WORKING PAPER 99/03

Putting Iterative Proportional Fitting on the Researcher's Desk

Paul Norman

School of Geography

University of Leeds

Leeds LS2 9JT

United Kingdom

e-mail p.norman@geography.leeds.ac.uk

October 1999

Abstract

‘Iterative Proportional Fitting’ (IPF) is a mathematical procedure originally developed to combine the information from two or more datasets. IPF is a well-established technique with the theoretical and practical considerations behind the method thoroughly explored and reported.

In this paper the theory of IPF is investigated with a mathematical definition of the procedure and a review of the relevant literature given. So that IPF can be readily accessible to researchers the procedure has been automated in Visual Basic and a description of the program and a ‘User Guide’ are provided.

IPF is employed in various disciplines but has been particularly useful in census-related analysis to provide updated population statistics and to estimate individual-level attribute characteristics. To illustrate the practical application of IPF various case studies are described. In the future, demand for individual-level data is thought likely to increase and it is believed that the IPF procedure and Visual Basic program have the potential to facilitate research in geography and other disciplines.

CONTENTS

List of Tables

List of Figures

1.0 INTRODUCTION

In Stan Openshaw's 'Census Users' Handbook' (1995), Philip Rees gave an account of the information provided to researchers by the 1991 Census. He suggested that 'Iterative Proportional Fitting' (IPF), a procedure originally developed to combine the information from two or more datasets (Bishop *et al.,* 1975), could be used to produce estimates of populations for years between censuses.

The application of IPF to update small area population estimates has proved useful to a local authority because previous reliance on 1991 data as if the population had not changed had 'undermined the case for government investment in regeneration schemes' (Simpson, 1998, p. 62). Given the updated information, council departments such as social services, planning, and housing, as well as the health authority and police were then able to re-evaluate their levels of resource and service provision (Norman, 1997b).

Many geographical analyses require large amounts of disaggregated data with detailed locational information for the observations. IPF can be used to integrate disaggregated data from one source with the aggregated data from another to estimate the characteristics of a population within given geographical units in a disaggregated format (Wong, 1992). For example, IPF has been used to estimate the distribution of conditional probabilities of household attributes at Enumeration District (ED) level by combining data from the Census 'Small Area Statistics' (SAS) and 'Sample of Anonymised Records' (SARs) (Ballas *et al.*, 1999).

In 1987, Mark Birkin discussed the mathematics that underpin IPF, summarised the outcome of the procedure and described the structure and execution of a computer program to carry out IPF written in FORTRAN that was developed at the School of Geography, University of Leeds. So that IPF can be readily accessible to researchers, especially as the use of spreadsheets is widespread, the procedure has recently been programmed in Visual Basic. In this paper the theory of IPF will be reported through a definition of the procedure and a review of the relevant literature. The method of utilising the IPF Visual Basic programme will be described and examples of IPF applications will be given.

2.0 IPF: THEORY

2.1 Description of the Procedure

'Iterative Proportional Fitting' (IPF) is a mathematical scaling procedure which can be used to ensure that a two-dimensional table of data is adjusted so that its row and column totals agree with constraining row and column totals obtained from alternative sources. IPF acts as a weighting system whereby the original table values are gradually adjusted through repeated calculations to fit the row and column constraints. The resultant table of data is a 'joint probability distribution' of 'maximum likelihood estimates' obtained when the probabilities are convergent within an acceptable (pre-defined) limit (Birkin, 1987; Bishop *et al.*, 1975). For example, Table 1 has initial population counts for marital status in various age-groups for 1957. These counts have then been updated to 1958 by IPF to be consistent with age-group and marital status totals for that year.

Table 1: Example of Using IPF

1957 Estimates of Female Population (1,000's) in England and Wales

Original Data	Single	Married	Widowed/	
Age			Divorced	1957 Totals
15-19	1306	83	0	1389
20-24	619	765	3	1387
25-29	263	1194	9	1466
30-34	173	1372	28	1573
35-39	171	1393	51	1615
40-44	159	1372	81	1612
45-49	208	1350	108	1666
50+	1116	4100	2329	7545
1957 Totals	4015	11629	2609	18253

1958 Estimates Produced using Iterative Proportional Fitting

Data Amended Using IPF	Single	Married	Widowed/	
Age			Divorced	1958 Totals
15-19	1325.27	86.73	0.00	1412.00
20-24	615.56	783.39	3.05	1402.00
25-29	253.94	1187.18	8.88	1450.00
30-34	165.13	1348.55	27.32	1541.00
35-39	173.41	1454.71	52.87	1681.00
40-44	147.21	1308.12	76.67	1532.00
45-49	202.33	1352.28	107.40	1662.00
50+	1105.16	4181.04	2357.81	7644.00
1958 Totals	3988.00	11702.00	2634.00	18324.00

Source: Bishop et al. (1975, Table 3.6-1, p. 98). The 1957 data have been amended to be consistent with data relating to 1958 using the IPF Visual Basic program.

2.2 Mathematical Procedure

IPF was first proposed by Deming and Stephan (1940, cited by Fienberg, 1970; Bishop *et al.*, 1975; and Wong, 1992) as a method for estimating cell probabilities in a contingency table based on observations subject to constraints from known and fixed marginal row and column totals. Fienberg (1970) has traced the subsequent development of the mathematical

procedures involved in IPF (through Stephan, 1942; Deming, 1943; Smith, 1947; El-Badry & Stephan, 1955; Brown, 1959; Friedlander, 1961; Bishop, 1967; Mosteller, 1968; Fienberg, 1968; and Fienberg & Gilbert, 1970) and, along with his 1977 paper, Fienberg is invariably cited in the literature.

The mathematical procedures involved in IPF have been widely explored and reported by Fienberg (1970 & 1977), Bishop *et al.* (1975) Birkin (1987) and Birkin and Clarke M (1988a). Also, as Paul Williamson of the University of Liverpool pointed out (pers. com. 12/2/99), 'the best, most up-to date and, in fact, only 'non-mathematical' critique of IPF [is] Wong's 1992 article in the *Professional Geographer'*. For a formal definition of IPF see Figure 1.

