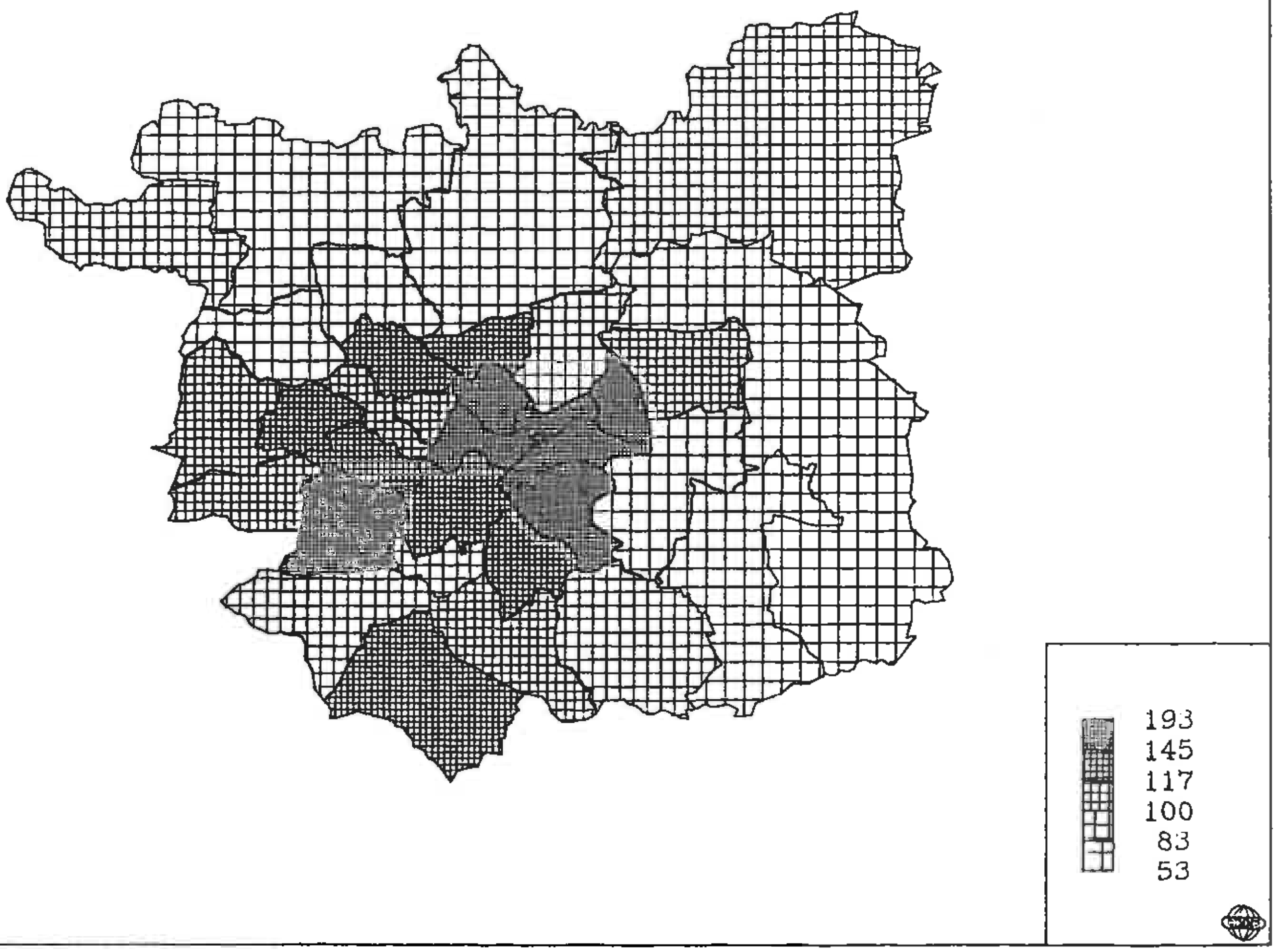

Figure 6. Household size by income: Leeds postal sectors.
(Layout as for Figure 3)

