



## **Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin**

### **Copyright statement**

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

### **Liability statement**

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

### **Access Agreement**

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Methodological considerations in the development of  
small area deprivation indices

A thesis submitted to the  
University of Dublin, Trinity College  
for the degree of  
Doctor of Philosophy

Conor Teljeur  
Department of Public Health & Primary Care  
August 2007

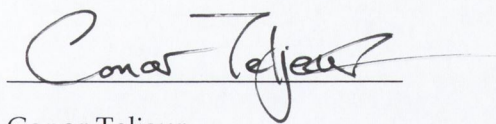
TRINITY COLLEGE  
P 21 NOV 2007  
LIBRARY DUBLIN

THESIS  
8200

# Declaration

I hereby declare that:

- (a) This thesis has not been submitted for an exercise for a degree at this or any other University.
- (b) This thesis is entirely the work of the author except where otherwise stated.
- (c) The Trinity College Library may lend or copy this thesis upon request.

A handwritten signature in black ink, reading "Conor Teljeur", is written over a horizontal line. The signature is cursive and includes a long horizontal flourish extending to the right.

Conor Teljeur

December 2006



## Acknowledgments

I would like foremost to thank my supervisor, Alan Kelly, for his help and guidance in completing this research. Alan has proven a remarkable resource of information and knowledge for which I am enormously grateful.

Funding was obtained from the Health Research Board, without which this study would not have been possible.

I would also like to extend my gratitude to the staff of the Department of Public Health & Primary Care for their encouragement, stimulation and tolerance of my continued presence. In particular I would like to thank Shane Allwright for her patience and suggestions.

A special thanks to Deirdre for her support, patience and confidence in my ability to complete this thesis.



“The man who really merits pity is the man who has been down from the start, and faces poverty with a blank resourceless mind”

George Orwell, 1933

“The poor remain poor. Someone has to work in Woolworth’s.”

Morrissey, 1995





## Summary

Health inequalities exist such that people with lower incomes and poorer social conditions experience poorer health. When individual level characteristics are aggregated to an area level, the socioeconomic status or deprivation of the neighbourhood also correlates with the health status and outcomes of the people who live in that area. The nature of these links between income and health vary across urban and rural areas reflecting the different social dynamics at play across the urban-rural continuum.

The aim of this research was to assess current deprivation index methodology and to propose improvements to the methodology. In addition, the issues surrounding urban-rural variation in deprivation indices were addressed.

To facilitate analysis of urban-rural deprivation differences, a small area classification was required. The previous urban-rural classification was based on a simple dichotomy which ignored the range of settlement and area types. Multiple data sources were used to develop a new multi-level urban-rural classification for small areas in Ireland. This classification provided better distinction between the variety of settlement types than the simple dichotomous classification and enabled a detailed analysis of regional bias.

The key stages of deprivation index development were identified as: indicator selection, shrinkage, data transformation, indicator combination and presentation. For each of these stages a number of methodologies were available and these were analysed with respect to their statistical characteristics and behaviour under different conditions.

Three methods of shrinkage were analysed in detail in order to understand the effects it can have on indicators. The Longford method was shown to be the most appropriate for application to deprivation indicators. Indicator combination methods were also analysed. Using the new multi-level urban-rural classification,

it was shown that regional bias can be increased during indicator combination. To reduce this effect, a new method called geographically weighted principal components analysis (GW-PCA) was developed and applied to Irish data. This method diminished regional bias and enhanced the understanding of regional variation in deprivation indicators.

A detailed sensitivity analysis was used to show the relative importance of choices made at each of the key stages of deprivation index development. It was found that indicator selection had the largest impact on the ranks of small areas while the impact of shrinkage was relatively small. Given the widespread use of deprivation indices in resource allocation and planning, these findings highlight the importance of performing sensitivity analysis to understand the effects of the choices made in deprivation index development.

In conclusion, this study has shown that the process of developing a deprivation index can be greatly improved by careful selection and justification of indicators, analysis of regional bias using a detailed urban-rural classification, the use of GW-PCA for data combination and by the application of sensitivity analysis to determine the impact of choices made.

# Contents

1	Background .....	1
1.1	Factors influencing health.....	1
1.2	Health inequalities.....	13
1.3	Deprivation.....	16
1.4	Urban-rural differences.....	21
1.5	Problems with existing methodology .....	29
1.6	Aims and objectives.....	35
2	Defining the urban-rural divide.....	37
2.1	What defines urban and rural? .....	37
2.2	Urban-rural divide in Ireland.....	39
2.3	Urban-rural measures .....	40
2.4	Combination methods.....	53
2.5	Proposed method of urban-rural classification for Ireland .....	66
2.6	Summary .....	88
3	Indicator selection and transformation .....	89
3.1	Index development.....	90
3.2	Shrinkage.....	92
3.3	Alternative methods of data transformation .....	118
3.4	Summary .....	119
4	Dimension reduction .....	123
4.1	Dimension reduction in existing indices .....	124
4.2	Some methods of dimension reduction .....	129
4.3	Principal Components Analysis .....	134
4.4	Potential sources of error in deprivation indices.....	141
4.5	Geographically weighted PCA .....	155
4.6	Outlier detection and influence functions.....	177
4.7	Summary .....	185
5	Sensitivity analysis.....	187
5.1	Key steps in deprivation index development.....	187
5.2	Assessing the choices.....	190

5.3	Example of a sensitivity analysis .....	192
5.4	Summary .....	206
6	Discussion.....	209
6.1	Urban-rural issues.....	209
6.2	Deprivation indices.....	212
6.3	The health context.....	226
6.4	General remarks .....	228
7	Conclusions & recommendations .....	231
7.1	Conclusions.....	231
7.2	Recommendations .....	234
8	References.....	1

## List of figures

Figure 1.1 Determinants of health status <sup>5</sup> .....	2
Figure 1.2 The field model of health <sup>6</sup> .....	3
Figure 1.3 Pathways between socioeconomic determinants of health <sup>172</sup> .....	16
Figure 1.4 Shaw's model of rural deprivation <sup>244</sup> .....	26
Figure 2.1 Rural-urban continuum .....	38
Figure 2.2 Ranked town populations in Ireland (towns > 5000 persons not shown) .....	42
Figure 2.3 The number of EDs classed as urban by cut-off to describe settlement as urban .....	43
Figure 2.4 Ranked ED population densities .....	46
Figure 2.5 Example of a donut shaped ED enclosing a town.....	48
Figure 2.6 Ranked ED access values .....	50
Figure 2.7 Percentage land use by ED (ranked by proportion built environment) .....	52
Figure 2.8 Scheme for classifying EDs by land use .....	77
Figure 2.9 Map of 8 category classification.....	80
Figure 3.1 Raw versus shrunk proportion unemployed (shrunk to national mean) .....	93
Figure 3.2 Shrinkage of 10,000 simulated datasets using three methods .....	100
Figure 3.3 Implications of a shift in the mean .....	104
Figure 3.4 Mean versus standard deviation for simulations where Longford and empirical Bayes could not be computed .....	106
Figure 3.5 Local means at Local Authority and district level .....	108
Figure 3.6 Spatial autocorrelation: raw vs. shrunk data .....	109
Figure 3.7 Local means from Monte Carlo exercise.....	111
Figure 3.8 Upper and lower bounds for unemployment generated by Monte Carlo for a subset of EDs.....	112
Figure 3.9 Effect on distribuion of natural log and empirical logit transformations .....	120
Figure 4.1 Proportion city and town population versus the percentage variance captured by the first principal component ( $R^2 = 0.72$ ).....	138

Figure 4.2 Proportion city or town population against eigenvector values for the first principal component for the five variables.....	140
Figure 4.3 The numbers of towns located in single ED by deprivation decile .....	143
Figure 4.4 Plots of variables by area type .....	144
Figure 4.5 Moran's I by spatial lag distance.....	147
Figure 4.6 Effect of changing boundaries on homicide rates .....	150
Figure 4.7 Comparison of distance decay functions ( $d_{\max} = 300$ ).....	157
Figure 4.8 Catchment based on a cut-off distance (10km).....	160
Figure 4.9 Catchment based on number of areas (nearest 20 areas) .....	161
Figure 4.10 Catchment based on maximum lags (5 lags) .....	162
Figure 4.11 Comparison of cut-offs.....	163
Figure 4.12 Correlation between unemployment and low social class with increasing distance using six different decay methods for Crookhaven ED ..	166
Figure 4.13 Regional variation in correlations for two pairs of variables .....	169
Figure 4.14 Variance explained by the first principal component.....	170
Figure 4.15 Eigenvector values for proportion unemployed .....	172
Figure 4.16 Eigenvector values for proportion low social class.....	172
Figure 4.17 Eigenvector values for proportion households with no car .....	173
Figure 4.18 Eigenvector values for proportion households living in Local Authority housing.....	173
Figure 4.19 Contribution of each variable to deprivation score (based on means).....	176
Figure 4.20 First principal component mapped for local and mean .....	177
Figure 4.21 Mahalanobis distance by area type .....	179
Figure 4.22 Empirical and theoretical influence function values .....	183
Figure 4.23 Mahalanobis distance by empirical influence.....	184
Figure 4.24 Empirical influence by area type .....	184
Figure 5.1 Flow chart of deprivation index calculation .....	187
Figure 5.2 Plots of scenario 1 ranks against ranks for scenarios 2 to 4 .....	193
Figure 5.3 Median and inter-quartile range of ranks for each ED.....	197
Figure 5.4 Plot of chi square against the percentage variance explained by the first principal component.....	204

Figure 5.5 Plot of two 1st principal components for 'optimal' results..... 205



Introduction	1
Chapter I	15
Chapter II	35
Chapter III	55
Chapter IV	75
Chapter V	95
Chapter VI	115
Chapter VII	135
Chapter VIII	155
Chapter IX	175
Chapter X	195
Chapter XI	215
Chapter XII	235
Chapter XIII	255
Chapter XIV	275
Chapter XV	295
Chapter XVI	315
Chapter XVII	335
Chapter XVIII	355
Chapter XIX	375
Chapter XX	395
Chapter XXI	415
Chapter XXII	435
Chapter XXIII	455
Chapter XXIV	475
Chapter XXV	495
Chapter XXVI	515
Chapter XXVII	535
Chapter XXVIII	555
Chapter XXIX	575
Chapter XXX	595
Chapter XXXI	615
Chapter XXXII	635
Chapter XXXIII	655
Chapter XXXIV	675
Chapter XXXV	695
Chapter XXXVI	715
Chapter XXXVII	735
Chapter XXXVIII	755
Chapter XXXIX	775
Chapter XL	795
Chapter XLI	815
Chapter XLII	835
Chapter XLIII	855
Chapter XLIV	875
Chapter XLV	895
Chapter XLVI	915
Chapter XLVII	935
Chapter XLVIII	955
Chapter XLIX	975
Chapter L	995

## List of tables

Table 1.1 Contribution to European mortality of various risk factors.....	4
Table 2.1 Minimum population sizes used by countries to define urban areas.....	41
Table 2.2 Criteria for MCC classification .....	55
Table 2.3 Count of EDs by class for each predictive method of classification.....	57
Table 2.4 Comparison of ED counts for classes and K-means clusters.....	58
Table 2.5 Comparison of ED counts for classes and hierarchical clusters.....	59
Table 2.6 Measures of fit for each classification method .....	63
Table 2.7 Rank and position of measures of fit against simulations .....	65
Table 2.8 Comparison of measures of fit for different numbers of classes identified using K-means clustering.....	66
Table 2.9 Calculation of median settlement size for Carrigtohill ED.....	68
Table 2.10 Comparison of measures of fit for three clustering methods.....	75
Table 2.11 Counts and minimum access scores for near and remote village and rural EDs.....	76
Table 2.12 Count of EDs in each land use category .....	77
Table 2.13 Class and sub-category structure (ED counts in brackets) .....	78
Table 2.14 Class and sub-category structure (percentage of total population in brackets).....	79
Table 2.15 Percentage population in each class and sub-class for county Carlow..	81
Table 2.16 The number of administrative counties in each of 6 classes.....	82
Table 2.17 Comparison of measures of fit for three clustering methods.....	84
Table 2.18 Counts and minimum access scores for near and remote village and rural EDs.....	84
Table 2.19 Count by year of EDs in each of 6 classes .....	85
Table 2.20 Population in each class as a percentage of the total population (population figures in brackets) .....	86
Table 2.21 Comparison of 2002 urban-rural classification based on 3,382 and 3,422 EDs.....	87
Table 2.22 Comparison with CSO urban-rural classification.....	87
Table 3.1 Indicators used in a range of deprivation indices.....	91

Table 3.2 Comparison of weights for two EDs with the same numerator .....	101
Table 3.3 Comparison of weights for four EDs with the same denominator .....	101
Table 3.4 Logit and inverse logit transformations for a range of numerator and denominator values .....	102
Table 3.5 Proportions that give approximately zero shrinkage for the unemployment data .....	103
Table 3.6 Moran's I for a selection of variables using different shrinkage techniques .....	113
Table 3.7 Comparison of correlations with baseline for different shrinkage methods using ranks and index values.....	115
Table 3.8 Comparison of indicator values using different shrinkage methods for Loughill ED .....	116
Table 3.9 Difference from baseline using different methods of shrinkage.....	117
Table 4.1 Numbers of indicators used in some indices of deprivation.....	124
Table 4.2 Examples of domains in deprivation indices.....	128
Table 4.3 Eigenvector values using subsets of EDs .....	148
Table 4.4 Percentage EDs in the least and most deprived deciles using PCA based on different subsets of EDs .....	148
Table 4.5 Change in selected indicator means from 1986-2002.....	153
Table 4.6 Comparison of results for different cut-off definitions .....	164
Table 4.7 Mean spatial autocorrelation by area type for a range of variables .....	165
Table 4.8 Count of EDs for which the first principal component explains 67.8% or more variance (percentage) by urban-rural class .....	171
Table 4.9 EDs classed as outliers by area type .....	179
Table 4.10 Eigenvector values for the first principal component using different robust estimates of the correlation matrix .....	180
Table 5.1 Percentage EDs in the least and most deprived deciles by scenario and area type .....	194
Table 5.2 Correlation matrix for the ten deprivation variables.....	196
Table 5.3 Example versus most probable decile .....	199

Table 5.4 EDs with most extreme differences between example and most probable decile ..... 201

Table 5.5 Variables, shrinkage and transformations applied for PCA result with largest eigenvalue..... 203

Table 5.6 Variables, shrinkage and transformations applied for optimal PCA result ..... 204



# 1 Background

A large body of evidence supports the view that both income and social inequalities give direct rise to health inequalities such that people with lower incomes and poorer social conditions experience increased morbidity and mortality.<sup>1</sup> Identifying and addressing the causes and consequences of these inequalities has become a key issue in public health research. Area level measures of poverty and deprivation have come to the forefront in identifying areas that require increased resources and attention in an attempt to diminish health inequalities. To have an understanding of the causes of these inequalities, it is important to recognize the factors that influence health and how these factors may vary with socioeconomic status. It is also important to be aware of area based measures of deprivation and how they may differ from individual level measures of deprivation.

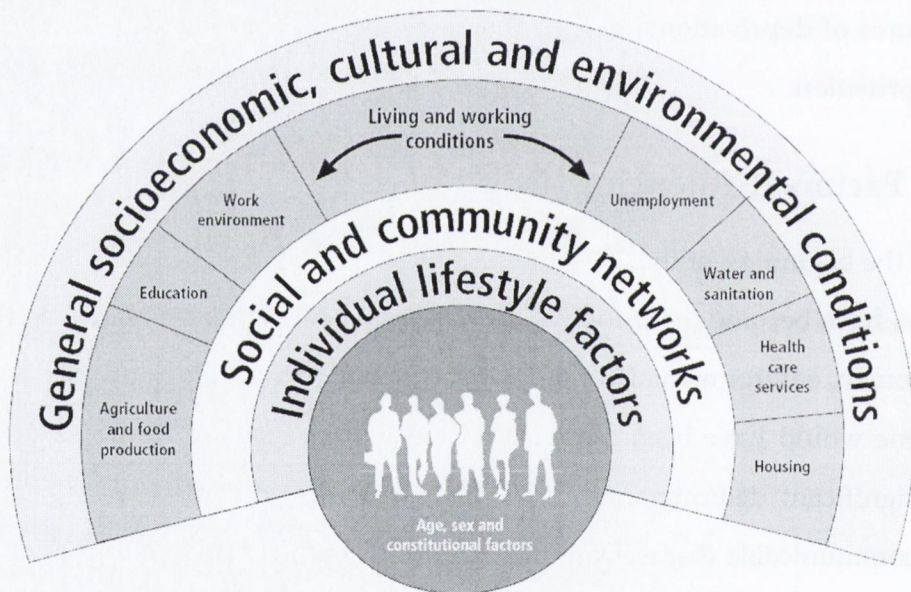
## 1.1 Factors influencing health

Since the beginning of the 20<sup>th</sup> century, unfavourable social conditions and lifestyle factors have become the principal determinants of health in the developed world.<sup>2</sup> Previously, environmental conditions such as poor housing standards and public hygiene would have been more significant determinants of health, and these are still significant determinants in the developing world.<sup>2</sup> In the developed world, non-communicable diseases are the main contributor to premature mortality.<sup>3</sup> The manner in which social factors affect health status are very complex, making the development of policies to address these issues quite difficult.<sup>4</sup>

A model of the main determinants of health was developed by Dahlgren and Whitehead<sup>5</sup> and is shown in Figure 1.1 below. At the centre are the most direct and unmodifiable factors linked to the individual – age, sex and constitutional or genetic factors. The next level of determinants includes the individual lifestyle factors such as smoking, physical activity and diet. The social and community networks constitute the next level of determinants: the linkages between

individuals that, when present, can provide support and access to resources and, when absent, lead to isolation and an inability to cope. The next level of determinants encompasses the living and working conditions experienced by an individual. These conditions include the workplace, local environment and access to important resources, such as clean water, health care and education, and access to amenities, such as green space. Beyond those factors, the wider socioeconomic climate impacts on the health of the individual, although the impact of these factors can have a more equal distribution across the population than the other factors mentioned.

Figure 1.1 Determinants of health status<sup>5</sup>

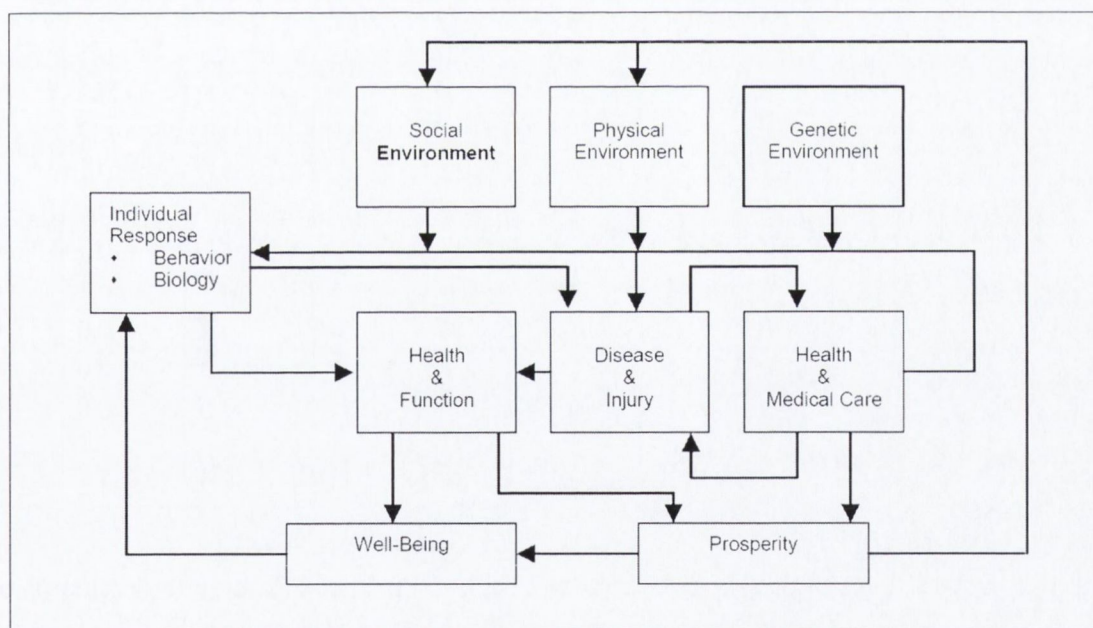


The determinants of health include factors relating to the individual and those relating to the community or social structure that the individual lives in.

The interactions between the various determinants of health are complex. Figure 1.2 shows a model of interactions between determinants as developed by the San Francisco Department of Public Health.<sup>6</sup> Some of the determinants affect themselves through a cycle of influences. For example, health affects well-being

which, in turn, affects individual behaviour which, in turn, affects health. These circular relationships suggest the role that positive and negative feedback can have on the health of an individual. Poor health can, by a complex sequence of interactions, lead to further poor health. It should be noted that the only determinant unaffected by other determinants after birth is the genetic environment. As with the model in Figure 1.1, where age, sex and constitutional factors are at the centre, genetic conditions are unalterable. While gender is, in terms of how it affects health, unmodifiable, like age it also affects health in relation to how societal structures may impact on health. For example, women may be expected to remain at home to raise children exposing them to different stresses not applicable to women who are at work.<sup>7-9</sup>

Figure 1.2 The field model of health<sup>6</sup>



A comprehensive study by Ezzati et al. sought to determine the burden of disease due to a range of selected risk factors by region.<sup>10</sup> A European region including Ireland was defined including countries with very low child and adult mortality.<sup>†</sup>

<sup>†</sup> The group of European countries with very low child mortality and very low adult mortality were: Andorra, Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Monaco,



Table 1.1 gives the number and percentage of deaths attributable to a range of major risk factors. It is important to note the prominence of lifestyle factors such as smoking, alcohol consumption, physical inactivity and diet (in the form of both high cholesterol and low fruit and vegetable intake). It is also interesting to note how alcohol has a net negative effect on male mortality and a net positive effect on female mortality.

Table 1.1 Contribution to European mortality of various risk factors

Risk factor	Count (1,000s of deaths)			Percentage of all deaths		
	Male	Female	Total	Male	Female	Total
High blood pressure	325	354	679	16.1	17.2	16.7
Tobacco	531	145	676	26.3	7.1	16.6
High cholesterol	265	282	547	13.1	13.7	13.4
High BMI	183	197	380	9.1	9.6	9.3
Physical inactivity	103	103	206	5.1	5.0	5.1
Low fruit & vegetable intake	95	75	170	4.7	3.7	4.2
Urban outdoor air pollution	12	11	23	0.6	0.5	0.6
Airborne particulates	17	2	19	0.8	0.1	0.5
Illicit drugs	11	6	17	0.5	0.3	0.4
Carcinogens	12	2	14	0.6	0.1	0.3
Unsafe sex	3	9	12	0.1	0.4	0.3
Lead	4	2	6	0.2	0.1	0.1
Iron deficiency	2	3	5	0.1	0.1	0.1
Risk factors for injuries	4	0	4	0.2	0.0	0.1
Childhood sexual abuse	1	1	2	0.0	0.0	0.0
Unsafe water, sanitation, and hygiene	0	1	1	0.0	0.0	0.0
Alcohol	65	-85	-20	3.2	-4.1	-0.5

The World Health Organization (WHO) defined a number of social determinants of health<sup>11</sup> which will be dealt with in more details in the following sections. The non-modifiable determinants, such as genetic predisposition, will not be addressed. Social gradients will be discussed in section 1.2 subsequently, as they are more relevant to inequalities in health.

---

Netherlands, Norway, Portugal, San Marino, Slovenia, Spain, Sweden, Switzerland and the United Kingdom.

### **1.1.1 Childhood**

Circumstances prior to birth and during childhood can have a significant impact on health during adulthood and later life. To a large extent, the circumstances in childhood that lead to poor adult health are socioeconomic in nature and reflect many elements of deprivation.<sup>12-14</sup> Poor socioeconomic conditions in childhood may be propagated into adulthood, e.g. Davey Smith et al. found that after controlling for adult socioeconomic status, low social class in childhood increased risk for mortality from stroke and stomach cancer.<sup>15</sup> Low birth weight is associated with increased risk of poor educational attainment,<sup>16</sup> high blood pressure in young adults,<sup>17</sup> hypertension and cardiovascular disease in adulthood,<sup>18</sup> and is a predisposing factor for metabolic abnormalities (such as atherosclerosis, renal disease and non-insulin diabetes mellitus), asthma, low IQ, obesity and psychological distress.<sup>19</sup> Globally, child and maternal underweight has been estimated to explain 9.5% of disability adjusted life years, making it the single most important cause of the global burden of disease, although within developed European nations it is not a significant factor.<sup>10</sup> Such findings suggest that childhood factors may influence later health independent of later socioeconomic status.

### **1.1.2 Stress**

Stress can occur at both the individual level and at an area level. For an individual, stress may be induced by pressures at work or in the living environment and are a function of the individual's interactions with others. At an area level, stress may exist due to crowding (i.e. high population density), conditions of threat, social disorganisation and a lack of access to opportunities.<sup>20 21</sup> Stress in either form impacts negatively on both mental and physical well-being although access to good social support and resources can reduce this impact.<sup>22</sup> A study by Elliott suggested that the protective effects of social support only operate in areas of higher socio-economic status.<sup>23</sup> The primary effects of stress are to increase the risk of mental health problems, such as depression, and cardiovascular morbidity and mortality.<sup>24-26</sup>

### 1.1.3 Social support and social exclusion

While social support can offer health benefits, social exclusion can give rise to health problems. Social support is often measured by social capital – itself a measure of the level of social organisation, mutual aid, interpersonal relations and trust within a community or area.<sup>27</sup> Social capital is said to affect health in three ways: the direct beneficial effects on individual attributes and activities; its effects on the wider social, economic, political and environmental factors; interaction with other determinants of health at the individual or group level.<sup>28</sup> The presence of strong social capital leads to greater well-being of individuals, better group cohesion and support for individuals and greater opportunities and economic success. Although social capital may be seen as a group attribute, the benefits can be observed at the individual level.<sup>29</sup> A study found that self-rated health was better and obesity lower in suburbs with greater social capital.<sup>28</sup> Work by Skrabski et al. found that social capital measures were significantly associated with middle-aged mortality in Hungary.<sup>30</sup> Measurement of social capital is generally based on the use of proxies that are frequently also proxies for deprivation, making it difficult to separate socio-economic influence from social capital influence.<sup>31</sup> Furthermore, social capital may tend to be better in areas of higher socio-economic status.<sup>32</sup>

Social exclusion can be defined as the process whereby an individual or group is excluded from participation in society and is a multi-dimensional concept that involves aspects of deprivation and disadvantage.<sup>33 34</sup> Schönfelder and Axhausen describe social exclusion as “a regular physical and social exclusion from the resources of a dignified life: an active labour market, good quality health care and consumption opportunities, and, finally, integration in the wider networks of civic life”.<sup>35</sup> Social exclusion can diminish an individual’s ability to cope with hardship and be detrimental to health.<sup>36 37</sup> It is also suggested that socially excluded individuals may adopt potentially health-damaging behaviour in the absence of social roles.<sup>38 39</sup>

### **1.1.4 Work**

Work related stress and occupational hazards can give rise to various health problems. Job insecurity can lead to health problems in the form of depression and stress.<sup>40 41</sup> Broom et al. found that some poor quality jobs could be as bad in terms of stress and health as being unemployed.<sup>42</sup> A British study of worker health found that occupations with low autonomy, rewards and security are associated with greater declines in health with age.<sup>43</sup> Aggression and bullying in the workplace can lead to depression and mental health problems.<sup>44-46</sup> Roberts and Lee showed that different occupations had different prevalence of depression, alcohol abuse and drug abuse.<sup>47</sup>

Different occupations also have differing risks of accident in the workplace. Rates of worker fatalities are published by the Health and Safety Authority showing the highest rates in the agriculture, hunting and forestry industry.<sup>48</sup> An Italian study of repeat accidents by occupation type found substantial variation in the accident rates across occupations.<sup>49</sup>

### **1.1.5 Unemployment**

The relationship between unemployment and health is complex and affected by economic context.<sup>50</sup> Unemployment causes increased financial strain and damage to an individual's sense of self.<sup>51</sup> The former is believed to primarily affect individuals of a lower socio-economic status while the latter is more applicable to those of a higher socio-economic status. Bartley<sup>52</sup> pointed to four mechanisms that needed to be considered in the relationship between unemployment and health: the role of relative poverty; social isolation and loss of self esteem; health-related behaviour; and the effect that a spell of unemployment has on subsequent employment patterns. Despite the above examples, it can be difficult to directly link unemployment to ill-health due to the many confounding factors associated with unemployed persons (e.g. increased smoking rates).<sup>53</sup> An assessment of research into unemployment and health by Mathers and Schofield concludes that there is consistent evidence linking unemployment to adverse health outcomes.<sup>54</sup>

Despite the difficulties in linking unemployment to ill-health, there is evidence to suggest that young adulthood unemployment can lead to health problems in later adulthood.<sup>55</sup> Individuals who experienced unemployment between the ages of 16 and 21 showed increased smoking and psychological problems. Young adulthood unemployment was also shown to be associated with decreased health capital, measured by body mass index, physical exercise, good diet and not smoking.<sup>56</sup>

However, the association between unemployment and health is affected by the wider social and economic context. In times of recession, for example, the relationship between employment status and health may change or be masked by other factors influencing behaviour.<sup>57 58</sup> Ruhm showed that during an economic downturn, rates of smoking and obesity decreased while rates of physical exercise increased.<sup>59</sup> Work by Bellaby and Bellaby found that increasing rates of unemployment impact on job stress while high levels of unemployment influence premature death and self assessed health.<sup>60</sup> A study by Williams showed that an unemployed individual was 'better off' in a poorer area than a more affluent area due to factors such as cost of living.<sup>61</sup> Not only is unemployment important in influencing health, but also the local community, job market and economy in general.

### **1.1.6 Education**

Individuals with better education tend to be healthier although uncertainty exists as to whether or not the link is causal.<sup>62</sup> It is possible that a better education leads to better employment, income and general circumstances compared to someone with a poor education, who is more likely to work in an unskilled manual occupation.<sup>63</sup> In some studies of mortality differentials by education level, education is treated as a socioeconomic marker rather than an independent factor.<sup>64</sup> A Dutch study concluded that material factors contribute more to educational differences in incidence of acute myocardial infarction (AMI) than behavioural factors.<sup>65</sup> The differences were most pronounced for individuals with only a primary education who tended to live in worse material circumstances. This suggests that what is

observed is a socioeconomic difference rather than an educational difference. A Danish study, however, found that both educational level and income independently affect mortality after AMI suggesting that education may have an influence on health irrespective of socioeconomic status.<sup>66</sup> A study of ischaemic heart disease in France and Northern Ireland showed a significantly lower prevalence among individuals with a higher education after controlling for smoking.<sup>67</sup> Individuals with lower levels of education were also shown to have higher rates of smoking and alcohol consumption. A European study observed large variations between 11 European countries in the effect of education on self-reported morbidity, although there was a universal trend for higher rates of less than good health and chronic illness among people with lower levels of education.<sup>68</sup> Studies in Europe and the US have also shown similar reductions of 6 to 8% in mortality for 35 to 54 years olds for each one year increase in education.<sup>69 70</sup> In other words, each additional year spent in education results in a reduction in mortality. Research by van Oort et al. noted that lifestyle risk factors such as smoking, excessive alcohol consumption and physical activity have a higher prevalence among people with a lower level of education, increasing the risk of poor health.<sup>71</sup> So while the reasons for the association between lower education and poorer health may not be clear, there is nevertheless a strong association that impacts on health.

### **1.1.7 Housing and living environment**

The indoor and outdoor living environments can have many influences on health. The indoor environment, or home, is where people tend to spend much of their time. Home ownership can confer considerable protection against poor general and mental health status through control over home environment as well as typically being associated with better housing conditions.<sup>72 73</sup> Housing tenure refers to ownership status – whether an individual owns or rents the home they live in. It has been shown that housing tenure in the form of rented accommodation is predictive of poor health which is linked to both the socioeconomic status of the individual and to the health hazards commonly found in and around rented

accommodation.<sup>74</sup> Such hazards are in the form of poor quality housing which may be lacking heating and subject to damp, and in the form of area characteristics such as crime and poor access to amenities. A study by Macintyre et al. found that virtually all of the adverse health effects associated with rented accommodation could be explained by factors such as housing problems, lack of access to garden, overcrowding, area poverty and lack of area amenities.<sup>75</sup> An estimated 1,500 to 2,000 excess deaths occur in Ireland every winter which is largely attributed to poor quality housing with no proper heating.<sup>76</sup> Excess winter mortality is generally associated with a lack of central heating combined with lowered external temperatures.<sup>77 78</sup> Indoor pollution, often from an exterior source although also through smoking and pets, is also a factor contributing to increased prevalence of asthma and other respiratory diseases.<sup>79-81</sup> Overcrowding leads to health problems in the form of mental disorders,<sup>82 83</sup> particularly amongst children, and the spread of infectious diseases such as meningitis,<sup>72</sup> scabies<sup>84</sup> and tuberculosis.<sup>85</sup>

The wider area within which an individual lives also impacts on health in a number of ways. Local amenities such as green space for recreation can have significant positive impacts on the health of the individual.<sup>86-88</sup> Predominantly disadvantaged neighbourhoods have higher rates of mental disorders,<sup>21 89 90</sup> crime,<sup>91</sup> drug dealing,<sup>92</sup> high risk behaviour,<sup>93</sup> early school leaving<sup>94</sup> and general ill health<sup>95-98</sup> – all of which directly or indirectly impinge on well-being and health. While many confounders exist, and the area level effects may be much smaller than the socio-economic status and behaviour of the individual, the neighbourhood can have an impact on health.<sup>99</sup>

### **1.1.8 Transport**

Transport can affect the health of the population in a number of ways: through increased pollution; road traffic accidents; sedentary lifestyle and social exclusion. As the latter is addressed separately in section 1.1.3, it will not be dealt with here. Traffic introduces particulate pollution into the atmosphere that adversely impacts on the health of individuals in the form of increased respiratory problems.<sup>100-102</sup> The

Vesta project, for example, found a significant association between childhood asthma and exposure to traffic exhausts.<sup>103</sup> Road traffic accidents (RTAs) also contribute to the negative impact of traffic. In the year 2000, a total of 12,458 hospitalisations occurred in Ireland due to RTAs, including 407 fatalities.<sup>104 105</sup> Elvik estimated the cost of road accidents to the economy for twelve countries and found that, on average, RTAs cost about 2.5% of the gross national product.<sup>106</sup> A study by Künzli et al. found that air pollution caused 6% of total mortality, with approximately half of that figure being attributable to air pollution due to motorised traffic.<sup>107</sup> Traffic pollution also contributed substantially to cases of chronic bronchitis, asthma attacks and person-days of restricted activities. Furthermore, the use of the car in preference to other modes of transport such as walking and cycling leads to a more sedentary lifestyle with increased risk of obesity and the consequent health risks such as diabetes and heart disease.<sup>108</sup> In essence, road traffic has a significant impact on population health.

### **1.1.9 Addiction**

Dependence on and abuse of substances such as tobacco, alcohol and illicit drugs leads to social problems and has adverse health effects. It is estimated that 26% of male deaths and 9% of female deaths in developed countries can be attributed to smoking – the single most important risk factor.<sup>109</sup> Peto et al. estimate that in the year 2000 in Ireland, 20.4% of male deaths and 15.9% of female deaths could be attributed to smoking.<sup>110</sup> This represents a serious burden of disease.

It is estimated that 4% of the global burden of disease is attributable to alcohol through contributions to certain cancers, neuro-psychiatric disorders, cardiovascular disorders, cirrhosis of the liver, and unintentional and intentional injuries.<sup>111</sup> Moderate alcohol consumption can, however, offer a protective effect against coronary disease and respiratory deaths and may even lead to a net reduction in mortality although the benefits primarily occur in the older population.<sup>112 113</sup> For younger members of the population, alcohol consumption is generally associated with poorer health outcomes.<sup>114</sup>



The 2001 hospital in-patient statistics for Ireland show that there were 2,326 cases of 'alcohol/drug use and alcohol/drug induced organic mental disorders', representing 0.43% of in-patient events.<sup>104</sup> Accidental and deliberate overdose, diseases contracted through sharing of needles and psychiatric disorders are just some of the health problems associated with problem drug use.<sup>115-119</sup> Although the burden is very small and is much lower than that for smoking, illicit drug use contributes significantly to health care utilisation indicating its contribution to health problems in general.<sup>120-122</sup> Furthermore, environmental chemical exposures, such as drugs, alcohol, and tobacco, contribute to neuro-developmental disabilities and disorders.<sup>123</sup>

### **1.1.10 Nutrition**

Bad diet can give rise to numerous problems at different stages of the life cycle: high blood pressure, poor dental health and a predisposition to infection in childhood; higher rates of dental caries and a predisposition to anaemia in adolescents; coronary heart disease, atherogenesis, stroke, peripheral vascular disease, cardiovascular disease, thrombosis and high blood pressure in adults; and osteoporosis, poor vision and weakened immune system in the elderly.<sup>124</sup> A poor diet can also lead to obesity which brings with it an increased risk of chronic health conditions such as high blood pressure, type 2 diabetes, high blood cholesterol, coronary heart disease and gallbladder disease.<sup>125-127</sup> Studies of the relative risk of excess mortality show that both underweight and obese individuals are at increased risk of excess mortality.<sup>128 129</sup> For obese individuals there is an increased risk of death from cardiovascular diseases, diabetes and digestive diseases in men.<sup>130</sup> It has also been shown that obesity is associated with increased odds of mood, anxiety and substance use disorders.<sup>131</sup>

### **1.1.11 Physical inactivity**

Evidence has been gathering since the 1950's to identify physical inactivity as a risk factor for cardiovascular disease and all-cause mortality.<sup>132</sup> The WHO defines diet

and physical exercise as two of the main factors for non-communicable disease and notes how “physical activity reduces blood pressure, improves the level of high density lipoprotein cholesterol, improves control of blood glucose in overweight people, even without significant weight loss, and reduces the risk for colon cancer and breast cancer among women.”<sup>133</sup> A study of exercise habits among civil servants in London found inverse associations between leisure time physical activity and mortality from all-causes, coronary heart disease, cardiovascular disease, all cancers, lung cancer, colorectal cancer and haematopoietic cancer.<sup>134</sup> Small improvements in physical health were associated with significantly lowered mortality risk amongst healthy middle-aged men.<sup>135</sup> Poor physical fitness in young adults has been shown to be associated with the development of increased cardiovascular risk factors such as hypertension, diabetes, metabolic syndrome and hypercholesterolemia in middle age.<sup>136</sup>

## **1.2 Health inequalities**

The social and environmental determinants discussed in the previous section do not affect all individuals equally: there exists a social gradient whereby poor social and economic circumstances adversely affect health throughout life.<sup>11</sup> Individuals of a lower socioeconomic status tend to have poorer health than those of a higher socioeconomic status. An investigation into inequalities in health Ireland in 2001 confirmed the presence of socioeconomic differences in health across a range of health measures including all cause and cause specific mortality, perinatal mortality, low birth weight, psychiatric admissions, depressive disorders, alcoholic disorders and treatment for drug misuse.<sup>137 138</sup>

Of the eleven broad determinants of health discussed in the previous section, socioeconomic differences are implicit in a number of them: work; unemployment; education; housing and living environment; and social support and social exclusion. Work and unemployment are indirect measures of income and direct measures of occupation, while education is itself a socioeconomic measure. Housing tenure and the prosperity of the area in which one lives are also markers

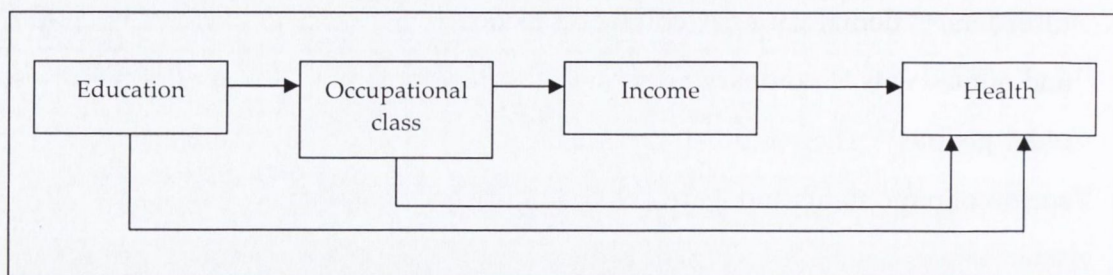
of socioeconomic position. Being able to own a house indicates a degree of wealth. Social support and social exclusion highlight social rather than economic inequalities, although these also contribute to socioeconomic differentials. In all five groups of determinants, what might be considered as 'poor' conditions are linked to poorer health and health outcomes, as outlined in the previous section. Of the remaining six determinants, each can be shown to display a socioeconomic gradient. Low birth weight, with its immediate and delayed consequences for health, is correlated with lower socioeconomic status as measured using the Townsend deprivation index.<sup>139</sup> Low socioeconomic status in childhood is also predictive of high blood pressure in later life.<sup>140</sup> Increased stress associated with skilled positions is a rare instance where health is adversely affected by a higher socioeconomic position. However, the risk for cardiovascular mortality is high when work demand and effort is high, but rewards and job control are low.<sup>25</sup> Thus the increased risk to skilled workers may be moderated by increased job control and pay. Area level stress due to fear of crime and violence is associated with disadvantaged neighbourhoods.<sup>97</sup> Transport has a greater effect on lower socioeconomic groups in terms of both pollution<sup>141</sup> and pedestrian accidents.<sup>142</sup> Illicit drug availability and usage is greater in disadvantaged neighbourhoods.<sup>92 143</sup> Physical inactivity is associated with low socioeconomic status at both an individual and an area level, the latter possibly being related to the availability of amenities for recreation.<sup>144</sup> People in lower socioeconomic groups consume more energy dense foods and fewer fruits, vegetables and high fibre foods.<sup>124</sup>

Socioeconomic differences in the prevalence of high blood pressure,<sup>140 145 146</sup> tobacco use,<sup>109 147 148</sup> high cholesterol,<sup>67 149 150</sup> high body mass index,<sup>124 151 152</sup> physical activity,<sup>97 153 154</sup> low fruit and vegetable intake,<sup>155-157</sup> and alcohol consumption<sup>47 158 159</sup> have been shown, with higher incidence amongst those in lower socioeconomic groups. These represent the seven most significant risk factors for mortality in Table 1.1 and account for an estimated 64.8% of deaths in Western Europe. Inequalities by socioeconomic status extend to a vast range of health measures, access to treatment and outcomes at all stages of life. Some examples are: diabetes in women,<sup>160</sup> stroke

incidence,<sup>161</sup> cancer survival,<sup>162</sup> <sup>163</sup> coronary heart disease,<sup>164</sup> common mental disorders,<sup>165</sup> dental caries,<sup>166</sup> childhood asthma,<sup>167</sup> road traffic injuries in children and adolescents,<sup>168</sup> coronary artery bypass graft survival,<sup>169</sup> and mortality amongst older people.<sup>170</sup> There is, in all cases, a clear and strong relationship between lower socioeconomic status and increased morbidity and mortality.