Figure 1: Mathematical Definition of IPF

$$p_{ij(k+1)} = \frac{p_{ij(k)}}{\sum_j p_{ij(k)}} * Q_i \qquad (1)$$

$$p_{ij(k+2)} = \frac{p_{ij(k+1)}}{\sum_i p_{ij(k+1)}} * Q_j \qquad (2)$$

Where $p_{ij(k)}$ is the matrix element in row i, column j and iteration k. Q_i and Q_j are the pre-defined row totals and column totals respectively. Equations (1) and (2) are employed iteratively to estimate new cell values and will theoretically stop at iteration m where:

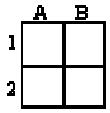
$$\sum_j p_{ij(m)} = Q_i \quad \text{and} \quad \sum_i p_{ij(m)} = Q_j$$

Source: after Wong, 1992, pp. 340-341

Noting that the procedure was originally developed for combining the information from two or more datasets; IPF's 'classical' use (p. 97), Bishop *et al.* (1975) described in detail the calculation of 'maximum likelihood estimates' by IPF. Given known row and column constraints, IPF can also estimate the maximum likelihood estimates of a two-dimensional matrix where the values are not known, if the initial table values are constant (e.g. all equal to one). Using 'cross-product ratios', an important property of IPF identified is the proof of the maintenance of interactions (or lack of them) in the initial values whilst the effects and interactions represented by the constraining values are fitted. The method prescribed by Bishop *et al.* (1975) can be used to check that the interaction pattern of the original table is preserved by examining the cross-product ratios (see Figure 2).

Figure 2: Calculating Cross-Product Ratios

To calculate cross product ratios, with a cell layout referenced:



the following formula can be used to calculate cross-product ratios:

$$\frac{A1 * B2}{B1 * A2}$$

Thus, the four cells in the top left-hand corner of the 1957 data in Table 1 give:

$$\frac{(1{,}306) * (765)}{(619) * (83)} = 19.45$$

and the equivalent cells in the 1958 data estimated using IPF give:

$$\frac{(1{,}325.27) * (783.39)}{(615.56) * (86.73)} = 19.45$$

Source: after Bishop et al. (1975, p. 98)

Bishop *et al.* (1975, pp. 85-102) also discussed the 'convergence of the procedure' and 'stopping rules'. Convergence is taken to have occurred and the procedure stops when no cell value would change in the next iteration by more than a pre-defined amount that obtains the desired accuracy. Convergence of the data will not occur if there are zero cells in the marginal row and column constraints, minus numbers in any of the data or a mismatch in the totals of the row and column constraints. 'Too many' zero cells in the initial matrix may prevent convergence through a 'persistence of zeros' (p. 101). 'Too many' was undefined but if a matrix contains evenly distributed zeros in more than 30% of the cells or are grouped closely together and comprise around 10% of the table, convergence may not occur.

2.3 IPF Utility

Birkin (1987 p. 2) identified the usefulness of micro-level population data in addressing aggregation problems reporting that IPF was an important 'weapon … in the micro-simulation armoury' (citing Birkin & Clarke M, 1986 for a 'defence of the method'). Birkin and Clarke M (1988b, p. 7) believed that IPF was robust, with maximum likelihood estimates produced for the data provided. A key feature noted was the ability to 'quantify explicitly the inter-dependencies between a very large number of attributes'. The outcome of the IPF procedure retains all of the information included in the constraints and this may subsequently be reaggregated (Birkin, 1987). Birkin and Clarke G (1995, p. 373) reported that 'no errors are introduced by the IPF process (i.e. we can estimate a complete set of joint probabilities which is completely consistent with all known constraints)'.

In addition to the mathematical definition of IPF, Wong (1992) investigated the reliability of using the procedure. He suggested that if an original matrix is derived from a sample, it is

only one of many possible matrices or random samples that could be extracted from the population. Since the sample's distribution may deviate from the 'real' distribution, random error may be introduced. If IPF is carried out on a matrix of sample observations to estimate the population, any random error will be propagated. Using sample populations from the Public-Use Microdata Sample (PUMS) and a statistical test of the total absolute error (after Upton, 1985), the random error effect was found to be influenced by the size of both the sample and the matrix. Wong (1992) noted that increasing the sample size improved the accuracy of estimates derived using IPF, but that this gain was more significant for larger matrices.

Wong (1992 p. 347) carried out a thorough evaluation of the factors which may affect the performance of the procedure and has identified the conditions under which IPF produces 'acceptable' estimates. He believed the procedure to be a feasible technique to create spatially disaggregated datasets by combining locational information from census reports with other disaggregated data. In his view, most geographers, 'with a few exceptions (e.g. Birkin & Clarke M, 1988a), have not recognised the utility and potential of IPF in generating individual level data from aggregated data for spatial analysis' (p. 341).

2.4 Applications

The IPF procedure has been applied under several names in various fields, including statistics, demography, spatial interaction modelling, 'Cross-Fratar' and 'Furness' methods in transportation engineering and 'RAS' in economics (Wong, 1992 p. 340). Most relevant to census-based applications, the use of IPF in geographical research and modelling has been demonstrated by Birkin and Clarke M (1988a).

Birkin and Clarke M (1988a) identified a deficiency of microlevel data regarding the characteristics of residents of small geographical areas. They described the development of 'SYNTHESIS' an updatable micro-database generated using IPF and Monte Carlo sampling from a number of different aggregate tabulations including the 1981 Census and Family Expenditure Service. Birkin and Clarke M (1988b) applied this method to generate incomes for individuals differentiated in terms of age, sex, occupation, industry and locational characteristics. They noted that, whilst both IPF and Monte Carlo sampling were well-established techniques, there were few applications in the spatial sciences.

Acknowledging that the 'plea for microdata has been partly met by the agreement to publish samples of anonymised records' in the 1991 Census, Birkin and Clarke G (1995 p. 364) believed there was still a lack of information for individuals or households. They reviewed microsimulation methods to synthesise census data including the use of IPF and described a method to estimate household water demand. Subsequently, in a collection of papers on microsimulation (Clarke G, 1996a) much of the same material was considered in greater depth (Clarke G, 1996b; Hooimeijer, 1996; Williamson *et al.*, 1996). Ballas *et al.* (1999) are currently using IPF as part of their research into microsimulation methodologies for the estimation of household attributes (see Section 4.3).

Rees (1994 & 1995) demonstrated a method for updating small area populations by adjusting age/sex census data to electors per ward using IPF. Similar in principle, 1996 ward age/sex/ethnicity population estimates were produced in Bradford using IPF to adjust local patient counts to ward-based electoral register data and district-wide age-group information (see Section 4.1) (Norman, 1997a & 1997b; Simpson & Norman, 1997; Simpson, 1998). Research in Bradford also saw the estimation of Labour Force Forecasts by age, sex and

ethnic group using IPF to fit 100% 1991 Census data for Bradford, the 2% SAR for West Yorkshire and full national information (Simpson, 1996). The work in Bradford used a macro written in Lotus 1-2-3 to carry out IPF (Norman, 1997a).

2.5 Literature Summary

It is interesting to note the geography of the literature summarised in Table 2. Work on the mathematical proof of IPF appears to have taken place in the USA, whereas for the majority of geography-related applications the researchers are located in Leeds and Bradford!