1	LS01	140	83	69	12	18	26
		54	58	28	22	13	20
		0	42	43	19	15	13

2	LS02	73	33	74	17	28	32	3	LS03	25	16	52	20	22	14
		29	31	10	12	9	13			5	13	7	1	12	15
		0	20	18	5	3	5			0	2	12	3	6	4
4	LS04	116	88	60	30	18	23	5	LS05	134	57	58	32	8	40
		53	48	21	23	10	15			47	58	28	27	15	23
		0	38	41	15	9	11			0	41	34	15	11	12
6	LS06	597	369	139	37	24	32	7	LS07	464	248	130	40	35	41
		231	236	152	95	37	46			171	210	117	55	47	48
		0	198	190	76	50	58			0	140	131	58	41	61
8	LS08	620	438	149	53	30	41	9	LS09	662	400	158	47	49	41
		223	323	168	118	93	82			187	255	126	88	65	53
		0	252	189	116	86	86			0	175	157	95	72	62
10	LS10	366	232	122	35	26	41	11	LS11	569	408	89	37	36	27
		104	188	103	71	54	48			177	221	129	75	57	43
		0	145	130	73	39	54			0	191	144	70	56	46
12	LS12	545	412	120	46	25	35	13	LS13	468	319	96	49	26	31
		200	318	153	81	61	36			159	240	103	87	61	42
		0	266	198	109	83	59			0	201	146	116	61	55
14	LS14	467	382	122	58	37	43	15	LS15	395	325	95	34	23	31
		153	293	135	83	80	67			122	311	114	85	52	39
		0	195	170	90	80	95			0	256	156	104	69	50
16	LS16	492	382	120	43	30	27	17	LS17	466	391	105	39	17	11
		160	343	140	90	60	47			147	362	121	99	60	27
		0	284	214	136	103	56			0	282	178	112	77	52
18	LS18	191	171	74	35	18	25	19	LS19	231	184	93	19	20	20
		77	130	65	55	39	18			71	167	87	58	37	19
		0	146	102	80	52	34			0	135	108	94	48	22
20	LS20	109	82	67	23	20	15	21	LS21	202	143	58	33	8	28
		19	79	40	26	21	17			64	127	46	39	33	18
		0	78	51	43	27	20			0	119	78	56	28	29
22	LS22	100	104	83	22	14	21	23	LS23	96	59	64	23	25	28
		26	82	35	32	35	12			25	96	45	25	14	14
		0	89	52	44	29	10			0	69	53	31	24	11
24	LS24	0	0	38	22	14	0	25	LS25	240	259	93	31	25	3
		0	0	5	6	4	0			74	263	128	100	65	22
		0	0	0	0	0	0			0	215	164	129	65	42
26	LS26	224	206	92	25	28	0	27	LS27	396	293	102	37	26	8
		70	181	87	86	48	30			111	287	126	94	47	32
		0	161	118	87	51	27			0	238	162	128	69	55
28	LS28	541	389	132	41	36	7	29	LS29	0	0	52	19	19	0
		160	327	177	134	56	36			0	0	3	4	12	0
		0	361	214	164	107	74			0	0	0	0	2	0
30	BD11	68	63	41	31	32	0	31	BD17	6	11	22	15	21	0
		27	52	26	22	15	7			0	7	12	10	7	0
		0	41	30	24	14	9			0	12	4	4	2	2
32	WF02	7	64	29	24	39	0	33	WF03	201	217	52	24	29	3
		2	3	2	5	9	1			64	131	62	42	33	10
		0	4	2	2	1	1			0	110	73	49	32	25
34	WF10	37	66	29	17	19	0	35	WF12	22	58	17	11	16	0
		12	33	19	13	7	4			4	14	11	6	8	1
		0	34	14	16	7	2			0	17	8	6	6	1

Figure 7. Profiles for 52 shopping centres in and around Leeds.