Mackenbach et al. compared socioeconomic inequalities in morbidity and mortality in eleven European countries.<sup>171</sup> Three measures of socioeconomic status were used: education, occupation and income. All three measures were found to give broadly similar results but it was also found that each measure was distinct and may capture a different element of socioeconomic status. Figure 1.3 shows a schematic of a conceptual model proposed by Lahelma et al. that marks the assumed pathways between three key socioeconomic indicators and health.<sup>172</sup> Each of the three indicators has a direct impact on health but there is also a hierarchy whereby education contributes to occupation which, in turn, contributes to income. Lahelma et al. argue that the three indicators are all independently and inter-dependently important when assessing inequalities in health with respect to socioeconomic differences. As income is frequently not measured at an individual or even small area level, the choice of indicator of socioeconomic status is frequently some measure of occupation which separates skilled from manual workers. The principal difficulty with occupation is that the groupings are quite broad. Furthermore, it was found in Ireland that the number of individuals labelled as "unknown" socioeconomic group was increasing and appeared to represent a group with worse health than those in the "unskilled manual" grouping.<sup>137</sup> While education data are available, education is less frequently used as a measure socioeconomic status.

Figure 1.3 Pathways between socioeconomic determinants of health<sup>172</sup>



The differing contribution of different indicators of socioeconomic status to health suggests the need for a multi-dimensional approach to measuring socioeconomic status. This is one of the motivations for the development of deprivation indices, which will be discussed in the following section.

### 1.3 Deprivation

Measures of socioeconomic status provide a method of grouping individuals of similar social and economic standing. However, a low socioeconomic status is not necessarily indicative of poverty – it is assumed that an unskilled manual labourer has a low income but they might be above the poverty line. Peter Townsend defined poverty in the following manner: “Individuals, families and groups in the population can be said to be in poverty when they lack the resources to obtain the types of diet, participate in the activities and have the living conditions and amenities which are customary, or are at least widely encouraged or approved, in the societies to which they belong. Their resources are so seriously below those commanded by the average individual or family that they are, in effect, excluded from ordinary living patterns, customs and activities.”<sup>173</sup>

In Townsend’s definition, people are labelled according to what resources they do not possess, rather than those they do possess. The resources need not only be income, education or good social support – they might include household goods such as a washing machine or television, clothes, and access to a social life. This definition also places poverty in terms of what resources and amenities the average individual expects to have access to, making poverty a relative measure. The difficulty with such a definition is that an individual may choose not to have an

item that the average person possesses. This definition of poverty is, in effect, a definition of deprivation – a state of being deprived of that which one should have access to according to the norms of society.

To measure poverty or deprivation in this manner requires both a list of the resources that people should have access to and a count of how many people lack each of the respective resources. Some cumulative score can then be generated which makes it possible to label an individual as deprived or not deprived. An example of such a deprivation measure would be that of Maitre et al. using the EU Statistics on Income and Living Conditions (EU-SILC) data.<sup>174</sup> Such individual level measures are difficult to generate as they require individual level data of the sort that is not routinely collected. Census data, for example, might identify households with no central heating and also households with no indoor toilet. However, as the data are provided in aggregate format is not possible to identify households lacking both central heating and an indoor toilet. As a result, composite measures using census data were developed to identify areas with high proportions for a number of deprivation indicators.

Early attempts at a composite index were made in 1972 by Craig and Driver in an attempt to identify small areas of adverse social conditions.<sup>175</sup> They chose a number of census variables that were seen as potential indicators of adverse social conditions – variables such as proportion of people with low social class, proportion of population under 15 and households with more than 1.5 persons per room. They highlight some of the possible methods of combining the indicators and present two indices using arbitrarily chosen weighting schemes. Subsequent work by Jarman in the UK resulted in the underprivileged areas score which identified small areas that are likely to have an increased primary care workload.<sup>176</sup> The weights for combining variables were derived by scores given to different variables by general practitioners responding to a questionnaire on social and service factors that contribute to increased workloads. There followed small area deprivation indices for the Northern region of England by Townsend<sup>177</sup> and for

Scotland by Carstairs and Morris.<sup>178</sup> In both cases, census proxies for deprivation were identified and combined to produce a continuous score that could then be presented in deciles. Thunhurst presents an overview of indices by Jarman, Scott-Samuel and Townsend et al. before presenting his own method of combining census and specific survey data to identify areas of poverty in Sheffield, again with a view to identifying small areas with increased need for primary care services.<sup>179</sup> Since then there have been numerous deprivation indices produced along similar lines in the UK,<sup>180</sup> Ireland<sup>181-185</sup> and elsewhere.<sup>186-189</sup>

More recently, there has been a move from a single index to domains of deprivation in England, Scotland, Wales and Northern Ireland.<sup>190-193</sup> These indices of multiple deprivation are intended to group variables that represent specific forms of deprivation such as housing, employment and physical environment. This makes it possible to analyse the relationships between ill-health and different aspects of deprivation.

### **1.3.1 Deprivation and health**

Deprivation measures are typically area based rather than individual based. It cannot be assumed that all individuals in an area experience the level of deprivation found for that area as a whole. It is a summary statistic and, depending on the homogeneity of the population in that area, it may be a misleading label for many of the people living in that area. However, it has been shown that when deprivation indices are calculated for reasonably small areas, the deprivation score for an area is predictive of deprivation for individuals living in that area.<sup>194</sup> In that case, reasonably small referred to UK enumeration districts (average population 450) rather than wards (average population 5,500). With increasing aggregation and hence increasing heterogeneity, the area label has decreasing likelihood of being representative of individuals in that area.

Like single variable measures of socioeconomic status, such as occupation or income, composite deprivation measures also show strong correlations with a

range of morbidity, mortality and health outcome measures. This is not surprising as deprivation indices typically contain one or more measures of socioeconomic status. However, associations that apply at an individual level do not necessarily apply at an area level and vice versa. In terms of the main risk factors associated with mortality, increased area deprivation has been shown to be related to high blood pressure,<sup>195</sup> higher smoking rates,<sup>196</sup> higher cholesterol,<sup>197</sup> high BMI,<sup>198</sup> higher physical inactivity,<sup>154</sup> lower fruit and vegetable intake<sup>199</sup> and increased alcohol consumption.<sup>114</sup>

Some examples of types of morbidity and mortality that have been shown to be associated with area deprivation are: depression;<sup>200</sup> angina;<sup>201</sup> irritable bowel syndrome;<sup>201</sup> cancer survival;<sup>202-204</sup> excess diabetes mortality;<sup>205</sup> infant mortality rate;<sup>13</sup> asthma admissions;<sup>206</sup> morbidity due to musculoskeletal diseases, angina, myocardial infarction, bronchitis and emphysema;<sup>207</sup> and mortality due to all causes, ischaemic heart disease, all cancers, lung cancer, and stroke.<sup>208 209</sup> As before, all measures show an increase with increased area deprivation with the exception of survival from a range of cancers, which decreases with increasing deprivation.

As an area level measure, it could be anticipated that associations between lower socioeconomic status and health might be less apparent. For example, the association between cholesterol and deprivation is not as compelling as at an individual level. As was mentioned previously, an area level measure is in effect a mean. The presence of very deprived individuals may be moderated by the presence of some affluent individuals resulting in a medium deprivation score. Furthermore, for most associations it is not assumed that the link is causal – for instance, area deprivation does not cause irritable bowel syndrome, but it is a good predictor of elevated incidence rates. For measures such as neighbourhood crime rates and the associated increased stress levels, they are linked to area characteristics for which deprivation is a more direct measure. In these instances, a causal relationship may well exist. Neighbourhood violent crime and unemployment have been shown to increase the risk of coronary heart disease



independent of individual factors.<sup>91</sup> Higher rates of obesity in deprived neighbourhoods have been linked to a greater density of fast food outlets in deprived areas.<sup>210</sup>

### **1.3.2 Service provision and resource allocation**

Although area level measures of deprivation remove the possibility of inferring associations between health and the individual, for health promotion, resource allocation and service provision, small areas have a greater utility than larger administrative areas such as counties.

The correlation between deprivation and increased mortality and morbidity makes it a proxy for health care need and as Julian Tudor Hart stated: “the availability of good medical care tends to vary inversely with the need for the population served.”<sup>211</sup> If Hart’s ‘inverse care law’ holds, then the availability of medical care will be lower for deprived populations. At this point the distinction between provision and availability or access is important. Living next door to a GP is not very useful if it is not possible to get an appointment due to limited availability, high demand or an overworked GP. In terms of health care services, there is evidence to suggest that access, if not provision, is sometimes lower in more deprived areas. For example, a study of general practices in Perth, Australia, found that although there were more practices in the vicinity of deprived areas the patients from the most deprived neighbourhoods were less likely to be able to see a GP at short notice or have access to a female GP.<sup>212</sup> Other examples associated with increased area deprivation include decreased likelihood of referral for bone densitometry,<sup>213</sup> longer waiting times for cardiac surgery,<sup>214</sup> and lower breast cancer screening uptake.<sup>215</sup> These studies suggest that increased deprivation may be linked to poorer access to services, even though provision may be good. It is not explained why differential treatment may be applied to more affluent patients but the association exists and effectively increases the disadvantage of deprived patients.<sup>214</sup> Goddard and Smith found that due to the difficulties in assessing the

causes of differences in access it was difficult to draw firm conclusions and make practical policy recommendations.<sup>216</sup>

Provision of services and, to some extent, access to those services is tied in to resource allocation. The underprivileged areas score developed by Jarman<sup>217</sup> was developed to identify areas with a predicted high primary care workload. Practices in underprivileged areas could then be targeted for increased funding to compensate for the higher workload. The benefit of using socioeconomic indicators rather than health outcomes such as mortality to predict workload is related to the notion of identifying the at-risk populations – a dead person is not a good predictor of future health care need. However, not all forms of morbidity are correlated with deprivation so resource allocation based on deprivation alone might not be a sensible approach.<sup>218</sup> Moore argued that making additional payments to GPs based on how many deprived patients they served may be of limited use without using the money to tackle the specific health needs of the population.<sup>219</sup> A further criticism by Connolly and Chisholm is that no deprivation index will perfectly identify the areas of highest need and hence resources, thus local knowledge should support decisions made using a deprivation index.<sup>220</sup> An important final point is that even in a highly deprived small area the majority of inhabitants are probably not living in poverty, so that when targeting additional resources at such a small area the majority of people who benefit are not actually in need.<sup>221</sup>

#### **1.4 Urban-rural differences**

The difference between rural and urban geography is important with regard to both health and deprivation. The structure of communities and behaviour of the population is markedly different in urban and rural areas with a consequent impact on health. Urban areas are often typified as having high population density, a more built environment and an industrialised economy while, in contrast, rural areas are seen to have more open space, a less stressful pace of life and less pollution. The effect of the urban-rural divide can be a significant predictor of health independent of socioeconomic status. However, due to the difference in

settlement patterns across different countries, results found in one country may not be applicable to another. In Ireland there is a trend for a dispersed population in rural areas, rather than the more clustered settlements found in countries such as England.

#### **1.4.1 Urban-rural health differences**

Some aspects of ill-health are more generally associated with hazards found in urban environments: substandard housing, crowding, air pollution, insufficient or contaminated drinking water, inadequate sanitation and solid waste disposal services, vector-borne diseases, industrial waste, and increased motor vehicle traffic.<sup>222</sup> Such environmental factors may give rise to higher rates of morbidity and mortality in urban areas. Air pollution is associated with increased risk of stroke,<sup>223</sup> asthma, and circulatory and respiratory mortality.<sup>224</sup> After controlling for smoking, lung cancer rates were still significantly higher in urban areas in Scotland which may be due to more air pollution, higher exposure to passive smoking or selective migration.<sup>225</sup> The concentration of ultrafine particles, which can have adverse health effects, is highest in urban areas.<sup>226</sup> Mental and physical health in children can also be adversely affected by overcrowding which is more commonly seen in urban areas.<sup>82 83</sup>

With regard to risk factors, few show a consistent difference between urban and rural areas. It is sometimes assumed that due to fewer opportunities or exposure, younger people may be less likely to smoke, drink excessively or use illicit drugs. A review of studies analysing risk behaviours such as smoking, drug use and alcohol consumption among adolescents found that the view that rural adolescents engaged in fewer risk behaviours was misleading.<sup>227 228</sup> An American study found increased risk for substance abuse among rural adolescents<sup>229</sup> while work by Levine and Coupey could not find an increased risk of substance abuse in urban areas.<sup>230</sup> The general conclusion is that for adolescents there is little difference between rural and urban areas in engaging in risk behaviour. Higher smoking rates

for adults were predicted for urban areas in Scotland, based on the socio-demographic profile of smokers.<sup>231</sup>

Haynes and Gale found significantly better than average health in rural areas of England and Wales after controlling for deprivation, although the relationship between deprivation and health was weak in rural areas.<sup>232</sup> Levin compared limiting long term illness (LLTI) in urban and rural areas and found the highest rates in urban areas.<sup>233</sup> While rural areas appeared to have lower rates, she found that this could be partly due to the heterogeneity of rural populations and that rural small areas should be sub-divided into more homogeneous communities to get a better picture of variations in health. A somewhat contradictory finding by Phillimore and Reading stated that when rural small areas are increased to have population sizes closer to urban small areas, the relationship between deprivation and health resembled that of urban areas.<sup>234</sup> While the gap between healthiest and poorest is generally smaller in rural areas, so is the gap between least and most deprived. They did find, however, that health in remote rural areas was better than in conurbations but could only speculate that this may be due to less pollution and slightly better social capital. Senior et al. showed that mortality differences between urban and rural areas could be partially explained using deprivation, depending on how deprivation was measured.<sup>235</sup> Judd et al. review a range of studies comparing psychiatric morbidity but show that there is little agreement on whether rates are higher in urban or rural areas.<sup>236</sup>

A Swedish study compared the health of farmers with urban and rural non-farmers.<sup>237</sup> The farmers and rural non-farmers had significantly lower morbidity and mortality rates than the urban non-farmers. Farmers also had better health than the rural non-farmers which was linked to the active and outdoor nature of their occupation. Work by Boland et al., however, found increased mortality and hospital admission rates in rural areas for unintentional injuries in Ireland.<sup>238</sup> Some of the excess mortality and morbidity was due to increased exposure to hazardous farm machinery.

Several reasons are put forward as to why there may be different rates of mortality and morbidity in urban and rural areas.<sup>239</sup> The differences may be due to spatial variations in behaviour and exposure to environmental factors. Alternatively, the differences may be due to selection due to migration. In the first hypothesis, a spatial concentration of poor health is due to increased exposure to risk factors such as air pollution, traffic, poor housing, drug abuse and physical inactivity. Health is a function of the social and physical environment. In the latter hypothesis, healthy people migrate, or remain together, to live in similar areas. People with similar health characteristics tend to end up living together. An example would be the movement of upwardly mobile individuals out of a deprived neighbourhood to be replaced by downwardly mobile, and typically less healthy, individuals. Boyle et al. compared the health of migrants in Scotland and found that those who moved a large distance were healthy while those who moved short distances tended to be unhealthy.<sup>240</sup> Short distance movers were often in social housing and did not have the opportunity or resources to move to a less deprived area. Verheij et al. found that people who migrated between urban and rural areas tended to be younger and healthier than those who had stayed in the same area type, although when demographic and socioeconomic factors were controlled for, this reversed the relationship.<sup>241</sup> The inconclusive findings suggest that to test the selection theory fully would require detailed information on the migration patterns and socioeconomic status of individuals over time.

#### **1.4.2 Rural poverty and deprivation**

It has become increasingly apparent that measures of poverty and deprivation may be biased towards an urban rather than a rural context.<sup>242</sup> If deprivation is to be defined by the lack of access to resources commonly available, any systematic spatial variation in what resources are defined as necessary will introduce difficulties in assessing the spatial variation in deprivation. For example, car ownership is a commonly used deprivation indicator. In an urban context, the lack of a car may be counterbalanced by access to frequent affordable public transport. In a rural setting, where public transport may be quite infrequent, the lack of a car

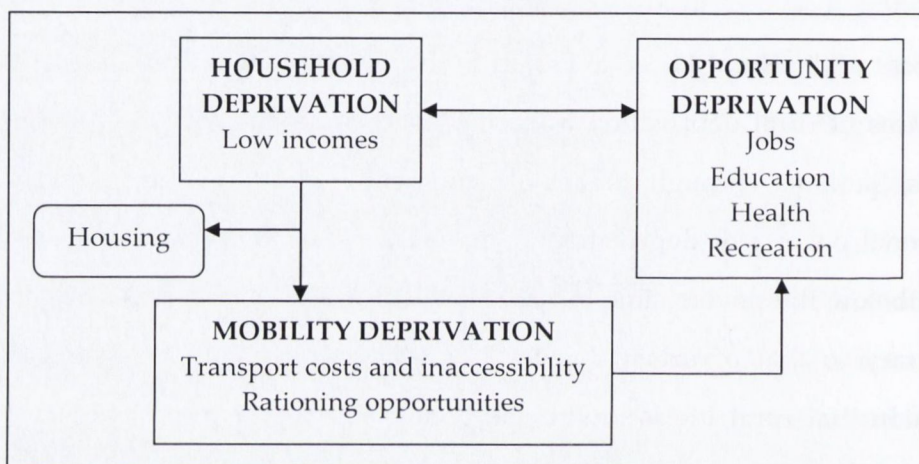
may be a much greater sign of deprivation. Lack of car ownership is a sign of poverty in rural areas while car ownership may be a sign of wealth in urban areas. Employment opportunities in rural areas may be far fewer than in urban areas, making it harder for an unemployed individual to get back into the workforce and out of a situation of poverty. While overcrowding is a distinctly urban problem, under-occupancy is a rural problem with disproportionate heating bills creating a financial burden. This lack of equivalence has implications for the suitability of many deprivation indices to capture poverty in both urban and rural areas.

On foot of the Rural Lifestyle Project in the UK, Woodward discussed some of the nuances of rural deprivation as seen by rural residents.<sup>243</sup> People living in rural areas, perhaps through a lack of anonymity, are unwilling to acknowledge personal poverty or deprivation. Some members of the population may be living well below the poverty line but accept this as their lot and make do in a manner contrary to that of urban dwellers. This is coupled with an overriding public opinion that rural life is idyllic and contented, free of the stresses and poverty associated with inner cities. Rural inhabitants, however, are faced with few options and often have limited access to resources and amenities that are taken for granted in cities, such as health care services, social settings and childcare facilities. Milbourne notes how in the rural English county of Wiltshire, with some of the least deprived areas in England, the majority of parishes lack basic services such as a shop, post office, daily bus service, a bank or cash dispenser, and a general practice.<sup>244</sup> For those with a car it is possible to access these services in the nearest town, but for the minority without a car they must use the infrequent bus service. The lack of a car results in much greater disadvantage than it would in an urban area where most of those services may be present.

Cloke et al. describe Shaw's model of rural deprivation in which there are three categories of deprivation: household, opportunity and mobility deprivation.<sup>245</sup> Figure 1.4 shows the diagram of Shaw's model. The problem of accessibility and transport is a category of deprivation in itself. Such a category may be of little

practical use when considering urban deprivation. It is apparent from Shaw's model that low income can lead to increased mobility deprivation which, by reducing opportunities, can hinder attempts to increase income. For example, someone on a low income may not be able to afford to get a job further away as they cannot pay the transport costs associated with taking that job. Such situations lead to persistent poverty.

Figure 1.4 Shaw's model of rural deprivation<sup>245</sup>



These categories of rural deprivation are reflected in the choice of indicators of rural disadvantage put forward by the UK Countryside Agency.<sup>246</sup> The indicators chosen include income, geographic availability of services, employment and mortgages. They also include educational and health disadvantage as pertinent measures. The Countryside Agency argue that the standard deprivation index does not adequately highlight rural disadvantage and the subset of indicators presented can be use to better distinguish between rural areas with high and low levels of deprivation.

Noble and Wright attempted to address deficiencies in a standard deprivation measure by collecting data on benefits for small areas as a proxy for low income households.<sup>247</sup> They used ordinary least squares regression to predict benefits with deprivation indicators in a subset of rural areas. They were able to produce a model which was better able to predict low income households in all rural areas

than the existing UK deprivation index. A selection of variables that predict low income well in rural areas might not be as good a predictor in urban areas, giving rise to the notion of different indices for urban and rural areas. Such a scheme retains the problems of lack of equivalence but at least gives more accurate measures of deprivation for specific areas. The incorporation of some measure of sparsity can act of a proxy for cost of transport and access to services which may enhance the utility of an income measure. A higher cost of transport adds to the cost of living in rural areas.

Nolan et al. investigated poverty in Ireland by area type in 1994 and found that 35.6% of households below the 60% income line lived in open country areas.<sup>248</sup> The incidence of poverty in open country areas dropped substantially between 1987 and 1994 while rates increased in Dublin city. The greatest risk of poverty was found in villages and towns with a population of less than 3,000 persons, where 46.5% of the population is at risk of being below the 60% income line. This contrasts with the 27.4% of persons at risk of being below the 60% income line in Dublin city. Commins refers to work by Frawley et al. to examine deprivation in low income Irish farm households.<sup>249</sup> Household items, such as strong footwear, that when lacking would be considered a sign of deprivation are occupational necessities for a farmer. These findings, coupled with those of Noble and Wright,<sup>247</sup> certainly suggest that income is a more appropriate measure of deprivation in rural areas than in urban areas, particularly if combined with some measure of cost of living or sparsity. In this context, sparsity acts as a measure of distance between people and also distance to services and employment.

### **1.4.3 Rural exclusion and access to health services**

While crowding and the associated stress may be a problem in urban areas, the converse is true in rural areas where the greater distances between houses can lead to isolation. Coupled with living away from social hubs such as towns, rural life can be synonymous with social exclusion. The concept of social exclusion refers to the "dynamic processes of being shut out, partially or fully, from any or all of



several systems which influence the economic and social integration of people into their society."<sup>249</sup> Social exclusion is a problem of both urban and rural environments, although geographic isolation is almost intrinsic to the definition of rural. Living a distance from community centres will inevitably lead to exclusion particularly if there are few options for travel. McDonagh notes how, in rural Irish counties, high rates of car ownership are indicative of there being no practical alternative for transport.<sup>250</sup> This has been exacerbated by the lack of investment in rural infrastructure and transport policy. This has also led to the increasing marginalisation of young and elderly people with no access to personal transport. Policies to reduce traffic congestion such as increased taxes on fuel tend to adversely affect those in rural areas who frequently do not have a viable alternative mode of transport. The sum effect is to increase exclusion from society for rural inhabitants.

The lack of transport options can have implications for timely or reasonable access to health services. A study in Ontario looked at repeated hospitalisations of children with chronic conditions.<sup>251</sup> Having to travel over larger distances to reach the hospital was found to strain family relationships. Jones et al. found that inaccessibility of acute hospital services may increase the risk of asthma mortality having accounted for deprivation.<sup>252</sup> Pan elli et al. investigated access to health services in rural New Zealand.<sup>253</sup> Some of the issues highlighted included the difficulty in getting appointments at a convenient time, work time lost travelling to and from an appointment, and the time to wait for an ambulance being unacceptably long. A study of access to general practices in a region of England found that most of the population lived within a short distance of a practice but for 5% of the population a longer distance was coupled with no weekday bus service.<sup>254</sup> For this rural subgroup, access to services is dependent on access to a car or the costly alternative of a taxi. Although it is not possible for the entire population to live within close proximity of a health service, those who have further to travel almost inevitably live in rural areas where transport options are

limited. The combination of low provision and poor access can have health consequences if timely treatment is not available.

## **1.5 Problems with existing methodology**

It is evident from the preceding sections that there are a large number of factors that influence health, many of which display a social gradient. These social gradients are also observed at an area level, although not necessarily in a causal relationship. Deprivation indices offer a methodology for representing the socioeconomic conditions present in an area. It has also been noted that spatial variation can be observed in health inequalities, and that this spatial variation may be linked to differences between urban and rural areas. These differences may be partly explained by environmental differences and partly through selective migration. It has also been shown that the notion of poverty and deprivation may be different in urban and rural areas as a consequence of differing opportunities, accessibility and demography. By virtue of the indicators used, some existing deprivation indices are criticised for being biased towards urban areas. It is therefore imperative that adequate measures of poverty and deprivation are used in the investigation of health disparities and for the purposes of policy development with a view to reducing inequalities. Some of the studies mentioned in this chapter have highlighted situations when deprivation indices may be inadequate or inappropriate for use in analysing health inequalities. A number of the principal problems with existing methodology for both urban-rural classification and deprivation indices are identified in the following sections.

### **1.5.1 Urban-rural classification**

The difference between urban and rural areas is often characterised as a simple dichotomy where urban areas have a high population density while rural areas have a low population density and are predominantly agricultural. This representation is convenient for simple comparative purposes but may ignore the gradient of settlement types that exist between dense metropolitan areas and sparse rural areas.

Approaches to defining urban and rural areas often begin with a definition of what constitutes urban with all remaining areas being labelled as rural. Many countries have adopted a simple cut-off for settlement size to distinguish urban from rural.<sup>255</sup> Such a method assumes that all settlements above the cut-off size have an inherent similarity and can be described as urban. Cut-offs can vary enormously from country to country depending on settlement patterns which can show marked differences across countries. Other methods include the use of population density,<sup>256</sup> accessibility<sup>257</sup> and multivariate techniques.<sup>258</sup> Some applications use a simple binary classification while others use a range of classes to distinguish between cities, towns, villages and dispersed rural populations.<sup>259</sup>

The urban rural classification in use in Ireland is a simple settlement size cut-off provided by the Central Statistics Office (CSO).<sup>260</sup> Although the classification is not provided at a small area level, with the data provided it is possible to determine the percentage population classed as urban and rural in each area. A small area that contains a town of 1,500 persons is considered the same as a small area at the centre of a city such as Dublin. In reality, these two small areas may be very different in terms of population density, access to services and typical land use. To better understand the health and socioeconomic differences between urban and rural areas, it is imperative that a suitably detailed small area urban-rural classification scheme is used rather than a simple dichotomy. At present no such classification exists for Ireland.

## **1.5.2 Choice of indicators and validation**

The choice of indicators for a deprivation index is partially driven by theory and partly by availability. The latter limitation is understandable given the potentially sensitive nature of the data required and the small area level at which it is needed to produce a sufficiently detailed picture of the spatial distribution of deprivation. In Ireland, for example, many useful indicators are only available at a county or Local Authority level. An index at such a geographic level would not be particularly useful for policy or research purposes. Routinely collected data such as

live register unemployment figures, medical card ownership, crime figures, hospital in-patient data – data that could be usefully incorporated into a measure of deprivation – are not routinely coded to small area. As there is no detailed postal code system in operation in Ireland, coding addresses to small areas is time consuming, expensive, often unreliable and may not be feasible due to issues of confidentiality. Although the census is conducted every five years, with the current climate of high immigration and rapid changes in demography the data are out of date and potentially misleading before the next census is conducted. Nevertheless census data are the only realistic source of deprivation indicators in Ireland despite the problem of timeliness.

Validation is the process whereby a deprivation index is assessed in relation to how well it measures deprivation which is essential in terms of the utility of an index.<sup>220 261</sup> This aspect of validation rarely extends beyond an assessment of the correlation matrix to confirm that all of the variables appear to indicate the same or a similar notion of deprivation. To assume that a deprivation index is a good measure of deprivation solely on the grounds that the variables were chosen on a sound theoretical basis would be unwise. In reality, there needs to be an analysis of the relationship between a deprivation index and the relevant health outcomes associated with deprivation. Gordon looked at validation using surveys to determine the likelihood of an individual to be deprived given some characteristic as recorded by the census.<sup>262</sup> The most popular measures of the outcomes of deprivation, however, are health related: mortality, morbidity and mental health. As outlined in section 1.3.1, extensive research has been conducted in the UK looking at the association between deprivation and health. Given the lack of small area health outcome data in Ireland it has been difficult to obtain suitable data for validation although mortality,<sup>182</sup> medical card ownership and disability have been used.<sup>183</sup> An Italian index has also been validated using mortality data.<sup>186</sup>

### 1.5.3 Data transformation

Once collated, deprivation indicators are frequently transformed in some way prior to combination into a single or smaller number of deprivation measures. Common types of transformation are log and logit transforms. Such transformations are generally used to improve the normality of the data.<sup>263</sup> Depending on the method of combination used, approximate normality of the data may be desirable, if not a prerequisite.

Another form of transformation that has emerged more recently is that of shrinkage. As deprivation indices are generally computed at a small area level where the denominator population may be quite small, a small fluctuation in the numerator may translate into a relatively large change in the observed proportion. The purpose of shrinkage is to move indicator values based on very small numbers closer to the mean for that indicator.<sup>264</sup> The degree of shrinkage is related to the standard error associated with the small area. The standard error is, in turn, related to the population of that small area such that a large population equates to a small standard error and vice versa. The technique of shrinkage is analogous to smoothing in that it reduces random fluctuations in the data. Shrinkage is not ideal as assumptions are made to associate the standard error with the population size. Criticisms have been levelled at the use of shrinkage on the grounds that in the subsequent stage of data combination, the small area values are no longer independent of each other.<sup>265</sup> In the Irish context, given the fact that the more populated small areas tend to be found in urban areas, shrinkage will tend to affect rural areas more than urban areas.<sup>185</sup> If rural areas are more affected by shrinkage they will tend to move closer to the mean and hence a more moderate deprivation score. A further point is that the properties of shrinkage and possible consequences are not fully described in relation to deprivation. It is generally assumed to be appropriate to apply shrinkage when an indicator displays large standard errors for some areas.

#### 1.5.4 Regional bias and spatial autocorrelation

An issue that is alluded to in deprivation index literature is the problem of indicators that may reflect a primarily urban or rural measure of deprivation. Spatial autocorrelation is a measure of the degree of similarity between neighbouring areas. High values indicate that geographic areas that are close in space tend to be similar, which in turn is indicative of systematic regional variation. Spatial autocorrelation can be quantified using metrics such as Moran's I.<sup>266</sup>

In their assessment of car ownership as a suitable proxy for deprivation in Wales, Christie and Fone<sup>267</sup> found that car ownership negatively correlated with the other Townsend indicators in rural areas. This was in contrast to positive correlations in urban areas and for all areas, suggesting that indicator correlations were driven by urban areas. Pacione looked at indicators of rural disadvantage in Scotland and noted how a number of the traditionally used deprivation proxies were more indicative of urban poverty than rural poverty.<sup>268</sup> Such analyses are unfortunately uncommon in the literature, as stated by Milbourne in his paper on the geographies of poverty.<sup>244</sup> He points to the dearth of research in the "local geographies of poverty" and the lack of understanding of spatial variation in the components of poverty. The contrast between urban and rural deprivation causes difficulties for a nationally calculated deprivation index.

One of the methods used to solve the problem of different urban and rural forms of deprivation has been the use of a range of indicators and retention of separate factors that appear to measure urban and rural deprivation.<sup>184 269</sup> In such cases the deprivation measure is calculated for the whole region of interest including both urban and rural EDs giving rise to urban EDs influencing the weights for a measure of rural deprivation and vice versa. In the case of Neylon,<sup>269</sup> who developed four indices of deprivation for County Clare, he shows Ennis Rural ED to be rurally deprived. This is despite the fact that 87.7% of the Ennis Rural ED population live in Ennis town.

### 1.5.5 Data combination

Deprivation scores are typically generated using a weighted sum of the indicators expressed in a standardised form. Numerous methods exist and have been used to derive the weights. These methods range from equal weights,<sup>177</sup> arbitrarily selected weights,<sup>180</sup> weights derived by survey,<sup>176</sup> Principal Components Analysis (PCA)<sup>187</sup>, and Factor Analysis (FA)<sup>184</sup>. The arguably simplistic approaches of equal and arbitrary weights selection have been replaced by the use of PCA and FA. This change may be partly explained by the advent of cheap high-power computers to facilitate calculation of PCA and FA. Both of these methods have been used for numerous deprivation indices and, although they tend to produce similar solutions, there is a fundamental theoretical difference between the two methods. While PCA is a straight arithmetic combination of the indicators, FA seeks one or more underlying factors. PCA does not account for differing levels of statistical accuracy or the imperfect measurement of the underlying factor.<sup>264</sup> Some forms of FA, such as Maximum Likelihood (ML) FA can distinguish between these forms of variance and take them into account. FA is founded on the notion that there are one or more underlying factors that can be identified from the indicators. It is at the discretion of the researcher who applies FA to determine how many underlying factors exist. That decision may be based on a sound theoretical justification or it may be determined by comparing the results from a range of choices of number of factors. A further set of options are available in both PCA and FA regarding rotation whereby a transformation can be applied to the results to make them easier to interpret.

In terms of deprivation index development, it has been argued that PCA is more appropriate than FA<sup>182</sup> and vice versa.<sup>264</sup> Without a consensus it is at the discretion of the researcher to decide which method is appropriate for the theory they adhere to. The choice between PCA and FA may have significant implications for the resultant index although no comparison is in evidence in deprivation literature.

## 1.6 Aims and objectives

The preceding sections have outlined some of the deficiencies of existing deprivation and urban-rural classification methodology. Given the extent to which both are used independently and in conjunction, it is important that the issues relating to both are dealt with in detail. The aim of this research is to assess deprivation index methodology and to address the issue of urban-rural variation in deprivation indicators. The specific objectives of the research are:

- To develop an urban-rural classification for Ireland
- To assess the characteristics of shrinkage
- To assess methods of combining indicators for deprivation scores with a view to accounting for urban-rural variations in deprivation indicators
- To identify the key problems and possible solutions associated with area-level deprivation measure methodology

There is no comprehensive rural-urban classification system for small areas in Ireland. It is proposed to develop such a classification using a range of data sources. Chapter 2 assesses methods for defining areas as urban and rural before developing a rural-urban classification for Ireland. In chapter 3, issues relating the now commonly used methodology of shrinkage are investigated. Chapter 4 looks at methods of combining indicators and dimension reduction for the development of deprivation indices. It is also proposed to develop a method for combining indicators such that urban-rural differences may to some extent be accounted for. A sensitivity analysis is conducted in chapter 5 to illustrate the impacts of different choices regarding data selection, transformation and combination. The discussion and conclusions are presented in chapters 6 and 7, respectively.

The methodological issues and suggested solutions in this study are applicable to small area deprivation measurement in any region.





## 2 Defining the urban-rural divide

It is evident from chapter 1 that there are geographic variations in both deprivation and health, and to some extent, these variations can be explained by the distinction between rural and urban areas. The differences in environment, lifestyle and access to essential services result in noticeable differences in health and poverty. To properly assess those differences, a classification system is required whereby areas can be labelled as urban or rural. This chapter sets out to define such a classification for Ireland. Section 2.1 contains a general discussion on urban and rural ideology which is followed in section 2.2 with a discussion of the Irish context. In section 2.3 a range of methods of urban-rural classification used internationally are applied for the first time to Irish data. Methods of data combination are discussed in section 2.4 and finally in section 2.5 a new urban-rural classification for Ireland is outlined.

### 2.1 What defines urban and rural?

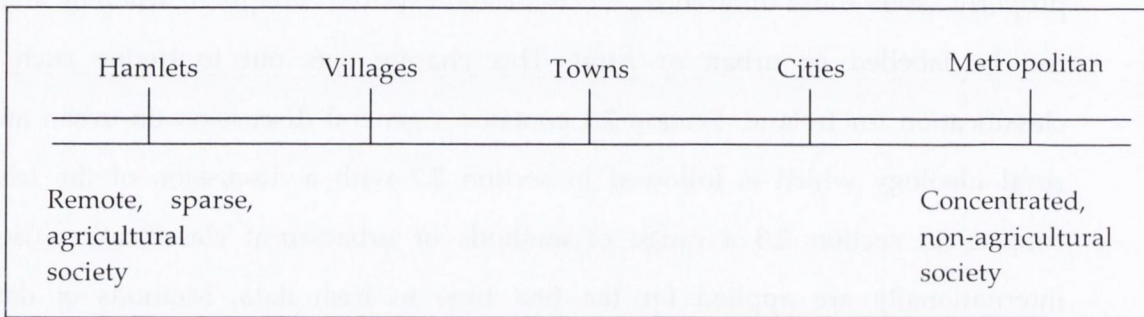
The definition of rural area is often constructed in a negative manner - defining what might constitute the urban area and then labelling all other areas as rural. As a result, the label 'rural' can be attached to a large variety of areas that might otherwise be considered as very different. There is an assumption that a clear distinction exists between urban and rural and that areas can be labelled as one or other however misleading that label might be.

'Urban' can generally be described as any area that is part of or has the characteristics of a city.<sup>255</sup> A city has a concentration of population with distinct employment patterns and lifestyle. One would expect a diversity of social, economic and cultural activity in a city. There should also be a variety of facilities, services and amenities in a city to cater for the large and varied population.

To say that everything else is 'rural' is to ignore the variety of settlement types and environments that exist outside of cities. The continuum from rural to urban is

shown in Figure 2.1 below. Within the extreme rural population there is a further distinction between agricultural and non-agricultural land. There are large areas of peat bog and native forestry with small isolated populations with no local means of farming. Coupled with little or no local commerce, inhabitants of these regions must travel long distances to work.

Figure 2.1 Rural-urban continuum



There is a further complication that a town may be in close proximity to a number of other towns or distant from other towns. This can be important as an isolated town will potentially be an important hub of activity in that region while a town in proximity to others may be relatively less important but the inhabitants may have greater opportunities and access to services such as medical care, policing and child care. With increasing house prices in Dublin, former villages are being expanded into satellite towns to accommodate the growing number of people working in Dublin city but unable to buy houses in the city. These satellite towns are generally quite immature and lack many of the services and amenities associated with urban centres and yet they do have substantial populations with medium to high population density.

In attempting to describe an area outside of a city, it is important to maintain information on the settlement size, local land use and proximity to other settlements. This is likely lead to a classification system with many levels which may result in small numbers of areas in some levels. However, this would be preferable to a simple dichotomy which would maximise loss of information and distinction.

## 2.2 Urban-rural divide in Ireland

Ireland, like almost any other country, is composed of a mix of urban and rural areas. While rural areas are typically sparsely populated with a predominantly agricultural economy, the urban areas are densely populated with a wide range of employment types. Rural areas can broadly be seen as dependent on the extraction of natural resources while urban areas process and sell services.

In the Republic of Ireland the term 'small area' generally refers to Electoral Divisions (EDs). There are 3440 EDs and they are the smallest output area for census data. Counties and other government constituencies are comprised of aggregations of EDs. Populations range from 55 to 24,404 in the 2002 census and areas range from 0.046 km<sup>2</sup> to 126.04 km<sup>2</sup>. If an individual can be identified in the census results for one ED, that ED will be merged with a neighbouring ED. As a result, there were only 3,422 output EDs in 2002. Town boundaries are not restricted to EDs so it is possible for a town may have parts in multiple EDs. It is also possible for multiple distinct towns to be in a single ED.

The definition of urban in Ireland is based on the town-dwelling population. Towns are comprised of those with and those without a legally defined boundary. Where a legally defined boundary exists, the town size is defined as the population living within that boundary. For towns without a legally defined boundary, there must be a cluster of 50 or more occupied dwellings. There must also be, within 800m of that cluster, a nucleus of either 30 occupied dwellings on both sides of the road or 20 occupied dwellings on one side of the road. Currently in Ireland, the population living in clusters of 1,500 or more persons is described as urban.<sup>270</sup> The rest of the population is termed rural.

Suburbs are defined based on a 200m criterion recommended by the United Nations<sup>270</sup> whereby a cluster is defined where no occupied dwelling is more than 200m from another occupied dwelling. Industrial, commercial and recreational buildings are not regarded as breaking the continuity of a built-up area. Suburbs

and environs are included in a town when counting the population resident in that town.

There is a further administrative distinction between urban and rural in Ireland – the urban and rural district boundaries. These are aggregations of electoral divisions (EDs). All EDs in town and city boundaries are classified as urban and the remainder are aggregated into 160 rural districts. These districts are intermediate in size between EDs and counties but are rarely used in research or governance. Mortality data published in the vital statistics aggregate rural districts by county to give data for 88 urban and rural districts.

## **2.3 Urban-rural measures**

A number of different methods of urban-rural classification have been identified in the literature. In the following section these will be described and briefly applied to Irish data to give an indication of the differences between the methods.

### **2.3.1 Population size**

In a number of countries the definition of urban relates directly to settlement size. A settlement is generally defined as a collection of houses where every house is within 200m of another house. The settlement size that constitutes “urban” varies from country to country, as can be seen in

Table 2.1 below.<sup>255</sup> The assumption is that above a certain population, a town can automatically be considered as “urban”. This is based on a critical mass of population having access to a number of essential services and that such a number of people living in relatively close proximity will automatically classify as urban. Different countries have very different notions as to what qualifies as urban and this is probably in part associated with population density and historical precedent. The choice may also be partly political as service provision or funding may be affected by the urban-rural status. The distinction between rural and urban would then have implications for government obligations as regards services.

Table 2.1 Minimum population sizes used by countries to define urban areas

Country	Minimum population
Sweden	200
South Africa	500
Australia*	1,000
Ireland	1,500
France	2,000
United States	2,500
Belgium	5,000
Spain	10,000
Japan	30,000

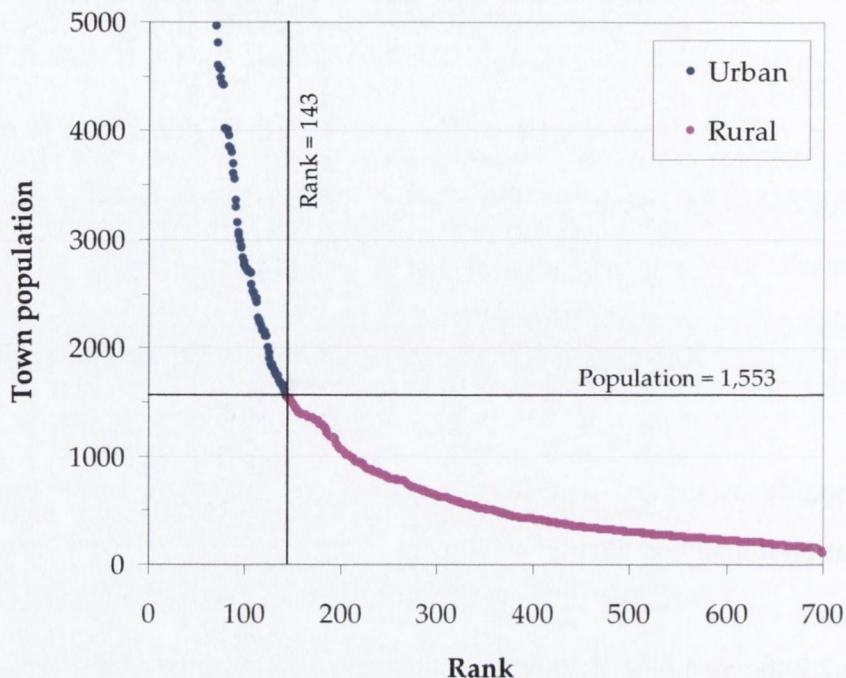
\* To qualify as urban Australia also stipulates that there must be a minimum population density of 400 persons/km<sup>2</sup>.

Figure 2.2 shows a plot of 700 Irish towns and villages ranked by size (towns with a population greater than 5,000 not shown). For illustrative purposes, it is intended to split towns into two groups: urban and rural. Due to a small number of towns and cities with extreme population sizes relative to other towns, the populations were log transformed.

Data can be grouped using k-means clustering. This is a method of clustering in which the user pre-defines the number of groups. The observations are then grouped so as to minimise the difference between observations in each group.