Table 2: IPF Literature Summary

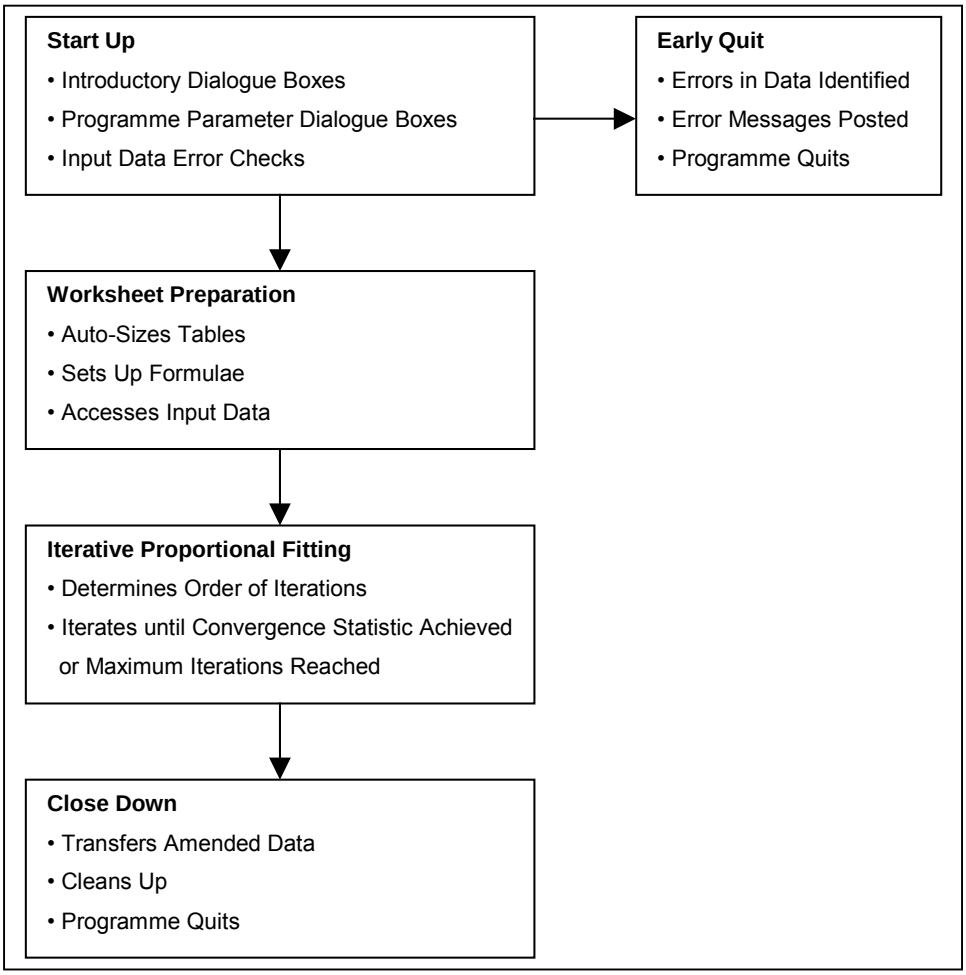
Author	Year	Context/Application	Researcher's Location
Fienberg SE	1970	Mathematical exploration of IPF	University of Chicago
Bishop Y *et al.*	1975	Mathematical exploration of IPF	Cambridge, Mass
Fienberg SE	1977	Mathematical exploration of IPF	University of Chicago
Birkin M	1987	IPF programmed in FORTRAN	University of Leeds
Birkin M & Clarke M	1987	IPF used in microsimulation	University of Leeds
Birkin M & Clarke M	1988a	IPF used in microsimulation	University of Leeds
Birkin M & Clarke M	1988b	IPF used in microsimulation	University of Leeds
Wong D W S	1992	Mathematical exploration of IPF	University of Connecticut
Rees P	1994	IPF used in population estimates	University of Leeds
Birkin M & Clarke G	1995	IPF used in microsimulation	University of Leeds
Rees P	1995	IPF used in population estimates	University of Leeds
Hooimeijer P	1996	IPF used in microsimulation	University of Utrecht
Simpson S	1996	IPF used to combine data sources	Bradford Council
Williamson P *et al.*	1996	IPF used in microsimulation	University of Leeds
Clarke G P	1996a	IPF used in microsimulation	University of Leeds
Clarke G P	1996b	IPF used in microsimulation	University of Leeds
Simpson S & Norman P	1997	IPF used in population estimates	Bradford Council
Norman P	1997a	IPF used in population estimates	Bradford Council
Norman P	1997b	IPF used in population estimates	University of Bradford
Simpson S	1998	IPF used in population estimates	Bradford Council
Wilson T & Rees P	1998	IPF used in population estimates	University of Leeds
Norman P	1999	IPF programmed in Visual Basic	University of Leeds
Ballas *et al.*	1999	IPF used in microsimulation	University of Leeds
Norman P	1999	IPF used to estimate household size	University of Leeds

3.0 IPF METHOD

So that the procedure can be readily accessible to researchers, IPF has been programmed in Visual Basic for Applications (VBA) and a self-contained ‘User Guide’ has been written (see below). The program runs in Excel 5.0 and above, on both PCs (Windows 3.1, 98 and NT) and Macs (System 7.0 or later) with minor variations in code depending on operating system and version of Excel.

Essentially an automated spreadsheet, the VBA program is structured in four main sections (see Figure 3) that each comprise short sub-routines: ‘Start-Up’ includes introductory and program parameter dialogue boxes as well as input data error checks and the opportunity to quit the program; ‘Worksheet Preparation’ in which the size of the input data is determined automatically and the spreadsheet formulae are set up; ‘IPF’ is carried out on the data until the stopping rules are satisfied and; ‘Close Down’ when the amended data are transferred to a results worksheet with a summary of the IPF procedure and the program quits.

Figure 3: IPF Visual Basic Program Structure



IPF: VISUAL BASIC PROGRAM USER GUIDE

Iterative Proportional Fitting (IPF)

- IPF is a mathematical procedure originally developed to combine the information from two or more datasets that can be used when the values in a table of data are inconsistent with row and column totals obtained from other sources. Acting as a weighting system, the values within the table are gradually adjusted to fit the fixed, constraining row and column totals. This occurs through repeated calculations when the figures within the table are alternatively compared with the row and column totals and adjusted proportionately each time with IPF keeping the cross-product ratios constant so that interactions are maintained.

- As the iterations potentially continue *ad infinitum* a 'Convergence Statistic' is set as a cut-off point when the fit of the datasets is considered close enough. The iterations continue until no value would change by more than the specified amount.

- An example of IPF and a mathematical definition are given by Philip Rees (1994), the mathematics are discussed in greater depth by Stephen Fienberg (1970) and the procedure is reviewed in an accessible manner by David Wong (1992). Practical uses of IPF have been reported by Paul Norman (1997a), Ludi Simpson (1996 & 1998) and Dimitris Ballas, Graham Clarke and Ian Turton (1999).