ZONE		PERS	EA	UNEMP	PENS	YOUNG	SC 1/2	%UNEMP	%PENS	%YOUNG	%SC12
1	Hemsworth	29	7	3	12	4	1	0.429	0.414	0.138	0.143
2	Normanton	34	15	3	2	9	1	0.200	0.059	0.265	0.067
3	Batley	44	18	5	6	8	3	0.278	0.136	0.182	0.167
4	Heckmondwike	32	12	3	3	7	1	0.250	0.094	0.219	0.083
5	South Elmsal	38	13	4	4	13	2	0.308	0.105	0.342	0.154
6	Featherstone	28	10	3	2	10	0	0.300	0.071	0.357	0.000
7	Ossett	44	21	2	6	11	1	0.095	0.136	0.250	0.048
8	Norbury	32	11	2	5	12	1	0.182	0.156	0.375	0.091
9	Dewsbury	180	66	6	40	38	7	0.091	0.222	0.211	0.106
10	Knottingley	25	13	2	4	2	1	0.154	0.160	0.080	0.077
11	Pontefract	452	211	16	45	107	34	0.076	0.100	0.237	0.161
12	Murfield	22	9	4	4	5	1	0.444	0.182	0.227	0.111
13	Castleford	1313	532	51	228	321	95	0.096	0.174	0.244	0.179
14	Durkhar	35	15	1	9	6	1	0.067	0.257	0.171	0.067
15	Wakefield	2684	1142	115	482	606	161	0.101	0.180	0.226	0.141
16	Tadcaster	17	5	0	10	1	0	0.000	0.588	0.059	0.000
17	Sherburn	155	70	4	24	39	12	0.057	0.155	0.252	0.171
18	Beeston	493	197	21	90	112	12	0.107	0.183	0.227	0.061
19	Moortown	142	51	4	47	19	16	0.078	0.331	0.134	0.314
20	Morley	1353	553	61	237	311	86	0.110	0.175	0.230	0.156
21	Oulton	179	74	6	31	49	7	0.081	0.173	0.274	0.095
22	Rothwell	146	62	2	27	34	9	0.032	0.185	0.233	0.145
23	Wetherby	1247	513	34	195	316	135	0.066	0.156	0.253	0.263
24	Ilkley	103	34	1	17	30	5	0.029	0.165	0.291	0.147
25	Otley	1449	616	35	247	348	133	0.057	0.170	0.240	0.216
26	Yeadon	456	200	17	62	112	39	0.085	0.136	0.246	0.195
27	Adel	352	149	16	46	78	31	0.107	0.131	0.222	0.208
28	Crossgates	2167	897	91	379	488	156	0.101	0.175	0.225	0.174
29	Halton	162	62	5	21	53	10	0.081	0.130	0.327	0.161
30	Bramley	436	182	28	80	84	13	0.154	0.183	0.193	0.071
31	Wortley	185	81	9	36	34	23	0.111	0.195	0.184	0.284
32	Armley	417	154	28	66	100	10	0.182	0.158	0.240	0.065
33	Sheepscar	139	45	9	25	34	5	0.200	0.180	0.245	0.111
34	Oakwood	705	268	42	135	148	35	0.157	0.191	0.210	0.131
35	Roundhay	255	97	7	53	63	30	0.072	0.208	0.247	0.309
36	Chapel All	762	289	54	134	153	48	0.187	0.176	0.201	0.166
37	Headingley	180	68	20	36	31	3	0.294	0.200	0.172	0.044
38	Meanwood	158	47	17	43	33	5	0.362	0.272	0.209	0.106
39	Horsforth	421	156	15	88	92	32	0.096	0.209	0.219	0.205
40	Pudsey	1388	617	56	226	300	106	0.091	0.163	0.216	0.172
41	Leeds	108901	44081	5438	19143	24455	6894	0.123	0.176	0.225	0.156
42	Silsden	35	16	3	6	8	1	0.188	0.171	0.229	0.063
43	Settle	27	15	1	5	5	2	0.067	0.185	0.185	0.133
44	Crosshill	26	11	0	6	5	0	0.000	0.231	0.192	0.000
45	Buttershaw	31	14	0	14	1	0	0.000	0.452	0.032	0.000
46	Skipton	53	19	4	9	16	1	0.211	0.170	0.302	0.053
47	Cleckheaton	30	15	2	2	7	3	0.133	0.067	0.233	0.200
48	Bingley	28	17	1	2	7	3	0.059	0.071	0.250	0.176
49	Greengates	20	5	0	5	5	1	0.000	0.250	0.250	0.200
50	Shipley	123	49	3	21	35	6	0.061	0.171	0.285	0.122
51	Keighley	22	7	0	3	7	1	0.000	0.136	0.318	0.143
52	Bradford	4545	1880	178	790	1052	287	0.095	0.174	0.231	0.153