Application of k-means clustering to log-transformed town populations to identify two clusters results in a cut-off at 1,553 persons. If we accept a cut-off of 1,553 persons, then there are 713 EDs with the majority of the population living in a town of 1,553 or more persons.

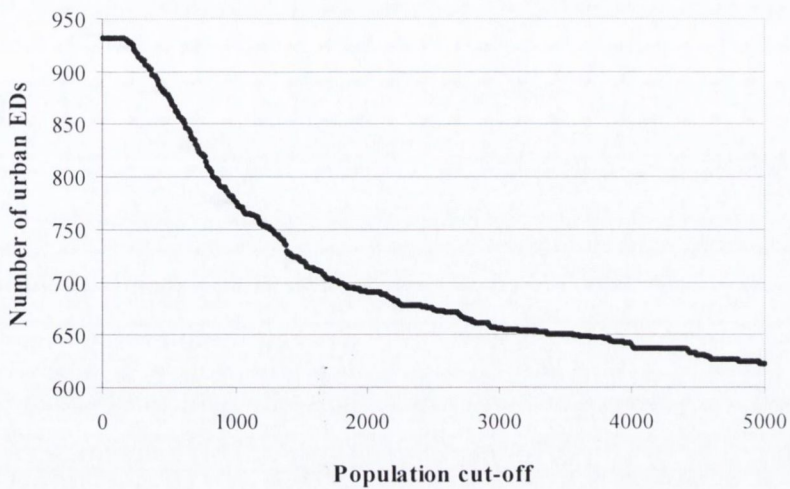
Figure 2.2 Ranked town populations in Ireland (towns > 5000 persons not shown)



Source: 2002 population data from CSO

The advantage of this method is that it is relatively easy to compute. The main difficulty with this method is that the choice of cut-off tends to be somewhat arbitrary. It also leads to a potentially misleading dichotomy – a town classed as rural may be reclassified as urban if the population increased by one person. This precise distinction is unrealistic and unreasonable. This method also ignores proximity to urban areas. A remote rural area is given the same classification as a town just below the cut-off size. This lack of distinction is also unreasonable. Figure 2.3 shows the number of EDs classed as urban by choice of urban population cut-off.

Figure 2.3 The number of EDs classed as urban by cut-off to describe settlement as urban



Source: 2002 population data from CSO

An alternative approach would be to classify a town based on the essential services available in that town. These may include emergency services, amenities and other facilities that may act as indicators of urban life. Collecting these data reliably may prove difficult and it leads to further problems. The most critical problem is that the location of most services is driven by the population distribution and market forces. This complicates matters where a town is in close proximity to other larger towns. For example, Portmarnock with a population of 8,376 does not have a Garda station. Meanwhile Donard, with a population of 201, does have a Garda station. It would be incorrect to label Portmarnock as rural and Donard as urban. Of course such a classification method would not be based on the presence or absence of a single service but on a range of amenities.

Furthermore, due to the sparser population in more rural areas, towns with small populations may have the services normally associated with larger towns in more urban regions. A small town in an agricultural region will be relatively urban in its context as it may act as a centre of commerce and interaction. An example would be Achill Sound, with a population of only 355, it has a supermarket, bank, post office and Garda station. It is a link between Achill Island and the bridge to the



mainland. Much larger towns close to major urban centres may essentially act as commuter towns. In terms of facilities they may be underdeveloped but the population lives a largely urban lifestyle. A town such as Portrane with 1,726 inhabitants does not have a supermarket, bank, post-office or Garda station and yet it has sizeable population and is considered urban under the definition applied in the Irish census.

So while the presence of certain amenities may point towards an urban environment it is not a reliable distinction. This information is perhaps more useful for differentiating between different types of settlement rather than their status as urban or rural.

### **2.3.2 Population density**

Instead of using the population of a town or area, it is possible to use the population density (i.e. persons per kilometre squared) for distinguishing rural and urban boundaries. High densities should only occur in urban areas where people tend to live close together. In rural areas, where people frequently live further apart, population density is lower.

This method shifts the problem from having to identify a suitable population cut-off to finding an appropriate population density cut-off. It also gives rise to the question of what area is the density being calculated for. Typically the area covered by water bodies such as lakes and sea are ignored in the density calculation. Perhaps it would then make sense to also exclude land above a certain height or any land that is otherwise uninhabitable. There may be instances of small areas where nearly all of the population lives in a small portion of the land in that area. Thus the inhabitants may experience a high population density but the calculation for the area would return a low density.

Determining what might constitute an urban level of population density is not straightforward. Both Australia and Canada use a cut-off of 400 persons/km<sup>2</sup> for

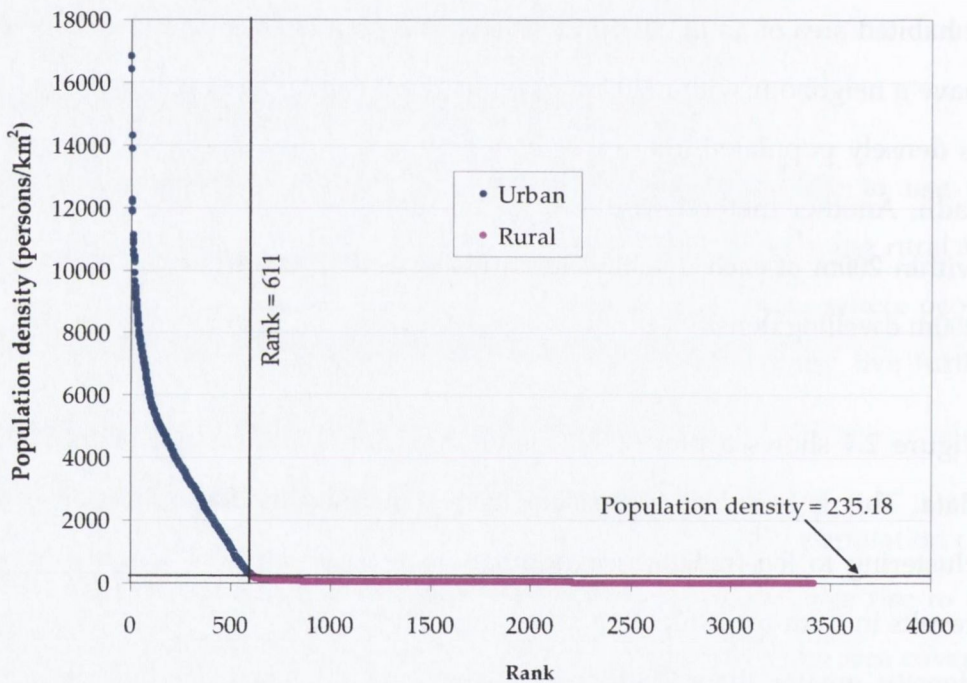
urban areas. This is a very arbitrary rule-of-thumb and it would make more sense to use data for undisputed urban areas to understand what might be a reasonable cut-off point. For example, the population densities for Dublin, Cork and Limerick cities are 4191.6, 3124.9 and 2610.3 persons/km<sup>2</sup>, respectively.

The definition of the area of an ED is important. It is acceptable to exclude areas of water from the calculation as they are uninhabitable. It is also arguable that if any portion of an ED is uninhabited, it is acceptable to ignore it in calculations. One solution is to define the area of an ED as the total area within 200m of a dwelling - the United Nations criterion for neighbouring houses. This can be referred to as the inhabited area of an ED. If an ED is sparsely populated, then most houses will not have a neighbour within 200m and thus the inhabited area will be large. If the ED is densely populated, there will be a large amount of overlap between the 200m radii. Another method would be to calculate the housing or population density within 200m of each dwelling in an ED and then determine the mean or median 200m dwelling density.

Figure 2.4 shows a plot of ED population density calculated using 2002 census data. This is based on a measure using inhabited land. Application of k-means clustering to log-transformed population density values to identify two clusters results in a cut-off at 676.48 persons/km<sup>2</sup>. This suggests that EDs with a population density greater than 676.48 persons/km<sup>2</sup> are distinct from those with a lower population density. If we assume that EDs above this figure are urban, this results in 611 of the 3,422 EDs being classed as urban. Using a measure of total land area less water bodies, the cut-off occurs at 235.18 persons/km<sup>2</sup>, which is somewhat lower than the cut-off used in Australia and Canada. This cut-off also results in 611 urban EDs, although there are differences in which EDs are labelled urban. The smallest town included in the urban areas in both instances has 1,064 persons. This suggests that the measure may indeed be capturing larger population centres. However, some EDs with a large population almost entirely situated in a city are classified as rural. For example, if classed by population density over the entire ED

area, Phoenix Park ED in Dublin city is classed with rural EDs. This ED contains a large public park with a very low population density. On the basis of inhabited area it is classed as urban. The opposite also occurs with Cabinteely-Loughlinstown ED being classed as rural using inhabited area, but urban using the entire area encompassed by the ED. These apparent discrepancies occur where the different definition of inhabited land leads to a substantial difference in computed population density.

Figure 2.4 Ranked ED population densities



Source: 2002 population data from CSO

In Ireland, the EDs comprising the five cities can safely be labelled as 'urban'. They are within the defined city boundaries so it is an acceptable assumption. They comprise of 332 city EDs with varying geographic sizes and population numbers.

Using the notion of inhabited area, an analysis of the city EDs shows that urban population densities range from 161.6 to 16,836.2 persons/km². On closer inspection, a number of the Waterford city EDs are transitional between the city

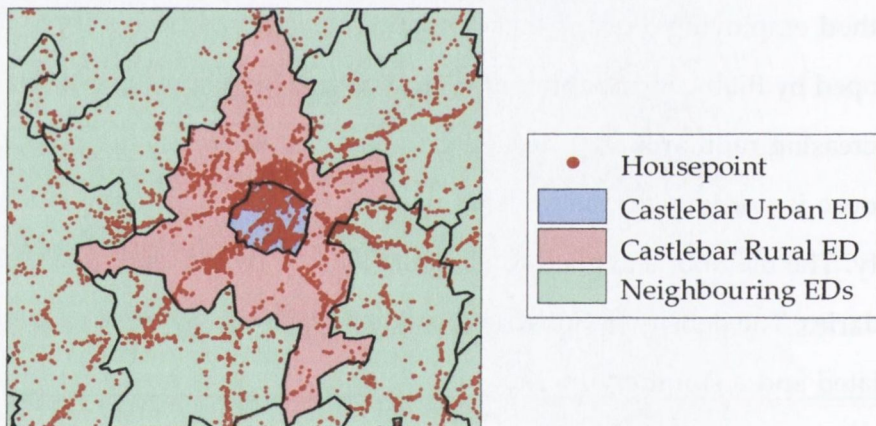
and the neighbouring rural areas. To account for these EDs, the city EDs have been ranked by population density and the bottom 1% have been ignored as effective outliers. Thus the minimum city population density is given as 693.1 persons/km<sup>2</sup>. If we apply this cut-off to the full set of Irish EDs, there are 609 urban EDs. Using the typical housing density around dwellings, between 673 and 680 EDs are classified as urban, using mean and median housing density respectively. There is a substantial difference in urban classification depending on whether population or housing density is used. Using mean housing density as a measure also results in small towns being classified as urban. A town with a population of only 502 would be classed as urban. As with population density, there are EDs with large city-based populations that are classified as rural, making this an unsatisfactory classification method.

A method employing a proxy for population density has been adopted in the UK developed by Bibby and Shepherd.<sup>271</sup> This method uses housing density calculated for increasing radii around 1 hectare grid squares. Although the exact population per house is not known, the housing density is a sufficient proxy for population density. The method is applied to all areas outside the officially recognised urban boundaries. The density of households within 10, 20 and 30km of each grid cell are calculated and a summary measure is calculated for each small area based on the grid cell values. The sparsest 5% of small areas are noted for each radius size to identify the areas that might be considered most sparse and therefore most rural. Housing densities are also computed for smaller radii of 200, 400, 800 and 1,600m with a view to identifying settlement types. For example, grid cells that show a sharp drop in housing density with increasing radius are presumed to be in a small village. It is not described how cut-offs are decided on to distinguish between town, village and hamlet based on the housing density profile.

As was the case for classification by settlement size, there is plenty of opportunity for misclassification. This is particularly evident for EDs on the edge of an urban ED. There are numerous instances of donut shaped EDs that enclose an

approximately circular urban ED, an example of which is shown in Figure 2.5 below. There are, in fact, 80 instances of EDs that enclose EDs containing towns. The outer ED is typically a combination of the outskirts of the enclosed town and the surrounding rural environment. There is generally a sharp transition from urban to rural landscape. If the majority of the population in the outer ED is living in the suburbs of the town, it would make sense to refer to the outer ED as being urban. However, using population density as the classifier can often result in the outer ED being classed as rural. The example in Figure 2.5 shows Castlebar Urban and Rural EDs. The entire population of Castlebar Urban ED lives in Castlebar town. Of the 5,882 people living in Castlebar Rural ED, 3,702 are defined as living in Castlebar town.

Figure 2.5 Example of a donut shaped ED enclosing a town



A further drawback to this method is that it does not take into account proximity to an urban area. Therefore everything that is not urban is automatically rural. As was discussed previously, this is not a very helpful classification method.

### 2.3.3 Access

The use of a gravity model approach allows for the combination of both population size and spatial location into a single measure (Equation 2.1). This formula essentially measures the spatial interaction between an origin ED and destination towns.

$$a_i = p_i \sum_{j=1}^N p_j d_{ij}^n \quad (2.1)$$

Where:  $a_i$  = access of area  $i$ ,  $i = 1, 2, \dots, N$

$p_j$  = population of town  $j$  (may be log transformed)

$d_{ij}^n$  = distance from  $i$  to  $j$  to the power of  $n$

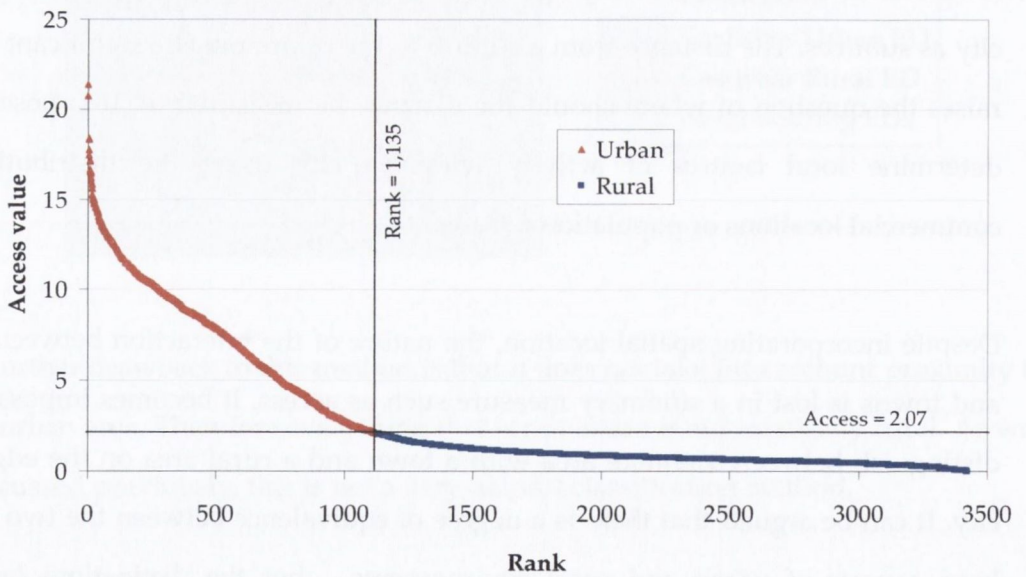
(where  $n$  is -2 by the inverse square law)<sup>272</sup>

There are some difficulties in the measurement of distance. It can be expressed as a simple distance either as-the-crow flies or along a road network. It can also be expressed as some form of cost distance such as travel time. It is also often in the form of a distance decay function so that influence declines rapidly in relation to proximity. For cities, they will generally have numerous centres. These might well have been suburbs or even outlying villages that have now been subsumed into the city as suburbs. The distance from a suburb to the centre may be significant which raises the question of where should the distance be measured to. It is possible to determine local centres of activity within a city using the distribution of commercial locations or population density.

Despite incorporating spatial location, the nature of the interaction between areas and towns is lost in a summary measure such as access. It becomes impossible to distinguish between a remote area with a town and a rural area on the edge of a city. It can be argued that there is a degree of equivalence between the two – both have aspects of urban and rural environments – but the distinctions between remote town and city edge are lost. A further problem is the bias towards the greater Dublin area, which includes counties Dublin, Kildare, Meath and Wicklow. This is an extensively developed area with many towns in addition to the major urban centre of Dublin city. Even the remote parts of Wicklow have relatively good access to towns when compared to western counties.

The access measure was determined for EDs to towns within a 48km radius with the distance decay set at  $d_{ij}^{-2}$ . The distance decay function is one often used in applications regarding population influence and interaction. The distance limit was chosen arbitrarily as it represents a typical travel time of between 45 minutes and 1 hour. Beyond that distance the influence of a town on daily life may be limited. It is assumed that opportunities and services more than one hour away are considered to be much less significant and so are not included in access calculations. The ranked values are shown in Figure 2.6 below. Using k-means clustering on log-transformed access values to define two clusters, a cut-off is identified at an access value of 2.07 which would have 1,135 EDs classed as urban. This includes many EDs that are small rural areas close to a number of urban centres. A different choice of distance cut-off and decay function will lead to different results but the current choice is justifiable in the context of this exercise.

Figure 2.6 Ranked ED access values



Source: 2002 population data from CSO

The advantage of this method is that it incorporates proximity to urban centres into the measure and results in a continuous, rather than binary, variable. That the variable is continuous also gives rise to the problem of how to classify the resultant

values. An area can have a high degree of access whilst being entirely rural. If it has high access to town and city areas, it is indicative that the population experiences a high degree of interaction with those urban areas. However, the population of that area lives in a rural environment rather than a built-up urban environment. Any classification of the access variable will group areas that are quite different, which needs to be avoided.

#### **2.3.4 Land use**

Using satellite imagery, land use can be mapped to a grid of relatively high resolution – sufficient for variation across a small area to be picked up. It can then be seen if an area is predominantly residential, industrial, agricultural or natural habitat. It provides a realistic representation of what an area is used for.

An example of land use data would be the Corine dataset maintained by the European Environment Agency. The Corine dataset is an inventory of land cover divided into 44 classes and is publicly available for the year 2000. The data can be broadly divided into the built environment, agricultural land and natural habitat. The latter includes natural forestry and peat bog. The proportion of land in each ED that falls into these three categories can be determined using GIS. It is possible to adopt a ‘majority rules’ type approach to classify areas. If the simple majority of the land is built, then the ED is classed as built, and so forth. Where there is no clearly dominant land use type, then a combination of land use types may be used (i.e. built-agricultural, mixed, etc).

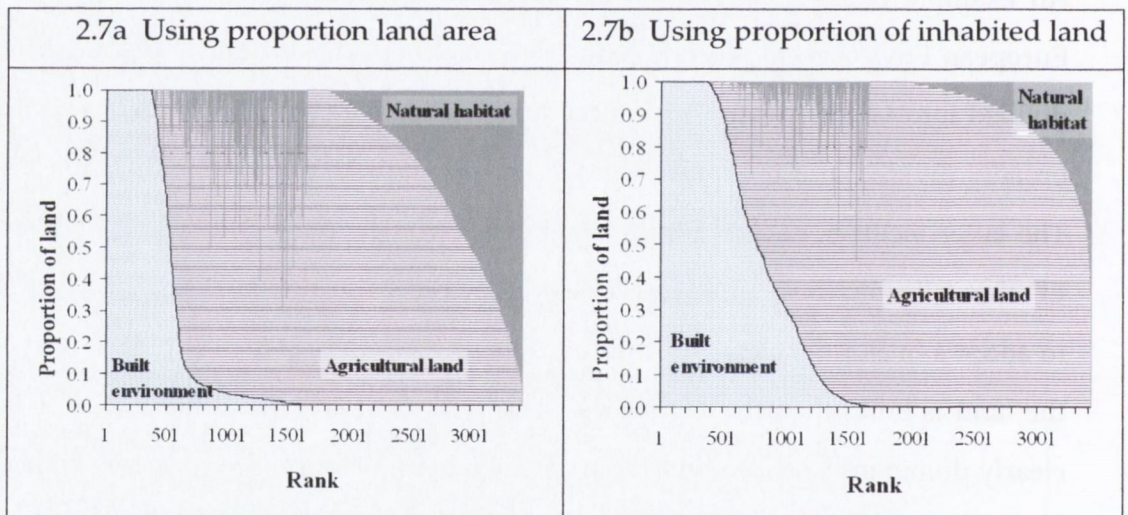
Classifying areas according to land use can be problematic. As with the definition of population density, the concept of inhabited land may be useful. An ED may be defined by the total area or by the area on which the population resides. For example, there may be the situation where the majority of the population lives on built land but the area is predominantly agricultural. In that case, the average individual experiences a built environment even though the average acre of land is



agricultural. It raises the question of whether the population or the land is being classified.

The graphs in Figures 2.7a and 2.7b show proportion land use by ED for all land and inhabited land respectively. EDs have been ranked by the proportion built environment, agricultural land and natural habitat, respectively. When using the proportion inhabited land, as shown in Figure 2.7b, almost all EDs are mainly comprised of built and agricultural land. This indicates that only a very small portion of the population live within 200m of land that is natural habitat. The most common situation is for the majority of an ED to be agricultural land, irrespective of whether the proportion of all land or only of inhabited land is used.

Figure 2.7 Percentage land use by ED (ranked by proportion built environment)



EDs were classified by both the proportion of built-land and the proportion of the houses on built-land. The former resulted in 530 urban EDs, the latter 763 EDs. Both methods give rise to different problems. EDs on the edge of a city are likely to contain substantial tracts of agricultural or non-built land. If classified by all land use they are generally labelled as rural. For example, 89% of the 18,624 persons living in Navan Rural ED are living in Navan town or its suburbs. Approximately 75% of the population lives on built land, which covers nearly 20% of land in the ED. Classified by proportion of land area it is rural, classified by inhabited land it is

considered urban. A second example is Stradbally ED in Kerry with a population of 230 living in a rural area. Just less than 6% of the land is built and yet 55% of the houses are on that land. According to land use it is clearly a rural ED but when it is classified by inhabited land it is labelled urban.

A small isolated population living in a predominantly built ED should not be classed as urban. Town size, population density and proximity to the nearest town are all ignored in this method and this gives rise to misclassification.

## 2.4 Combination methods

Data derived from the above methods (e.g. population size, population density, access and land-use) can be combined using factor analysis, principal component analysis or other multivariate variable combination techniques. These methods have the benefit that all useful data can be incorporated into a single derived variable. Such methods have been used by Cloke,<sup>273</sup> Cloke and Edwards,<sup>258</sup> and McDade and Adair.<sup>274</sup> Cloke used census indicators relating to population structure, occupancy, household amenities, occupation, migration and distance to urban centres. McDade and Adair used a large set of indicators including demography, infrastructure, household and neighbourhood amenities. The principal drawback is that if a method is used that results in reduction to a single continuous variable, it may be difficult to compare values as is the case with using a measure of access. Some combination methods look for clusters across a number of dimensions which enables retention of more information and grouping EDs with similar attribute values. As has already been discussed, it is imperative that a rural ED close to an urban centre is distinguishable from an ED containing an isolated town.

Combination methods of classification can be divided into supervised and unsupervised methods.<sup>275</sup> The latter type typically attempts to seek either convenient breaks in the data or some sort of structure which can be used to

delineate clusters in the data (e.g. cluster analysis). Supervised techniques require some form of prior knowledge about the classification.

### **2.4.1 Supervised classification**

For supervised classification it is required to have some form of prior knowledge about the classification structure. Frequently there will be a training dataset for which the classes are known and a number of variables are recorded for each observation. A model can be constructed to predict the classes using the variables and that model applied to a complete dataset for which the classes are not known. The methods used in that instance are predictive methods. In some cases classes are not known for any observation in which case there is no training dataset available. If there is a theoretical basis for developing classes then it is possible to use a supervised method such as multi-criteria classification.

#### **2.4.1.1 Multi-criteria classification**

One approach to supervised classification is multiple criteria classification (MCC)<sup>276</sup> whereby qualification criteria are specified for each class and each observation is tested to see which class it qualifies for. The criteria can incorporate numerous variables. For example, an ED might be classed as urban if the majority of the population is in a town of 1,500 or more persons and either the population density is greater than 693 persons/km<sup>2</sup> or more than 50% of the land is built. Criteria can be developed that guarantee that a point can only be eligible for a single class. The problem of choosing cut-off points between classes becomes an issue once again as the criteria must be well specified.

The MCC approach has been implemented using a number of the variables and cut-offs suggested in the previous sections. For example, if 50% or more of the population lives in a city, then the ED is labelled as urban. In consideration of different settlement types, six classes were defined for the analysis, four of which contained predominantly settled population (urban, town, near village and remote village). The remaining two classes encompass rural land divided into near rural

which is less than 15 minutes from a town, and remote rural, which is more than 15 minutes from a town. The urban class is defined by city population, where the five cities of Dublin, Cork, Limerick, Galway and Waterford are the only cities included.

Table 2.2 Criteria for MCC classification

Class	Count of EDs	% living in a city	% in town of 1,553 or more persons	% living in a settlement	Distance to the nearest town of 1,553 or more persons
Urban	472	$\geq 0.5$	$\geq 0.5$	$\geq 0.5$	< 15
Town	247	< 0.5	$\geq 0.5$	$\geq 0.5$	< 15
Near village	137	< 0.5	< 0.5	$\geq 0.5$	< 15
Remote village	75	< 0.5	< 0.5	$\geq 0.5$	> 15
Near rural	1803	< 0.5	< 0.5	< 0.5	< 15
Remote rural	688	< 0.5	< 0.5	< 0.5	> 15

A problem with a simplistic MCC approach, such as the one outlined above, is that based on the indicator means, EDs may be closer to a class other than the one they have been assigned to. In other words, some EDs may bear more similarity to the EDs of a group other than the one they have been classed in.

#### 2.4.1.2 Predictive methods

For the following methods, the MCC classification can be given as a function of several variables other than those used to construct the classification. It is then possible to predict the class of each ED given the observations in each class. For example, the urban-rural classification may be defined using the MCC approach and variable selection outlined in Table 2.2 previously. The class of each ED could then be predicted given the classification provided and a different set of variables, such as access and land-use. An ED that might be classed as urban may be re-classed as town if the values for the access and land-use variables were found to be more similar to town EDs than to urban EDs.

A number of these methods result in a probability of an ED being in each class. The class with the highest probability is then the predicted class. The manner in which the probabilities are calculated depends on the method used. The five methods considered are as follows:

- Logistic regression<sup>277</sup> – a method for predicting the probability of a binary dependent variable using a set of independent variables
- Discriminant analysis<sup>278</sup> – a method that examines the set of predictors and uses similarities and differences to assign each observation to one of a set of classes
- Classification tree<sup>277</sup> – a method used to predict membership of cases or objects in the classes of a categorical dependent variable from their measurements on one or more predictor variables
- Partitioning<sup>279</sup> – a method to recursively partition data according to a relationship between the categorical dependent variable and the set of independent variables
- Neural network<sup>279</sup> – a method to predict response variables from a flexible network of non-linear functions of input variables

For the five methods listed above, the independent variables can typically be a combination of continuous and categorical variables. All five methods were applied using the MCC classification along with three log-transformed predictor variables: median town size, population density and access to settlements within 48km. The median town size was calculated as the size of the settlement that contains 50% of the cumulative ED population when settlements are ranked by population numbers (see page 68 for an illustrative example). Two statistical packages, JMP 5.0.1<sup>280</sup> and SPlus 6.0,<sup>281</sup> were used to perform all of the calculations.

In Table 2.3 the numbers of EDs in each class for the different methods are shown. The most apparent disparity occurs for the partition method, which results in quite a different classification from the other methods.

Table 2.3 Count of EDs by class for each predictive method of classification

Class	MCC	Discriminant analysis	Logistic regression	Classification tree	Neural network	Partition
1	472	446	491	470	473	474
2	247	238	302	261	246	245
3	137	119	127	135	164	212
4	75	128	11	32	48	1,565
5	1,803	1,455	2,312	2,038	2,225	737
6	688	1,036	179	486	266	189

This is not exactly the correct application of these methods as they are intended to predict known classes using observed data – thus allowing classification of unclassified datasets with the same observed variables. In this case, the method attempts to identify EDs that, based on the observed attributes, are misclassified when compared to other EDs within the same class. If the MCC was based on the same three variables as the other methods, then the predictive approaches would be able to perfectly predict the MCC classification which would defeat the purpose of the exercise.

#### 2.4.2 Unsupervised classification

A number of unsupervised clustering methods exist for dealing with multi-dimensional data. One option is to use k-means type clustering.<sup>278</sup> In this method the user specifies the number of clusters they want to identify. For example, let there be  $m$  variables and we wish to specify  $n$  clusters. The first step is to select  $n$  random points representing cluster centroids in the  $m$ -dimensional space. Each data point is allocated to the nearest cluster centroid. The centre of gravity is calculated for each cluster and that becomes the new cluster centroid. The last two steps are repeated iteratively until the centroids show negligible change between iterations. The main drawback of this method is that the user specifies the number of clusters even though it will probably not be known in advance how many clusters there are or need to be identified.

Given that six classes were developed using the MCC approach, k-means clustering was applied to the same six variables to allocate EDs to six clusters. The

results are shown in Table 2.4 where the clusters are compared with the MCC classes. A number of urban EDs are classified with the majority town EDs. Similarly, a number of town EDs are classified with the villages for which near and remote are not distinguished. This leaves three rural categories, the first of which includes EDs that have some settlement but not sufficient to be labelled as village.

Table 2.4 Comparison of ED counts for classes and K-means clusters

Class	K-means cluster						Total
	1	2	3	4	5	6	
Urban	366	106	0	0	0	0	472
Town	0	150	97	0	0	0	247
Near village	0	0	137	0	0	0	137
Remote village	0	0	75	0	0	0	75
Near rural	0	0	0	544	1,181	78	1,803
Remote rural	0	0	0	79	437	172	688
Total	366	256	309	623	1,618	250	3,422

A variation on k-means called self-organising maps (SOMs)<sup>282</sup> is also available for classification applications in JMP.<sup>280</sup> In SOMs the clusters have a grid structure which can aid interpretation of the clusters. Essentially it generates a two dimensional output where clusters that are close in multivariate space are shown close together in the SOM grid.

An alternative to iterative techniques are agglomerative methods of clustering, such as hierarchical clustering.<sup>283</sup> In these methods all data points are initially individual clusters. At each step of the process the two clusters closest to each other in multidimensional space are combined into a single cluster. This process continues until all points have been combined into a single cluster. There is no search for the optimum number of clusters so it is possible to subjectively select how many clusters will be defined. There are numerous methods for measuring the distance between two clusters so depending on the metric used the results will be different.

As with the k-means clustering, the hierarchical clustering is compared to the MCC classes in Table 2.5 below. The first two clusters represent urban areas, giving 490 urban EDs, with the third cluster representing towns. Again the near and remote villages are merged along with a number of the EDs classed as town using the MCC approach. There are two rural clusters but the distribution between them is quite different as the separation is dictated by the access scores rather than distance to the nearest town.

Table 2.5 Comparison of ED counts for classes and hierarchical clusters

Class	Hierarchical cluster						Total
	1	2	3	4	5	6	
Urban	289	180	3	0	0	0	472
Town	0	21	138	88	0	0	247
Near village	0	0	0	137	0	0	137
Remote village	0	0	0	75	0	0	75
Near rural	0	0	0	0	439	1364	1803
Remote rural	0	0	0	0	58	630	688
Total	289	201	141	300	497	1994	3422

### 2.4.3 Comparing classifications

There are methods for comparing cluster allocations which can provide a basis for choosing one technique over another or, indeed, to select the choice of how many clusters to use. One such measure is the goodness of variance fit (GVF),<sup>284</sup> outlined in equation 2.2 below. Values of GVF range from 0 for the very poorest fit to 1 for a perfect fit. In the case of a single class the GVF will be 0 whereas the GVF will equal 1 when N areas are allocated to N classes.

$$GVF = 1 - \frac{\sum_{h=1}^k \sum_{i=1}^{N_h} \sum_{j=1}^m (x_{hij} - \bar{x}_{ij})^2}{\sum_{i=1}^N \sum_{j=1}^m (x_{ij} - \bar{x}_j)^2} \quad (2.2)$$

Where:  $x$  = indicator value



$\bar{x}$  = mean indicator value

$k$  = number of clusters

$M$  = number of indicators

$N$  = number of areas

A similar measure is the tabular accuracy index (TAI)<sup>284</sup> which differs from the GVF in that it uses Manhattan rather than Euclidean distance. Another useful measure is Akaike's Information Criterion (AIC)<sup>285</sup> which is defined in 2.3 below. For the AIC, a lower value represents a better fit.

$$AIC = N \cdot \ln \left( \frac{\sum_{h=1}^k \sum_{i=1}^{N_h} \sum_{j=1}^m (x_{hij} - \bar{x}_{ij})^2}{N} \right) + f(k, N) \quad (2.3)$$

Where:  $x$  = indicator value

$\bar{x}$  = mean indicator value

$k$  = number of clusters

$m$  = number of indicators

$N$  = number of areas

$f(k, N)$  = penalty function

The AIC includes a penalty function which is a function of  $k$ , the number of clusters, and  $N$ , the number of areas. The AIC, Bayesian Information criterion (BIC), Hannan & Quinn's criterion (HQC), and the Generalized Cross Validation criterion (GCVC) penalty functions are shown below.<sup>286</sup> In all cases the natural logarithm is generally used.

$$AIC, f(k, N) = \frac{2k}{N} \quad (2.4)$$

$$BIC, f(k, N) = \frac{k \cdot \log(N)}{2} \quad (2.5)$$

$$HQC, f(k, N) = k \cdot \log(\log(N)) \quad (2.6)$$

$$GCVC, f(k, N) = -N \cdot \log\left(1 - \frac{k}{N}\right) \quad (2.7)$$

The intention of the penalty function is to counter the improvement in fit afforded by an increased number of clusters. If too many clusters are identified there is the risk of developing a classification that is either too unwieldy or contains classes with too few members to be of real use. Depending on the penalty function chosen, the optimal number of clusters may vary. The AIC is more typically used to identify the optimal number of parameters to be included in a model but has also been used in classification problems.<sup>285</sup>

A small increase or decrease in  $k$  does not have a substantial impact on the penalty function when  $N$  is large, such as in the current case where  $N$  is 3,422. Given the data being used in the current classification problem, the penalty functions outlined above result in values that are too small to identify a benefit for smaller numbers of classes. The penalty functions listed above generate values under 100 for  $N = 3,422$  and  $k$  less than 25. In the current context, the AIC values are in the order of 1,000s rendering the above penalties functions ineffective for identifying an optimal number of classes. It is possible to develop a penalty function that will suit the size of the data and the following function is proposed for the current case:

$$f(k, N) = \frac{k \cdot N}{\log(N)} \quad (2.8)$$

A difficulty with the AIC is that depending on the data being used, the upper and lower bounds change so it is not possible to know what the best possible solution is without searching for the optimal result. For example, with the GVF and TAI

measures it is known in advance that allocating  $n$  areas to  $n$  classes will result in an perfect classification which will return a GVF and TAI of one. Due to the penalty function, the allocation of  $n$  areas to  $n$  classes should not result in a maximum AIC value. When comparing a number of methods of allocating  $n$  areas to  $k$  classes, it is not possible to quantify 'how good the best solution is' unless the best and worst possible solutions have been found.

The measures of fit outlined above have two purposes: they make it possible to compare different classification methods and to compare different numbers of classifications. The GVF and TAI are most useful for comparing different methods as an increased number of classes will always return an improved measure, making them unsuitable for the latter purpose. The AIC can be used for both purposes.

The three measures of fit have been calculated for all of the multivariate classification methods outlined above. As this analysis was only for comparative purposes only three variables are used: median town size, population density and access to settlements within 48km. The variables were log transformed for all methods other than MCC. In all cases the computations are for 3422 EDs classed into six groups. The figures are given in Table 2.6 and are sorted by GVF in descending order. Hierarchical clustering was performed using five different distance metrics: ward, average, centroid, complete and single.

Table 2.6 Measures of fit for each classification method

Method	GVF	TAI	AIC
K-means	0.982	0.879	942.0
Hierarchical (ward)	0.980	0.878	1304.3
Hierarchical (average)	0.979	0.865	1579.8
Discriminant analysis	0.976	0.860	1910.2
Hierarchical (complete)	0.976	0.860	1996.7
Partitioning	0.975	0.858	2109.1
Self Organising Map	0.974	0.869	2200.7
Hierarchical (centroid)	0.972	0.845	2452.1
Neural network	0.972	0.845	2502.3
MCC	0.967	0.836	3006.4
Logistic regression	0.948	0.803	4610.5
Hierarchical (single)	0.938	0.796	5198.2

The ranking of methods are the same for GVF and AIC and slightly different for TAI. The k-means method achieves the best fit of the unsupervised techniques, whereas the best ranked of the supervised classifications is the discriminant analysis. That k-means has the best fit is to be expected given that it attempts to minimise the variance within clusters and the measures of fit all incorporate variance to some degree. Hierarchical clustering using ward distances provides a similar fit to k-means clustering.

It is necessary to decide whether k-means clustering offers a classification that is advantageous over the other methods. This can be achieved by means of a simulation exercise. If there are  $k$  classes, it is possible to randomly select  $k$  data points and label them as centres. Each data point can then be allocated to the nearest centre to produce a classification for which the measures of fit can be evaluated. If a suitable number of simulations are evaluated, such as 1000, it would be possible to determine whether a given classification is better than one produced by picking random class centres. If the given classification does not offer an improvement over classifications produced by picking random class centres, then the utility of the classification is highly questionable.

A simulation exercise was performed using the three variables used to produce the unsupervised classifications. Centres for six classes were chosen at random from the 3422 data points with each ED allocated to its nearest centre. Euclidean distances were used in the calculations. The GVF, TAI and AIC were recorded for each of 999 simulations. The corresponding measures for each of the 13 classification methods were ranked in turn with the simulation results, to give each method a ranking within a thousand classifications. The position of the fit on the scale of worst to best fit is also given alongside the ranking in Table 2.7 below. The rankings are the same for GVF and AIC due to the similarities of the two formulae.

From Table 2.7 it can be seen that k-means clustering provides the best results based on all three measures of fit. Each method is ranked against 999 randomly generated classifications. From this rank it is possible to calculate the probability that a particular method is significantly better than a random classification. To be considered as significant at the 0.05 level a method would have to be ranked in the top 50 classifications. Only the k-means and ward distance hierarchical clustering methods are significant for all three measures of fit and none of the supervised classification methods are significant for any of the measures.

Table 2.7 Rank and position of measures of fit against simulations

Method	GVF		TAI		AIC	
	Rank	% of best	Rank	% of best	Rank	% of best
K-means	1	100.0	9	99.6	1	100.0
Hierarchical (ward)	10	99.8	11	99.5	10	97.6
Hierarchical (average)	30	99.6	62	97.7	30	95.4
Discriminant analysis	66	99.3	100	96.9	66	92.9
Hierarchical (complete)	76	99.2	103	96.9	76	92.2
Partitioning	83	99.1	125	96.5	83	91.4
Self organising map	89	99.0	48	98.1	89	90.7
Hierarchical (centroid)	120	98.8	318	94.6	120	88.7
Neural network	125	98.7	317	94.6	125	88.4
MCC	250	98.2	366	93.3	250	84.5
Logistic regression	395	95.7	448	88.5	395	72.2
Hierarchical (single)	430	94.5	451	87.4	430	67.7
Classification tree	448	93.4	451	87.5	448	64.1

The magnitude of the GVF values is perhaps misleading. It has been noted that in the case of a single class, the GVF will equal zero. However, the GVF, TAI and AIC are 0.8777, 0.7197 and 6007.8 respectively when k-means clustering is applied to allocate EDs to two groups. These values are therefore a better indication of what is a poor fit than the single class case. The GVF, TAI and AIC values for k-means clustering to a range of numbers of classes is shown in Table 2.8 below. There is a substantial improvement in fit when the number of classes is increased from three to four, after which the benefit of additional classes diminishes. Based on the AIC using the penalty function described in equation 2.8, the optimal number of classes is six. Using GVF and TAI, 3,422 classes would result in a perfect fit (i.e. the GVF and TAI would both be equal to 1). It can be seen that GVF values of over 0.99 are obtained using only 10 classes. The values for TAI, however, increase at a slower rate reflecting the different distance metric used in the formula.

Table 2.8 Comparison of measures of fit for different numbers of classes identified using K-means clustering

Classes	GVF	TAI	AIC
2	0.878	0.720	6007.8
3	0.887	0.743	6701.5
4	0.967	0.840	3510.4
5	0.971	0.852	4026.4
6	0.982	0.879	3299.3
7	0.984	0.889	3880.0
8	0.988	0.904	3743.0
9	0.989	0.909	4447.8
10	0.991	0.913	4982.4
...	...	...	...
15	0.994	0.927	8571.4
...	...	...	...
20	0.995	0.935	12521.1
...	...	...	...
25	0.996	0.941	16654.7
...	...	...	...
30	0.996	0.944	21047.9

Given the measures of fit observed, the previous analysis suggests a choice between k-means clustering (unsupervised classification) and discriminant analysis applied to MCC classification (supervised classification). The principal advantage of using the supervised approach is that there is control over the definition of each class. While it can be argued that k-means clustering may produce a better fit than discriminant analysis, it is not a substantial improvement and it might not generate meaningful classes. Furthermore, cross-temporal comparisons may not be appropriate when classes have been generated independently using unsupervised classification. For the generation of an urban-rural index, it is proposed that discriminant analysis applied to MCC is the most suitable methodology.

## 2.5 Proposed method of urban-rural classification for Ireland

In generating an urban-rural classification for Ireland, a number of factors need to be considered. To classify areas based on an urban-rural dichotomy would disregard the variety in area types that occur outside the main urban centres. It is therefore appropriate that multiple classes are used. The use of a single indicator

such as population density for classification purposes will undoubtedly lead to poor discrimination between classes. Multiple indicators allow for some degree of validation and hence should reduce the amount of misclassification. A spatial component should be included in the indicators as the location of an area relative to others is important. Proximity to a major urban centre, for example, can have major implications for the lifestyles of inhabitants of that area. The classification should be independent of deprivation measures if it is to be used to compare poverty levels in different area types.

A hierarchical or collapsible class system is preferable so that related area types can be merged to reduce the number of classes without overly compromising distinction between the merged classes. For example, it should be possible to combine near and remote rural areas to obtain a single rural class. This may be particularly useful when considering subsets of EDs where the numbers in any one class may be too small to be of use. It is also preferable that any classification method be applicable to multiple geographic scales. This would mean that it could be applied at ED, district, county or even regional level. At higher geographic levels, a simple two or three class system might be more appropriate for which the notion of collapsible classes is also important.

It should also be possible to perform valid temporal comparisons. If classifications for two time periods are obtained in a different manner then it will not be possible to investigate the change in rural-urban patterns.

### **2.5.1 The choice of indicators**

A range of indicators has been selected that are obtainable and comparable across a range of time periods. The indicators have been chosen on the grounds that they each provide a different method of potentially discriminating between different types of urban and rural area. The following set of indicators is used for classification:



- *Population living in each settlement by ED*

The population living in each settlement size by ED is derived from the 1986, 1991, 1996 and 2002 censuses. It is not a complete listing of settlements in the country but it does include all settlements that fulfil a number of criteria regarding layout and size. The definition of a settlement was given in section 2.2 previously.