Program Instruction Summary

a) Open an Excel workbook and save it as ‘ipfdata.xls’

b) On a worksheet insert the original data and name the ranges ‘table’, ‘rows’ and ‘cols’

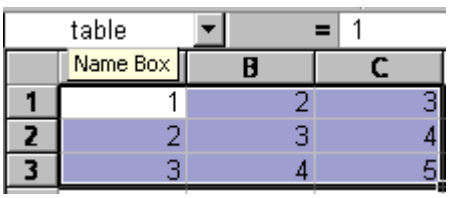
c) With ipfdata.xls still open, open ‘ipfprog.xls’ to run the program

d) Save the amended data that has been added to a new worksheet in ‘ipfdata.xls’

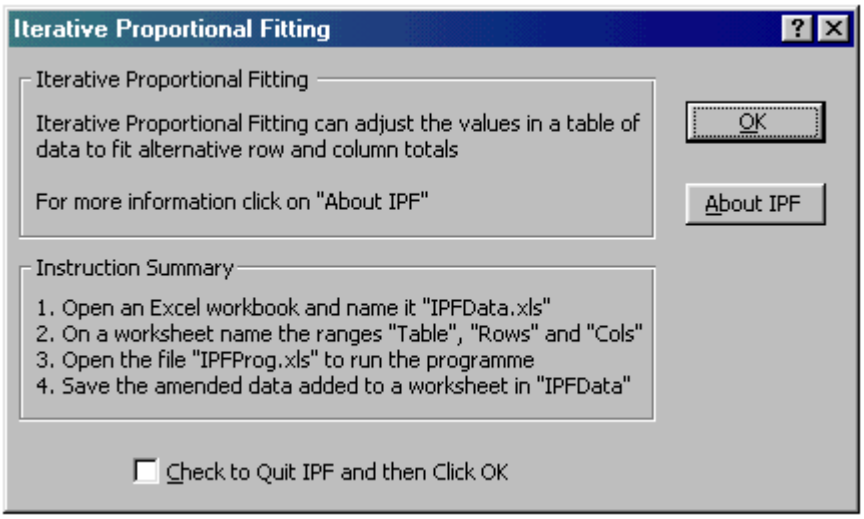
Detailed Instructions

- The data to be used by the IPF program must be entered into an Excel workbook named ‘ipfdata.xls’. Data can be real numbers, probabilities or percentages, but the row and column constraints must be consistent. Within the limitations of Excel spreadsheets the data may be of any size.

- The data need to be named as the ranges ‘table’, ‘rows’ and ‘cols’. To achieve this shade the data using the mouse and enter the name in the ‘Name Box’ and press the ‘enter’ key or use the ‘Insert’ menu then ‘Name’ and ‘Define’ (for assistance see below or use Excel’s Help: search for ‘naming cells in a workbook’).

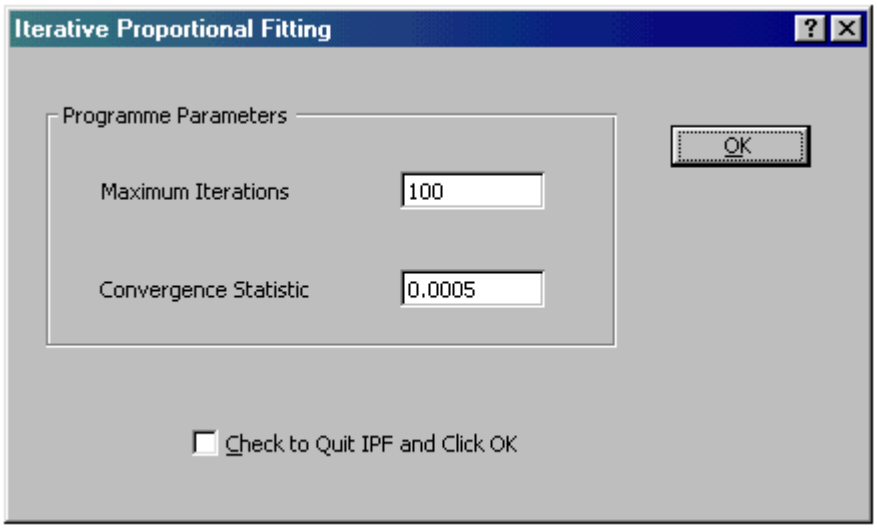


- With the ipfdata.xls file still open, open the Excel file ‘ipfprog.xls’ to run the IPF program. Click on ‘enable macros’ (Excel 97 and 2000 only) and the following dialogue box appears:



- Click on ‘OK’ to continue the program (or ‘About IPF’ to display further dialogue boxes which contain text similar to this guide).

- A dialogue box is then displayed so that values for the maximum number of iterations and the convergence statistic may be set (default values 100 and 0.0005 respectively):



- The program checks for errors in the input data that would prevent the calculation from being carried out successfully. If errors are identified the user is informed of their nature and may leave the program to make amendments.

- Error checks include:-

  a) The named ranges table, rows and cols are available

  b) The sum of rows equals the sum of cols (to within 0.0000000001)

  c) The table is equivalent in size to the rows and cols

  d) There are no blank cells in the table, rows or cols

  e) There are no zero cells in the rows or cols

  f) There is no text in the table, rows or cols (blank cells may be interpreted as text)

  g) There are no minus numbers in the table, rows or cols

- If no errors are found the program uploads the named ranges and the IPF procedure is carried out. Excel's 'Status Bar' will indicate the operation that is being carried out.

- The IPF program ends when either the convergence statistic is achieved or the maximum number of iterations is reached.

- A worksheet is added to the 'ipfdata.xls' file containing the amended data and a summary of the IPF operation including the number of iterations that occurred and the convergence statistic set.

4.0 IPF: EXAMPLES

Three case studies will be reviewed that have each utilised the IPF procedure. The first is the methodology used to update 1991 ward populations in Bradford to 1996, the mid-census year. Related to this, the second is an estimation of household overcrowding for 1996 in Bradford and the third is a method of estimating the head of household attributes age, sex and marital and employment status for EDs in Leeds. Since only an outline of each application and methodology is given, researchers are encouraged to pursue the references at the end of each section and/or to contact the authors.

4.1 1996 Ward Population Estimates Using IPF

Users of population statistics in Bradford including various council departments, the health authority and the police service indicated that their quality of service and resource planning would benefit from population estimates updated from 1991 Census data for geographical areas smaller than the Metropolitan District.

In addition to the 1991 Census, three main sources of data were used: the Bradford District Population Forecast, the Electoral Register and the Family Health Service Authority (FHSA) Patient Register. Developed by Bradford Council's Research Section, two methods were devised to fully utilise the information that each data source included. Mid-year 1996, ward-based age/sex/ethnic group population estimates were an average of these.