Figure 8. Modified ranked shopping centre profiles for Leeds M.D.

Rank	Centre	Inflow	Centre	Unempt	Centre	SC 1/2
1	41 Leeds	107561	37 Headingley	0.2941	35 Roundhay	0.3093
2	52 Bradford	4487	33 Sheepscar	0.2000	31 Wortley	0.2840
3	15 Wakefield	2661	36 Chapel All	0.1869	23 Wetherby	0.2632
4	28 Crossgates	2144	32 Armley	0.1818	25 Otley	0.2159
5	25 Otley	1433	34 Oakwood	0.1567	27 Adel	0.2081
6	40 Pudsey	1370	30 Bramley	0.1538	39 Horsforth	0.2051
7	20 Morley	1338	41 Leeds	0.1234	26 Yeadon	0.1950
8	13 Castleford	1285	31 Wortley	0.1111	13 Castleford	0.1786
9	23 Wetherby	1228	20 Morley	0.1103	28 Crossgates	0.1739
10	36 Chapel All	754	27 Adel	0.1074	40 Pudsey	0.1718
11	34 Oakwood	701	18 Beeston	0.1066	17 Sherburn	0.1714
12	18 Beeston	484	28 Crossgates	0.1014	36 Chapel All	0.1661
13	26 Yeadon	450	15 Wakefield	0.1007	29 Halton	0.1613
14	11 Pontefract	445	39 Horsforth	0.0962	11 Pontefract	0.1611
15	30 Bramley	431	13 Castleford	0.0959	41 Leeds	0.1564
16	39 Horsforth	414	52 Bradford	0.0947	20 Morley	0.1555
17	32 Armley	413	9 Dewsbury	0.0909	52 Bradford	0.1527
18	27 Adel	349	40 Pudsey	0.0908	24 Ilkley	0.1471
19	35 Roundhay	252	26 Yeadon	0.0850	22 Rothwell	0.1452
20	31 Wortley	182	21 Oulton	0.0811	15 Wakefield	0.1410
21	37 Headingley	178	29 Halton	0.0806	34 Oakwood	0.1306
22	9 Dewsbury	177	19 Moortown	0.0784	50 Shipley	0.1224
23	21 Oulton	177	11 Pontefract	0.0758	33 Sheepscar	0.1111
24	29 Halton	160	35 Roundhay	0.0722	38 Meanwood	0.1064
25	38 Meanwood	157	23 Wetherby	0.0663	9 Dewsbury	0.1061
26	17 Sherburn	150	50 Shipley	0.0612	21 Oulton	0.0946
27	22 Rothwell	145	17 Sherburn	0.0571	30 Bramley	0.0714
28	19 Moortown	139	25 Otley	0.0568	32 Armley	0.0649
29	33 Sheepscar	138	22 Rothwell	0.0323	18 Beeston	0.0609
30	50 Shipley	123	24 Ilkley	0.0294	37 Headingley	0.0441

+ Note. Centres attracting less than 100 households are ignored.