- *Median settlement size*

The median settlement size is calculated using information on the settlements in an ED as recognised in the census. The method of calculation is best described with an example using the Carrigtohill ED in county Cork. In 2002 the ED had a population of 3,507 of whom 1,477 lived outside a defined settlement and the remainder living in one of three settlements. The details are provided in Table 2.9 in which the figures are sorted by settlement size. The population living outside a defined settlement is given a nominal settlement size of one. Alternatively the average household size for that ED could be used. This figure is merely used to denote a single household settlement. The size of the settlement that contains 50% of the cumulative ED population is given as the median settlement size. In this example that is Carrigtwohill, giving a median settlement size of 1,411 to the ED.

Table 2.9 Calculation of median settlement size for Carrigtohill ED

Area	ED population in settlement	Total settlement population	% ED Pop. in settlement	Cumulative % ED pop. in settlement
Non-settlement	1,477	1	42.1	42.1
Carrigtwohill	1,411	1,411	40.2	82.3
Middleton	20	7,957	0.6	82.9
Cork City	599	186,239	17.1	100.0

The above formulation gives a more accurate summation of settlement size than using either a simple average or population weighted average of the

settlements. Similarly, using a simple median of the settlement sizes does not take into account the numbers living in each settlement.

- *Population density*

The population density is calculated as the total population divided by the total inhabited area. The inhabited area is the total area not including land over 300m or water bodies such as lakes. This is notionally the population density of the inhabitable land area.

- *Access to essential services*

The access indicator for a given ED is restricted to settlements within 48km of the ED centroid. The indicator is the sum of ratios of log transformed settlement population to distance squared, given in equation 2.1 previously.

- *Land use*

Land-use values are divided into a number of categories: artificial, agricultural crops, agricultural pasture, natural habitat and wetlands. For each category, the proportion land area is recorded by ED. The values are derived from the 2000 Corine dataset.<sup>287</sup> The degree of land use change between 1990 and 2000 is deemed to be small enough that the 2000 patterns are representative of the 1990 patterns with the exception of urban fringes, where the man-made environment has impinged on the surrounding rural areas.

## 2.5.2 The method of combination

It is proposed to use a combination of supervised classification and rule based partitioning to classify areas. It has been shown above that a combination of MCC and discriminant analysis can produce a classification that has similar fit to the data as when k-means clustering is used. Further sub-division of classes using a simple rule-based system will produce a classification that will facilitate intuitive class aggregation.

The application of MCC requires a prior knowledge of the class structure. For this purpose, it is necessary to outline, in advance, the classes and which EDs should be members of those classes. The following section will describe the classes that EDs are to be allocated to. It is intended to have four main classes, each of which can be further sub-divided based on additional criteria.

**City** – this will describe EDs that are in the five cities or have the same characteristics as city EDs on the basis of population density, access and settlement size. Cities are distinct in that they are commercial and industrial hubs of activity and are largely self-sufficient.

**Town** – this class will apply to larger settlements that do not qualify as cities but have substantial populations and have high population densities and relatively high access. Towns are a step down from cities although they can also be hubs of commercial activity. They tend to have good connectivity to other towns and cities but lack some of the economic drivers available to larger urban centres.

**Village** – this class corresponds to EDs that are predominantly settled but do not qualify as city or town EDs. These EDs can be further divided into near and remote, to distinguish those that are near and far respectively from towns and cities.

**Rural** – all EDs that are not predominantly settled will be classed as rural. As this grouping of areas is anticipated to have a large degree of diversity, it will be divided based on primary land use and proximity to towns and cities. Land use can be sub-categorised as agricultural crops, agricultural pasture, natural habitat and wetlands. Each land use type has implications for earning potential and general utility. The division of rural areas into near and remote will, as for villages, be based on access to towns and cities in the neighbouring area. Thus the rural EDs can potentially be sub-divided into eight classes. The eight groups have a number

of potential aggregations such as farmed and non-farmed or simply near and remote, depending on the application.

The method of allocating EDs to a class is initially hierarchical to generate the four main groups. Once EDs have been given the main designation, additional information is used to give further breakdowns within each main class.

### **Step 1. Identify city EDs**

All EDs where 50% or more of the population live within one of the five cities or suburbs thereof as legally defined by the government are designated as urban.

### **Step 2. Identify the town EDs**

The census identifies 75 towns with legally defined boundaries. Due to expanding populations in many towns, the defined boundaries often do not encompass the entire area occupied by the town. In the collection of census statistics, the population residing within a town but outside the defined boundary are included in the statistics for that town.

The group of EDs for which 50% or more of the population live in a town with legally defined boundary was subsetted. The minimum median town size, population density and access scores were determined for these EDs. From the full set of EDs, any ED not already identified as a city ED and for which the median town size, population density and access scores are all greater than the minimum identified was designated as a town ED.

### **Step 3. Identify village EDs**

All EDs with 50% or more of the population living in settlements and not already identified as city or town were designated as village EDs. Using k-means clustering, the village EDs were split into two classes based on the access scores. This discriminates between near and remote village EDs.

#### **Step 4. Designate to rural**

All EDs not designated as city, town or village were designated as rural. The rural EDs were split into near and remote by applying k-means clustering to the access scores.

#### **Step 5. Apply discriminant analysis**

Discriminant analysis was applied to predict the 4 classes using median town size, population density and access scores. This effectively smoothes the figures by moving borderline cases to a class that is closer in multi-dimensional space.

#### **Step 6. Identify remote EDs**

The village and rural EDs can be divided into near and remote EDs, based on their proximity to settlements. An ED that is distant from a city or town may be considered remote. Identification of remote EDs can be achieved using access scores, which measure both the proximity to settlements and the size of those settlements. A high score can indicate close proximity to many small to medium sized settlements or proximity to a single large settlement such as a city. By performing k-means clustering on access scores it is possible to divide EDs into near and remote groups.

#### **Step 7. Identify land use types**

City and town EDs are considered separately from village and rural EDs. As city and town areas are assumed to be predominantly built-up, the concern is with the type of man-made environment rather than the natural environment.

For city and town areas, three dominant kinds of land use were identified: commercial, industrial and residential. Commercial buildings were identified by the building use designation in the Irish GeoDirectory database. The city and town EDs were split into commercial and non-commercial by applying k-means clustering to the proportion commercial buildings looking for two clusters. Of those that were predominantly non-commercial, they were then divided into

industrial and residential, based on the predominant land use pattern derived from the Corine database.

For village and rural areas, four land use types are identified: agricultural crops, agricultural pasture, natural habitat and wetlands. For each ED, the proportion of land in each category is calculated. If agricultural crops and pasture cover more than 50% of the land, then the ED is designated as crop or pasture according to which is the dominant type of the two. Otherwise the ED is designated as natural or wetland, depending on which is dominant. There are cases where none of the four land use types dominate, in which case the choice between agricultural and non-agricultural is important. Agricultural land provides inhabitants with a means of earning from the land, while areas of natural habitat frequently do not.

### **2.5.3 Temporal calculations**

Some of the data used for the further breakdown of classes in the previous section were not available prior to the 2002 census. This applies most particularly to the Irish GeoDirectory data, used to identify commercial EDs in city and town areas.

The proposed method for cross-temporal analysis is to concatenate the data across four censuses for identifying cut-off points for town EDs. This also applies to the k-means clustering to distinguish between near and remote village and rural areas.

### **2.5.4 Application and results**

The methodology described above has been applied separately to the 2002 data and the combined 1986 to 2002 data. The reasons for the distinction are that the common set of EDs for 1986 to 2002 merges some EDs that are quite distinct and that additional information is available for 2002 which further enhances the classification.

#### **2.5.4.1 Urban and rural areas 2002**

The first step was to classify EDs as city based on the population residing in the five cities. A total of 488 EDs have all or some population residing in a city. In 474

EDs the simple majority of the population lived in a city and these EDs were classified as city.

The next step was to identify town EDs based on the census definition of towns with legally defined boundaries. Of the 248 EDs with a portion of the population living in one or more of the 75 defined towns, 168 had a predominantly town dwelling population. The minimum values for median town size, population density and access score (all log-transformed) were 6.921, 5.128 and 0.188 respectively. All EDs where all three attribute values were above the minima defined and were not previously classed as city were classed as town. This gave 297 town EDs.

All EDs with more than 50% of the population living in settlements and not previously classed as city or town were classed as village. This gave 160 village EDs. The remaining 2,491 unclassified EDs were labelled rural. For all of the rural EDs less than 50% of the population resides in a defined settlement.

Application of discriminant analysis using the four categories given above and with the log-transformed median town size, population density and access scores as covariates resulted in the reclassification of 79 EDs. Seven EDs classified as urban, all in Waterford city, were reclassified as town. One village EDs was reclassified as town and 71 town EDs were reclassified as village.

The reclassification of seven Waterford EDs from city to town appears problematic, as the government has legally defined it as a city equal in status to Dublin, Cork, Limerick and Galway. It is, however, closer in size to the towns of Drogheda, Dundalk and Bray than to the next smallest city, Galway. Although it has city status, it is more similar to a town than any of the other cities in terms of population and services. As such, it is not unreasonable to reclassify Waterford city EDs as town EDs. The more difficult problem is that not all of Waterford has been reclassified, only those EDs with lower population densities and access scores have

been reclassified. Due to its smaller population, its access scores will be more like those of Dundalk and Drogheda and less like those of the other cities. For the fringe EDs, the combined impact of lower population density and lower access is to give them the attributes of a town ED rather than a city ED. Given the size of the city it is not unreasonable to say that it bears similarities to both a town and a city, and that different parts of the same settlement may be classed differently.

Unlike county councils, city boundaries are not defined by ED boundaries. Therefore the legally defined boundary of a city might include only part of an ED. It is possible for an ED to be part of a city council area but not have the attributes of a city ED. This can arise if the city only impinges on a small part of the ED with the majority of the population living outside of the town.

The next question is: does the reclassification using discriminant analysis provide a significantly improved fit to the data over the MCC classification? A comparison of the GVF, TAI and AIC measures for the MCC, discriminant analysis and a simple four class k-means are compared in Table 2.10 below. The table also includes the rank within 999 simulations of random clustering using methodology described in section 2.4.3 above.

Table 2.10 Comparison of measures of fit for three clustering methods

Clustering method	GVF	TAI	AIC	Rank GVF	Rank TAI
MCC	0.9637	0.8291	1597.3	77	117
Discriminant analysis	0.9663	0.8356	1345.9	21	60
K-means	0.9665	0.8396	1319.5	16	28

The difference in the GVF and TAI measures are small. The AIC measure, however, reveals a more significant difference with the discriminant analysis being much closer to the k-means than the MCC result. Both the k-means and discriminant classifications are significant at the 0.05 level for GVF, but only the k-means is



significant for TAI. Given the improvement in fit offered by the discriminant analysis, it is chosen over the MCC classification.

The next step is to analyse the village and rural EDs to establish cut-offs to distinguish between near and remote EDs. The log-transformed access scores were split into two groups using k-means clustering. Village EDs will generally have better access values than rural EDs because they contain at least one settlement, giving them a good access to at least one settlement and the services and opportunities available in that settlement.

Table 2.11 Counts and minimum access scores for near and remote village and rural EDs

Class	Near		Remote	
	Count	Minimum access score	Count	Minimum access score
Village	159	1.192	71	-1.029
Rural	1,301	-0.279	1,190	-3.939

The land use types for village and rural EDs are crops, pasture, natural habitat and wetlands. The land uses are initially grouped into crops & pasture and natural & wetland to decide what the dominant land use is, then the sub-classification is decided based on which aspect of the dominant land use is more abundant. A land use type is measured as the proportion of land used for that purpose in an ED. The land use types for city and town EDs are commercial, industrial and residential. Commercial status is determined using GeoDirectory data whilst all other land use types are evaluated using the Corine dataset. Industrial and residential land are measured as proportions of total ED land area while commercial is measured as the proportion of buildings categorised as commercial or commercial-residential mix. The proportions were then split into commercial and non-commercial using k-means clustering.

The scheme for designating land use is given in the flow chart in Figure 2.8 below. Each ED is given one of seven land use categories. In Table 2.12 the number of EDs in each category is given.

Figure 2.8 Scheme for classifying EDs by land use

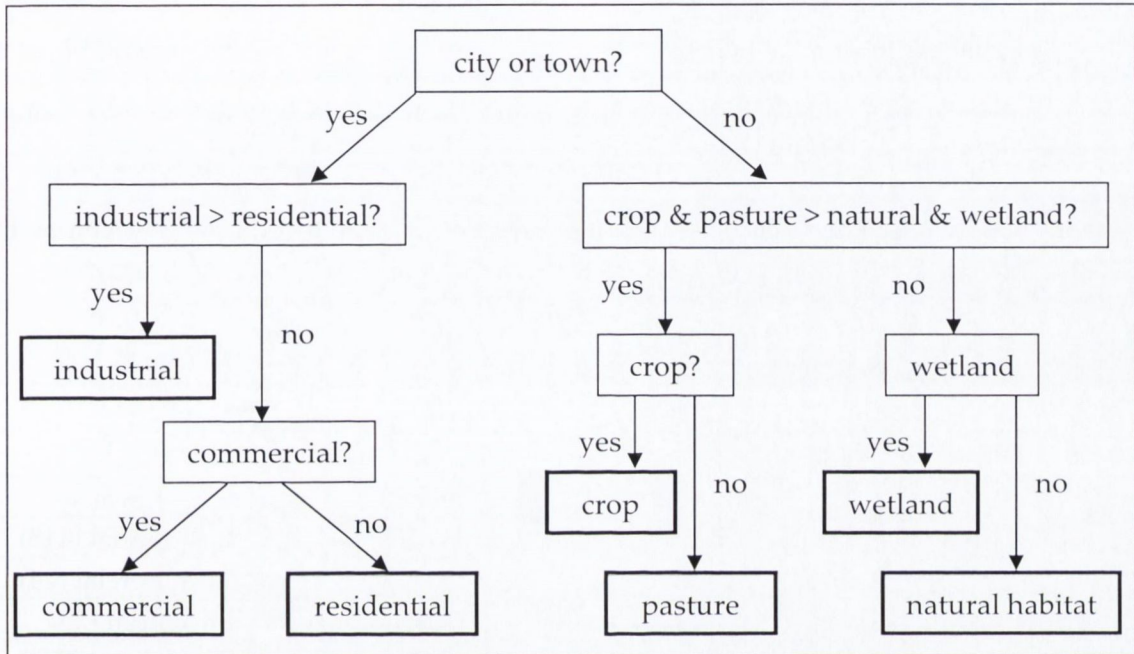


Table 2.12 Count of EDs in each land use category

Land use	Count of EDs
Commercial	29
Crop	421
Industrial	42
Natural	207
Pasture	1,752
Residential	630
Wetland	341

By combining the information on the four main groups, remoteness and land use categories, it is possible to categorise all EDs into one of 22 groups. The count of EDs in each group is given in Table 2.13 below. Moving from left to right across the table, any individual category can be further broken down into sub-categories which are labelled levels. It is not required for all classes to be broken down to the

same level of detail. For example, an index could be constructed with city and town EDs at level 1, village EDs at level 2 and rural EDs at level 3. There is no specific equivalence within any one level. As no city or town ED can be labelled remote, there is one level less for these EDs. As many of the level 4 groups are very small, they may prove to be of little practical use in most applications. This applies particularly to town and village EDs. There is a potential for a level 0 to provide a binary classification. For this classification city and town would be combined to form the urban class leaving the village and rural classes to combine into a single rural class. However, the distinction between different settlement types would be lost and this is not desirable. For this reason, the four level scale is given as the highest level of class aggregation.

Table 2.13 Class and sub-category structure (ED counts in brackets)

Level 1	Level 2	Level 3	Level 4	
City (467)		Residential (418)	Residential (418)	
		Other (49)	Commercial (24) Industrial (25)	
Town (234)		Residential (212)	Residential (212)	
		Other (22)	Commercial (5)	
			Industrial (17)	
Village (230)	Near (159)	Crop & pasture (148)	Crop (26)	
			Pasture (122)	
		Natural & wetland (11)	Natural (5) Wetland (6)	
	Remote (71)		Crop & pasture (49)	Crop (8)
				Pasture (41)
			Natural & wetland (22)	Natural (6) Wetland (16)
Rural (2,491)	Near (1,301)	Crop & pasture (1,178)	Crop (273)	
			Pasture (905)	
		Natural & wetland (123)	Natural (55) Wetland (68)	
	Remote (1,190)		Crop & pasture (798)	Crop (114)
				Pasture (684)
			Natural & wetland (392)	Natural (141) Wetland (251)

The percentage population in each class is given in Table 2.14 below. Outside the residential city and town EDs, the largest class in terms of population is the near rural crop and pasture EDs.

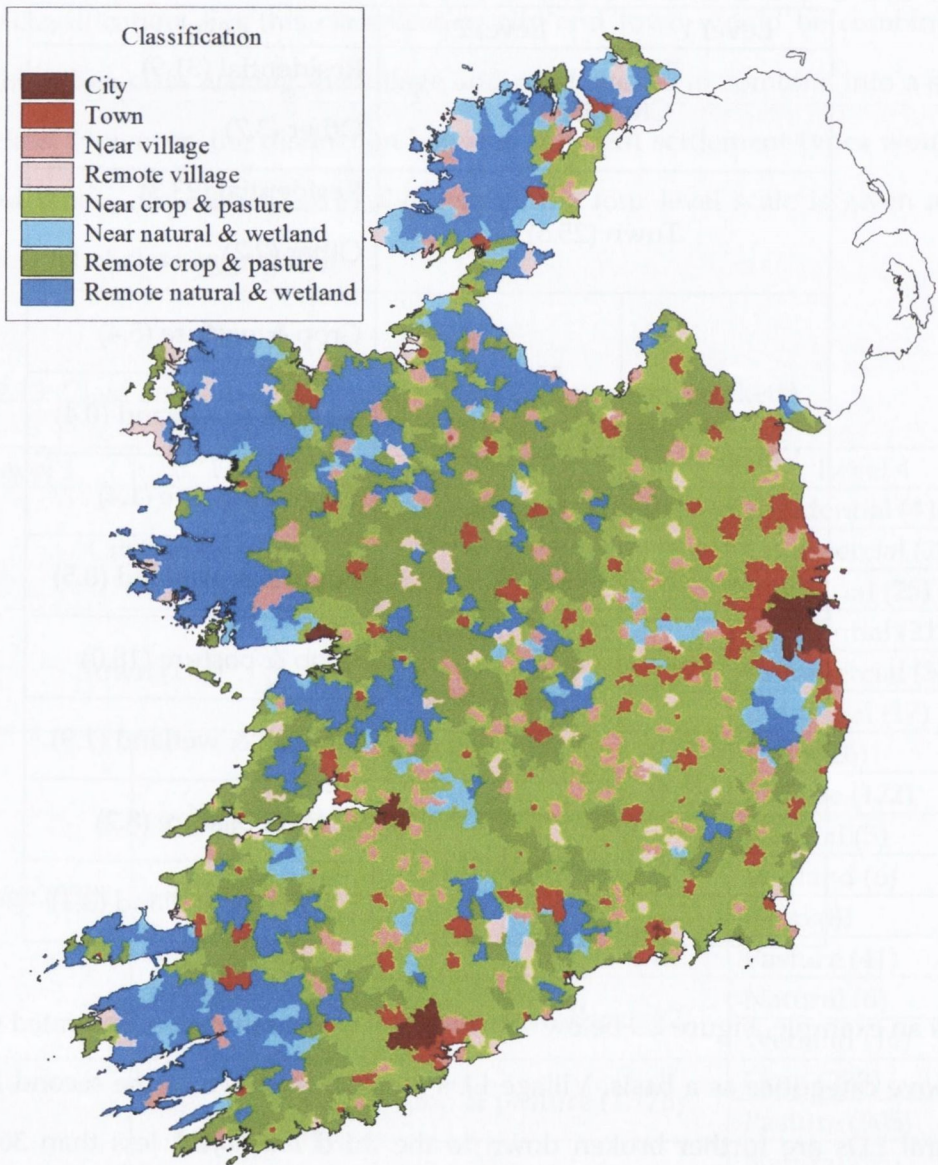
Table 2.14 Class and sub-category structure (percentage of total population in brackets)

Level 1	Level 2	Level 3
City (35.6)		Residential (31.9)
		Other (3.7)
Town (25.5)		Residential (23.3)
		Other (2.2)
Village (7.7)	Near (5.8)	Crop & pasture (5.4)
		Natural & wetland (0.4)
	Remote (1.9)	Crop & pasture (1.4)
		Natural & wetland (0.5)
Rural (31.2)	Near (19.9)	Crop & pasture (18.0)
		Natural & wetland (1.9)
	Remote (11.3)	Crop & pasture (8.2)
		Natural & wetland (3.1)

As an example, Figure 2.9 below shows an 8 level classification generated using the above categories as a basis. Village EDs are broken down to the second level and rural EDs are further broken down to the third level. Just less than 36% of the population live in city EDs with a further 25.5% living in town EDs. A further 18.0% of the population live in near crop & pasture areas. This means that 79.0% of the population lives in a town or city area or the agricultural areas close to those towns and cities. Only 5.1% of the population live in predominantly natural & wetland areas although considering that this is marginal land with little farming

prospect, that is not surprising. Of the 7.7% of the population living in village EDs, only a quarter are in remote village EDs. The remote villages tend to be larger but more isolated than near villages.

Figure 2.9 Map of 8 category classification



This leaves the question of how to derive a class for any grouping of EDs. For example, how might the ED level classification be aggregated to produce a county level classification? There are two possible approaches which are best described using the example of county Carlow. Table 2.15 gives the proportion of population

in each class and subclass for county Carlow. It is also required to decide on the level of classification to be used. For this example county Carlow will be classified to level two (i.e. one of six classes).

Table 2.15 Percentage population in each class and sub-class for county Carlow

Level 0	Level 1	Level 2	Level 3	Level 4
Urban (49%)	City (0%)	City (0%)	City (0%)	City (0%)
	Town (49%)	Town (49%)	Residential (49%)	Residential (49%)
Rural (51%)	Village (5%)	Near (4%)	Crop & pasture (4%)	Pasture (4%)
		Remote (1%)	Crop & pasture (1%)	Crop (1%)
	Rural (45%)	Near (41%)	Crop & pasture (41%)	Crop (27%)
			Pasture (14%)	
		Remote (4%)	Crop & pasture (3%)	Pasture (3%)
			Natural & wetland (1%)	Wetland (1%)

The first approach would be to focus on the column for the desired level and identify the class with the greatest percentage population. For level two in this case that would be town. At level 2, the most representative class is town.

The alternative method, which is hierarchical, would be to start at the lowest level (i.e. level zero), and identify the predominant class – in this case rural. Moving up a class, one would only consider the subgroups of the predominant class in the previous level. Thus at level one, the classes at levels one and two would be rural and near rural. The advantage of this method is that it is consistent – the class at any given level will not contradict the class found at a different level.

The prime distinction between the two methods is that the first bases the class on a snapshot taken at the desired level of classification, while the second method arrives at the class by navigating through the levels of classification. Whether you classify at level 1, 2 or 3, the first method will class Carlow as town (or residential town at level 3). Using the second method it will be classed as rural, near rural, near rural crop & pasture or near rural crop, depending on the level of

classification chosen. Only 27% of the population live in near rural crop EDs while 49% live in residential town EDs. Classifying 49% of the population based on location of 27% of the population might be too misleading. While the majority of the population live in rural areas, at level 1 or above the predominant class is town.

Given the implications of the two methods of combining classes for aggregations of EDs, it is recommended to first decide on the desired class structure and then classify by calculating the simple probability of the aggregation and choosing the class with the highest probability. For this, the probability for a class would be calculated as the proportion of the population in that class.

Due to the small numbers of EDs in some classes, it is entirely possible that some of those classes may contain no EDs after aggregation. This is particularly the case for the village class, where village EDs tend to have much smaller populations than city or town EDs. This means that in any aggregation, the village class has less influence. As an example, the 34 administrative counties have been classified using six groupings: city, town, near village, remote village, near rural and remote rural. The numbers of counties in each class are given in Table 2.16 below. The most common class is town, with no counties classed as near or remote village.

Table 2.16 The number of administrative counties in each of 6 classes

Class	Number of counties
City	8
Town	12
Near village	0
Remote village	0
Near rural	9
Remote rural	5

It is interesting to note that county Donegal classifies as near rural. It is generally considered to be a sparsely populated rural county. While it is both sparsely populated and rural, only 20% of the population actually live in remote rural EDs.

Some of these EDs are very isolated with land that is unsuitable for agriculture. When summarised to county level, the presence of isolated populations is lost, suggesting that a four class system is probably more appropriate at this level of ED aggregation.

#### **2.5.4.2 Urban and rural areas 1986 – 2002**

A similar methodology was applied to generate urban-rural classifications for the 1986, 1991, 1996 and 2002 censuses using the 3,382 ED boundaries for which data is equivalent for all four censuses. The data were concatenated into a single file containing 13,528 EDs.

As before, the first step was to identify city EDs on the basis of the majority of population living in one of the five cities. The following step was to identify the EDs with the majority of the population in a town from the list of 75 towns with legally defined boundaries. The minimum log-transformed values of 6.921, 3.442 and 0.1334 were recorded for these EDs for median town size, population density and access score respectively. Any ED which was not already classed as city and where all three values were above the recorded minima were classed as town. From the remaining unclassified EDs, those with 50% or more population in a settlement were classed as village and the rest as rural.

Discriminant analysis was applied to reclassify EDs that were closer to a neighbouring class than the class they had been allocated to. The number of EDs reclassified in 1986, 1991, 1996 and 2002 were 74, 69, 65 and 68 respectively. The changes are predominantly to town EDs being reclassified as village. As can be seen in Table 2.17 the application of discriminant analysis improves the fit of the classification so that it is significantly better than a random classification.



Table 2.17 Comparison of measures of fit for three clustering methods

Clustering method	GVF	TAI	AIC	Rank GVF	Rank TAI
MCC	0.954	0.810	2552.2	105	118
Discriminant analysis	0.958	0.818	2219.0	35	73
K-means	0.959	0.824	2125.7	9	13

The near and remote EDs then had to be distinguished for the village and rural classes. The two classes were each split into two groups using k-means clustering applied to log-transformed access scores. Table 2.18 shows the minimum values and counts of EDs for near and remote groupings. The counts include all four years, hence the large numbers.

Table 2.18 Counts and minimum access scores for near and remote village and rural EDs

Class	Near		Remote	
	Count	Minimum access score	Count	Minimum access score
Village	568	1.178	302	-1.148
Rural	4,960	-0.316	5,049	-3.977

Land use patterns are applied as before. However, due to the assumption that land use has remained largely unchanged from 1986 to 2002, the only changes in land use occur in EDs that change from rural or village to town or city. As such, levels three and four of the classification may not be as useful.

A six level classification is shown in Table 2.19 with the number of EDs in each class given by year. There have been increases in both city and town EDs while the number of village EDs has remained quite stable. Due primarily to the increased number and size of towns, access has improved and this is reflected in the reduced number of village and rural EDs classed as remote in 2002 compared to previous time periods.

Table 2.19 Count by year of EDs in each of 6 classes

Class	Count of EDs			
	1986	1991	1996	2002
City	463	466	468	469
Town	193	191	195	206
Near village	141	138	140	147
Remote village	78	77	76	71
Near rural	1,193	1,194	1,214	1,359
Remote rural	1,314	1,316	1,289	1,130

The changes in population within EDs are also of interest. Table 2.20 shows the population in each class as a percentage of the total population in that time period. The most remarkable change is that of town populations, which have increased from 22.7% to 26.0% of the total population with a corresponding decrease in the percentage of rural inhabitants. The population as a whole has increased by 10.6% between 1986 and 2002 and this has been driven by the increases in town populations.

The relatively small changes in the city population may be indicative of reduced affordability of houses in cities and increased housing development in commuter towns drawing young families out of the cities. It is also possible that the cities have reached saturation for the traditional low density/low rise pattern of development and a move to high density/high rise development is required to increase city populations.

Table 2.20 Population in each class as a percentage of the total population (population figures in brackets)

Class	1986	1991	1996	2002
City	35.9 (1,271,091)	36.2 (1,276,310)	36.4 (1,319,896)	35.7 (1,398,441)
Town	22.7 (803,726)	22.9 (807,390)	24.1 (873,887)	26.0 (1,018,473)
Near village	5.3 (187,654)	5.2 (183,337)	5.0 (181,304)	5.2 (203,695)
Remote village	2.3 (81,435)	2.3 (81,092)	2.1 (76,148)	1.9 (74,427)
Near rural	19.4 (686,885)	19.4 (683,989)	19.1 (692,583)	20.4 (799,109)
Remote rural	14.4 (509,853)	14.0 (493,601)	13.3 (482,270)	10.7 (419,141)

There are many potential implications for these shifts in population, not least for the provision of services. As towns potentially vie for a bigger slice of resources, more isolated rural areas may be left behind. While the actual population of rural and village EDs has increased at a rate well below that experienced by the rest of the country, it still represents a significant portion of the population.

A final point is to compare the urban-rural classifications generated for the 2002 data. One is based on a full set of 3,422 EDs while the other is based on the 3,382 ED set containing merges for the cross-temporal classification. The number of EDs and proportion population in each class for each method is given in Table 2.21 below. It is expected that there will be differences in the distinction between near and remote EDs, as the cross-temporal analysis applied k-means clustering across four datasets simultaneously. The other apparent difference is in the town population. The explanation for this is that the cross-temporal dataset required the aggregation of a number of town EDs and their neighbouring village or rural EDs to maintain consistent census data. A town ED merged with a village or rural ED will typically combine to form a town ED, as the town will contribute a higher population to the merged ED. This results in a slight over-count of town populations and an under-count of rural populations. Overall the classifications are

very similar which suggests that the cross-temporal classification is certainly acceptable.

Table 2.21 Comparison of 2002 urban-rural classification based on 3,382 and 3,422 EDs

Class	Count		% Population	
	3382 EDs	3422 EDs	3382 EDs	3422 EDs
City	469	467	35.7	35.6
Town	206	234	26.0	25.5
Near village	147	159	5.3	5.8
Remote village	71	71	1.9	1.9
Near rural	1,359	1,301	20.4	19.9
Remote rural	1,130	1,190	10.7	11.3

### 2.5.4.3 Comparison to CSO urban-rural classification

To compare the classification developed here to that of the CSO is difficult as the CSO do not provide an ED level classification. In the CSO method, populations living in towns of 1,500 or more persons are classified as urban with the remaining population designated as rural. For comparative purposes, a CSO classification has been produced by classifying EDs as urban if 50% or more of the population lives in a town of 1,500 or more persons. In Table 2.22 below the CSO classification is compared with a level 1 classification as developed earlier in this chapter.

Table 2.22 Comparison with CSO urban-rural classification

Area type	Estimated CSO class		Total
	Urban	Rural	
City	467	0	467
Town	229	5	234
Village	23	207	230
Rural	0	2,491	2,491
Total	719	2,703	3,422

As the CSO consider towns as urban the CSO definition of urban should encompass both city and town EDs. This is apparent in the previous table. Overall, there is good agreement between the two classification methods. It can be seen,

however, that 23 EDs are defined as urban that actually bear more resemblance to village EDs. This is due to the simple cut-off used by the CSO based on town size alone. Similarly five town EDs are classified as rural using the CSO method. These differences highlight the fact that the CSO method ignores other variables, such as population density, that give additional information about the population distribution within an ED.

Due to the application of a simple dichotomy by the CSO, there is no distinction between village and rural. In reality, there may be marked differences in EDs depending on whether the majority of the population live in settlements or are dispersed. These differences can affect social interaction and the availability of basic services, which are more likely to be based in a village than in a sparsely populated area.

## **2.6 Summary**

In the course of this chapter a number of existing methods of urban-rural classification were applied to Irish data for the first time. While these methods may have merits in certain contexts, they do not adequately differentiate between the distinct area types that exist in the urban-rural continuum. A new methodology was described and applied to Irish data to produce an urban-rural index for Ireland. A variety of data sets were employed to produce an index that is sensitive to population density, land use and access to population centres. The index has a hierarchical construction which may be presented at a chosen level of detail depending on the context or application.

There were no small area health data available to analyse possible differentials in health outcomes between urban and rural areas.

### 3 Indicator selection and transformation

Deprivation is generally defined after Townsend whereby an individual is deprived if they lack the resources “which are customary, or at least widely encouraged and approved, in the society in which they belong”.<sup>288</sup> As early as 1972 the notion of a composite index was illustrated by Craig and Driver.<sup>175</sup> They combined five census-derived indicators using a simple unweighted arithmetic mean. Since then a number of indices have been developed, primarily in the UK, to distinguish between affluent and deprived areas. The general methodology revolves around the identification of suitable proxies for deprivation and then combining those proxies into a single index or score. More recently there has been a move to so-called domains of deprivation.

There is a further useful distinction between indices in that some are individual-based and others area-based. The latter are far more common and relate to calculations being conducted at an area level, which is an aggregation of individuals. Data for areas tend to be routinely collected in the census and available for research whereas individual level data must be collected by survey and rarely has national coverage. The primary disadvantage of area level indices is that they may overlook small pockets of deprivation in an otherwise non-deprived area. This degree of averaging is related to the heterogeneity of an area and can be difficult to quantify using census variables. As area-based indices are more typical and given the difficulty in collecting appropriate data for individual level indices, only area level indices will be discussed in this chapter.

This chapter will initially discuss indicator selection before an in-depth comparison of three methods of shrinkage. This comparison is accomplished by applying shrinkage to Irish data to assess the properties of these methods of shrinkage. How these properties can impact on the subsequent deprivation index will be discussed. Recommendations will then be made on the best method for shrinkage. Other methods of data transformation will also be discussed briefly.

## **3.1 Index development**

The majority of indices combine a number of census-based proxies for deprivation into a single score. Combination is generally achieved either by a simple arithmetic summation with weights derived either by Principal Components Analysis (PCA) or Factor Analysis (FA). In some cases, more than one component or factor is extracted, either because the first component does not account for sufficient variance or because the authors hypothesize that there should be more than one component.

### **3.1.1 Indicator selection**

Indicators are typically census-based variables expressed as a proportion or rate. There has been a move towards incorporating non-census data into the UK indices which is partly driven by the fact that their census is only every ten years, as opposed to once every five years in Ireland. As the next census approaches, the data from the previous census becomes increasingly out of date. The utility of an index with 10 year old data is highly questionable, particularly during a period of marked demographic or economic change. Non-census data can include live register unemployment figures and other government issued welfare benefits. These data sources have the advantage of being current and electronically stored. The difficulty with these sorts of measures is that they often rely on denominator data that is either census-based or estimated in some way using other information on population changes. A further problem is that not all data are available at the same level of disaggregation, meaning that some indicators have to be estimated for lower levels of aggregation. Table 3.1 lists the indicators used in eleven different deprivation indices. The indicators are not always expressed in the same manner (e.g. the proportion unemployed might be expressed as proportion employed) and the definition of the denominator may not be identical but they tend to be approximately equivalent. It should be noted that the Canadian index utilised two income indicators.

Table 3.1 Indicators used in a range of deprivation indices

Indicator	England (Townsend) <sup>177</sup>	Genoa, Italy <sup>186</sup>	Ireland (Haase) <sup>184</sup>	Ireland (Howell) <sup>181</sup>	Ireland (SAHRU) <sup>182, 183</sup>	New Zealand <sup>187</sup>	Quebec, Canada <sup>188</sup>	Scotland (Carstairs) <sup>178</sup>	Spain <sup>189</sup>	UK (DoE) <sup>180</sup>	UK (Jarman) <sup>176</sup>
Unemployment	•	•	•	•	•	•	•	•	•	•	•
Overcrowding	•		•	•	•	•		•	•	•	•
Low social class			•	•	•			•	•		•
Single parents		•	•			•	•			•	•
Education		•	•	•		•	•				
Car ownership	•		•	•	•	•		•			
Owner occupied	•	•	•	•		•					
Income support				•		•					
Lone pensioners										•	•
New commonwealth										•	•
Age dependency			•								
Children <5											•
High social class			•								
Illiteracy								•			
Income							•				
Lack amenities										•	
Council housing			•		•						
Medical card				•							
No bathroom		•									
One year migrants											•
Persons living alone							•				
Separated, divorced or widowed							•				
Small farming			•								
Telephone						•					

The only indicator that is used universally is unemployment. It is well established as both an indicator of deprivation and also poor health and risk of mortality.<sup>289</sup> Some indicators, such as the 'new commonwealth' variable, are very country specific and may not have an equivalent in other jurisdictions. Replicating an index that contains such a variable is not possible with the data available in another country. With the exception of income in the Quebec index which is given as an average income and overcrowding, all of the indicators listed in Table 3.1 are



expressed as proportions. Indicators generally highlight the deprived portion of the population and so are positively correlated with deprivation. This is, however, not a prerequisite and indicators can highlight an affluent portion of the population.

The selection of indicators is driven primarily by availability. In Ireland and the UK there is no income data collected at small area level that could be incorporated into an index. Clearly income is a proxy measure of deprivation although cost of living varies regionally and would have to be taken into account. In the absence of an income measure, alternative proxies have to be used. Indicators tend to fall into one of three categories: those that highlight people deprived of certain material goods (e.g. car, house, telephone); those that are in receipt of welfare benefits (e.g. unemployed, lone parents); and those that highlight people otherwise unaccounted for who are at higher risk of experiencing poverty or health problems (e.g. migrants, lone pensioners).

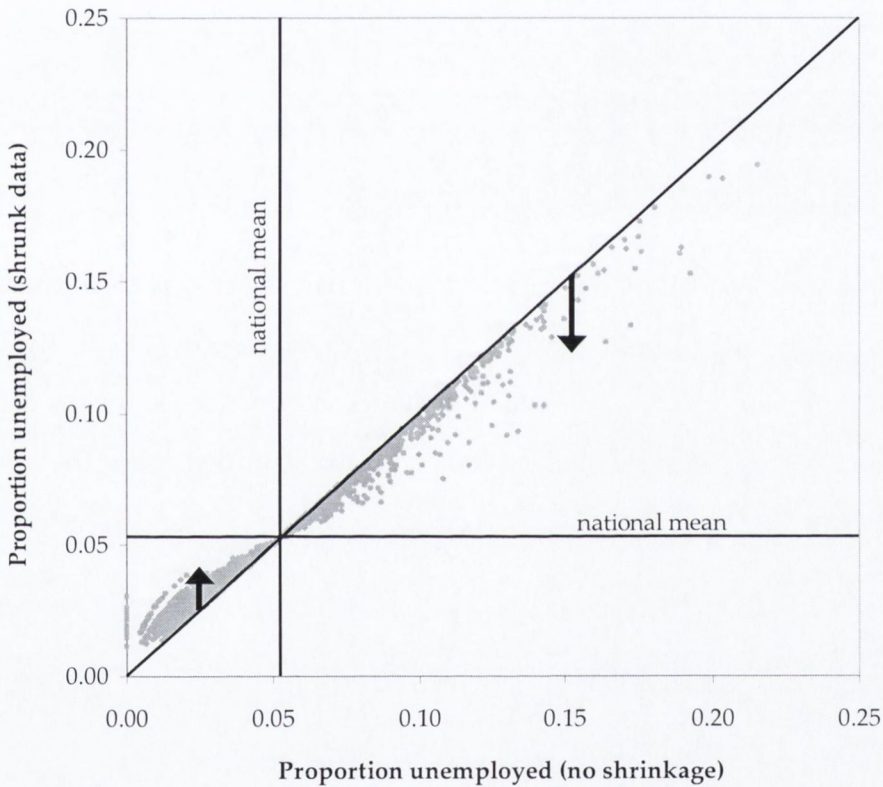
Indicators are just that: indicators, and as such, they are not ideal measures of deprivation. The choice of indicators should be driven by the intended end-use of the index. Many of the first indices were developed in the UK to identify areas with an increased need for health care resources (e.g. Jarman<sup>217</sup>). The indices were primarily used for resource allocation in the health services and could be validated by comparison with health measures. Some subsequent indices have been developed to identify impoverished areas for structural funding, investment and other forms of local government intervention. The latter purpose is less well defined and more difficult to validate.

### **3.2 Shrinkage**

Variables used for estimating deprivation scores are collected at a small area level, mostly through the census of population or similar survey type data. When the denominator is very small, the indicator can be more susceptible to seemingly large shifts by chance. To improve the reliability, an indicator can be adjusted towards a

central value in proportion to its standard error. The mean can be calculated locally or nationally. In areas where the standard error is very low, which suggests that the indicator is known with a higher precision, the amount of shrinkage should be negligible. The graph in Figure 3.1 shows how the data shrink towards the mean, the amount of shrinkage for any area being proportional to its standard error.

Figure 3.1 Raw versus shrunk proportion unemployed (shrunk to national mean)



### 3.2.1 Methods of shrinkage

Three techniques have been identified: a method used in the Noble index of deprivation,<sup>290</sup> a method proposed by Longford<sup>291</sup> and an Empirical Bayes method.<sup>275</sup> These methods are outlined below and applied to variables expressed as proportions. In all cases, the subscript  $i$  refers to the small area. The symbols  $r_i$ ,  $n_i$  and  $\hat{x}_i$  are the numerator, denominator and shrunken estimate for area  $i$

respectively. The number of observations or small areas being considered is denoted by  $k$ .

### 3.2.1.1 Shrinkage proposed by Noble

Shrinkage of indicators was introduced into the 1999 UK Index of Local Deprivation by Noble et al.

A simple logit transformation takes the following form:

$$f(x) = \ln\left(\frac{x}{1-x}\right) \quad (3.1)$$

Where  $x$  is a proportion

For many datasets this is not practical due to small areas with 0 cases. In those instances,  $f(x)$  is not computable, hence the use of the empirical logit. By adding 0.5 to the numerator and the denominator the transformation is always computable. Both the logit and empirical logit can be back transformed using the inverse logit transformation:

$$f(x) = \frac{e^x}{1 + e^x} \quad (3.2)$$

The logit transform is used in the Noble approach to shrinkage shown below which has implications that will be discussed later in this chapter.

$$\hat{x}_i = \frac{\exp(x_i^*)}{1 + \exp(x_i^*)} \quad (3.3)$$

Where:  $x_i^* = x_i w_i + (1 - w_i) \bar{x}$

$$\bar{x} = \ln\left(\frac{\sum r_i + 0.5}{\sum n_i - \sum r_i + 0.5}\right)$$

$$x_i = \ln\left(\frac{r_i + 0.5}{n_i - r_i + 0.5}\right)$$

$$w_i = \frac{1/s_i^2}{1/s_i^2 + 1/t^2}$$

$$s_i^2 = \frac{(n_i + 1)(n_i + 2)}{n_i(r_i + 1)(n_i - r_i + 1)}$$

$$t^2 = \frac{1}{k-1} \sum (x_i - \bar{x})^2$$

Given that deprivation indicators are generally expressed as proportions, the sum of squared differences that appear in  $t^2$  tends to be a small value. Coupled with a large  $k$ ,  $t^2$  will typically be a small value.