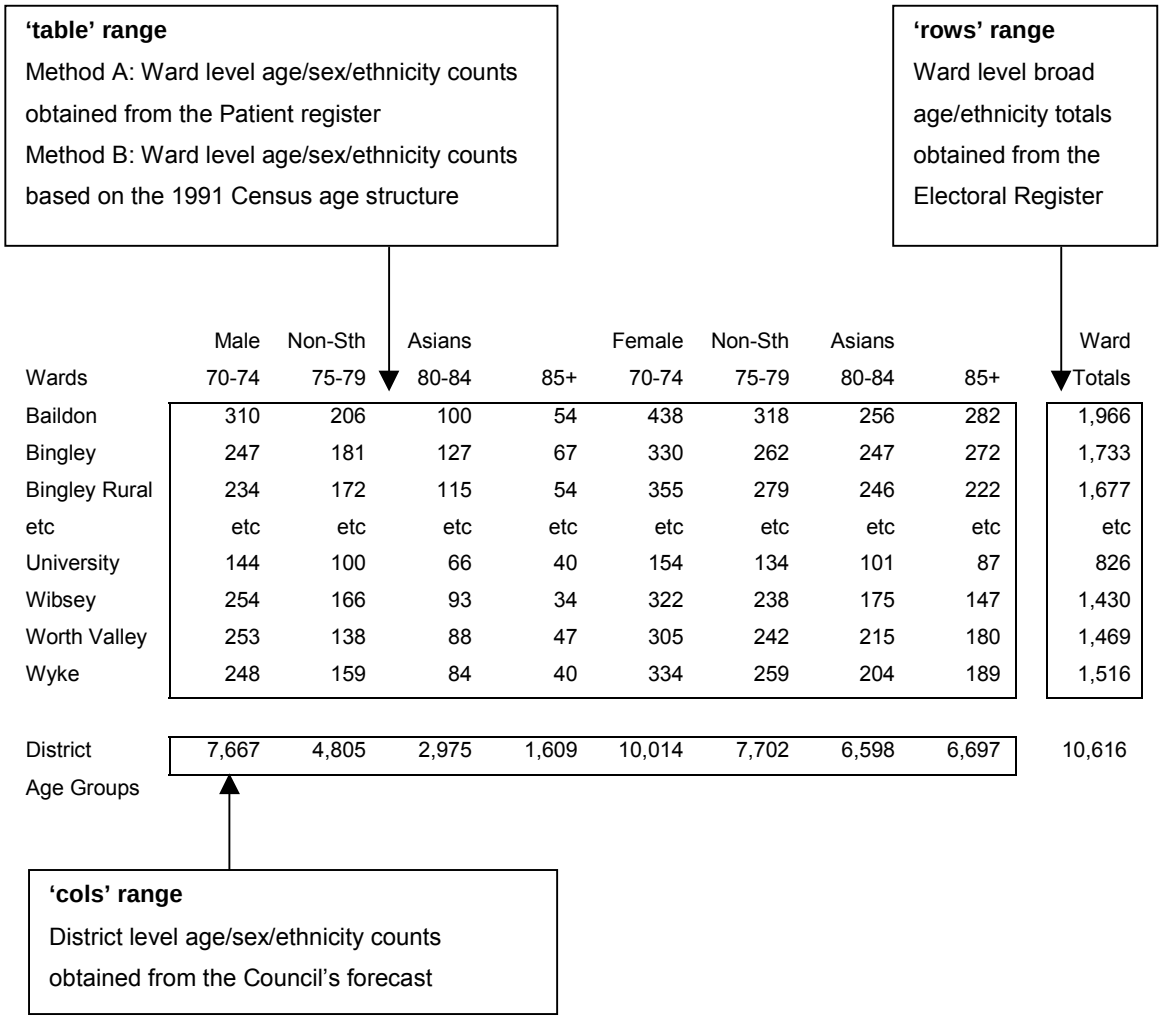
4.1.1 Method A: Estimate Based on the Patient Count, Electoral Register and District Forecast

The principle was to take the Patient Register, a detailed dynamic record of population that suffers from some inaccuracy, and use IPF to scale it to fit with more reliable, but less detailed, evidence from the District Forecast and Electoral Register. The 'table' of data for the IPF procedure was provided by an anonymised Patient Register categorised by age, sex and broad ethnic group (South Asian and Non-South Asian) and allocated to a ward using postcodes.

The columns 'cols' constraint was age-group totals (quinquennial with extra divisions at 15, 16 & 17 to allow subsequent aggregation versatility) by sex and ethnic group from the Council's District Forecast for 1996. Using the 1991 Census as its base population, the Forecast was calculated by 'cohort survival' method taking into account births, deaths and recorded migration and including assumptions about trends in fertility, mortality and migration for the most recent and future years (Simpson, 1994 & 1995).

The 'rows' constraint was provided by the Electoral Register (ER). Adult ward level totals can be estimated by aggregating postcoded individual data into wards in broad age bands (18-69 and 70+) and ethnic groups (South Asian and Non-South Asian) after allowances for time (a weighted average of the 1995 and 1996 ERs) and incompleteness (scaling for non-registration and ineligibility). Under 18 ward totals were estimated by scaling the patient count of children by the ratio of the adult patient count to the ER-based adult ward totals. The 'table' of patient counts was then constrained to the District 'cols' totals and the ER-derived ward 'rows' totals using IPF (see Figure 4).

Figure 4: 1996 Ward Population Estimates in Bradford using IPF



	Male				Female				Ward
Wards	Non-Sth 70-74	Non-Sth 75-79	Asians 80-84	Asians 85+	Non-Sth 70-74	Non-Sth 75-79	Asians 80-84	Asians 85+	Totals
Baildon	310	206	100	54	438	318	256	282	1,966
Bingley	247	181	127	67	330	262	247	272	1,733
Bingley Rural	234	172	115	54	355	279	246	222	1,677
etc	etc	etc	etc	etc	etc	etc	etc	etc	etc
University	144	100	66	40	154	134	101	87	826
Wibsey	254	166	93	34	322	238	175	147	1,430
Worth Valley	253	138	88	47	305	242	215	180	1,469
Wyke	248	159	84	40	334	259	204	189	1,516
District Age Groups	7,667	4,805	2,975	1,609	10,014	7,702	6,598	6,697	10,616

4.1.2 Method B: Estimate Based on the 1991 Census Age Structure, 1996 Electorate and District Forecast

The 'rows' and 'cols' constraints were the same as in Method A, but for Method B the 'table' was based on the 1991 Census age structure. After allowances for student residence and undercount (Simpson, 1993), 1991 Census data were used as a base population. For recent births, those aged 0-4 in the Census data were replaced with counts from the Patient Register. Those aged 5-39 were taken as the same as in 1991 (on the assumption that young communities tend to remain fairly static). To allow for older communities ageing *in situ*, those aged 40+ were replaced by the 1991 count of persons aged five years younger. The 1991 Census-based 'table' was adjusted by IPF to the District age-group 'cols' and ER-based ward 'rows' totals.