For large values of the denominator  $n_i$ , the value of  $s_i^2$  is small which in turn, due to inversion, makes it dominant in the calculation of  $w_i$ . The value of  $t^2$  is fixed for all observations so tends to be less influential on the weights. As a consequence, for large values of  $n_i$ ,  $w_i$  is large and tends towards 1 whereas for small  $n_i$  the value of  $w_i$  is low. It should be noted that  $s_i^2$  is dependent on the value of  $r_i$ .

### 3.2.1.2 Shrinkage proposed by Longford

A univariate shrinkage method described by Longford<sup>291</sup> is outlined below. This method does not use logit transformation.

$$\hat{x}_i = \frac{\sigma^2 x_i + v_i \bar{x}}{\sigma^2 + v_i} \tag{3.4}$$

Where: 
$$\sigma^2 = \frac{S_b - (\bar{x}(1 - \bar{x})(k - 1))}{N - M - k + 1}$$

$$S_b = \sum n_i (x_i - \bar{x})^2$$

$$N = \sum n_i$$

$$M = \frac{\sum n_i^2}{N}$$

$$\bar{x} = \frac{\sum x_i n_i}{N}$$

$$v_i = \frac{\bar{x}(1 - \bar{x})}{n_i}$$

The formula can be rewritten in the following manner:

$$\hat{x}_i = x_i w_i + (1 - w_i) \bar{x} \quad (3.5)$$

Where:  $w_i = \frac{\sigma^2}{\sigma^2 + v_i}$

For large  $n_i$  the value of  $v_i$  is small with  $w_i$  tending towards 1. Therefore, for large  $n_i$  the shrinkage is small.

### 3.2.1.3 Empirical Bayes shrinkage

In Bayesian statistics prior knowledge about parameters as well as observed data are taken into account when estimating the values of the parameters. For empirical Bayes techniques, the prior distribution can be based on global aspects of the data being used. In this instance, the global aspect of the data would be the observed national mean. The amount of shrinkage is dictated by precision in the observed rate.

The empirical Bayes formula as described by Bailey & Gattrell<sup>275</sup> is an approximate one-step estimation method rather than the more complete iterative maximum likelihood estimation. The method is as follows:

$$\hat{x}_i = \bar{x} + (x_i - \bar{x}) \left( \frac{\hat{\psi}}{\hat{\psi} + \frac{\bar{x}}{n_i}} \right) \quad (3.6)$$

Where:  $\bar{x} = \frac{R}{N}$

$$R = \sum r_i$$

$$N = \sum n_i$$

$$x_i = \frac{r_i}{n_i}$$

$$\hat{\psi} = \frac{\sum \psi_i}{N} - \frac{\bar{x}}{\bar{n}}$$

$$\psi_i = n_i (x_i - \bar{x})^2$$

$$\bar{n} = \frac{N}{k}$$

In the above formulae,  $\hat{\psi}$  is the estimated variance based on a weighted sample variance of observed rates about the mean,  $\bar{x}$ .

Again, this formula can be rewritten in the following form:

$$\hat{x}_i = x_i w_i + (1 - w_i) \bar{x} \quad (3.7)$$

Where:  $w_i = \frac{\hat{\psi}}{\hat{\psi} + \frac{\bar{x}}{n_i}}$

As before, for large  $n_i$  the weight  $w_i$  tends towards 1 resulting in negligible shrinkage.

### 3.2.2 Comparisons of shrinkage techniques

To understand the various merits of these three methods of shrinkage and which might be the most appropriate for the present purpose, this section looks at some of their properties and relative strengths and weaknesses.

#### 3.2.2.1 Shrinkage of simulated data

To compare the behaviour of the three shrinkage methods, a simulation exercise was performed. To ensure that random variables were used that had similar distributions to real variables, an existing dataset was used and then modified using a number of random processes. Proportions for each simulated variable were produced and shrinkage performed on the variable. The mean and standard deviation for each random variable was recorded along with the amount of shrinkage produced by the three shrinkage methods. Each simulation involved the following steps:

1. Use the distribution of an existing variable, such as unemployment, as a starting point
2. Rank the EDs by population
3. Randomly decide whether or not to invert the data (i.e.  $x_i^* = 1 - x_i$ )
4. Randomly select 1,000 EDs and for each assign a random value in the range of variable values for the 20 EDs closest in population size
5. Shift the mean by adding a randomly generated constant to all  $x_i$ . The constant must be selected so that all  $x_i$  remain in the range 0 to 1.
6. Change the standard deviation by applying the following formulae where  $\alpha$  and  $\beta$  are randomly generated constants between 0 and 1:

$$\begin{aligned}x_i < \bar{x} &\Rightarrow x_i^* = \bar{x} - \alpha(\bar{x} - x_i) \\x_i > \bar{x} &\Rightarrow x_i^* = \bar{x} + \beta(x_i - \bar{x})\end{aligned}\tag{3.8}$$

7. Record the mean and standard deviation of ED proportions
8. Calculate the shrinkage transformation of the proportions using the three methods

9. Calculate the amount of shrinkage using the formula:

$$s = 1 - \frac{\sum |\hat{x}_i - \bar{x}|}{\sum |x_i - \bar{x}|} \quad (3.9)$$

If no shrinkage occurs,  $s = 0$ . If all areas are fully shrunk to the national proportion, then  $s = 1$ .

10. Steps 3 to 7 are repeated 10,000 times.

Figure 3.2 shows the amount of shrinkage that occurs across a range of means and standard deviations for the simulated data, using the three shrinkage methods. The three methods are correlated with the standard deviation although the Empirical Bayes method is correlated with the mean at higher values. The values of all variables must be between 0 and 1, so the mean and standard deviation are related as each variable is a function of the binomial distribution. At extreme means, the standard deviation must necessarily be small, even when the data are substantially skewed. The largest standard deviation possible with uniformly distributed data ranging from 0 to 1 is approximately 0.332 but as these data are more normally distributed due to the large number of observations, the standard deviations will tend to be lower.

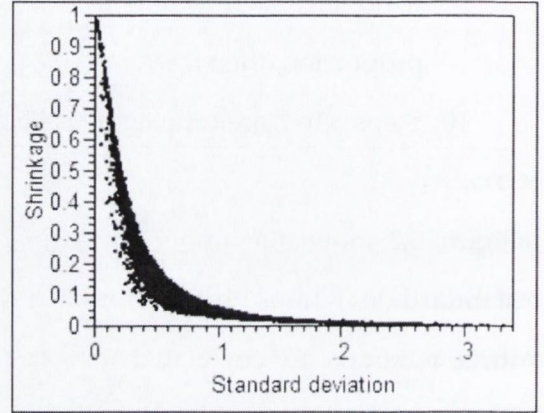
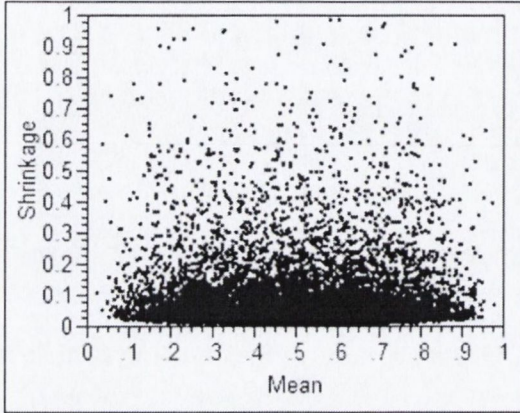
A shrinkage value of 0 indicates no shrinkage and 1 indicates complete shrinkage to the mean. The plots on the left-hand side show the relationship between dataset mean and the amount of shrinkage. The Noble and Longford methods show approximate symmetry about a mean of 0.5 while for the Empirical Bayes method there is increased shrinkage when the mean is high.

For all three methods, the graphs on the right-hand side indicate that for datasets with high standard deviations the amount of shrinkage is small. For low standard deviations, the degree of shrinkage is high. The Empirical Bayes method exhibits high shrinkage at higher standard deviations than the other two methods.

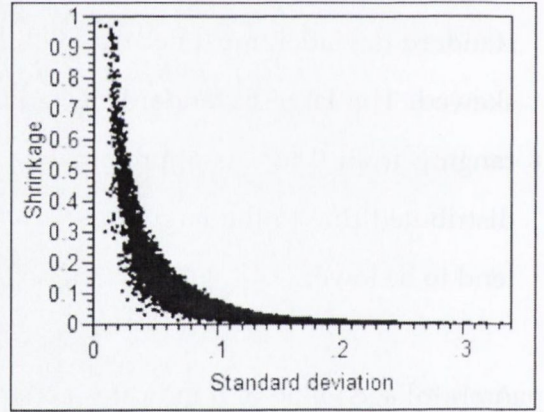
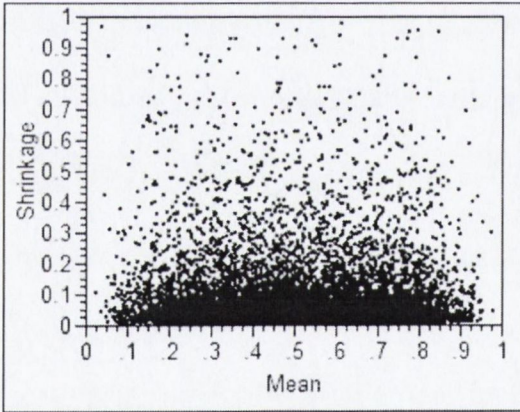


Figure 3.2 Shrinkage of 10,000 simulated datasets using three methods

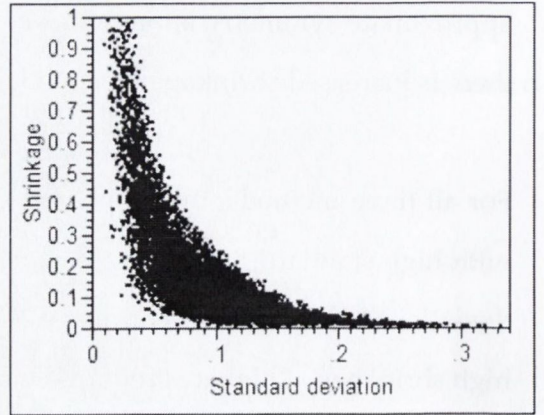
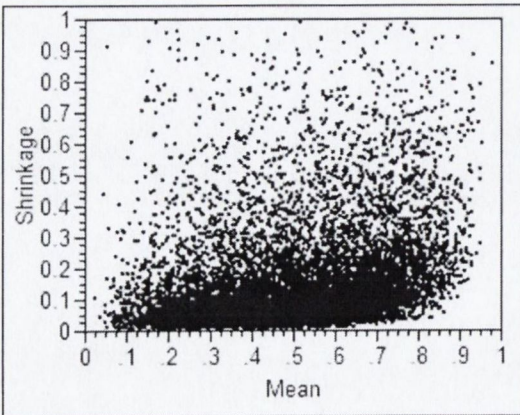
3.2a Noble



3.2b Longford



4.2c Empirical Bayes



### 3.2.2.2 Weighting in relation to numerator and denominator values

In each of the methods being considered, the intention is to apply greater shrinkage to areas with small denominators that are deemed to be ‘unreliable’. The difference between the Noble method and the other two is that the numerator is the driving force for the choice of weight, rather than the denominator.

Table 3.2 shows the weights for two EDs with the same numerator and differing denominators. As can be seen they have very different proportions but similar weights using the Noble method. The other two methods have weights directly proportional to the denominator.

Table 3.2 Comparison of weights for two EDs with the same numerator

ED	Numerator	Denominator	Proportion	Weight, $w_i$		
				Noble	Longford	Empirical Bayes
A	3	55	0.05454	0.58684	0.45190	0.43731
B	3	453	0.00662	0.60603	0.87164	0.86488

In Table 3.3 there are four EDs with the same denominator but differing numerators. The ED with the largest numerator has the largest weight using Noble. The weights for Longford and empirical Bayes methods remain the same irrespective of the numerator value.

Table 3.3 Comparison of weights for four EDs with the same denominator

ED	Numerator	Denominator	Proportion	Weight, $w_i$		
				Noble logit	Longford	Empirical Bayes
C	2	73	0.02739	0.52488	0.52251	0.50776
D	4	73	0.05479	0.64158	0.52251	0.50776
E	5	73	0.06849	0.67922	0.52251	0.50776
F	6	73	0.08219	0.70883	0.52251	0.50776

The effect of larger numerators leading to larger weights is that an ED with a large numerator will have larger weight than an ED with equivalent population but

smaller numerator. When an area has a large weight, the shrunken estimate will be close to the original proportion whereas for a small weight, the shrunken estimate will be close to the overall mean. The Noble method results in EDs with low proportions being raised closer to the mean and EDs with large proportions remaining high. This is despite the fact that these proportions may be much larger than the mean. In the event of the mean being greater than 0.5, the reverse is true.

### 3.2.2.3 Shrinking to the mean

The Noble method of shrinkage employs an empirical logit transformation as shown in Equation 4.3 above. This can then be back transformed using an inverse logit. In the case of the standard logit transformation this provides an exact inversion. For the empirical logit it does not and depends on the magnitude of the original values. Table 3.4 shows a range of numerator and denominator values, all with the same proportion. The table also shows the empirical logit and subsequent inverse logit for each set of values. For proportions below 0.5, the inverse will always be an over-estimate while for proportions over 0.5 the inverse logit will always be an under-estimate. As the denominator increases, the inverse tends towards the original proportion. This property has the knock-on effect that for a national mean less than 0.5, the transformed mean is larger than the real mean and the data will be shrunk towards a value larger than the real mean. However, as the denominator at a national level is very large in the case under consideration, the difference is negligible and not of concern.

Table 3.4 Logit and inverse logit transformations for a range of numerator and denominator values

Denominator	Numerator	Proportion	Empirical logit	Inverse logit
20	1	0.05	-2.5649	0.0714
40	2	0.05	-2.7344	0.0610
80	4	0.05	-2.8332	0.0556
100	5	0.05	-2.8544	0.0545
500	25	0.05	-2.9257	0.0509
1000	50	0.05	-2.9350	0.0504
2000	100	0.05	-2.9397	0.0502

This change in value using the Empirical logit is apparent when one looks at values close to the mean. If, for example, there was an ED whose proportion was equal to the mean it could be expected that there would be no shrinkage, irrespective of the denominator size. The purpose of shrinkage is to bring the values closer to the national mean, thus if the value is already at the mean then shrinkage will have no effect. This is not the case with the Noble logit method. The formula shrinks to a point other than the mean.

This point can be found empirically by taking an ED and searching for the numerator value which results in the shrunken proportion being equal to the original proportion. In other words, at the point when  $x_i = \hat{x}_i$ . Tests were carried out on the previous data for proportion unemployment to establish at what value zero shrinkage was occurring. A range of population (denominator) sizes were tested and the results are shown in Table 3.5 below. The proportion tends towards 0.0606 as the population increases. Given that the mean is 0.0516, this indicates that some EDs with values below the mean will be shrunk to a value greater than the mean.

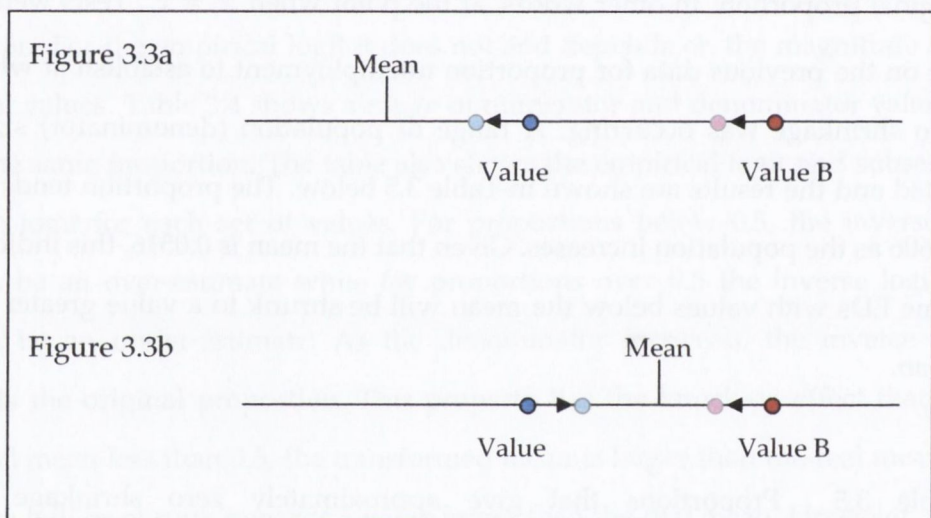
Table 3.5 Proportions that give approximately zero shrinkage for the unemployment data

Numerator	Denominator	Proportion
3	50	0.0600
6	100	0.0600
15	250	0.0600
30	500	0.0600
46	750	0.0613
61	1000	0.0610
303	5000	0.0606
606	10000	0.0606
1212	20000	0.0606

The impact of this property is questionable. It can be argued, for instance, that so long as all values for a given variable are shrunk to the same mean, does it really matter what that mean is? If the relative differences are maintained it should not

matter. As can be seen from Figure 3.3, the relative distances are not maintained if the mean is shifted. In Figure 3.3a, both values are to the same side of the mean and thus shrunk in the same direction, maintaining the relative distance. Naturally the difference will change depending on the weights attached to each ED. However, if the mean is shifted so as to be between the two values, they will be shifted towards each other, reducing the relative difference. This can be seen in Figure 3.3b below. Therefore, a shift in the mean affects the relative differences between the ED indicator values.

Figure 3.3 Implications of a shift in the mean



If the purpose is to shrink to a mean but the formula then shrinks to a value other than the mean, it is difficult to justify the use of that method.

### 3.2.2.4 Computational failure of the Longford and Empirical Bayes methods

While the Noble method ensures that  $w_i$  is between 0 and 1, this is not the case for the Longford and empirical Bayes methods. Under certain conditions these two methods produce weights greater than 1 or less than 0.

In the case of the Longford method, as long as  $0 \leq \bar{x} \leq 1$  holds then  $v_i$  is positive. Thus for  $w_i$  to be positive  $\sigma^2$  must be positive. By inserting the formula for  $S_b$  into the formula for  $\sigma^2$  the following is obtained:

$$\sigma^2 = \frac{\sum n_i (x_i - \bar{x})^2 - \bar{x}(1 - \bar{x})(k - 1)}{N - M - k + 1} \quad (3.10)$$

For  $\sigma^2$  to be positive, the following condition must hold:

$$\sum n_i (x_i - \bar{x})^2 > \bar{x}(1 - \bar{x})(k - 1) \quad (3.11)$$

Similarly for the empirical Bayes method, as long as  $0 \leq \bar{x} \leq 1$  holds then  $\bar{x}/n_i$  is positive. For  $w_i$  to be positive,  $\hat{\psi}$  must be positive. By inserting the formulae for  $\bar{n}$  and  $\psi_i$  into the formula for  $\hat{\psi}$  the following equation is obtained:

$$\hat{\psi} = \frac{\sum n_i (x_i - \bar{x})^2 - \bar{x}k}{N} \quad (3.12)$$

For  $\hat{\psi}$  to be positive, the following condition must hold:

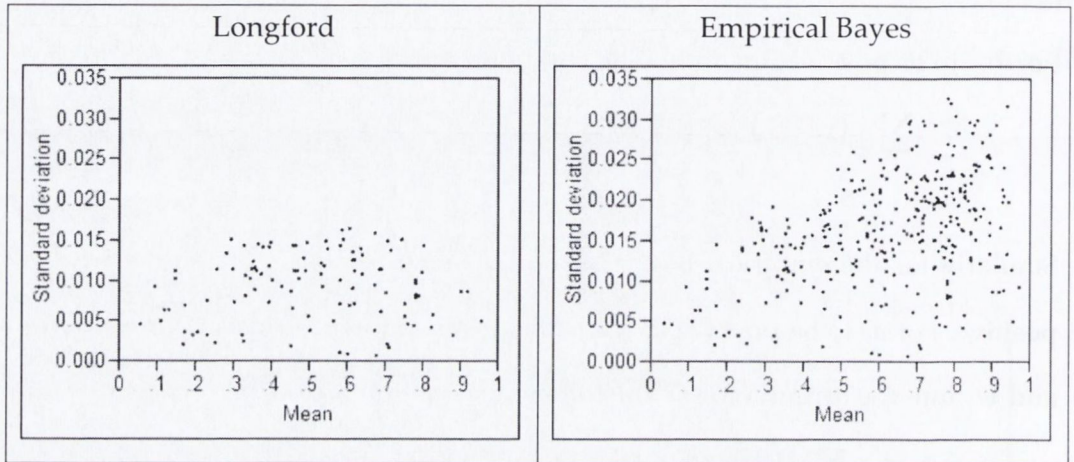
$$\sum n_i (x_i - \bar{x})^2 > \bar{x}k \quad (3.13)$$

In the case of equation 3.11, at  $\bar{x} = 0.5$  the right hand side reaches a maximum value. The left hand side is small if the area has a small population and if the standard deviation is small. For equation 3.13, the right hand side is directly proportional to the mean and increases linearly with the mean. By convention,  $\hat{\psi}$  is set to zero when a negative value is obtained.<sup>275</sup> This adjustment makes the formula computable but it still important to understand when this adjustment might be required.

In the 10,000 simulations carried out previously, there were a number of variables generated for which the Longford and empirical Bayes methods could not be applied as they generated weights outside the range of 0 to 1. The means are plotted against the standard deviations for those instances in Figure 3.4 below. For

the 10,000 simulations, the number of failed computations for the Longford and empirical Bayes methods were 124 and 348 respectively.

Figure 3.4 Mean versus standard deviation for simulations where Longford and empirical Bayes could not be computed



The largest standard deviation for which the Longford and empirical Bayes methods could not be computed in the simulations were 0.017 and 0.033 respectively. These values were dependent on the data being tested and cannot be applied as a general rule. The success of the empirical Bayes method is dependent on the combination of mean and standard deviation. For example, the unemployment data have a mean and standard deviation of 0.052 and 0.027 respectively. These data can be successfully shrunk using the empirical Bayes technique. However, if the data are inverted so that the mean is 0.9484 and the standard deviation remains the same, the empirical Bayes technique fails.

Datasets with very low standard deviations may not be suitable for shrinkage. By the very fact that the standard deviations are low, these data show little variation and the extreme values that we wish to adjust may not be present.

### 3.2.3 National or local shrinkage

An important aspect of shrinkage is the choice of mean to shrink to. The examples used thus far have been shrunk to a national mean. It is possible, however, to

subset areas into regions and to calculate shrinkage for that region using the local mean. This section will look at a number of ways of shrinking based on sub-national groupings of areas.

### **3.2.3.1 National: shrinkage to an urban mean**

Due to concentrations of population in urban areas, 61.9% of the population live in just 21.0% of the EDs, which cover only 6.8% of the land area. For any variable the national mean will be driven by that urban 21% of EDs. If the variable shows a marked urban-rural difference then the majority of EDs will be shrunk towards what is essentially an urban mean. Furthermore, as urban EDs tend to have larger populations than rural EDs, their weights will be larger, resulting in less shrinkage for urban EDs and relatively more shrinkage for rural EDs. The net effect is that small rural EDs are shrunk to an urban mean while the more populace urban EDs remain relatively unchanged.

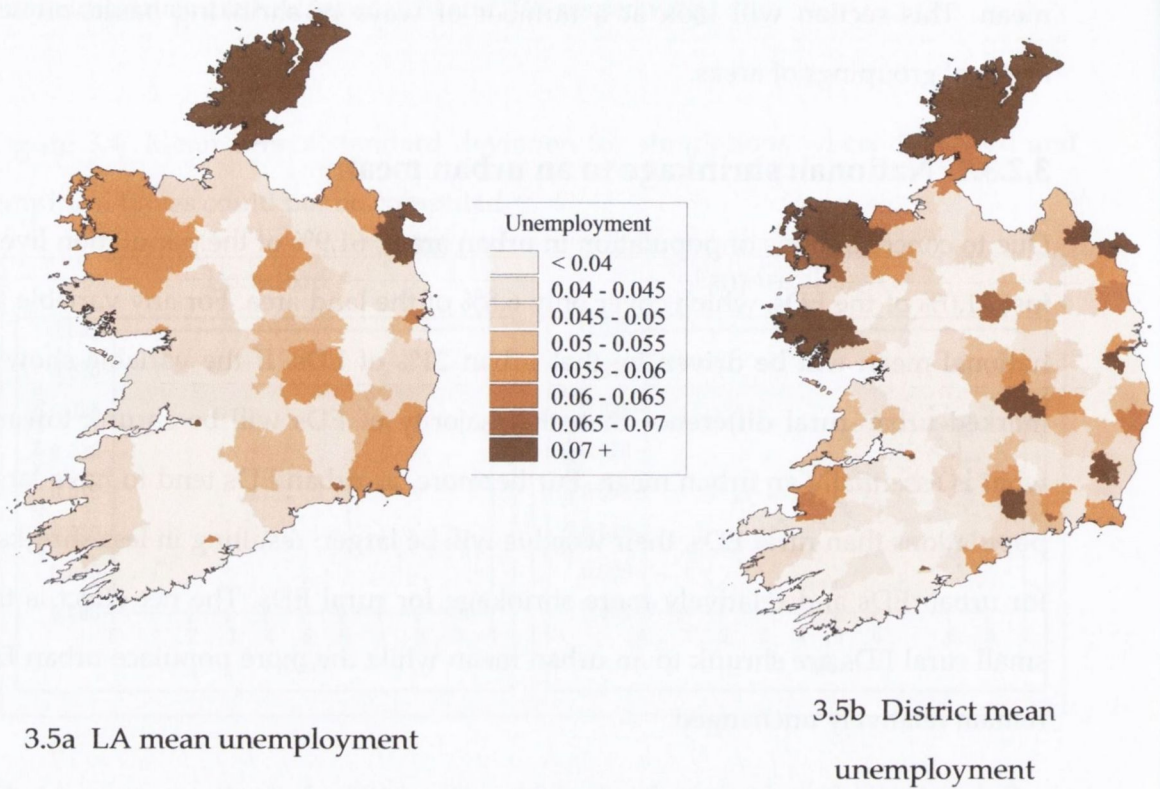
### **3.2.3.2 District shrinkage: guilt by association**

A method that has been used in the UK deprivation index is shrinkage to a district mean. In the UK, the average district contains 23 wards. The average ward population in England and Wales is 5,927 persons so a typical district contains approximately 138,000 persons. This means that a UK district is roughly equivalent to a Local Authority (LA) in Ireland in terms of population. The shrinkage is not quite comparable as UK wards are larger than Irish EDs in terms of population so they should have larger weights and consequently less shrinkage.

The district boundaries have been defined by rural and urban districts with further subdivisions applied in the larger urban centres of Dublin, Cork and Limerick. There are 157 districts with a mean population of 24,950. As such they are much smaller than the UK districts although this reflects the lower population density in Ireland. The maps in Figures 3.5a and 3.5b show the mean unemployment by LA and district respectively. Both are prone to sharp changes in neighbouring areas.



Figure 3.5 Local means at Local Authority and district level

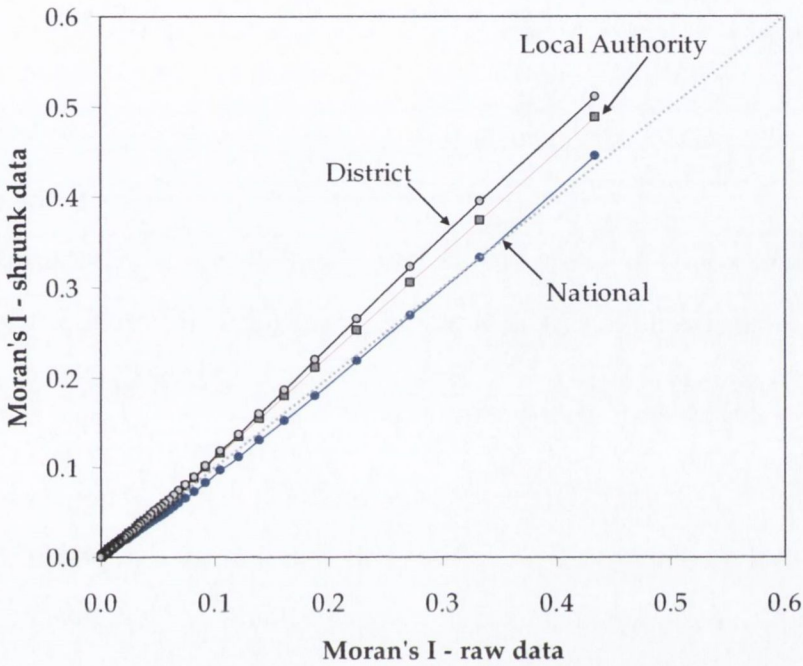


Spatial autocorrelation gives a measure of the similarity of geographic areas that are near each other in spatial terms. One formula for measuring this is Moran's I with values ranging from -1 to 1 where positive values indicate positive spatial autocorrelation (i.e. nearby areas are similar) and negative values indicate negative spatial autocorrelation (i.e. nearby areas are dissimilar). A related term is 'spatial lag' which refers to contiguity between neighbouring areas. If two neighbouring EDs share a common border they are said to have first order contiguity and a lag distance of 1. If the areas do not share a common border but are both contiguous with a third area they have second order contiguity. The spatial lag is simply the order of contiguity between two areas.

The graph in Figure 3.6 plots spatial autocorrelation for the raw unemployment data against that for the shrunk data. Shrinkage to three different means is shown: national, LA and district. Observations represent the Moran's I values at a given

lag distance, with the largest Moran's I values occurring at the shortest lag distances.

Figure 3.6 Spatial autocorrelation: raw vs. shrunk data



It can be seen from Figure 3.6 that shrinkage to a local mean increases the spatial autocorrelation, particularly at short lag distances.

Small area boundaries are often delineated by natural boundaries, such as rivers and lakes, or man-made boundaries, such as motorways. As a result, neighbouring EDs can be quite different and an affluent ED can border a deprived ED. A well-off street can be separated from a less affluent estate by the width of a river or a dual carriageway. Imagine a small but affluent ED surrounded by densely populated deprived EDs. With shrinkage to a highly localised mean, that affluent ED will effectively be rendered more deprived than it is due to its proximity to deprived EDs. That in itself is a compelling argument not to use district means. However, use of a national mean may ignore an underlying regional variation in the mean that may be important.

A further point is that district boundaries are administrative areas that are typically invisible to those living within them. In that sense, they are somewhat arbitrary and a different choice of boundaries may result in a very different set of means and subsequent shrinkage. This is an instance of the modifiable areal unit problem (MAUP) which recognises that small area boundaries are arbitrary and modifiable.<sup>292</sup>

### 3.2.3.3 Localised shrinkage using Monte Carlo methods

An alternative is to establish smaller districts around each ED, based on distance or population, and apply shrinkage to that district, recording the shrunken value for that ED. Thus neighbouring EDs would have slightly different districts but should be shrunk to similar means.

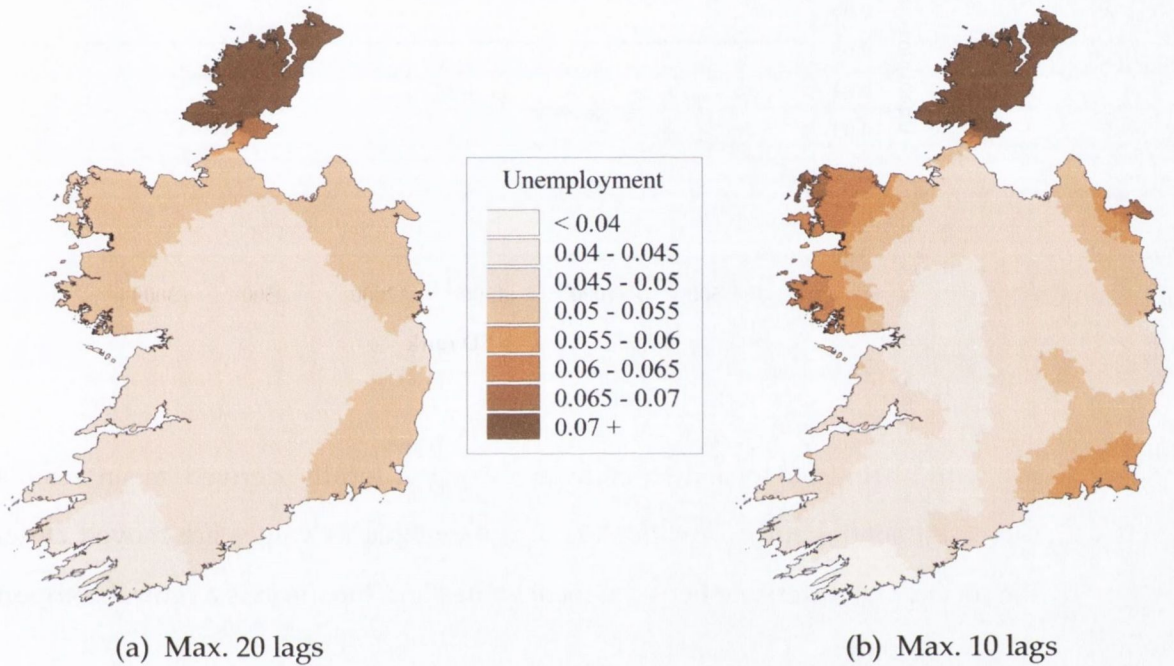
A method is proposed in which multiple district configurations are tested giving each ED a range of shrunken values. This may limit the impact of MAUP on the final results. A simplified algorithm for such a method is as follows:

1. Randomly select an ED,  $i$
2. Randomly select a number of spatial lags,  $L$ , to include in the district
3. Create a district including all EDs within  $L$  lags of  $i$
4. Apply shrinkage to the district
5. Save the shrunken values and district mean for all EDs in the district
6. Repeat steps 1 to 5 until all EDs have been included in 1000 districts

The output of this method is 1000 shrunken values and district means for each ED, for which a mean and standard deviation can be calculated. This gives the option of quoting a mean shrunken value and upper and lower bounds for this value. If the maximum number of lags allowed is very large then the mean shrunken value will tend towards the value obtained by using the national mean for shrinkage. The two maps in Figure 3.7 show the average district means for two Monte Carlo runs using maximum lags of 20 and 10 respectively. As this is a mean value for an ED it

does not represent an actual district mean so much as the average mean for districts including that ED. When compared to the maps of Figure 3.5 they show much more gradual shifts in the underlying mean.

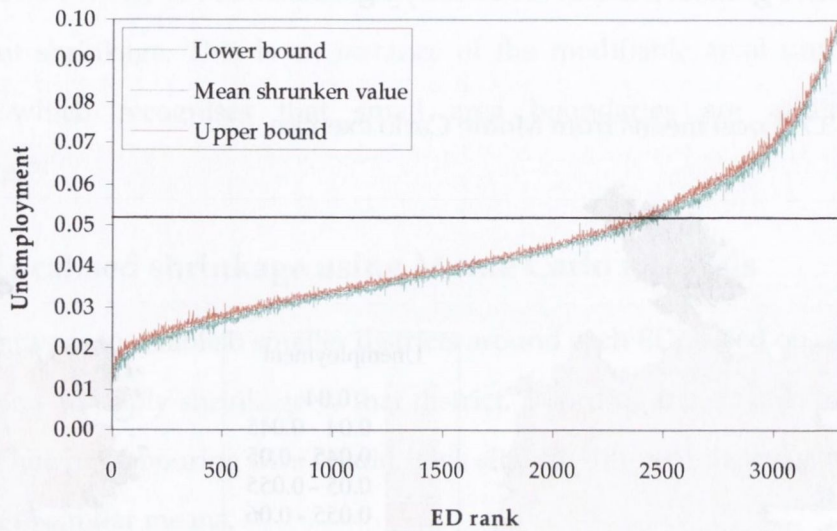
Figure 3.7 Local means from Monte Carlo exercise



The number of lags required to encompass all EDs varies across the country. Centrally located EDs can span the country in as few as 40 lags whilst peripherally located EDs can require up to 80 lags.

By generating 1,000 shrunken values for each ED, it is possible to calculate the standard deviation from which upper and lower confidence bounds can be calculated (from  $\bar{x} \pm 1.96$  standard errors). This could be used as part of a sensitivity analysis when indicators are being combined. The indicator values for a particular ED could be replaced by the upper and lower bounds to see to what extent it would change its deprivation ranking. Figure 3.8 shows the upper and lower bounds for a subset of EDs.

Figure 3.8 Upper and lower bounds for unemployment generated by Monte Carlo for a subset of EDs



As with shrinkage to a district mean, using a locally derived mean results in increased spatial autocorrelation. This is inevitable as values are moved closer to the mean. The difference between local values and the mean is a central component in the calculation of Moran's I.

Table 3.6 shows the spatial autocorrelation measured by Moran's I for several variables using a number of shrinkage techniques: no shrinkage, shrinkage to the national mean, Monte Carlo with a maximum of 20 lags allowed, Monte Carlo with a maximum of 10 lags allowed, shrinkage to the LA mean, and finally, shrinkage to the district mean. Depending on the maximum number of lags allowed, the Monte Carlo method results in lower increases in spatial autocorrelation than shrinkage to either the LA or district mean.

Table 3.6 Moran's I for a selection of variables using different shrinkage techniques

Shrinkage	Moran's I			
	Unemployment	Low social class	Car ownership	LA housing
None	0.4343	0.4364	0.6962	0.3189
National mean	0.4453	0.4437	0.7042	0.3195
Monte Carlo (20)*	0.4650	0.4524	0.7119	0.3229
Monte Carlo (10)*	0.4816	0.4638	0.7194	0.3275
Local Authority mean	0.4878	0.4662	0.7212	0.3296
District mean	0.5105	0.4892	0.7338	0.3368

\* Shrinkage applied using the Monte Carlo technique with maximum catchment sizes of 20 and 10 spatial lags respectively.

The problem with a stochastic approach, such as the Monte Carlo method, is that it is not very transparent and given that the shrunken variables will be used in a deprivation index which is intended for policy and funding, transparency is preferable. This must be weighed against the advantage of being able to produce bounds for the shrunken value and avoid specification of arbitrary boundaries for districts.

There have been suggestions that small areas could be grouped according to certain attributes such as socio-demography and topology, and shrinkage could be applied to these groupings. This will result in non-contiguous groups of small areas which may span the entire country. This would, like shrinkage to a national mean, diminish the presence of any regional variation. It also leads to difficulties in defining similarity for grouping purposes. It also runs the risk of merely increasing the deprivation of already deprived areas and reducing it in more affluent areas.

### 3.2.4 Choice of shrinkage method

A number of properties were discussed earlier in section 3.2 relating to each of the shrinkage methods. The Noble method of shrinkage is not preferable due to the strong influence of the numerator. While it is acknowledged that the Longford and

Empirical Bayes methods can fail for datasets with very low standard deviations, it is a valid argument that shrinkage should not be applied to such datasets. Finally, the Longford method is superior to the Empirical Bayes method as the latter method results in greater shrinkage for datasets with high means. It is my opinion that the Longford method of shrinkage is the most appropriate for use on deprivation indicators. The principal drawback of this method is the failure to compute under certain conditions of low standard deviations. With a standard deviation below 0.017 this method is likely to fail, although failure is also dependent on the mean of the data under consideration. However, as has been discussed, data with a low standard deviation is not suitable for shrinkage as there is insufficient variation in the data to warrant it.

For the choice of mean to shrink to, the use of a Monte Carlo method with multiple districts tested is preferable. It takes into account regional variation and enables the calculation of upper and lower bounds. It is recommended that the largest district size tested should certainly include more than a quarter of the total small areas in the country. With a large maximum, the results will be broadly similar to applying shrinkage at a national level.

However, should the Monte Carlo method not be deemed transparent enough for a deprivation index that affects public funding, it would be preferable to shrink to the national mean or some very large aggregation of small areas. The use of small districts gives rise to potentially misleading results as evidenced by the increases in spatial autocorrelation. If the populations of the small areas are sufficiently large so that the weights associated with them are also large, this is less of a concern.

### **3.2.5 Implications of choice of shrinkage**

It is important to investigate what impact the method of shrinkage has on final deprivation index values. A simple test is to generate an index using a number of indicators with various forms of shrinkage applied. The variables used in previous examples are used again: proportion unemployment (UE), proportion low social

class (SC), proportion households with no car (NC) and proportion of households in Local Authority housing (LH). The four indicators have been shrunk using the national means, district means, and the Monte Carlo method. The results are compared to an index generated using the raw non-shrunk variables. The non-shrunk results are used as the baseline for comparisons. The index values are calculated as deciles of the first principal component with 1 being the least deprived and 10 the most deprived.

Table 3.7 below shows the correlations between a deprivation index calculated using standard PCA with no shrinkage and a range of methods of shrinkage. The correlation coefficients for ranks comparisons are high with the only lower values appearing where shrinkage to a district mean is used. The obvious conclusion is that with the exception of shrinkage to a district mean, there is little or no impact from using alternative methods of shrinkage. However, the correlation coefficients for index values show a much greater variation. The low  $R^2$  values reflect the fact that the index values are restricted to a ten level scale. Even relatively small changes can lead to a substantial reduction in the correlation coefficient. To say that the different means lead to essentially the same results would clearly be incorrect.

Table 3.7 Comparison of correlations with baseline for different shrinkage methods using ranks and index values

Shrinkage	$R^2$	
	Ranks	Index values*
None	1.000	1.000
National mean	0.994	0.805
Monte Carlo	0.991	0.769
District mean	0.960	0.614

\* The  $R^2$  for two indices is the portion of the total uncertainty attributed to the model fit calculated in JMP.

A comparison of the effects of shrinkage was carried out as part of an evaluation of the Scottish indices of deprivation.<sup>293</sup> The comparison consisted of correlations



between ranks and scores using different methods of shrinkage. The correlations were very high, generally of the order of 0.998 for the single overall deprivation score “indicating that none of the alternatives has an excessive influence on the domain or multiple deprivation scores or rankings”. The problem is that when looking at 6,505 data zones, the correlation coefficient may not be sensitive to a small number of more substantial changes. In any case, as the index is finally presented in the form of a 1 to 10 scale, it is more pertinent to look at the correlations of the index values rather than scores or ranks.

The impact of differing methods of shrinkage can be seen by looking at the indicator values for a single ED. The figures for Loughill ED in county Sligo are given in Table 3.8 below. This is one of two EDs that showed a difference of four index values between calculation without shrinkage and with shrinkage to a district mean. This is an ED with a population of only 76, making it one of the smallest EDs in the country in terms of population. As can be seen in the table, the value for proportion low social class (SC) varies dramatically depending on the type of shrinkage used. With no shrinkage, the ED is classed in the most affluent decile. With shrinkage to a district mean, the ED is classed 5 corresponding to a middle deprivation level.

Table 3.8 Comparison of indicator values using different shrinkage methods for Loughill ED

Variable	Shrinkage type			
	None	To national mean	Monte Carlo	To district mean
UE	0.018	0.036	0.040	0.045
SC	0.077	0.109	0.132	0.179
NC	0.182	0.191	0.189	0.187
LH	0.000	0.026	0.038	0.051

In Table 3.9 the numbers of EDs being assigned different index values are shown. The difference in value is calculated as the index value using no shrinkage minus the index value using the specified form of shrinkage.