To lessen the impact of large errors in any one data source, Methods A and B were averaged and disseminated for males and females aggregated into seven broad age bands and the two ethnic groups. All figures were rounded to the nearest ten persons.

For further details on the above methodology see Norman (1997a & b), Simpson and Norman (1997) and Simpson (1998).

4.2 Estimating Household Overcrowding for an Intercensal Year

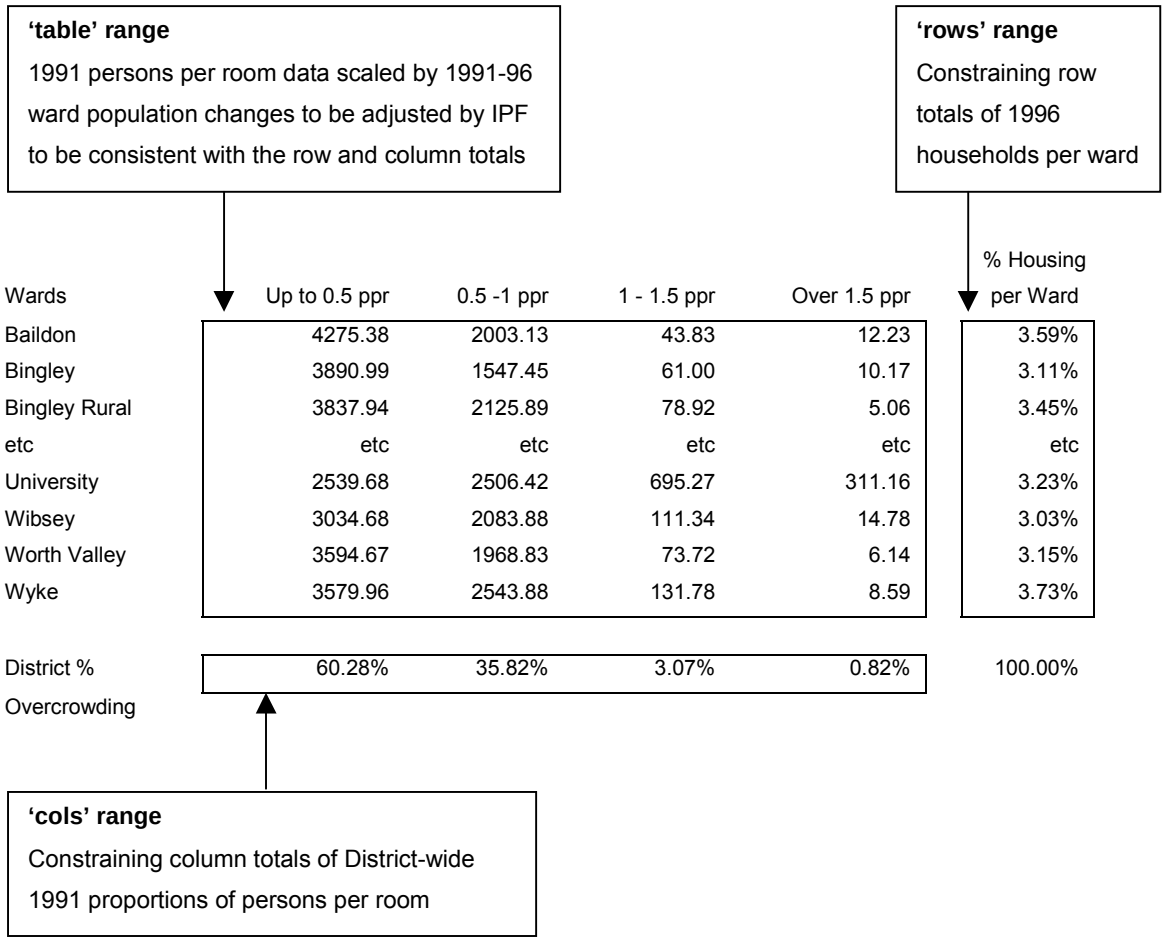
The aim was to estimate household overcrowding in Bradford Metropolitan District for 1996, the mid point between the decennial censuses, using evidence relevant to that year. The principle was to take the number of persons per room as indicated by the 1991 Census and bring forward this information using IPF to fit 1996 ward-level populations whilst taking into account the number of new houses built in Bradford between 1991 and 1996. The output of the IPF model would be an estimate of the number of households in each persons per room category for 1996.

The 1991 Census persons per room information from Local Based Statistics (LBS) table L23 was used to provide a table of data to be updated. The persons per room categories were scaled by the percentage change in each ward population between 1991 and 1996 as indicated by Bradford Council's mid-year 1996 ward population estimates (see Section 4.1). This represented the effect of ward-specific changes in pressure on housing through the increase or decrease in population and provided interaction data for the relationships between wards. The housing completions in each ward between 1991 and 1996 (Walton, 1999) were added to the number of households as shown by the 1991 Census in LBS L23. The total number of households in Bradford District for 1996 was divided into the four persons per room categories in the same proportions as indicated for the District as a whole by table L23.

To estimate levels of household overcrowding for 1996 various assumptions were made: all new houses were occupied; the same ratio of 1991 residents in households to persons in institutions existed in 1996 and; the same proportions of each 1991 Census persons per room category existed across the District as a whole in 1996.

Prior to running the IPF program the data were entered into an Excel worksheet in a workbook named 'ipfdata.xls'. As the Visual Basic program accesses data in named 'ranges' the ward-specific persons per room data were defined as a range called 'table', the constraining ward household totals as the range 'rows' and the District proportions of persons per room categories as the range 'cols' (see Figure 5). The IPF Visual Basic program was run and numbers of overcrowded and severely overcrowded households were calculated from the output and expressed as percentages of the 1996 number of households.