Table 3.9 Difference from baseline using different methods of shrinkage

Shrinkage	Counts of EDs by difference from baseline index								
	-4	-3	-2	-1	0	+1	+2	+3	+4
None	0	0	0	0	0	0	0	0	0
National mean	0	0	5	224	2,959	234	0	0	0
Monte Carlo	0	1	6	281	2,838	296	0	0	0
District mean	2	16	62	464	2,264	585	28	1	0

The example of Loughill ED raises the question of whether or not shrinkage is necessary at all. In this instance, there is no LA housing in the ED. This is symptomatic of rural EDs where the LA is less likely to buy or build housing stock. A rural ED is rendered less deprived on the grounds of LA policy. The application of shrinkage generates a non-zero value for the proportion of local authority housing (LH) variable that effectively makes the ED more deprived. This could also be seen as a reflection of the choice of indicators and the use of shrinkage may merely mask the fact that some of the indicators display substantial spatial autocorrelation and are perhaps misleading indicators of deprivation. Worse still, they may actually bias the results in favour of one area type, such as urban or rural.

It is important to point out that shrinkage is intended to improve estimated rates that have been produced using survey or sample data. When employing census data, this should not be as important an issue. In the case of EDs with small populations where there is a real possibility of zero or inflated values for indicators, the application of shrinkage with minimal effect may be useful to prevent extreme deprivation scores on the grounds of a single outlying value. However, it is more important that indicators are carefully chosen and that shrinkage is not used to smooth a variable that may be inappropriate for regional comparison.

### **3.3 Alternative methods of data transformation**

Shrinkage is not the only method of data transformation available. It is possible to use other functions such as log transforms. Frequently the primary purpose of data transformation is to improve the normality of the data.<sup>263</sup> However, in principal components analysis the distribution of the variables is not an issue. Therefore the application of a log, logit or other transform is not a prerequisite for analysis. The application of a transformation can also lead to greater ambiguity in interpretation of the resulting index which is undesirable.

As an example, three variables have been transformed using natural log and empirical logit transforms. The natural log transform cannot be computed for zero although if applied to data after shrinkage, there should be no zero values. The empirical logit produces a very similar distribution to the standard logit and can be computed for zero values. The histograms in Figure 3.9 show the effect of natural log and empirical logit transformations on three variables: proportion unemployment, proportion households with no car and proportion households in Local Authority housing.

Clearly the transformations improve the normality of the data, particularly for the Local Authority housing variable. The utility of transforming the data will depend on whether assumptions of normality or approximate normality are required for the chosen method of combination.

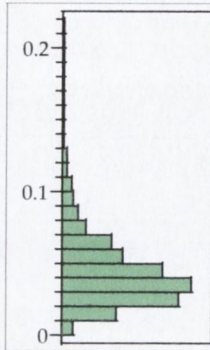
### **3.4 Summary**

Three methods of shrinkage were outlined and applied to Irish data in this chapter. While resulting in similar levels of shrinkage, each method has distinguishing properties that can lead to markedly different levels of shrinkage under certain conditions. For example, the Empirical Bayes method results in increased shrinkage at higher means. The Longford method of shrinkage was recommended as the best when applied to deprivation indicators.

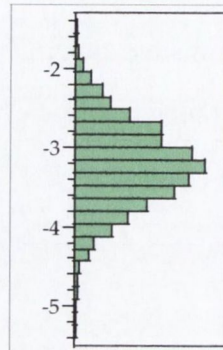
The issue of shrinking to a district mean rather than a national mean was also discussed. An analysis of spatial autocorrelation revealed that shrinkage to a district mean, particularly if districts are small, leads to indicator values for EDs being brought substantially closer to the local mean than if a national mean is used. The national mean, on the other hand, might not be representative of a regional mean. A Monte Carlo approach to district delineation was proposed and applied as a trade-off between a national and a district mean.

Figure 3.9 Effect on distribuion of natural log and empirical logit transformations

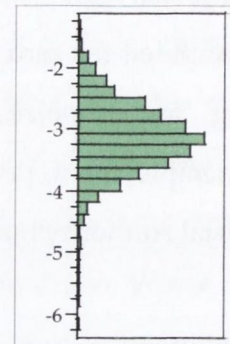
(a) % Unemployment



No transform

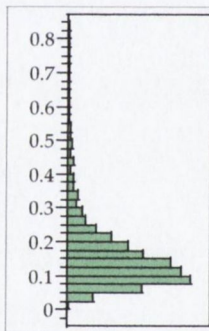


Natural log transform

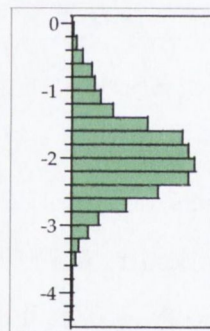


Empirical logit transform

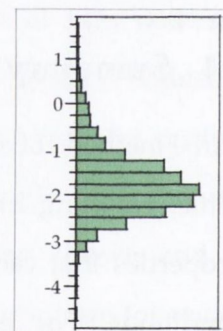
(b) % Households with no car



No transform

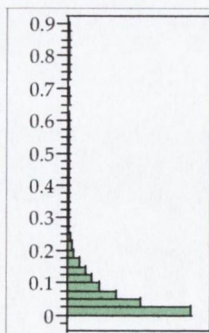


Natural log transform

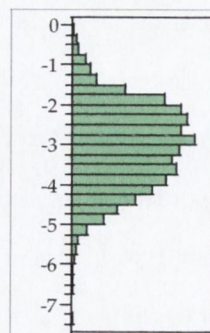


Empirical logit transform

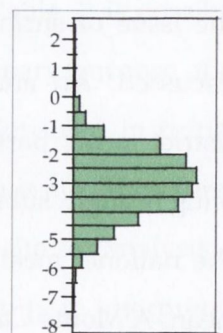
(c) % Households in Local Authority housing



No transform



Natural log transform



Empirical logit transform

While the different choices of method and mean may appear to have a big impact on indicator values, the impact on a resultant deprivation index has not been assessed. In chapter 5 a sensitivity analysis is described and applied to Irish data to assess the impact of different choices regarding shrinkage.

Figure 1 shows the effect of the different types of treatment on the mean values of the different variables. The results show that the different types of treatment have a significant effect on the mean values of the different variables. The results are presented in Table 1.



Figure 1. Effect of different treatments on the mean values of the different variables.



Figure 2. Effect of different treatments on the mean values of the different variables.



## 4 Dimension reduction

Measures of deprivation are generally comprised of a number of indirect measures or proxies for deprivation combined in some manner to produce a single variable. It is generally accepted that no single indirect measure will fully capture deprivation but that a composite value will come closer to giving an accurate distinction between areas of high and low deprivation. Furthermore, these composite measures typically quantify deprivation in relative rather than absolute terms. To measure deprivation in an absolute sense there must be accepted thresholds for when an area or person may be described as deprived. With proxy measures, such as unemployment, it is not possible to identify such a threshold unless, for example, it is assumed that anyone who is unemployed is automatically classed as deprived. On the other hand, to measure deprivation in relative terms, no thresholds need to be defined. The composite measure is given as a continuous variable which can then be divided into deciles whereby the most deprived ten percent can be easily identified. In this way an area can only be identified as deprived relative to other areas.

It is arguable that proxies for deprivation should not be combined as the inevitable loss of information may mask the existence of deprivation in some areas. Additionally, some indicators may be geographically biased and be typically higher in urban areas than in rural areas or vice versa. However, a composite measure is intended to be just that: a composite of various indicators that can highlight areas that are relatively deprived in multiple aspects.

The focus of this chapter is on methods of dimension reduction in relation to deprivation indices. Section 4.1 is a discussion of dimension reduction as used in a number of existing indices from a number of countries. Five methods of dimension reduction are compared in section 4.2 with a more detailed description of principal components analysis given in section 4.3 by applying the method to Irish data in a similar fashion to the SAHRU index.<sup>183</sup> Sources of error in deprivation indices,



including regional variation in indicators, are discussed in section 4.4. In section 4.5 a new method of principal components analysis is developed and illustrated with a worked example using Irish data. Finally, in section 4.6, outlier detection and robust analysis are discussed.

## 4.1 Dimension reduction in existing indices

The number of indicators used in different indices varies and is partly dictated by methodology and partly by data availability. With large numbers of indicators there is an increased chance of poor correlation between some indicators to the extent that groups of indicators emerge where inter-group correlation may be poor but intra-group correlation is high. If PCA or FA is used with a large number of indicators it is likely that more than one component or factor will need to be retained, resulting in a multidimensional index.

### 4.1.1 Simple indices

Table 4.1 shows the number of indicators, method of weights derivation and number of dimensions used in a number of indices of deprivation.

Table 4.1 Numbers of indicators used in some indices of deprivation

Index	Variables	Weights derivation	Dimensions
Australia <sup>294</sup>	56	PCA	4
England (Townsend) <sup>177</sup>	4	Equal weights	1
Genoa, Italy <sup>186</sup>	4	Equal weights	1
Ireland (Haase) <sup>184</sup>	13	FA	3
Ireland (Haase) <sup>185</sup>	10	FA	3
Ireland (Howell) <sup>181</sup>	8	Equal weights	1
Ireland (SAHRU) <sup>182 183</sup>	5	PCA	1
New Zealand <sup>187</sup>	9	PCA	1
Quebec, Canada <sup>188</sup>	6	PCA	2
Scotland (Carstairs) <sup>178</sup>	4	Equal weights	1
South Africa <sup>295</sup>	13	PCA	1
Spain <sup>189</sup>	4	FA	2
UK (DoE) <sup>180</sup>	6	Arbitrary	1
UK (Jarman) <sup>176</sup>	8	Survey	1
US <sup>296</sup>	16	FA	3

The advantages of including many variables are unclear. It is possible that if a large number of indicators contribute to the score, any one variable can only have a limited influence on the computation. With more indicators, there is a greater possibility of balancing regionally biased variables (i.e. urban-centric indicators could be balanced by including rural-centric indicators). With a small number of indicators an area has only a limited number of ways in which to register as deprived. It could be argued that with many variables an index may be more sensitive to levels of deprivation.

On the other hand, with an increasing number of indicators it becomes very difficult to comprehend on what grounds an area is deemed deprived or affluent. There are increased opportunities for an averaging effect whereby an area may be very deprived in some respects but not deprived or even affluent in others, resulting in a mid-range score.

The Australian index retains four components in the PCA based on Catell's scree test<sup>297</sup> on a plot of the eigenvalues. However, given the plot of eigenvalues they provide they should have retained at least 5 components on the basis of a scree test. Using the Kaiser-Guttman rule and parallel analysis they would have retained 10 and 9 components, respectively. With such a large number of variables this is almost inevitable and a nine- or ten-dimensional index would not be interpretable for general usage. Four components were retained for the purpose of rotation but only the first component was used for the deprivation index. Prior to rotation, the first component explained only 28.9% of the variance and the cumulative variance explained by the first four components was 55.7%. Given these characteristics the utility of the index is questionable. Furthermore, for any given small area, it is very difficult to comprehend what variables have contributed to its deprivation score and comparison between areas with similar scores becomes ambiguous.

With the exception of the Haase indices,<sup>184 185</sup> the other cases of multiple components are found by determining how many components to retain. When more than one component or factor is retained, the distinction between what those components reflect is subject to the interpretation of the researcher. In the case of Haase, the numbers of components are pre-defined and FA is used to determine the factor loadings. No tests are documented to ascertain whether fewer components would have been adequate or if more components were required. Having completed a factor analysis and retained three factors, Haase allocates pre-determined labels to the factors based on the variable loadings.

There is a criticism of using PCA and FA on the grounds that they give too much weight to variables that are well correlated. A legitimate deprivation indicator that happens to correlate poorly with the other variables will receive a lower weight when it is arguable that it is equally important, hence the argument for equal weights. There is, however, an inherent assumption that all of the variables chosen are good indicators of deprivation which is not necessarily the case. An indicator may, in theory, be a good proxy for deprivation but either may be subject to substantial regional bias or might simply not support the theory. The only validation generally applied to indicators is a review of the correlation matrix, which is self-serving if PCA or FA is subsequently used.

#### **4.1.2 Domain based indices**

In the UK in the late 1990's there was a move from a single deprivation score to a number of domains of deprivation, each being comprised of a number of domain specific indicators and an associated index.<sup>190</sup> The approach of Noble et al. reflected their "view that multiple deprivation is a combination of different, though clearly inter-related, deprivations". The justification behind this approach was an attempt to capture different aspects of deprivation rather than resorting to a single catch-all index. This method allows for areas to register as deprived in certain domains and not others, rather than having an average score in a single index. The alternative of retaining additional components to describe different dimensions runs the risk of

including potentially counter-intuitive variables in a given dimension. With a system of domains it is possible to ensure that only appropriate variables will be incorporated into a particular domain based on a theoretical model of indicators and domains. This use of specific domains also removes the need to subjectively interpret factors and attach *post hoc* labels to define dimensions based on the observed loadings.

Furthermore, a selection of domains allows for the use of a possibly more appropriate index for a given application. For example, if a piece of research was investigating school funding in relation to educational needs, the use of a domain of educational deprivation may be more relevant than the single universal deprivation score. Once the domain system of deprivation was introduced in England, similar indices followed in Scotland,<sup>298</sup> Wales<sup>191</sup> and Northern Ireland.<sup>193</sup> Many of the indicators used are based on routinely collected data, rather than census variables which would only be collected once every ten years. This enables the creation of a more current set of indices although it has other problems as outlined in section 3.1.1 previously.

It should be noted that the terms 'domain' and 'dimension' are used interchangeably in the UK deprivation index literature. The predominant term is domain so that will be used in preference here.

### **4.1.3 Types of domain**

Seven domain-based indices have been identified from a number of countries and the types of domains considered are given in Table 4.2 below. The variables used in each domain are not universal across indices although they bear many similarities. In most cases only a small number of variables, typically 3 to 5, are used in each domain. This is partly due to the difficulty in gathering sufficient indicators and also due to the specific nature of domains. In each domain, depending on the number and nature of variables, they are combined using either PCA or FA with only a single component or factor retained. Given the narrow scope of domains, it

would be unlikely that a second or even third component would be required. In some cases, such as the income and employment domains in England,<sup>264</sup> only one variable is used so only shrinkage is applied prior to standardisation.

The domains generally encompass aspects of both social and material deprivation. The crime and access to services domains, for example, could be considered as measures of social deprivation while income and employment are more closely related to material deprivation.

Table 4.2 Examples of domains in deprivation indices

Domain	England <sup>264</sup>	Scotland <sup>299</sup>	Wales <sup>300</sup>	Northern Ireland <sup>301</sup>	South Africa <sup>302</sup>	South Africa <sup>303</sup>	USA <sup>304</sup>
Income	•	•	•	•		•	
Employment	•	•	•	•		•	•
Education	•	•	•	•		•	•
Health	•	•	•	•	•	•	
Access to services	•	•	•	•			
Housing		•	•				•
Housing & services	•						
Physical environment	•		•	•		•	
Crime	•			•			
General					•		
Policy					•		
Poverty							•
Residential stability							•
Occupation							•

#### 4.1.4 Combining domains

Although domains of deprivation can give a greater insight into specific forms of deprivation, it may still be desirable to have a single overall estimate of multiple deprivation. Some of the indices have accomplished this by combining the domain indices into a single value.<sup>264 301 303</sup> Noble notes that there are a number of methods

by which weights can be determined<sup>264</sup>: theoretical determination using research evidence; empirical methods such as survey, PCA or FA; policy relevance or in proportion to public expenditure; consensus of experts and/or policy makers; and arbitrary choice such as equal weights.

The use of PCA or FA might not be appropriate in the context of combining domain indices as some of the indices may correlate poorly with each other given that they expressly account for different aspects of deprivation. A policy relevance approach may be biased by political influence (i.e. it may be led by what might be a 'popular' issue at the time the index is being developed). A consensus of experts requires the identification of appropriate "experts" which, given the potentially large number of interested parties, may be difficult to achieve and to strike a balance between the different domains. This essentially leaves the options of theoretical determination and arbitrary choice. Thus far a theoretical approach has been used combining evidence from literature and a consultation process<sup>264 301</sup> with arbitrary equal weights used in one South African index.<sup>303</sup> Given the large number of indicators that comprise the final composite index, interpretation may be difficult. In the case of the English index,<sup>264</sup> a total of 28 indicators are eventually combined into a single index. Some indicators are shrunk, some indicators are combined by FA, while some are combined using equal weights. Prior to combination into a single index, all domain indices are standardised and an exponential transformation applied. Understanding how any single individual indicator contributes to the final index is far from clear. For such a broad composite index there is a trade-off between transparency and comprehensiveness. With more indicators there is more information but it becomes more difficult to interpret.

## **4.2 Some methods of dimension reduction**

There are numerous methods for taking two or more variables and reducing the information into a smaller set of variables. For the purposes of measuring deprivation, it is often preferred to reduce the data to only one or two composite measures where possible. If variables are combined that are indicators for disparate

forms of deprivation, then a multi-dimensional composite is probably required. The following sections outline a small number of dimension reduction techniques available.

#### 4.2.1 Combining z-scores

A simple composite measure is to standardise all of the variables prior to combining them with a simple summation. Equation 4.1 gives the formula for standardisation. The standardisation avoids problems of combining variables of differing scales. During combination, weights can be applied to the variables in accordance with a decision regarding the relative importance of the variables being used. Without a clear reasoning for choosing a set of weights, the choice is entirely arbitrary.

$$z_i = \frac{(x_i - \bar{x})^2}{\sigma} \quad (4.1)$$

Where:  $x_i$  = observation for area  $i$

$\bar{x}$  = mean

$\sigma$  = standard deviation

This method is used to produce the Carstairs<sup>178</sup> and Townsend<sup>288</sup> deprivation indices, where equal weights are used for all variables. Weighted combinations of z-scores have been used in two other UK indices of deprivation with arbitrarily chosen weights used by the DoE<sup>180</sup> and empirically derived weights used by Jarman<sup>176</sup>. In the latter three indices, some or all of the variables are also transformed using either a natural log or arcsine function.

#### 4.2.2 Multidimensional scaling (MDS)

In MDS, variables are combined on the basis of a statistical distance matrix. It is possible to define how many dimensions will be used in the final solution and the result is found through an iterative procedure. The end result is a 'map', called an

ordination, in which observations that are similar in terms of their profiles across variables are near each other and dissimilar observations are far apart.

The resulting ordination can be quite difficult to interpret and, depending on the data being used, it is possible that more than three dimensions are required to adequately separate groups of similar areas. A further difficulty is that the ordination is a conceptual map which is not intuitive. MDS can facilitate the labelling of areas indicating the level of deprivation but it is not an ideal methodology.

### **4.2.3 Clustering**

A number of clustering methods, such as hierarchical and k-means clustering, were detailed in a previous chapter. The appeal of clustering methodology is that it will result in a label, in this case the cluster identifier, which can be attached to each area.

One of the characteristics of many clustering algorithms is the requirement to define the number of clusters in advance of analysis. In the context of deprivation it is unclear how many clusters there might be. Furthermore, the interpretation of the clusters and their relative positions may be difficult. It is unlikely that the clusters can be neatly ordered on a one-dimensional scale. Odoi et al. used k-means clustering and principal components analysis of socioeconomic variables to group census tracts in Hamilton, Canada.<sup>305</sup> Using 18 variables, the clustering reduced the variables to five clusters while principal components returned a five-dimensional index.

### **4.2.4 Principal components analysis (PCA)**

In PCA  $n$  new uncorrelated variables called principal components are generated from the original  $n$  correlated dependent variables using an orthogonal transformation. The components are ordered so that the first accounts for the largest proportion of the variation in the original data. It is typically hoped that a



small number of components might account for sufficient variation to give an adequate summary of the original data, thus the  $n$  variables can be reduced to a smaller number of components. The components are continuous variables that can, for example, be divided into deciles. Each component is a linear combination of the original  $n$  variables.

It is a requirement of PCA that there is a certain amount of correlation between the variables. If this is not the case then as many components as original variables will be required, rendering the PCA purposeless. PCA is a purely mathematical technique and does not have an underlying statistical model and no assumptions are made about the distributions of the variables used.

The components that result from the PCA can be difficult to interpret. A visual inspection of the weights associated with each variable can be used to develop an interpretation although with less significant components that becomes increasingly difficult. Additionally, no error structure is implied in PCA.

#### **4.2.5 Factor analysis (FA)**

FA bears similarities to PCA in that it derives new variables from the set of supplied variables. The new variables are called factors and they are estimated as latent variables in that the factors are assumed to be indirect measures of the unmeasured independent variables. Unlike PCA, FA produces error terms for the variance that is unexplained by the factors. FA is also regarded as having a conceptual model, in contrast to PCA. FA is referred to as confirmatory analysis – that is, the analysis may be used to support a predefined theory.

When using FA it is assumed that there is a prior knowledge of how many factors exist. Typically a range of factor numbers are tested but the factor loadings can vary substantially depending on how many factors are chosen. This is further complicated by the fact that rotation can be applied to the factors to produce a

different set of factors. It can therefore be tempting to generate a post hoc justification for the number of factors and the method of rotation chosen.

#### **4.2.6 Comparison of dimension reduction techniques**

Combining z-scores is a simple procedure and has been used to produce deprivation indices in the past. As has been mentioned, however, the question of weights arises. In the case where all of the variable pairs have very similar correlations, the resultant un-weighted combination will be very similar to the first component of a PCA analysis. Any choice of weights may prove highly subjective and open to criticism. It is possible to use survey data as a basis for weight development. The procedure does not give any indication as to how much variance has been accounted for by the sum of z-scores.

While multidimensional scaling and clustering techniques may result in well defined groupings, those groups may not be readily explicable or meaningful in the context of deprivation. Clustering may help to show if area A is similar or dissimilar to area B, but understanding if it is more or less deprived might prove quite difficult. For any given observation, the profile of the variable values will dictate which cluster it is assigned to. This can result in very small clusters for which only a handful of areas share the same profile. With a large set of deprivation indicators it may be difficult to discern a useful set of clusters.

Both PCA and FA develop weights associated with each variable based on how it correlates with the other variables. As a consequence, no prior knowledge of which indicators are more or less important is required. A possible drawback to FA is that it assumes that there are underlying factors in the first place. PCA does not make such an assumption - it merely combines the variables into new variables. This fact leads to a further problem with FA which is that the results are dependent on the choice of the number of factors and the method of rotation used. Two researchers working with the same dataset could potentially find evidence to back up two very different theories based on how they extracted the factors.

In contrast to combining z-scores, PCA and FA provide a basis for deciding objectively what weights should be applied to each variable. While multidimensional scaling and clustering may link similar observations, they do not produce a continuous variable or relative scale whereby areas can be compared. Principal components may be difficult to interpret but this is also true of estimated factors in FA. As pointed out in Velicer and Jackson's paper,<sup>306</sup> despite the supposed differences between PCA and FA, they both produce surprisingly similar results. This is debated by Bentler and Kano<sup>307</sup> who state that the similarity is diminished when the number of variables is small although they concede that when dimension reduction is the primary concern then PCA is as useful as FA. In the current context, given that FA is open to manipulation to achieve particular results, PCA is the preferable technique for dimension reduction.

### 4.3 Principal Components Analysis

PCA was first used by Karl Pearson in 1901 and was further developed by Hotelling and others in the 1930s to arrive at the methodology described in the following section.<sup>308</sup> PCA has been used in climate studies, financial data analysis, information retrieval and pattern recognition amongst other disciplines.<sup>309 310</sup> Most of the major statistical and mathematical software packages including R, SAS, SPSS, S-Plus, Mathematica, Stata and JMP contain functions to calculate PCA.

#### 4.3.1 Method

This section outlines the mathematical details of PCA following Feinstein.<sup>277</sup> Given a set of  $n$  variables with  $m$  observations supplied in a matrix  $\mathbf{Y}$  such that:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1n} \\ \cdots & & \cdots \\ y_{1m} & \cdots & y_{nm} \end{bmatrix} \quad (4.2)$$

To determine the principal components of  $n$  variables, the first step is to calculate the  $n \times n$  covariance or correlation matrix. In the case of a correlation matrix, the

variables are first standardised so that the correlations are independent of the measurement units. The correlation between two variables,  $\mathbf{X}$  and  $\mathbf{Y}$ , is defined as:

$$r = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m (x_i - \bar{x})^2 \sum_{i=1}^m (y_i - \bar{y})^2}} \quad (4.3)$$

The correlation matrix,  $\mathbf{R}$ , is an  $n \times n$  matrix with the diagonal elements equal to 1. The next step is to determine the  $n$  eigenvalues,  $\lambda_k$ , and respective eigenvectors,  $\mathbf{e}_k$ , of  $\mathbf{R}$ , where  $k = 1, \dots, n$ . The eigenvectors form an  $n \times n$  matrix,  $\mathbf{U}$ .

The matrix of principal components,  $\mathbf{F}$ , can be calculated by multiplying a matrix of standardised observations,  $\mathbf{S}$ , by the eigenvectors,  $\mathbf{U}$ . The matrix  $\mathbf{S}$  is an  $n \times m$  matrix where each observation is standardised so that:

$$s_{ij} = \frac{y_{ij} - \bar{y}_j}{\sigma_j} \quad (4.4)$$

$$i = 1, 2, \dots, m; j = 1, 2, \dots, n$$

Thus,  $\mathbf{F} = \mathbf{S}\mathbf{U}$ . The contribution of each principal component to the model is given by:

$$c_k = \frac{\lambda_k}{\sum_{i=1}^n \lambda_i} \quad k = 1, 2, \dots, n \quad (4.5)$$

### 4.3.2 Choosing how many components to retain

For perfect reconstruction, all  $n$  principal components must be used, presuming that no two variables have a perfect correlation. A number of rules for deciding

how many components to retain have been developed. The most basic rule is the Kaiser-Guttman rule which states that all components with an eigenvalue greater than 1 should be retained.<sup>297</sup> A component with an eigenvalue less than 1 effectively explains less variance than is explained by one of the original variables. A similar but slightly more stringent rule is the Jolliffe criterion which states that at least 70% of the variance should be explained by the retained components.<sup>311</sup> Also available is the scree graph which plots the eigenvalue of each component developed by Catell.<sup>297</sup> The number of components retained is equal to the number of eigenvalues included at the point where an approximately horizontal straight line is reached on the graph.

All of the above criteria are relatively arbitrary and tend to lead to over-retention of components.<sup>306 312</sup> More sophisticated methods such as Velicer's Minimum Average Partial Method (MAP)<sup>313</sup> and Parallel Analysis (PA)<sup>314</sup> exist. The former method partials out the main component on successive iterations and seeks the point of minimum average correlation to indicate the number of components to retain. The latter method uses Monte Carlo simulation of random datasets with the same number of variables and observations as the real data to generate correlation matrices for which the eigenvalues are determined. Typically the 95<sup>th</sup> percentile of each eigenvalue is used as a cut-off value and if the corresponding eigenvalue from the original data is larger then the corresponding component is retained. It is time consuming to calculate PA repeatedly although attempts have been made to replace the simulation with a simple regression<sup>315</sup> or with pre-computed tabulated values from Monte Carlo simulation<sup>316</sup>.

Despite the computational burden, PA is recommended as the most appropriate methodology for deciding how many components to retain. All subsequent analysis in this study will use PA to determine how many components to retain.

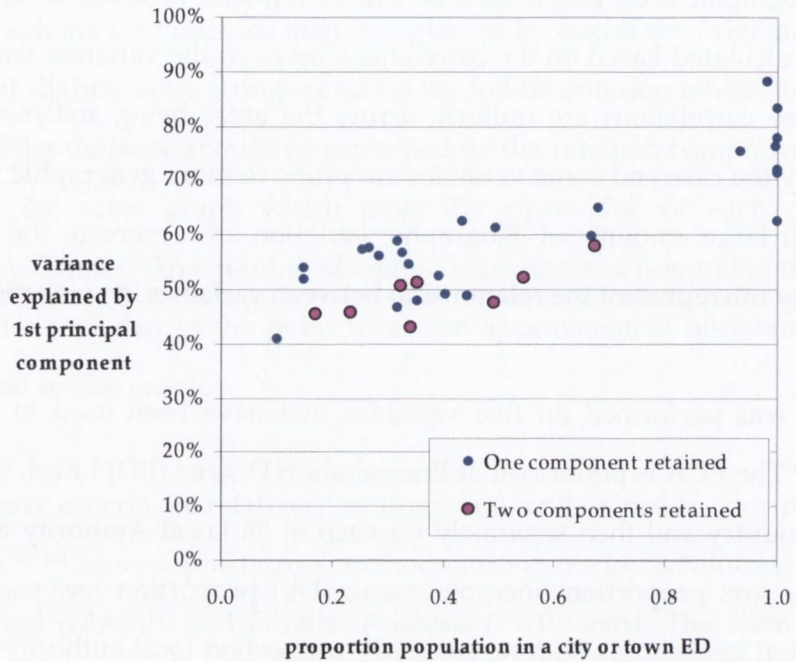
### 4.3.3 Problems with regional variation

When PCA is applied to area level data, where each observation represents values for a geographic area, problems arise due to regional variation in the data. As the PCA is calculated based on the correlations between the variables used, it assumes that those correlations are uniform across the areas being analysed. This is not generally the case and some variables are prone to more geographic variation than others. If large amounts of geographic variation are observed, the results of the PCA may misrepresent the relationship between variables in some regions.

A PCA was performed on five variables that have been used in a deprivation index.<sup>183</sup> The PCA is performed at Enumeration District (ED) Level, initially for the whole country and then separately for each of 34 Local Authority areas. The five variables are: proportion unemployment, (UE) proportion low social class (SC), proportion households with no car (NC), proportion local authority housing (LH) and overcrowding (OC).

Of the 34 Local Authorities (LAs), 8 require a second principal component to be retained. In Figure 4.1 the proportion population living in a city or town ED is plotted against the percentage variance explained by the first principal component. The LAs with two principal components retained are distinguished from those with only one retained.

Figure 4.1 Proportion city and town population versus the percentage variance captured by the first principal component ( $R^2 = 0.72$ )



It is clear from Figure 4.1 that the fit of the first principal component is better for areas with a higher proportion urban population. This indicates that the model is more applicable in predominantly urban LAs and that perhaps the variables do not adequately capture deprivation in the more rural areas. However, it does not follow that this relationship holds at an ED level.

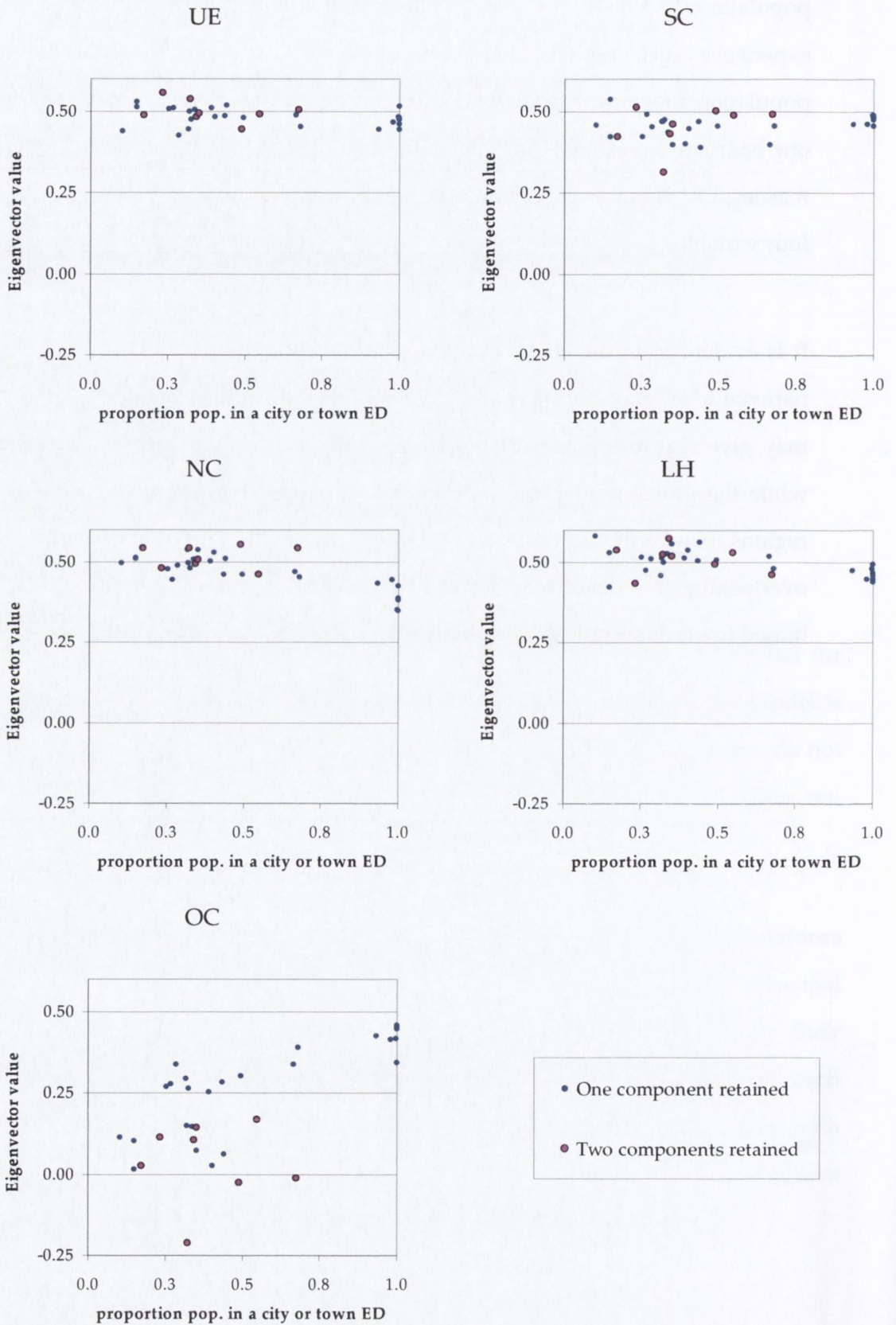
Figure 4.2 shows the plot of proportion urban population against the eigenvectors for the first principal component for each of the five variables. As the principal component value is the sum of the standardised variables multiplied by their respective eigenvector values, the eigenvector gives a relative weight for each variable. A very low value indicates that the variable does not correlate well with the other variables whilst a negative value indicates predominantly negative correlations with the other variables.

For some areas, notably more rural LAs, OC is negatively correlated with the other variables. It also has a significant positive correlation with the proportion urban population ( $R^2 = 0.468$ ,  $p < 0.0001$ ). There are many reasons why this variable may experience such regional variation: housing is cheaper in rural areas; the population movement from rural to urban areas; the proliferation of studio and one-bedroom apartments in city centres. Irrespective of the possible underlying reasons, OC is not as consistent a proxy for deprivation with respect to the other four variables.

It is evident from the above example that the correlations between variables at a national level may not hold at a regional level. Assuming a single global model may give rise to a model that performs poorly in certain regions. Furthermore, while the global model may be reduced to a single principal component, some regions may only be reduced to two or more principal components. Again, overlooking this detail may result in a misleading model of deprivation that is biased towards specific geographic areas.



Figure 4.2 Proportion city or town population against eigenvector values for the first principal component for the five variables



## **4.4 Potential sources of error in deprivation indices**

There is an inherent assumption that a deprivation index is a reasonable reflection of the actual distribution of deprivation in an area. In the process of generating a deprivation index, there are numerous instances where the final index may be affected to a greater or lesser degree by a single step. These range from assumptions about the validity of indicators to regional bias introduced by the choice of indicators. Some of these problems will be dealt with in this section.

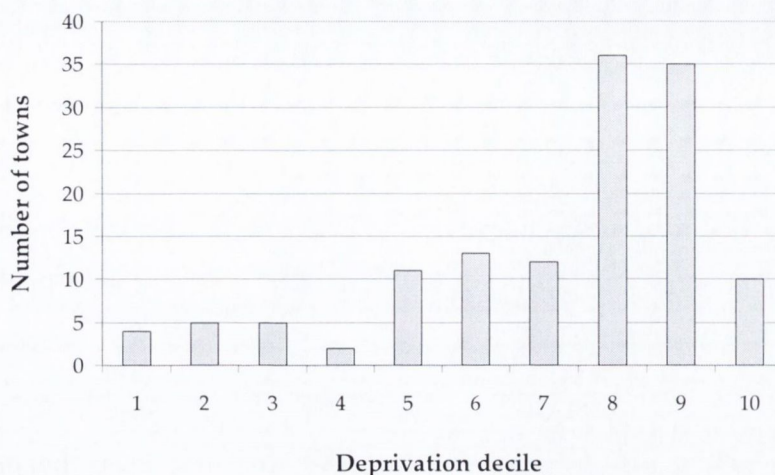
### **4.4.1 Heterogeneity in small areas**

As a deprivation score is typically calculated at a small area level, it results in a label being attached to an area as it is not possible to label individuals. It should be noted, however, that not every individual living in a deprived area is deprived, and not every individual living in an affluent area is affluent. The demographic heterogeneity present in small areas means that there will always be a mix of individuals in an area experiencing different levels of personal deprivation. As a consequence of this, it is possible that there may be 'pockets' of deprivation in an area that is considered non-deprived or even affluent. The characteristics of one part of the community in the ED effectively mask the presence of the other part of the community by way of an averaging effect.

From a deprivation measurement perspective, the boundaries of small areas would ideally be delineated to maximise the homogeneity of the community living in that area. Naturally this does not happen and without individual level statistics it is not possible to assess the degree of heterogeneity present in a given small area. An assessment of heterogeneity in New Zealand small areas by Salmond and Crampton looked at the population comparing individual-level and area-level deprivation.<sup>317</sup> They found that 14.0% of individuals with few or no deprivation characteristics lived in areas in the three most deprived deciles. They also noted that if resources were only targeted at areas in the three most deprived deciles, these areas would miss 13.9% of deprived individuals.

In Ireland, there is substantial variation in the population size of small areas. In the 2002 census, ED populations ranged from 55 in Branchfield ED in Sligo to 24,404 persons in Blanchardstown-Blakestown in Dublin. Pringle pointed out that large EDs tend to be more socially heterogeneous, although he goes on to say that EDs in cities tend to be more socially homogeneous.<sup>318</sup> This can be explained by the fact that a rural ED with a large population will tend to cover a large geographic area and thus encompass a greater range of communities while a city ED will, due to higher population densities, cover a smaller area and hence a more homogeneous community. Pringle contends that this problem of rural heterogeneity is particularly problematic in medium sized towns where the entire town may fall within a single ED. In such instances a heterogeneous population is almost inevitable and the consequent averaging effect guarantees a deprivation score close to the average, although he does not perform any analysis to back up this assertion. As a simple test using 2002 census data, 133 towns and villages were identified, each of which is located in a single ED and comprised more than 50% of the ED population. The 133 towns and villages ranged in size from 233 to 18,288 persons, and were labelled with the 2002 SAHRU deprivation decile<sup>183</sup> of the ED they occupied. The frequency of each deprivation decile is given in Figure 4.3 below. For Pringle to state that many of the towns fall into EDs that have scores close to the national average appears to be incorrect. Of course, this is partly dependent on the deprivation index used and an index with urban bias may tend to generate high deprivation scores for EDs with a predominantly town population.

Figure 4.3 The numbers of towns located in single ED by deprivation decile



One method for dealing with heterogeneity would be to use indicators that cover different ends of the spectrum, the intent being to identify areas that show signs of both deprived and non-deprived communities. For example, the proportion of households with no car could be compared with the proportion of households with 2 or more cars. By using measures of both deprivation and affluence, it may be possible to highlight areas with possibly disparate communities where some masking of deprivation may occur. However, even if such areas could be highlighted, there is little that could be done to adjust for heterogeneity other than possibly reviewing the appropriateness of the chosen indicators.

#### 4.4.2 Regional bias and spatial autocorrelation

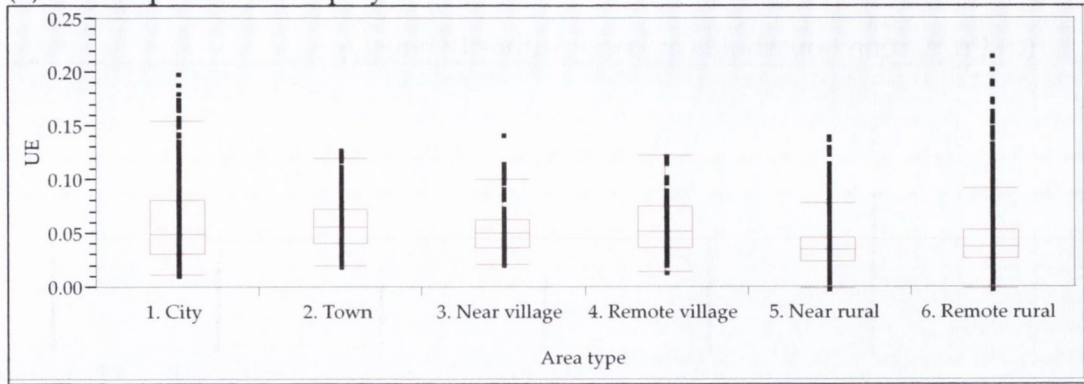
None of the deprivation indices outlined in section 4.1 have included either an analysis of rural-urban variation or an assessment of spatial autocorrelation in the indicators used. Both are simple analyses to conduct and would assist in the identification of variables that might introduce regional bias. Spatial autocorrelation gives a measure of how similar observations are to nearby observations and can be quantified using Moran's I.<sup>266</sup> A variation on Moran's I is the Local Indicator of Spatial Autocorrelation (LISA)<sup>319</sup> which facilitates the analysis of correlations locally. While Moran's I provides a single measure across all areas, LISA can be used to assess correlation at each small area.

The context in Ireland is complicated by two main factors: the large variation in ED population and the proportionately small number of urban EDs. If EDs had a uniform population size, the majority of EDs would be urban as the majority of the population live in urban areas. As this is not the case, the majority of observations in a PCA or FA are rural EDs. For variables that show a strong urban bias this means that this bias is only increased by PCA or FA. This can be illustrated with the variable for proportion no car ownership which has some of the lowest values in the correlation matrix using data for urban areas only. As a consequence, in an urban only index the proportion no car ownership variable receives a lower weight, making it a less important variable. For rural EDs, the proportion car ownership has relatively higher values in the correlation matrix resulting in a higher weight. When all EDs are used, the weight for no car ownership is higher due to the dominance of rural EDs in the calculations. However, this also means that when coupled with the higher proportion no car ownership observed in urban areas, the urban EDs appear even more deprived.

As an example, a deprivation index was constructed with the five variables defined in section 4.3.3 previously: proportion unemployed, (UE) proportion low social class (SC), proportion households with no car (NC), proportion local authority housing (LH) and overcrowding (OC). The graphs in Figure 4.4 (page 144) plot the five variables by area type. The NC and LH variables appear to decrease with increasing rurality. For NC this is not surprising as more urban areas have greater public transport options which impact on car ownership. For LH, it is likely that local authority estates will be built in cities and towns rather than sparsely populated rural areas. This is a policy restriction rather than the choice of people living in LA housing. OC does not appear to correlate with area type. What is certain is that it does not correlate well with other deprivation indicators in rural areas suggesting that it is, at best, a regionally inconsistent measure of deprivation. While these three variables all undoubtedly capture some elements of deprivation, they are regionally biased.

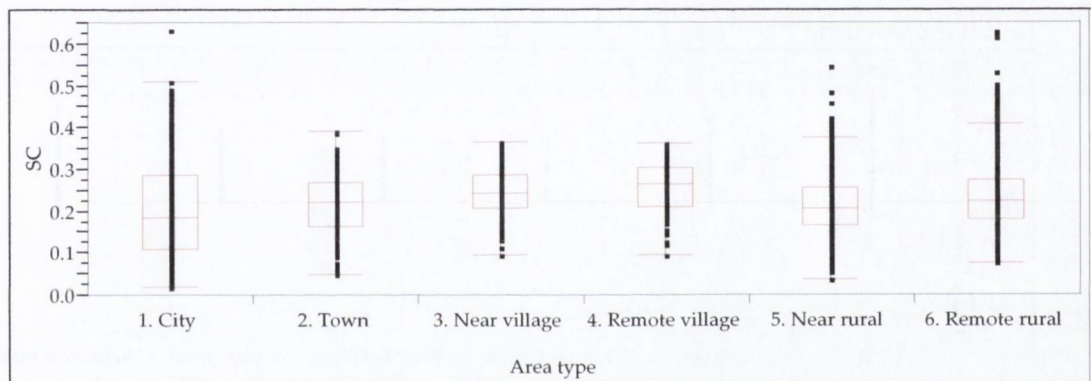
Figure 4.4 Plots of variables by area type

(a) Proportion unemployed



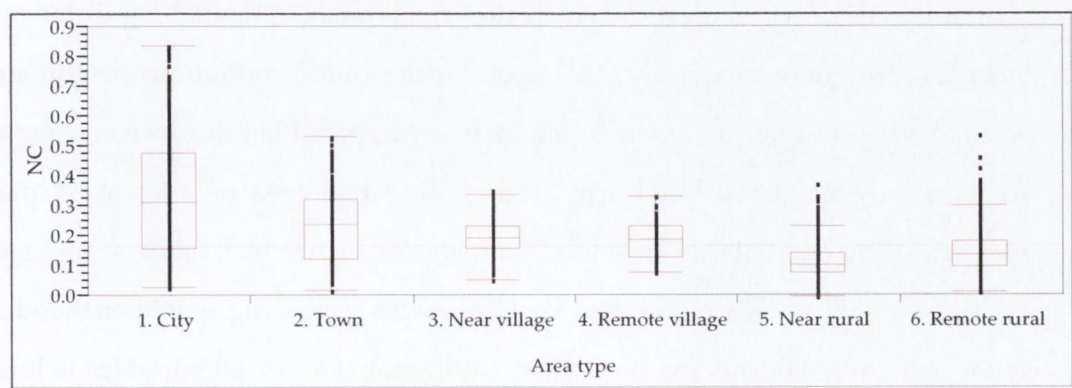
(Kruskal-Wallis:  $\chi^2 = 333.7$ ,  $df = 5$ ,  $p < 0.0001$ )

(b) Proportion low social class



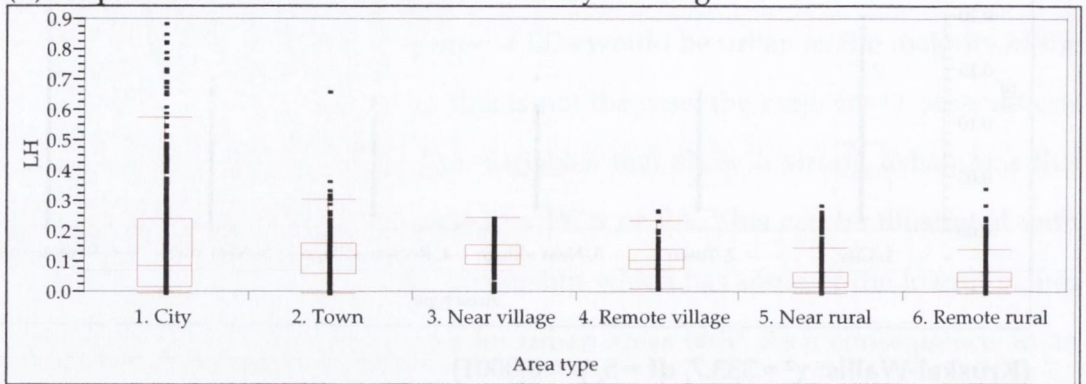
(Kruskal-Wallis:  $\chi^2 = 118.2$ ,  $df = 5$ ,  $p < 0.0001$ )

(c) Proportion households with no car



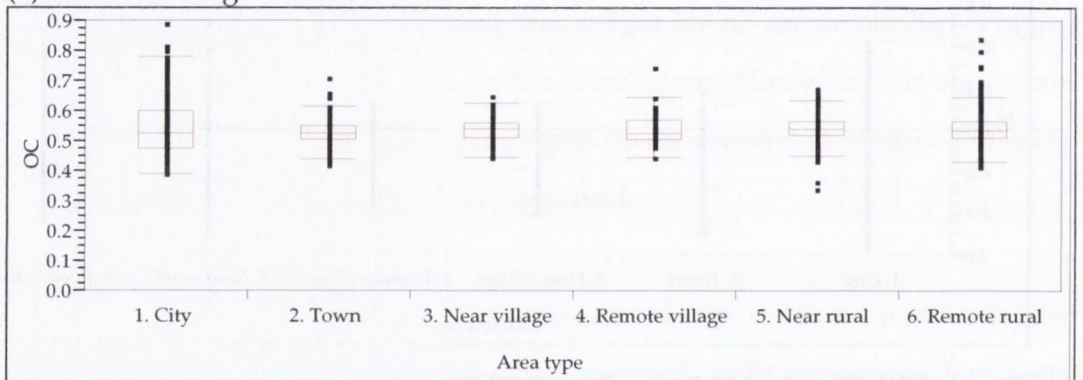
(Kruskal-Wallis:  $\chi^2 = 891.3$ ,  $df = 5$ ,  $p < 0.0001$ )

(d) Proportion households in Local Authority housing



(Kruskal-Wallis:  $\chi^2 = 584.8$ ,  $df = 5$ ,  $p < 0.0001$ )

(e) Overcrowding

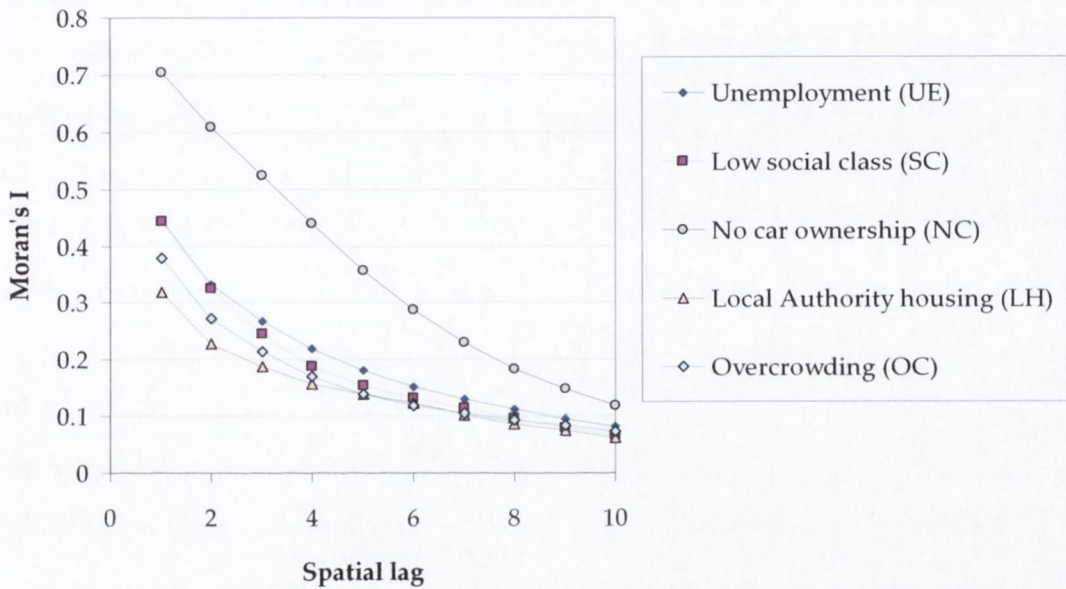


(Kruskal-Wallis:  $\chi^2 = 33.0$ ,  $df = 5$ ,  $p < 0.0001$ )

Spatial lag distance relates to neighbouring small areas. A spatial lag distance of 1 includes contiguous neighbours. A lag of 2 includes the contiguous neighbours of areas at a lag distance of 1, and so on. Increasing spatial lag distance encompasses an increasing neighbourhood size around the small area of interest. A plot of Moran's I by spatial lag distance, shown in Figure 4.5, shows the spatial autocorrelation of each of the five variables with increasing neighbourhood size. With large neighbourhoods, the spatial autocorrelation for all variables is low. At short lag distances, however, very high spatial autocorrelation is observed for the car ownership variable. Further evidence for this can be seen in

Table 4.7 (page 165), where the mean LISA values were given by area type. All of the variables have high mean LISA values in city EDs suggesting that neighbouring EDs tend to have similar levels of unemployment, etc. in urban areas. In rural areas, EDs tend to be more varied and neighbouring EDs are more likely to have different characteristics suggesting that there is less of a neighbourhood effect.

Figure 4.5 Moran's I by spatial lag distance



A PCA is conducted using the five variables for three groupings of EDs: all EDs, city and town EDs and finally, village and rural EDs. The eigenvector values for the first principal component are shown in Table 4.3 below. The car ownership variable has a lower weight based on an urban subset of EDs than on a rural subset while the converse applies to overcrowding.



Table 4.3 Eigenvector values using subsets of EDs

Indicator	All EDs	City & town EDs	Village & rural EDs
UE	0.4891	0.4785	0.4946
SC	0.4403	0.4691	0.5060
NC	0.4430	0.4047	0.4672
LH	0.4963	0.4659	0.4719
OC	0.3526	0.4125	0.2416

To understand the implications of the different sets of weights, each set was applied to all EDs to derive three deprivation indices. For convenience, city and town has been labelled ‘urban’ and village and rural has been labelled ‘rural’. In Table 4.4 the percentage EDs in the least and most deprived deciles is shown by area type. As a decile contains 10% of EDs, the expectation would be that for any given area type 10% of the EDs would be in any specific decile. An interesting and counter-intuitive result is that an index based on correlations in urban EDs increases the number of rural EDs in the most deprived decile. This is due to lowered weights for NC and LH, both of which tend to have lower values in rural areas, and greatly increased weight for OC which has a slightly higher mean in rural areas. Conversely, an index based on correlations in rural areas leads to an increased number of urban EDs in the most deprived decile.

Table 4.4 Percentage EDs in the least and most deprived deciles using PCA based on different subsets of EDs

Area type	Least deprived decile			Most deprived decile		
	All	Rural	Urban	All	Rural	Urban
1. City	21.0	18.4	21.8	36.6	40.3	34.7
2. Town	6.0	6.8	6.0	18.8	20.9	18.4
3. Near village	2.5	3.1	3.1	6.3	7.5	6.9
4. Remote village	2.8	1.4	2.8	16.9	14.1	15.5
5. Near rural	11.6	12.8	11.1	2.7	1.9	2.7
6. Remote rural	6.1	5.7	6.2	6.0	5.0	6.8

These findings are based on an index produced with a specific set of indicators and a different selection may produce quite different results. But this is a salient point – the choice of variables can have numerous impacts on a deprivation index, some of

which will introduce bias but in a potentially counter-intuitive manner. There are sufficient tools available to identify variables that might bias a deprivation index towards urban or rural areas and yet researchers do not appear to address this issue. The Irish SAHRU index was recalculated at a regional level, acknowledging that significant regional variation existed.<sup>183</sup>

### **4.4.3 Issues relating to small area boundaries and definitions**

Choosing the level of spatial aggregation for an analysis can have a number of consequences for the results. Two of the main processes affecting the results are the modifiable areal unit problem (MAUP) and the ecological fallacy. Each will be dealt with in this section.

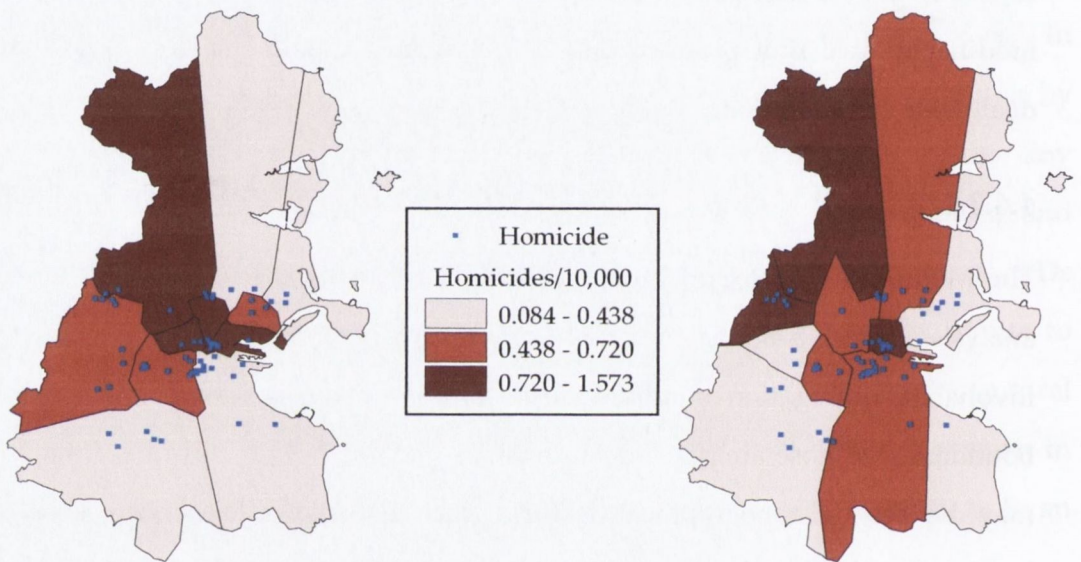
#### **4.4.3.1 MAUP**

The Modifiable Areal Unit Problem (MAUP) relates to the use of areal units for analysis. Unless a spatial analysis is conducted at an individual level, it will involve the aggregation of individuals to some defined set of geographic areas. The boundaries of those areas, or areal units, are generally quite arbitrary. Boundaries may be set by topographical features such as rivers, lakes, roads or field boundaries. If the boundaries are redrawn then individuals are aggregated differently and the results of a spatial analysis may be quite different. This is, in essence, the MAUP. The MAUP will occur whenever data are aggregated into areal units and this is the case with deprivation indices, which rely on census and other small area data sources. It is desirable to analyse at the lowest level of aggregation possible which is frequently some form of enumeration district.

The impact of the MAUP can be shown through a simple example. The point locations of homicides in Dublin from January 2004 to August 2006 were obtained from [dublincrime.com](http://dublincrime.com)<sup>320</sup> and mapped. The homicides were aggregated into two different sets of arbitrarily chosen boundaries. Figure 4.6 shows the rate per 10,000 persons for the two different choices of boundaries. The change of boundaries affects which areas appear to have higher rates.

The MAUP cannot be adjusted for *per se*, it can only be taken into account in the form of caveats. It was suggested by Openshaw that some form of optimisation procedure be used when aggregating data into zones.<sup>266</sup> For example, if the hypothesis was that homicide rates were correlated with unemployment then zones would be delineated to maximise the correlation between unemployment rate and homicide rate.

Figure 4.6 Effect of changing boundaries on homicide rates



Raw data source: <http://www.dublincrime.com>

In the process of generating a deprivation index it is not desirable to aggregate areas unless the populations are so small as to make the indicators unreliable. In the Irish context this is not really a problem as the EDs are generally quite large. In the UK, where post code level data are sometimes available, it is generally required to aggregate data either for confidentiality reasons or to make the data compatible with other area-level data. Aggregation may also be required if boundaries change between censuses, as happens in both UK and to a much lesser degree in Ireland.

#### 4.4.3.2 The ecological fallacy

An issue that is linked to MAUP is that of the ecological fallacy.<sup>266</sup> Ecological fallacy arises when a relationship observed at one level of aggregation is assumed to hold

at a different level of aggregation. It generally occurs with a relationship being observed at a high level of aggregation, such as county level, being assumed to hold at a lower level of aggregation, such as ED or even individual level. As Pearce points out, however, it also happens in the reverse direction producing the individualistic fallacy.<sup>321</sup> An example of the ecological fallacy in terms of deprivation indices would be the inference that if an ED is deprived then all of the inhabitants of that ED must be deprived. As was shown in New Zealand this is rarely, if ever, the case.<sup>317</sup>

Research by Lancaster and Green examined the induced bias due to ecological fallacy in studies linking deprivation and ill-health.<sup>322</sup> They looked at how effectively individual-level and area-level deprivation indices explained rates of limiting long term illness. They found area-level effects distorted the relationship between health and the deprivation indicators compared to the individual-level analysis. Latent variable models were improved by accounting for the interaction between age and deprivation. They conclude that inclusion of age effects into a deprivation score produces results that are more consistent with an individual level analysis. Latent variable models were also used by Hewson when investigating deprivation and child pedestrian accidents.<sup>323</sup> Hewson also suggests using Council Tax valuation band as a reliable individual-level indicator of deprivation.

Salway and Lakefield discussed bias in ecological studies of non-rare events.<sup>324</sup> They conclude that bias is much more difficult to characterise in studies involving non-rare events. This has implications for deprivation indices as the indicators used typically affect larger numbers of persons than rare disease events or even more common causes of morbidity and mortality. Waldron noted that to minimise the effects of the ecological fallacy, studies should use smaller area aggregations to maximise the homogeneity of the populations within those areas.<sup>325</sup> This recommendation cannot be generalised and it is argued that the appropriate area size is context specific – sometimes more might be better than less.<sup>326</sup> The use of

very small areas can introduce problems of sampling error that may be diminished at higher levels of aggregation.

MacRae pointed out that although the ecological fallacy is a problem, substantial evidence exists to show that correlations between deprivation and ill-health do exist at an individual-level as well as an area-level.<sup>327</sup> Therefore, he concludes, the ecological fallacy is not a legitimate criticism of studies correlating deprivation and ill-health. While Ben-Shlomo and Davey Smith argue for both individual-level and area-level data to be collected and analysed simultaneously, they concede that frequently the availability and quality of individual-level does not allow for this.<sup>326</sup>

These recommendations, although useful, are also somewhat conflicting. While latent variable models may offer some form of solution, without individual-level data for comparison it is not possible to assess whether the resultant measure of deprivation is 'better'. Numerous studies linking socioeconomic status or deprivation to health and living conditions have made reference to the ecological fallacy but only as a caveat to interpreting the results.<sup>166 208 328-331</sup> In conclusion, it is difficult to account for the ecological fallacy without access to individual-level data. Furthermore, individual- and area-level indicators have independent effects on health,<sup>326</sup> so that the interpretation is necessarily different.

#### **4.4.4 Temporal comparisons**

Having computed a deprivation index and identified areas of disadvantage, it is reasonable to assess how areas may have changed over time in terms of both ranking and disadvantage. Is an area better or worse off than ten years ago in both relative and absolute terms? The former is not so difficult to address – the change in ranking of an area will indicate whether it has improved or disimproved relative to the other areas. To measure the change in absolute terms is more difficult. For any given indicator it is possible to compare absolute change temporally. For a composite indicator, however, such a comparison is not straightforward. Inherent in the process of combination, be it by PCA or FA, is standardisation of the

variables. Many of the deprivation indicators show changes over time in both their mean and standard deviation. Due to this standardisation, scores from two different indices cannot be meaningfully compared. To illustrate the changes in mean, Table 4.5 shows the national proportions for four different indicators across four censuses. The changes between 1986 and 1996 are small but from 1996 to 2002 the changes were much greater. This is particularly pronounced in the unemployment and early school leaver variables. An ED with average values in 1986 would be considered quite deprived in 2002. Despite this, not all EDs improved in that time period. In fact, 3% of EDs had higher unemployment rates in 2002 than they did in 1986 even though the national rate more than halved.

Table 4.5 Change in selected indicator means from 1986-2002

Indicator	Year			
	1986	1991	1996	2002
Dependency ratio	0.398	0.381	0.351	0.323
Unemployment	0.179	0.169	0.148	0.088
Low socioeconomic group	0.138	0.147	0.195	0.155
Early school leavers	0.245	0.233	0.200	0.111

One option would be to standardise all of the variables simultaneously using a global mean and standard deviation which are either calculated across all datasets or set using a baseline such as the first or last year of data. In theory the scores would then be comparable.

A further problem is that if scores are calculated for a number of time periods, the weights associated with each variable will be different as correlations can be expected to change from year to year. The argument is that using different weights renders the scores incomparable and that the same weights must be applied to the data from each time period. Such an approach ignores the fact that the variables that contribute to the notion of deprivation change over time and this change should readily be accounted for, just as contributions change regionally. If the data are combined using PCA on correlations then the analysis is both scale-independent and the sum of squared weights sum to one. In a FA this is typically

not the case so that the same distribution of deprivation scores would simply not be possible from one time period to the next. Haase and Pratschke use structural equation modelling and FA with fixed weights between time periods to compare deprivation across time.<sup>185</sup> They do not indicate if variables were standardised across or within time periods.

A final problem with cross-temporal comparisons is that of indicator and boundary changes over time. For example, to facilitate changes in occupation structure the social class definitions in Ireland changed from 1991 to 1996 rendering back-comparison to 1986 impossible.<sup>332</sup> In Ireland, alternate censuses are more narrowly defined and do not collect what might be considered the full dataset. This means that some indicators that might be considered useful in a deprivation context are not available for every census. Boundaries also change to facilitate confidentiality in small areas and town boundary increases in rural towns. In small EDs where confidentiality may be compromised by publication of the results, the EDs will be merged with neighbouring EDs to increase population size. In some town EDs, part of the town occupies the neighbouring rural ED. In some census years, the town population in the rural ED is given as part of the town ED and other times not. The lack of consistency is unexplained but causes problems for comparability. In such cases, the town and rural EDs must be merged. Although undesirable, these merges must be made to each dataset to ensure common areas for valid comparison.

The difficulty with any approach to temporal deprivation comparisons is that it will undoubtedly ignore that what might be considered a deprived status has changed over time. Not owning a car might not have been considered a sign of deprivation 20 years ago although now it is. It can be seen from Table 4.5 that major changes in the Irish economy took place in the latter half of the 1990's and that expectations of what resources an individual might have access to will have changed drastically in that time. This suggests that temporal comparisons might be most effective in periods of relative economic stability when the definition of

deprivation might be reasonably constant. An assessment of absolute changes in deprivation might not be valid over a longer period of time and that relative change might be the most appropriate measure of change.

## 4.5 Geographically weighted PCA

A possible solution to the regional variation in the correlations of variables is to develop the notion of a geographically weighted PCA (GW-PCA). The details of how this can be achieved are detailed in this section.

### 4.5.1 Calculation of GW-PCA

PCA is performed for each of  $m$  small areas in turn, with the observations inversely weighted by their geographic distance from the area being considered. The weights are determined using some distance decay function such that  $w_{ir} = f(d_{ir})$  where the weight and distance between areas  $i$  and  $r$  are  $w_{ir}$  and  $d_{ir}$  respectively. Subscript  $r$  refers to the area being considered.

The correlation matrix  $\mathbf{R}$  is now comprised of weighted correlations so that:

$$r = \frac{\sum_{i=1}^m w_{ir} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^m w_{ir} (x_i - \bar{x})^2 \sum_{i=1}^m w_{ir} (y_i - \bar{y})^2}} \quad (4.6)$$

Where:  $\bar{x} = \frac{\sum_{i=1}^m w_{ir} x_i}{\sum_{i=1}^m w_{ir}}$

$$\bar{y} = \frac{\sum_{i=1}^m w_{ir} y_i}{\sum_{i=1}^m w_{ir}}$$

As in normal PCA, the eigenvalues and eigenvectors of  $\mathbf{R}$  are calculated based on  $n$  variables. The eigenvectors are stored for area  $r$  as the  $n \times n$  matrix  $\mathbf{U}_r$ . Standardised



scores are calculated in the same manner as for the standard PCA. The matrix of standardised scores is stored for each area as a  $1 \times n$  matrix,  $S_i$ . The principal components for area  $i$  are calculated as  $F_i = S_i U_i$ .

If all of the weights are equal to 1 (i.e.  $w_{ir} = d_{ir}^{-0}$ ), then the GW-PCA will return the same result as the standard PCA.

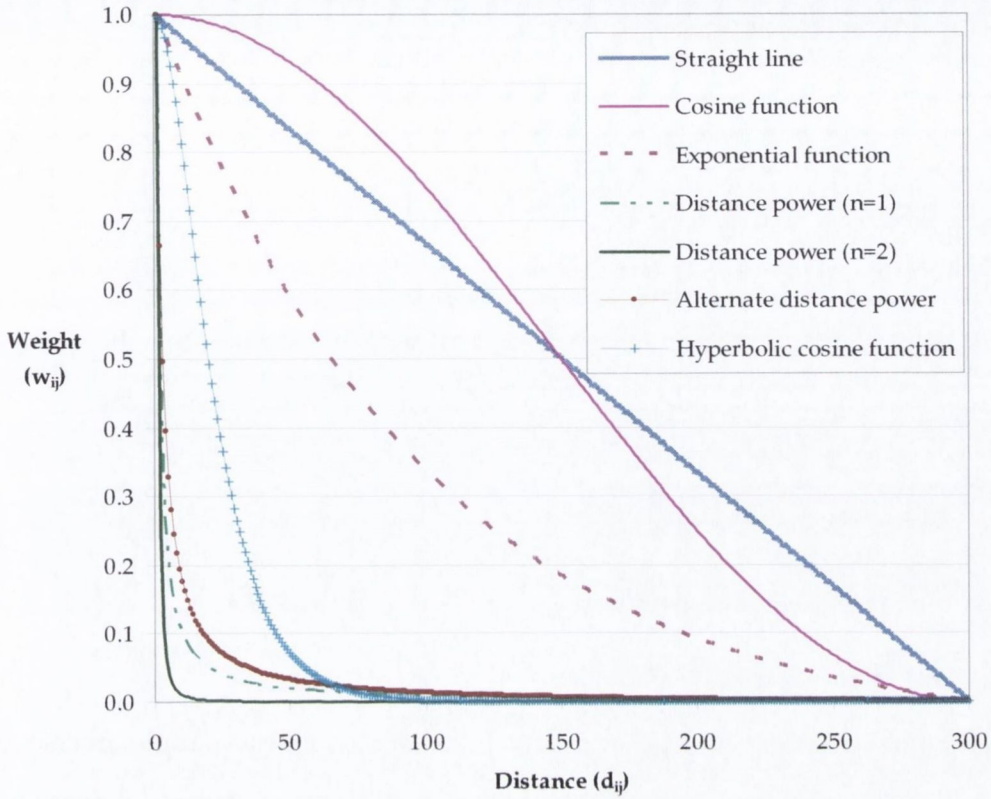
## 4.5.2 Distance decay

The distance decay function dictates the rate at which the influence of a small area diminishes with increasing distance. It is also possible to specify a maximum distance beyond which areas will be given a weight of zero. This maximum can be given as an actual distance or as a percentage of areas. For example, a maximum distance of 100km could be used, with all EDs more than 100km from the ED of interest being given an influence weight of zero. Alternatively, it could be specified that 10% of areas will be included in the analysis. With 3,422 EDs this entails that other than the nearest 342 EDs, all other EDs will have a weight of zero.

### 4.5.2.1 Different distance decay functions

A variety of distance decay functions have been tested. These functions have largely been derived from cooling functions used in simulated annealing.<sup>333</sup> In the following formulae,  $d_{max}$  refers to the largest distance between areas  $i$  and  $j$  such that all areas beyond that distance have a weight of zero. The weights generated by the decay functions are shown with increasing distance in Figure 4.7 below.

Figure 4.7 Comparison of distance decay functions ( $d_{max} = 300$ )



#### 4.5.2.1.1 Straight line

In this decay function, there is a direct linear relationship between weight and distance so that at a distance of zero from the ED being estimated the weight is one and at the maximum distance the weight is zero. In this way, an observation at half the maximum distance receives half the maximum weight.

$$w_{ij} = 1 - \frac{d_{ij}}{d_{max}} \tag{4.7}$$

#### 4.5.2.1.2 Cosine function

The incorporation of the cosine function into this formula produces a weighting scheme that initially gives similar weights before influence diminishes more rapidly. Up to half of  $d_{max}$  the weights are larger than for the straight line function

while beyond that distance the weights are smaller than for the straight line function.

$$w_{ij} = \frac{1}{2} \left( 1 + \cos \left( \frac{\pi \cdot d_{ij}}{d_{\max}} \right) \right) \quad (4.8)$$

#### 4.5.2.1.3 Exponential function

This function produces weights are always below those for the straight line function. The choice of parameters results in a moderate rate of decay.

$$w_{ij} = 20 \cdot e^{-\frac{1}{d_{\max}} \ln(20) \cdot d_{ij}} - 1 \quad (4.9)$$

#### 4.5.2.1.4 Distance power

The standard function used in many applications applies a simple inverse distance decay. The choice of power is typically a small integer value as large values will lead to a very rapid decay. The consequence of this rapid decay is that within a very short distance the weights drop to near zero which results in all but the very closest observations having little or no influence.

$$w_{ij} = d_{ij}^{-n} \quad (4.10)$$

Where:  $n$  is a positive real number, typically 2

#### 4.5.2.1.5 Alternate distance power

The previous function does not incorporate the value of  $d_{\max}$  and therefore tends to zero only at very high distances or for large values of  $n$ . This alternative formulation produces a similarly shaped curve that tends to zero for the chosen value of  $d_{\max}$ .

$$w_{ij} = \frac{A}{d_{ij}} + B \quad (4.11)$$

Where:  $A = \frac{(d_{\max} + 1)}{d_{\max}}$

$$B = 1 - A$$

#### 4.5.2.1.6 Hyperbolic cosine function

This function utilises the hyperbolic cosine, or cosh, function. The rate of decay is less rapid than either the distance power or alternate distance power functions although the weights do tend to zero relatively rapidly.

$$w_{ij} = \frac{1}{\cosh\left(\frac{20 \cdot d_{ij}}{d_{\max}}\right)} \quad (4.12)$$

#### 4.5.2.1.7 Comparison of distance decay functions

The curves in Figure 4.7 show the rate at which weights decrease for seven different decay functions. The distance power and alternate distance power curves results in very rapid decay of weight by distance. The straight line, cosine and exponential functions lead to very gradual decay by distance and generally give less regional distinction unless a restrictive cut-off is used. The hyperbolic cosine function has a gradual initial decay but tends toward zero by approximately a half of the cut-off or maximum distance.

In more rural areas, where EDs are generally larger in area, the distances between EDs are usually larger than in urban areas. With a very rapid distance decay function, all but the nearest EDs will have low weights close to zero. In that event, only a small number of EDs will influence the PCA.

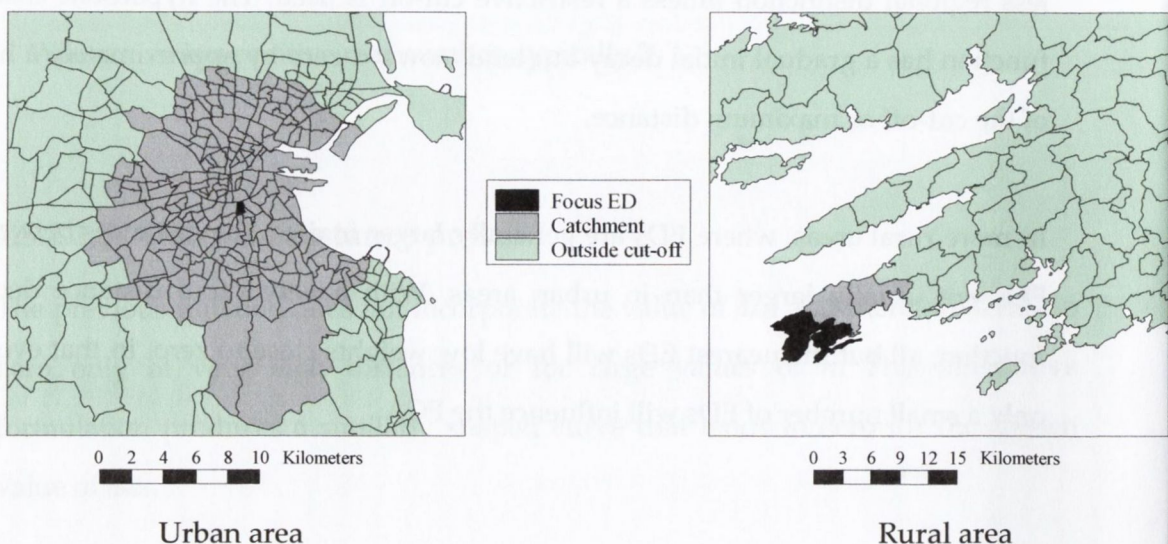
### 4.5.3 Cut-off specification

Choosing a cut-off distance for weighting is recommended where any regional variation may be lost by including all areas. Coming back to the first law of geography, areas that are close in space are typically quite similar in their attributes. By extension, areas that are distant in space may be quite dissimilar. The purpose of employing distance decay functions is to ensure that the attributes of nearby areas are given the greatest weight. The use of a cut-off distance can further ensure that areas that are distant do not unduly influence the findings. The region that falls within the cut-off is called the catchment area for the given ED of interest.

#### 4.5.3.1 Specifying a maximum distance

In this case, a maximum distance is specified beyond which areas are given a weight of zero. In urban areas, where EDs tend to be smaller in area due to the higher population density, a given maximum distance may encompass many more EDs than in a rural region, where EDs typically cover more area. An urban catchment will contain many times the number of EDs as a rural catchment with the same cut-off distance, as is shown in Figure 4.8 below. The numbers of EDs, including the focus EDs (Rathmines West B and Crookhaven, respectively), within 10km of the urban and rural focus EDs are 223 and 3 respectively.

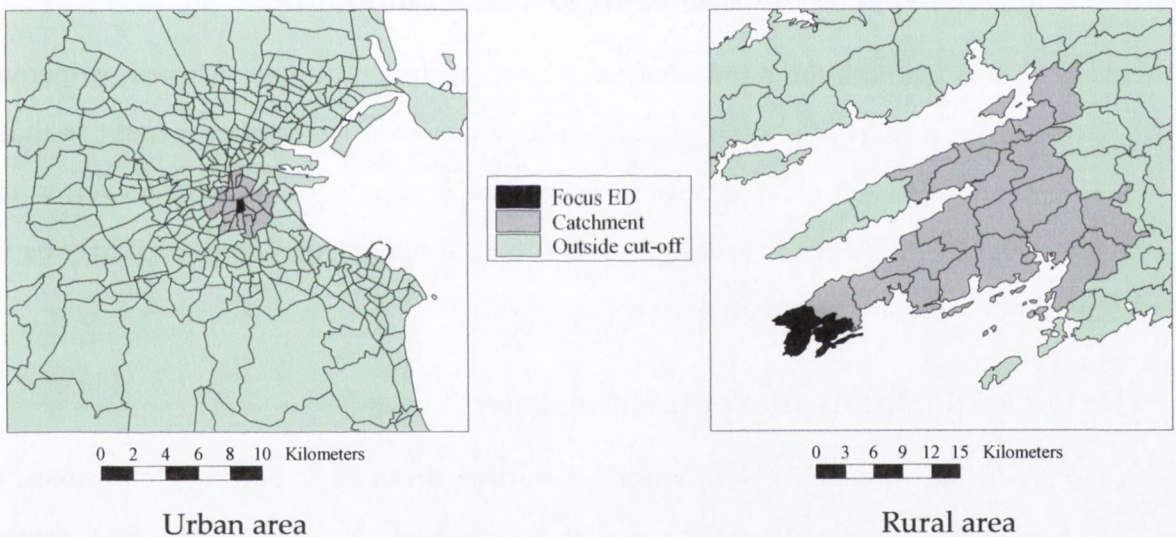
Figure 4.8 Catchment based on a cut-off distance (10km)



### 4.5.3.2 Specifying a number of areas to include

Rather than selecting a distance, a fixed number of areas, ranked by distance, were chosen. This overcomes the problem of differing ED sizes in urban and rural regions. It does, however, result in differing maximum distances covered. This is particularly evident for rural coastal areas. EDs in the centre of the country will have an approximately circular catchment while EDs in coastal areas or on peninsulas will have catchments with quite a distorted shape, as can be seen in Figure 4.9 below. The catchments containing 20 nearest neighbours for the urban and rural focus EDs span 2.86km and 35.96km respectively.

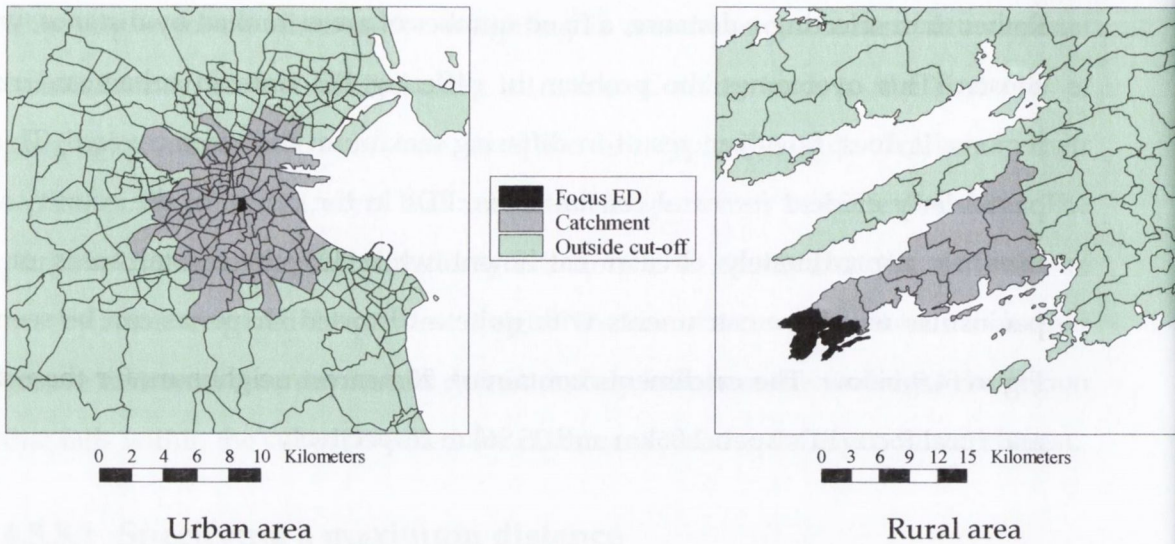
Figure 4.9 Catchment based on number of areas (nearest 20 areas)



### 4.5.3.3 Specifying a maximum lag distance

The benefits of this method are that it takes into account the fact that rural EDs tend to be further apart. Unlike using a fixed number of areas, this method also retains an approximately circular catchment shape. Catchments that encompass both urban and rural areas are still likely to be distorted as area sizes will not be homogeneous. Figure 4.10 shows catchments for an urban and rural ED respectively based on 5 lags. The urban catchment contains 127 EDs and spans 8.08km while the rural catchment contains 10 EDs and spans 24.50km.

Figure 4.10 Catchment based on maximum lags (5 lags)



#### 4.5.3.4 Specifying a percentage of the population

In this method, the number of neighbours included in the catchment is increased until a certain minimum population is included. The primary benefit is that all catchments will have an approximately equal population base. However, in rural areas the number of EDs required to fulfil the catchment requirement may be very large compared to an urban area.

#### 4.5.3.5 Finding the optimal distance

A further possibility is to specify a starting distance or percentage of areas, and increase the cut-off until a model is achieved where only the first principal component is retained. The cut-off may be increased by either one ED or one spatial lag at a time. Although very time consuming computationally, it should lead to results that use the minimum cut-off necessary for a good model.

#### 4.5.3.6 Recommendations for cut-off choice

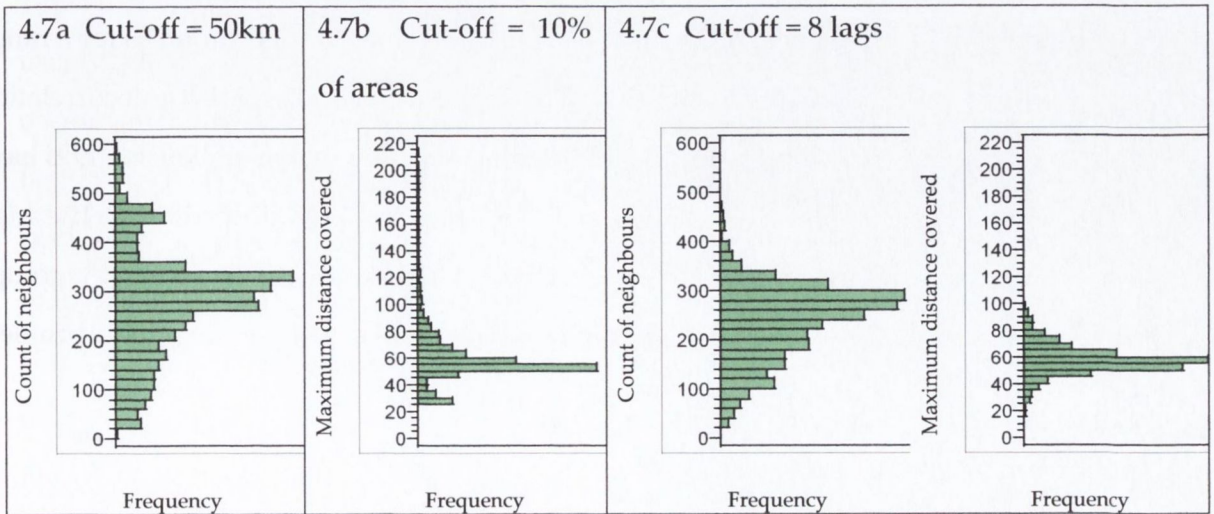
The use of a cut-off distance results in urban catchments having many more EDs than rural catchments. Including a percentage of areas is useful for ensuring the number of neighbours used in the PCA but leads to quite distorted catchment shapes. The use of lags to define the cut-off has the benefits of both the distance

and percentage areas methods. The use of percentage population has been discounted as the unit of observation is the ED, not the population.

In finding the optimum distance for each individual ED, the final results may not be comparable. Two neighbouring EDs may need significantly different radii to achieve acceptable models depending on how homogeneous the data are and the severity of the distance decay. It also makes it difficult to interpret regional variations in the relative weights associated with the variables used.

The first three methods were applied to a four variable, ED-level GW-PCA for Ireland. There are 3422 EDs in the country. The histograms in Figure 4.11 compare the frequency of results using three different methods of cut-off definition. The histogram in 4.7a shows the frequency of neighbour counts. The counts range from 14 to 588 within 50km of the focus ED, with a median of 285. The histogram in 4.7b shows the frequency of distances covered when the cut-off is based on 10%, or 342, of the areas. The distances range from 26.19km to 210.20km, with a median of 55.12km. Finally, the two histograms in 4.7c show the frequency of catchment size and distance respectively. The sizes range from 27 to 467, median 250, while the distances range from 17.90km to 98.95km, median 56.71km.

Figure 4.11 Comparison of cut-offs





Each cut-off is approximately equivalent, as can be seen from the median values. However, they result in widely varying ranges as can be seen in Table 4.6, where the ranges are shown for the three methods. The method using spatial lags results in smaller ranges and is thus a good hybrid of the two other methods.

Table 4.6 Comparison of results for different cut-off definitions

Cut-off method	Distance (km)			Percentage areas (%)		
	10%	Median	90%	10%	Median	90%
Distance	-	50	-	3.4	8.5	12.8
Number of areas	37.9	55.1	86.3	-	10.0	-
Lags	46.4	56.7	72.0	3.6	7.5	9.4

Comparison with the optimal distance method is not as straightforward as it is also dependent on the geography, the data and the distance decay function used while the other three methods are dependent on geography alone. Furthermore, use of the optimal distance method can lead to neighbouring EDs having very different sized catchments and potentially markedly different eigenvectors. This would certainly complicate interpretation of the variation in eigenvector values.