Figure 5: 1996 Household Overcrowding Estimates in Bradford using IPF



Wards	Up to 0.5 ppr	0.5 -1 ppr	1 - 1.5 ppr	Over 1.5 ppr	% Housing per Ward
Baildon	4275.38	2003.13	43.83	12.23	3.59%
Bingley	3890.99	1547.45	61.00	10.17	3.11%
Bingley Rural	3837.94	2125.89	78.92	5.06	3.45%
etc	etc	etc	etc	etc	etc
University	2539.68	2506.42	695.27	311.16	3.23%
Wibsey	3034.68	2083.88	111.34	14.78	3.03%
Worth Valley	3594.67	1968.83	73.72	6.14	3.15%
Wyke	3579.96	2543.88	131.78	8.59	3.73%
District % Overcrowding	60.28%	35.82%	3.07%	0.82%	100.00%

Whilst it was believed there was an inherent logic in a method that used location-specific data about changes in population and housing stock to estimate their effect on levels of household overcrowding, a criticism of the approach could be that, although the housing completions were direct evidence of change, the method used 1996 populations that were themselves estimated (albeit from 1996 evidence) increasing any uncertainties. More fundamentally, the assumption that the same proportions of each 1991 census persons per room category existed across the District as a whole in 1996 may not be true. Ideally the columns constraint should have contained direct evidence relating to 1996.

For further details about the above methodology see Norman (1999) or contact the author at the School of Geography, University of Leeds.

4.3 Exploring Microsimulation Methodologies for the Estimation of Household Attributes

Microsimulation methodologies aim to build large-scale datasets based on the attributes of individual persons, households, firms or organisations. As a result, through the simulation of economic, demographic and social processes, the policy impact on micro-units can be analysed (Orcutt *et al.*, 1986; Birkin & Clarke G, 1995; Clarke, 1996).

In the context of using microsimulation techniques to build population microdata sets for small geographical areas, Ballas *et al.* (1999) are currently exploring various methodologies for the estimation of household attributes. Relevant to IPF, the first stage of the microsimulation process is to estimate attribute conditional probabilities which are then used in the second stage, when detailed micro-level populations are created.

Acknowledging that there are several ways to compute conditional probabilities for spatial microsimulation modelling (including linear programming models, discrete choice models and balancing factors in spatial interaction models), Ballas *et al.* (1999) have used the IPF VBA program to estimate the joint probability distribution for the following head of household attributes: Age, Sex, Marital Status and Employment Status. The probabilities for economic activity and employment status were derived using data from the 1991 Census SAS Table 8. From SAS Table 39, probabilities of the head of household being male or female and married or single-widowed-divorced (SWD) conditional upon age and location at ED level were calculated (see Table 3).

Table 3: Probability Rates for an ED Beeston, a Ward in Leeds

1991 Census SAS Table S08

Employment Status	Economically Active	Employee	Self-Employed	On Gov. Scheme	Un-Employed
Male Aged 16–29	0.054054	0.689189	0.027027	0.027027	0.202703

1991 Census SAS Table S39

Marital Status	SWD	Married
Males Aged 16–29	0.825	0.175

With the initial matrix values set to one (since the original distribution was unknown and there were no interactions to maintain), IPF was used to estimate the joint probability distribution by adjusting the matrix so that its row and column sums match the row and column constraints (see Table 4). In this way, IPF can be applied to combinations of data from more SAS tables to estimate joint probability distributions for a larger number of household attributes and for all EDs in an area.

Table 4: Matrix and Constraints for Beeston, Before and After IPF

Household Attributes	Economically Active	Employee	Self-Employed	On Gov. Scheme	Un-Employed	Row Constraint
SWD	1.000000	1.000000	1.000000	1.000000	1.000000	0.825000
Married	1.000000	1.000000	1.000000	1.000000	1.000000	0.175000
Column Constraint	0.054054	0.689189	0.027027	0.027027	0.202703	∑ = 1

Household Attributes	Economically Active	Employee	Self-Employed	On Gov. Scheme	Un-Employed	Row Constraint
SWD	0.044595	0.568581	0.022297	0.022297	0.167230	0.825000
Married	0.009459	0.120608	0.004730	0.004730	0.035473	0.175000
Column Constraint	0.054054	0.689189	0.027027	0.027027	0.202703	∑ = 1

Ballas *et al.* (1999) also used IPF to create spatially disaggregated conditional probability distributions by combining data from the SAS and the SARs. The data used refer to the employment status of a subset of the population who are male, head of household, SWD and

aged 16-19 years. The row constraints were derived from the SARs and depict the employment status percentages of this population subset for Leeds Metropolitan District as a whole. The column constraints are derived from the SAS depicting the percentages of the subset for each ED in Leeds. The cells in the table contain the subset's employment status probability distribution with data for each ED from the SAS with their relative weights reflecting the interactions (see Table 5).

Table 5: SAS and SAR-Based Probability Matrix and Constraints

Household Attributes/ED	Economically Active	Employee	Self-Employed	On Gov. Scheme	Un-Employed	SAS Row Constraint
DAFA01	0.40000	0.00000	0.00000	0.000000	0.600000	0.014621
DAFA02	0.47368	0.00000	0.00000	0.000000	0.526315	0.058484
DAFA03	0.63636	0.00000	0.00000	0.090909	0.272727	0.043863
DAFA04	0.61538	0.00000	0.00000	0.076923	0.307692	0.058484
DAFA05	0.22220	0.00000	0.11111	0.000000	0.666667	0.021931
DAFA06	0.34375	0.00000	0.00000	0.03125	0.625000	0.007310
etc	etc	etc	etc	etc	etc	etc
DAGK45	0.90909	0.00000	0.00000	0.000000	0.909090	0.014621
SARs Column Constraint	63.0541	9.11330	0.985222	13.38259	13.46470	∑ = 100

IPF can be applied to the matrix so that its row and column totals will be equal to the row and column constraints and with the original cell interactions maintained. The result is a joint probability distribution for the head of household characteristics: Age, Sex, Marital Status, Employment Status and Location (ED). Based on this set of probabilities, the second stage of the microsimulation procedure, using Monte Carlo simulation to create a sample of individuals, can then proceed.

For more details of this ongoing work, see Ballas *et al.* (1999), or contact the authors at the School of Geography, University of Leeds.

5.0 CONCLUSIONS

Iterative Proportional Fitting is a well-established technique with the theoretical and practical considerations behind the method thoroughly explored and reported. The procedure is employed in various disciplines but has been shown to be particularly useful in practical applications to provide updated population statistics and to estimate individual-level attribute characteristics. Analyses based on spatially aggregated data have been criticised for leading to ecological fallacies and research into the 'modifiable areal unit problem' (MAUP) has demonstrated that using areal data for drawing statistical inference is not appropriate (Openshaw, 1984; Fotheringham & Wong, 1991; Wong, 1992). Analyses using individual level data about persons or households for small areal units such as census EDs 'can yield more conservative results. IPF is a procedure that can be used to produce this disaggregated data' (Wong, 1992, p.342).

To use IPF, it is fundamental to think through the implications of combining the information in the datasets. It is necessary to include the maximum amount of information in the constraints and to organise the data into the necessary 'table', 'rows' and 'cols' ranges prior to running the VBA program. The marginal constraints for IPF can be real numbers or a set of probabilities, but the input data must be consistent (Birkin, 1987). Despite the program's data error checks, as Birkin (1987, p. 6) noted, 'there is no substitute for careful checking of data at the input stage'.

Openshaw and Clarke (1996, p. 31) reported ‘an inverse relationship between the degree of mathematical complexity present in a spatial analysis method and the associated level of applied geographical usefulness’. Statistical and GIS packages tend to hide the underlying complexity thereby enabling their use by non-specialists but not necessarily guaranteeing sensible analyses or meaningful results. The IPF Visual Basic program is intended to be very easy to use and it is not necessary to fully understand how the mathematics of the IPF procedure works. Similar to Birkin’s (1987) view, it is possible to treat the program as a ‘black box’ by inputting a set of marginal constraints and obtaining a full, fitted, joint distribution. The skill is in the choice of data, data preparation and the interpretation of the program output.

Wong (1992) believed that the demand for individual-level data will increase in the future and that IPF has the capability to generate disaggregated data to meet the demand. He concluded that the ‘IPF procedure has immense potential to facilitate research in geography and other disciplines’ (p. 348). Hopefully, the IPF procedure programmed in Visual Basic can make a valuable contribution.

REFERENCES

Ballas D, Clarke G & Turton I (1999) Exploring microsimulation methodologies for the estimation of household attributes, paper presented at the 4th International Conference on GeoComputation, Mary Washington College, Virginia, USA, July 1999

Birkin M (1987) Iterative Proportional Fitting (IPF): Theory, Method and Examples Computer Manual 26, School of Geography, University of Leeds

Birkin M & Clarke G (1995) Using microsimulation methods to synthesise census data, Chapter 12 in Census Users' Handbook editor Stan Openshaw GeoInformation International; Cambridge

Birkin M & Clarke M (1988a) SYNTHESIS - a synthetic spatial information system for urban and regional analysis: methods and explanations *Environment and Planning A*, 1988, volume 20, pp. 1645-1671

Birkin M & Clarke M (1988b) The Generation of Individual and Household Incomes at the Small Area Level Using Synthesis, Working Paper 500, School of Geography, University of Leeds

Bishop Y, Fienberg S & Holland P (1975) Discrete Multivariate Analysis: Theory and Practice MIT Press; Cambridge, Massachusetts

Clarke G P editor (1996a) European Research in Regional Science 6 Microsimulation for Urban and Regional Policy Analysis Pion Ltd; London

Clarke G P (1996b) 'Microsimulation: an Introduction' in European Research in Regional Science 6 Microsimulation for Urban and Regional Policy Analysis editor G P Clarke Pion Ltd; London

Deming W E & Stephan F F (1940) On least square adjustment of sampled frequency tables when the expected marginal totals are known, *Annals of Mathematical Statistics* Vol. 6, pp. 427-444

Fienberg SE (1970) An Iterative Procedure for Estimation in Contingency Tables, *The Annals of Mathematical Statistics,* Vol.41, No. 3, 907-917

Fienberg SE (1977) The Analysis of Cross-Classified Categorical Data, MIT Press; Cambridge, Massachusetts

Fotheringham A S & Wong D W S (1991) The modifiable areal unit problem in multivariate statistical analysis *Environment & Planning A,* 23 pp. 1025-44

Hooimeijer P (1996) 'A life-course approach to urban dynamics: state of the art in and research design for the Netherlands' in European Research in Regional Science 6 Microsimulation for Urban and Regional Policy Analysis editor G P Clarke, Pion Ltd; London

Norman P (1997a) Small Area Population Updates Research Section, City of Bradford Metropolitan District Council; Bradford

Norman P (1997b) An Investigation into the Need for Updated Small Area Population Estimates in Bradford Metropolitan District and the Development of a Model of Estimation, Work Placement Report, Department of Environmental Science, University of Bradford

Norman P (1999) An Investigation into Household Overcrowding in Bradford in the 1990s, MA GIS Dissertation, School of Geography, University of Leeds

Openshaw S (1984) The Modifiable Areal Unit Problem, in Concepts and Techniques in Modern Geography, No. 38 Geo Books, Norwich, England

Openshaw S & Clarke G (1996) 'Developing Spatial Analysis Functions Relevant to GIS Environments', Chapter 2 in Spatial Analytical Perspectives on GIS, edited by Manfred Fischer, Henk J Scholten & David Unwin, Taylor & Francis, London, pp. 21-37

Orcutt G H, Mertz J & Quinke H, editors (1986) Microanalytic Simulation Models to Support Social and Financial Policy, Amsterdam, North-Holland

Rees P (1994) Estimating and Projecting the Populations of Urban Communities *Environment & Planning A*, 1994, Volume 26, pp. 1671-1697

Rees P (1995) 'Putting the census on the researcher's desk', Chapter 2 in Census Users' Handbook editor Stan Openshaw, GeoInformation International; Cambridge

Simpson S (1993) Measuring and Coping with Local Under-Enumeration in the 1991 Census, Research Section, City of Bradford Metropolitan District Council, Bradford

Simpson S (1994) Population Forecasts for Bradford District Technical Report (incomplete), Research Section, City of Bradford Metropolitan District Council, Bradford

Simpson S (1995) User Guide to Program Forecast Research Section, City of Bradford Metropolitan District Council, Bradford

Simpson S (1996) Labour Force Forecasts, 1996 Round Technical Report Research Section, City of Bradford Metropolitan District Council, Bradford

Simpson S (1998) Case Study: apportionment using counts of patients and electors, section III.6 in Making local population estimates: a guide for practitioners, Edited by Stephen Simpson for the Estimating with Confidence project, Local Authorities Research and Intelligence Association, Wokingham

Simpson S & Norman P (1997) 1996 Population Estimates for Wards in Bradford District: Methodology, Research Section, City of Bradford Metropolitan District Council, Bradford

Upton G J G (1985) Modelling cross-tabulated regional data, in Measuring the Unmeasurable, edited by P Nijkamp, H Leitner and N Wrigley, Martinus Nijhoff Publishers, Amsterdam, pp. 197-218

Walton T (1999) Deriving the Ward Housing Completion Data, Transportation and Planning Division, City of Bradford Metropolitan District Council, Bradford (pers. com. 6/7/99)

Williamson P, Clarke G P & McDonald A T (1996) 'Estimating small-area demands for water with the use of microsimulation' in European Research in Regional Science 6 Microsimulation for Urban and Regional Policy Analysis editor G P Clarke, Pion Ltd; London

Wilson T & Rees P (1998) Look-up Tables to Link 1991 Population Statistics to the 1998 Local Government Areas, Working Paper 98-5, School of Geography, University of Leeds, Leeds

Wong D W S (1992) The Reliability of Using the Iterative Proportional Fitting Procedure *Professional Geographer*, 44 (3), 1992, pp. 340-348