Groups of urban EDs typically produce markedly different correlations between variables than groups of rural EDs. This may be due to the greater homogeneity that appears to occur within urban areas. The figures in Table 4.7 show the mean Local Index of Spatial Autocorrelation (LISA)<sup>334</sup> values for five variables for a range of area types from urban to rural. The highest levels of spatial autocorrelation occur in city EDs. A high positive LISA value indicates that neighbouring EDs have very similar values while a high negative value indicates dissimilarity. The high average values in city EDs indicate a strong neighbourhood effect whereby groups of EDs tend to be quite homogeneous and have similar socio-demographic profiles.

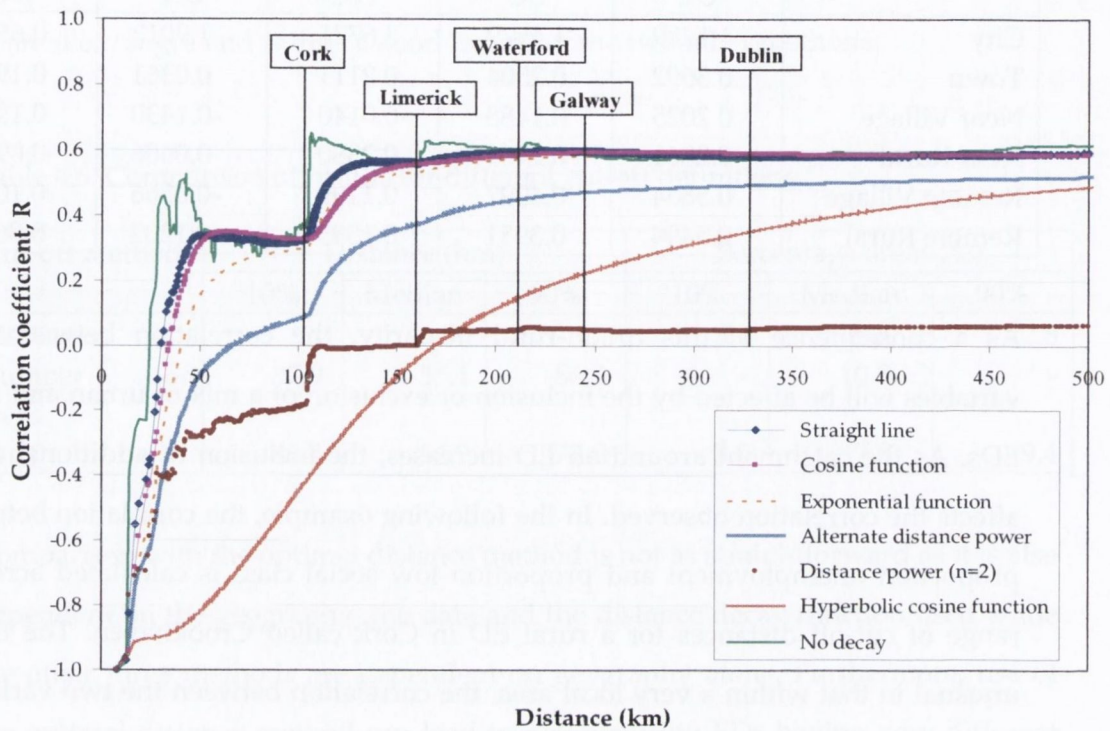
Table 4.7 Mean spatial autocorrelation by area type for a range of variables

Area	Mean LISA index value				
	UE	SC	NC	LH	OC
City	1.1289	1.4579	3.6771	1.9012	0.6574
Town	0.3692	0.2504	0.2111	0.0361	0.1909
Near Village	0.2025	0.1488	-0.0140	-0.1430	0.1025
Near Rural	0.2861	0.2409	0.2380	0.0606	0.0923
Remote Village	0.5834	0.4397	0.1199	-0.0566	0.1034
Remote Rural	0.5494	0.3951	0.1395	0.1202	0.0696

As a consequence of this urban-rural disparity, the correlation between two variables will be affected by the inclusion or exclusion of a mix of urban and rural EDs. As the catchment around an ED increases, the inclusion of additional areas affects the correlation observed. In the following example, the correlation between proportion unemployment and proportion low social class is calculated across a range of cut-off distances for a rural ED in Cork called Crookhaven. The ED is unusual in that within a very local area, the correlation between the two variables is negative. The change in correlation with increasing catchment size is shown in Figure 4.12 below. The graph also shows the distance at which each of the five cities impacts on the correlations.

It is clear that other than the hyperbolic cosine curve, all of the decay curves are strongly affected by the inclusion of Cork city and to a lesser extent Limerick city. The cosine and distance power curves are quite susceptible to changes and the inclusion of a single additional ED can have quite an impact on the correlation. The distance power function tends to zero at large distances but at short distances, the furthest ED within the catchment may still have significant influence on the correlation. For the other decay functions, the weights tend to zero at the cut-off point, leading to smoother transitions. The  $\cos()$  and straight line curves, also produce quite sudden changes in the correlation. The  $\cosh()$  curve produces a much smoother curve and does not show the sudden changes of the other curves. As the  $\cosh()$  curve tends towards zero at half the specified cut-off distance, Cork does not begin to affect the curve until the cut-off is over 200km.

Figure 4.12 Correlation between unemployment and low social class with increasing distance using six different decay methods for Crookhaven ED



Another important attribute of the decay curves is that the straight line,  $\cos()$  and exponential functions all tend towards the national correlation when all EDs are included. That means that if no cut-off is specified, then these decay curves will produce a local PCA that is very similar to the global PCA. The other decay curves retain some local information even when all EDs are included.

The above example suggests that the distance power curve should not be used in conjunction with a cut-off distance but only in cases where all EDs are included. The results of the GW-PCA would be overly sensitive to the choice of cut-off distance if a distance power curve was used. The straight line,  $\cos()$  and exponential function curves tend towards the national correlation at larger distances and should therefore be applied with a more stringent cut-off distance. Where variables are used that display strong spatial autocorrelation or urban-rural differences, it should be noted that the results of the GW-PCA may be quite

sensitive to the choice of cut-off. The pseudo power and cosh() curve methods retain more of the local information in the correlation. However, when the local correlation is opposite to what is seen nationally these decay curves may lead to a GW-PCA with a poor fit in terms of the first principal component. As such, it would be recommended to employ a large catchment size when using these decay functions.

The example given above raises questions about instances where correlations are observed that are the inverse of those found nationally. The correlation between unemployment and low social class is 0.617 nationally. This is not surprising as one would expect unemployment to be highest amongst the semi-skilled and unskilled labour force. A negative correlation suggests that unemployment decreases with an increased proportion of the population in a low social class. In the case of Crookhaven ED, this inverse correlation holds for the ED itself and the nearest 3 neighbours. This is coupled with the fact that the distances between EDs in the vicinity of Crookhaven are large so that even with a less severe distance decay model, the nearest EDs have a strong influence on the correlation coefficient. Indeed, if no decay function is used, once the nearest nine EDs are included the correlation coefficient is positive rather than negative. It should be noted that Crookhaven and its neighbouring EDs have small populations and that even with shrinkage, unusual characteristics may remain. This means that the negative correlation observed locally may simply be an artefact caused by a combination of small numbers, the choice of distance decay model and chance. One method of guarding against this is to use a decay and cut-off combination that ensures that all EDs have correlations that agree with those observed nationally. There is a danger that a genuinely anomalous local correlation may be removed although a strong negative correlation will generally turn into a weak positive correlation if a sufficient catchment is used. Of Crookhaven and its 20 nearest neighbours, only two have above average unemployment while 11 have above average proportion low social class. When looking at Crookhaven, rather than give a negative weight to unemployment or low social class, a weak positive weight would indicate that

the variable in question is not very influential locally. It would not make sense to say that greater unemployment leads to reduced deprivation in Crookhaven. It would be more appropriate to say that deprivation is not greatly influenced by unemployment in that local area.

#### **4.5.4 Implementation of GW-PCA**

As GW-PCA is an entirely new methodology, it is not available in any commercially available software. For the purposes of this thesis a software package has been developed in Visual Basic .Net<sup>335</sup> incorporating all of the elements discussed in the preceding section. As is required for a standard PCA, the indicator values for each small area are required as input. In addition, the neighbourhood structure and coordinates of each small area centroid are also required. The software allows the user to pick parameters relating to the preferred distance decay function and cut-off distance. The user can select the number of eigenvalues to output and whether to calculate Moran's I and local Moran's I (LISA) values. The output also indicates the number of components required for each small area based on the parallel analysis method and the local weights associated with each variable.

#### **4.5.5 Example application of GW-PCA**

To illustrate the utility of GW-PCA, an example of a deprivation index will be given. This is a four variable index including proportion unemployed (UE), proportion in a low social class (SC), proportion households with no car (NC) and the proportion households living in a LA house (LH). The overcrowding variable, OC, has been omitted given the large number of negative correlations it produces locally as was seen in Figure 4.2 previously.

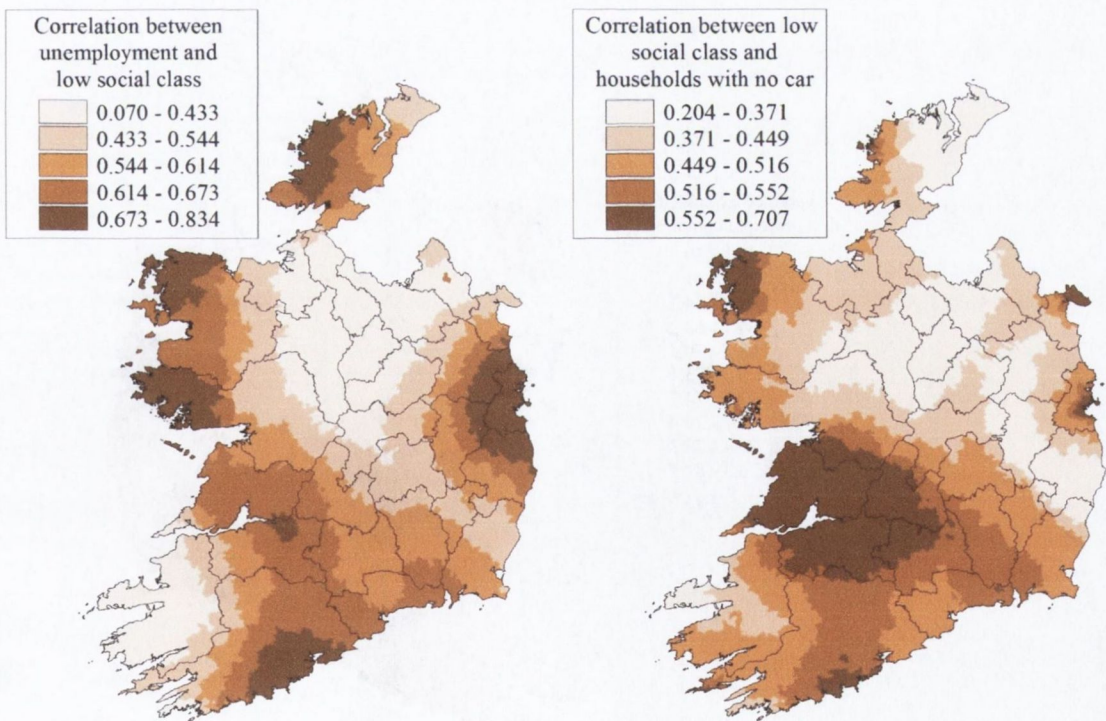
##### **4.5.5.1 Correlations between variable pairs**

There is regional variation in the correlations between pairs of variables. Figure 4.13 shows two maps to illustrate the extent of the variation. The correlation between unemployment and low social class ranges from 0.070 to 0.834 locally while it is 0.617 for a global PCA. Similarly, the correlation between low social class

and car ownership ranges from 0.204 to 0.371 locally compared to a correlation of 0.408 for a global model. The relationship that exists at a national level does not apply to all areas locally.

Correlations are typically greater in city or city fringe areas. This is partly due to the greater homogeneity that exists in cities. Due to smaller populations and greater distances between EDs in rural areas, correlations are sometimes poor in rural areas. The current example used a maximum lag distance of fifteen with the exponential distance decay function. The average number of neighbours was 737.0 and the average maximum distance encompassed was 106.8km. This meant that the districts used were quite large and many rural EDs would have had some urban EDs included when computing the localised PCA.

Figure 4.13 Regional variation in correlations for two pairs of variables



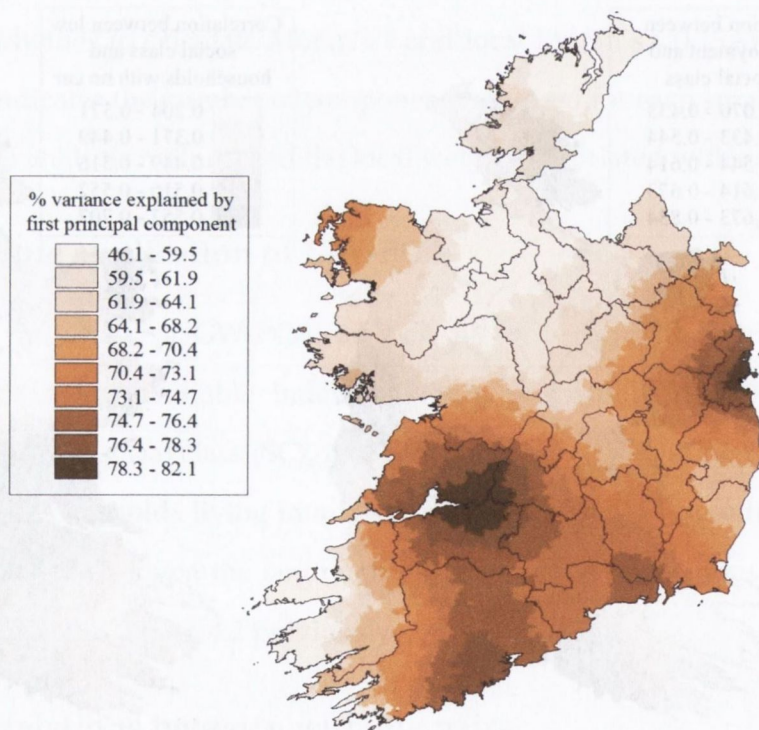
#### 4.5.5.2 Variance explained by the first principal component

As part of the GW-PCA, all eigenvalues are calculated for each ED. By employing a large maximum lag distance and the exponential distance decay function, only one

principal component had to be retained in all cases. All EDs could be adequately described by the first principal component, leading to a more easily interpretable index. However, the amount of variance explained by the first principal component still varied substantially by region. The map in Figure 4.14 shows this variation by ED. For the global model, 67.8% of variance was explained by the first principal component. Locally, this figure ranges from 46.1% to 82.1% indicating that for some areas, the first principal component is a poor fit.

All EDs also had positive correlations between all variable pairs. Again, this was achieved by using a relatively large maximum lag distance and a slack curve distance decay function. The benefit of this is that there are no counter-intuitive correlations.

Figure 4.14 Variance explained by the first principal component



To identify where these areas of poorer fit occur, the percentage of EDs with more variance explained by the first principal component than in the global model is given by urban-rural class in Table 4.8 below.

Table 4.8 Count of EDs for which the first principal component explains 67.8% or more variance (percentage) by urban-rural class

Urban-rural class	Frequency	Count of EDs for which first principal component explains 67.8% or more variance (%)
City	467	445 (95.3)
Town	234	160 (68.4)
Near village	159	91 (57.2)
Remote village	71	38 (53.5)
Near rural	1301	949 (72.9)
Remote rural	1190	410 (34.5)
Total	3422	2,093 (61.1)

It is evident that the first principal component tends to have a poorer fit for remote rural EDs than for any other class. This certainly suggests that the relationships between the deprivation variables are quite different in rural areas compared to city and town areas. It may also point to the resultant deprivation score being more descriptive of deprivation in urban areas than rural areas.

#### 4.5.5.3 Eigenvector values

The eigenvector values obtained by the PCA process are used as weights for the standardised variables. The weight for a given variable is influenced by the correlation between that variable and the other variables. As the sum of squared weights must equal 1, a low weight for one variable is compensated by higher weights for at least one of the other variables and vice versa.

The maps in Figures 4.11 to 4.14 show the weights for each of the four variables by ED. The weights indicate the relative importance of the variables in calculating the deprivation score. The interrelationships between the variables are complex and vary regionally, making interpretation difficult. Not all rural areas have the same correlations between variables or underlying means. It is also clear that Dublin, being the capital and significantly larger than the other cities, has a wide influence which extends to the edges and beyond of the Greater Dublin area of Dublin, Kildare, Meath and Wicklow.



Figure 4.15 Eigenvector values for proportion unemployed

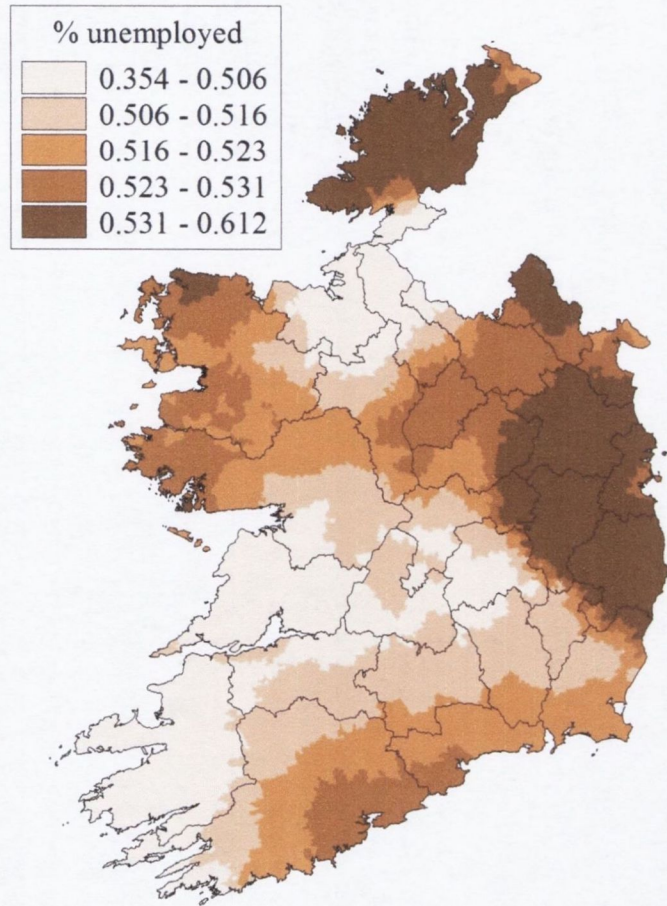


Figure 4.16 Eigenvector values for proportion low social class

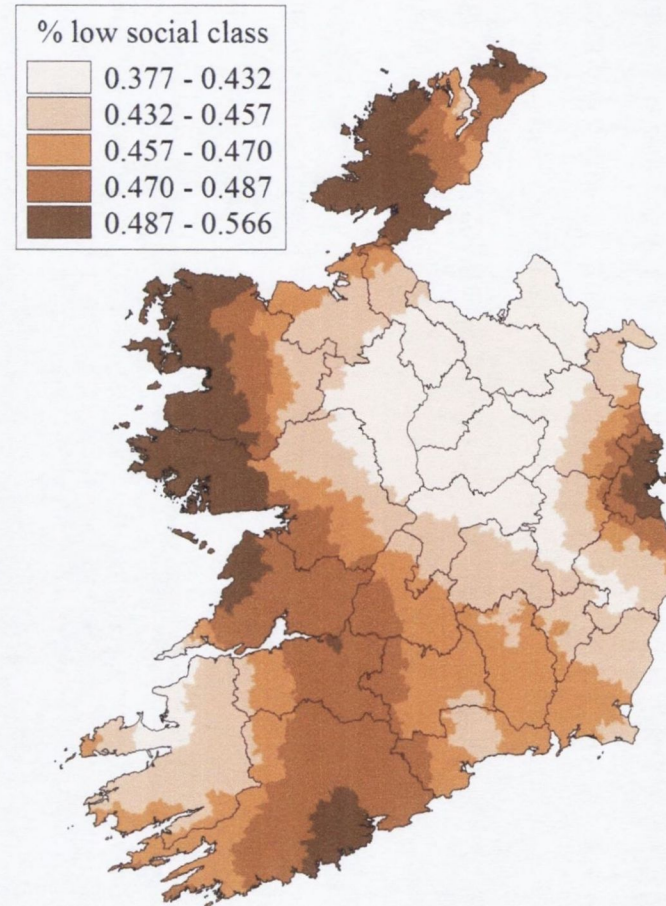


Figure 4.17 Eigenvector values for proportion households with no car

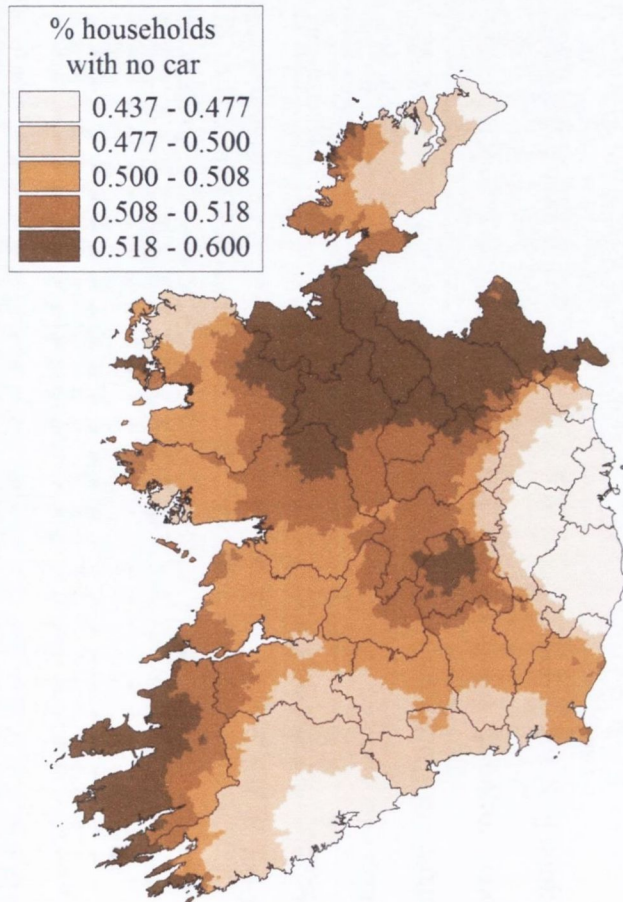
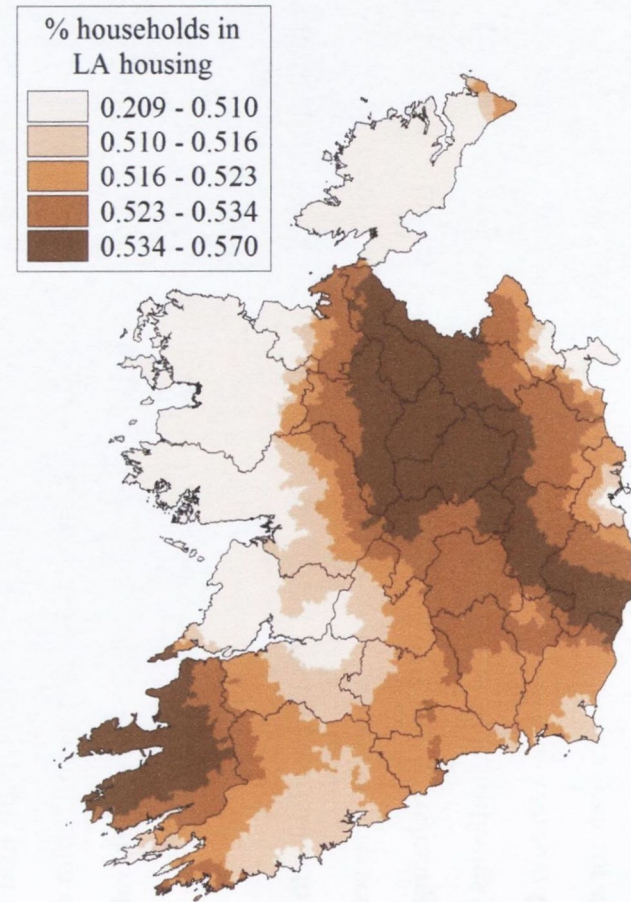


Figure 4.18 Eigenvector values for proportion households living in Local Authority housing



The amount of variability captured in the model is largely dependent on the choice of distance decay function. If a rapid decay with a small maximum distance is used, much greater variation is captured. As a less severe decay function and large maximum distance was used in this example, less variability in the weights is observed.

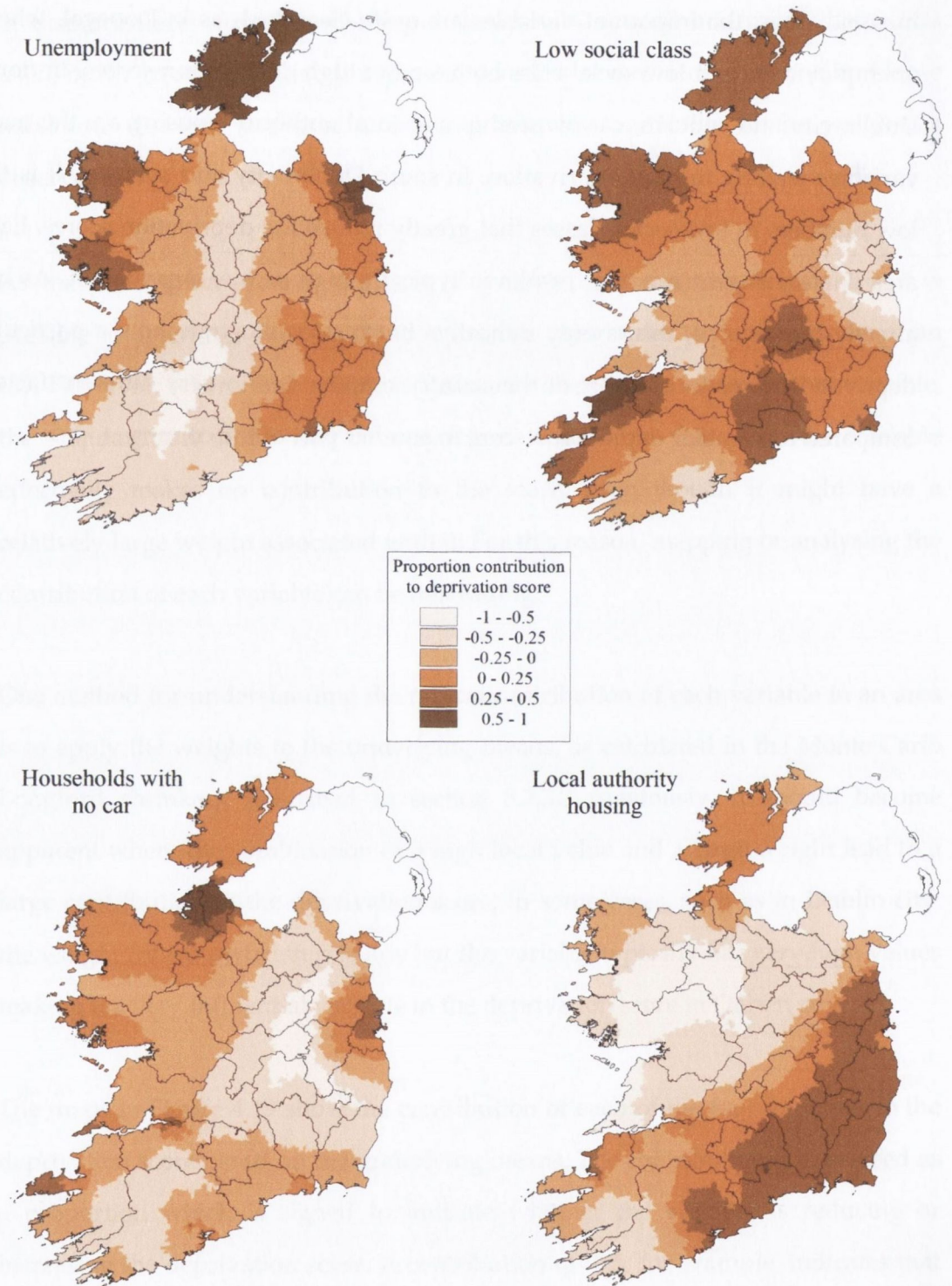
As the deprivation score is the sum of weights by standardised variables, it is possible to determine how much each variable contributes to the final score of an ED. However, where the variable is arbitrarily close to the mean for that variable, the standardised value will be very close to zero. In that case, the variable effectively makes no contribution to the score, even though it might have a relatively large weight associated with it. For this reason, mapping or analysing the contribution of each variable can be misleading.

One method for understanding the typical contribution of each variable in an area is to apply the weights to the underlying means, as calculated in the Monte Carlo Longford shrinkage described in section 3.2.3.3 previously. It should become apparent where the combination of a high local value and a large weight lead to a large contribution to the deprivation score. In some cases, such as in Dublin city, the weight for car ownership is low but the variable typically has very high values making it a very influential variable in the deprivation score in Dublin city.

The maps in Figure 4.19 show the contribution of each of the four variables to the deprivation score based on the underlying means. The contribution is expressed as a proportion which is signed to indicate whether the variable is reducing or increasing the deprivation score. A contribution of -0.5, for example, indicates that the variable in question contributes half of the score and is below average, thus reducing the score. As the calculation is based on the underlying mean for each variable, it is merely indicative of the contribution as the local variation will lead to differing contributions in each ED. It does, however, suggest which variables give rise to higher and lower deprivation scores in different parts of the country.

In some areas, the important variables are quite clear such as in Donegal, where unemployment and low social class both lead to high deprivation scores. In north Dublin city and suburbs, car ownership and local authority housing are the main variables leading to high deprivation. In south Dublin city and suburbs, it is the low numbers in low social classes that greatly reduce the deprivation scores. Each area has a different mix that results in typically high or low deprivation. As has already been stated, this is only indicative but it does illustrate how a particular deprivation score in one part of the country may be due to very different factors compared to a similar deprivation score in another part of the country.

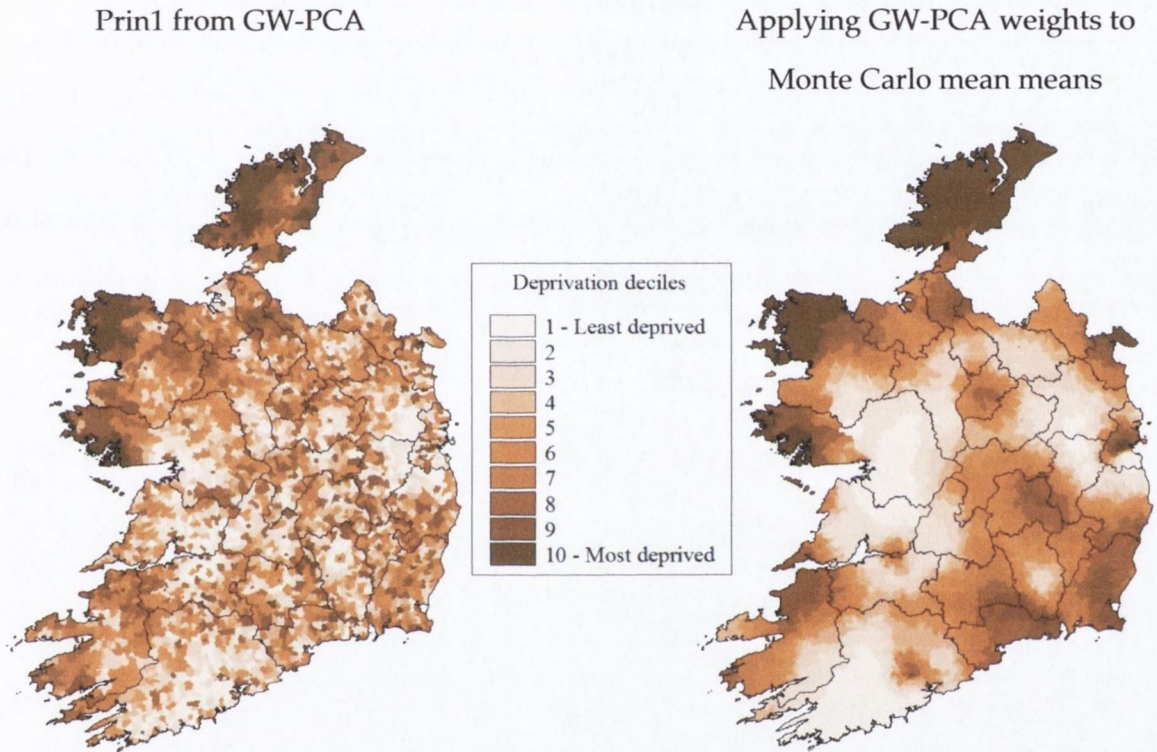
Figure 4.19 Contribution of each variable to deprivation score (based on means)



A final step is to map the deprivation scores. This is shown in Figure 4.20 in two maps, both showing the scores by decile. The left-hand map shows the scores from the GW-PCA analysis while the right-hand map shows the score calculated using the underlying means applied to the weights from the GW-PCA. The latter map

affectively smoothes out locally unusual EDs with small populations, giving an indication of the general deprivation of an area. This sort of map may be misleading as it removes genuine local variation although it is useful for a rapid overview.

Figure 4.20 First principal component mapped for local and mean



It is apparent that there is substantial variation in the relative importance of different variables when they are combined into a deprivation index. This variation can be observed and analysed using GW-PCA. It can also be seen that the fit of the first principal component varies regionally, indicating that in some areas, a particular choice of variables may be less appropriate than in others.

#### 4.6 Outlier detection and influence functions

An outlier can be defined as an observation that is different or inconsistent with the remainder of the data.<sup>297</sup> Outliers can, but not as a rule, have undue influence on a multivariate method such as PCA or FA. So an outlier can be an influential

observation but an influential observation is not necessarily an outlier. Being able to identify influential observations and limit the effect of outliers are useful techniques in both PCA and FA.

#### 4.6.1.1 Outlier detection and robust calculations

PCA itself can be used for outlier detection due to its dimension reduction qualities. It is possible to identify outliers based on their inconsistency with the rest of the data, particularly on 2<sup>nd</sup>, 3<sup>rd</sup> and subsequent principal components.<sup>297</sup> A number of robust PCA methods exist that use a robust estimate of the covariance or correlation matrix to diminish or remove the effect of outliers.<sup>336-341</sup> These methods frequently incorporate Mahalanobis distance or some similar measure of statistical distance from the multivariate mean. The Mahalanobis distance for observation  $i$  is defined in Equation 5.1 below.

$$d_i = \sqrt{(\bar{x}_i - \bar{\mu})' S^{-1} (\bar{x}_i - \bar{\mu})} \quad (5.1)$$

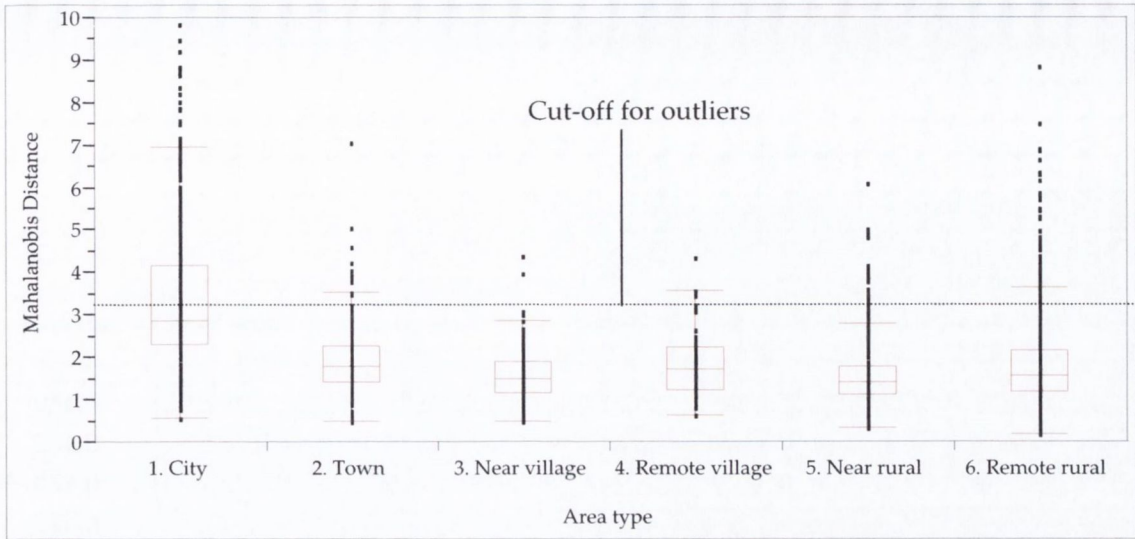
Where:  $\bar{x}_i$  =  $p$ -dimensional observation  $i$ ;  $i = 1, 2, \dots, n$

$\bar{\mu}$  =  $p$ -dimensional mean

$S$  = covariance matrix

The cut-off Mahalanobis distance is calculated with reference to a Chi-square distribution, specifically the critical value is given by  $\sqrt{\chi_{p,0.975}^2}$  where  $p$  is the number of variables.<sup>338</sup> Using the five variables outlined in section 4.4.2, the Mahalanobis distances were calculated and are shown in Figure 4.21 by area type. For the example using five variables, the cut-off is 3.58 above which an observation is considered an outlier. This method classifies 249, or 7.3%, of the 3,422 observations as potential outliers.

Figure 4.21 Mahalanobis distance by area type



From Figure 4.21 it would appear that many of the outliers are city EDs. Summarising the numbers of outliers by area type in Table 4.9 shows that a disproportionate number of outliers are city EDs. This is probably due to the fact that, as was seen in Figure 4.4, city EDs can experience extreme values in some variables such as car ownership and local authority housing.

Table 4.9 EDs classed as outliers by area type

Area type	EDs	Outliers	
		Count	Percentage
1. City	467	156	33.4
2. Town	234	10	4.3
3. Near village	159	2	1.3
4. Remote village	71	1	1.4
5. Near rural	1,301	13	1.0
6. Remote rural	1,190	67	5.6

A number of the robust correlation matrix estimation methods have been implemented in the Robust package as part of S-Plus.<sup>281</sup> Table 4.10 shows the eigenvectors of the first principal component using each of four robust estimators: MCD, M, Pairwise QC and Pairwise GK. As these methods tend to operate by identifying and either dropping or down-weighting outliers, it can be expected that in this instance a robust method will diminish the impact of city EDs. This can be



seen in Table 4.10 in the fact that the eigenvectors bear more similarity to that derived from rural only EDs shown in Table 4.3 previously.

Table 4.10 Eigenvector values for the first principal component using different robust estimates of the correlation matrix

Indicator	Robust estimation method				
	None	MCD	M	Pairwise QC	Pairwise GK
UE	0.489	0.469	0.491	0.491	0.489
SC	0.440	0.506	0.483	0.481	0.483
NC	0.443	0.471	0.456	0.476	0.473
LA	0.496	0.486	0.511	0.511	0.513
OC	0.353	0.258	0.236	0.202	0.201

The intention of robust analysis is to remove or diminish the effect of outliers. How well this is achieved depends on how well the outliers are identified. A third of city EDs are identified as outliers and, as a consequence, the city EDs have less impact on the correlation matrix. This, in turn, results in a correlation matrix that has stronger resemblance to the correlation matrix produced by only considering rural EDs. This consequence is predictable given the fact that the multivariate distance is computed in relation to a vector of means. As the majority of EDs are rural, the means vector will be dominated by rural observations and, if there is any significant urban-rural difference, the urban EDs will be more likely to be considered as outliers.

When robust estimation is applied to the covariance or correlation matrix, the resultant eigenvalues and eigenvectors are considered robust. However, as the principal components are computed using the standardised variables with the eigenvectors, the outlying observations will most likely have outlying principal component scores. In other applications, such as image analysis, it is possible to effectively ignore the identified outliers or to estimate new values for the outlier based on neighbouring observations.<sup>310 342</sup> Such an approach would not be acceptable in the development of a deprivation index as each area should be

represented and, if census data are used, it is arguable that while values may be extreme they were correctly measured and thus legitimate.

#### 4.6.1.2 Influential observations

It has already been stated that an outlier need not be influential and an influential observation need not be an outlier. Distance metrics are useful for identifying outliers but are not so useful for influence measurement. Influence functions were described by Critchley,<sup>343</sup> and subsequently developed further for application in robust PCA by Croux and Haesbroeck.<sup>336 344</sup> Brooks used influence functions to assess the influence of individual observations on the eigenvalues and eigenvectors.<sup>345</sup> Equation 5.2 gives the theoretical influence function for observation  $i$  and eigenvalue  $k$ .

$$I(\bar{x}_i; \lambda_k) = \alpha_k' (\bar{x}_i - \bar{\mu})(\bar{x}_i - \bar{\mu})' \alpha_k - \alpha_k' \Gamma \alpha_k \quad (5.2)$$

Where:  $\bar{x}_i = p$ -dimensional observation  $i$ ;  $i = 1, 2, \dots, n$

$\bar{\mu} = p$ -dimensional mean

$\lambda_k =$  eigenvalues for  $k^{\text{th}}$  component

$\alpha_k = k^{\text{th}}$  principal component;  $k = 1, 2, \dots, p$

$\Gamma =$  correlation matrix

An alternative to the theoretical influence function is to calculate an empirical influence function using jackknife methods.<sup>345</sup> The jackknife approach essentially involves leaving out each observation in turn and re-computing the PCA each time. Although computationally intensive, it is possible to look at the impact of each observation on the eigenvalues and eigenvectors of each principal component by calculating the difference between the eigenvalue with and without the  $i^{\text{th}}$  observation. It is also possible to measure the change in angle of the eigenvector of

a principal component with and without the  $i^{\text{th}}$  observation. Equation 5.3 defines the empirical influence function.

$$I(\bar{x}_i; \theta) = (n - 1)(\hat{\theta}_{(i)} - \theta) \quad (5.3)$$

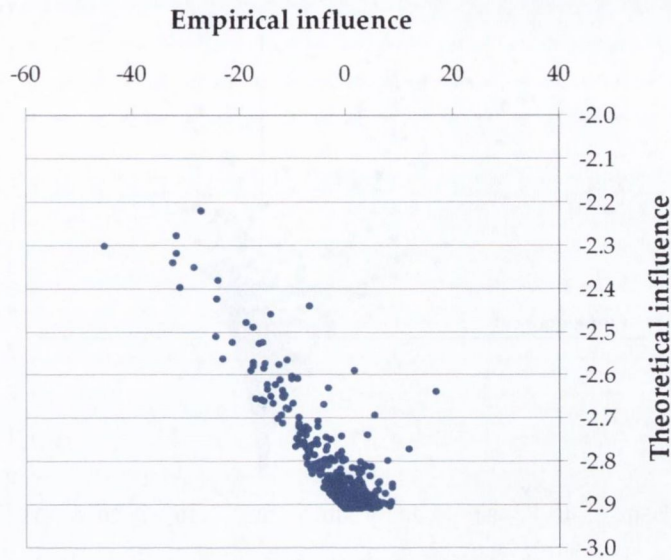
Where:  $\bar{x}_i = p$ -dimensional observation  $i; i = 1, 2, \dots, n$

$\hat{\theta}_{(i)} =$  eigenvalue computed without observation  $i$

$\theta =$  eigenvalue computed with all  $n$  observations

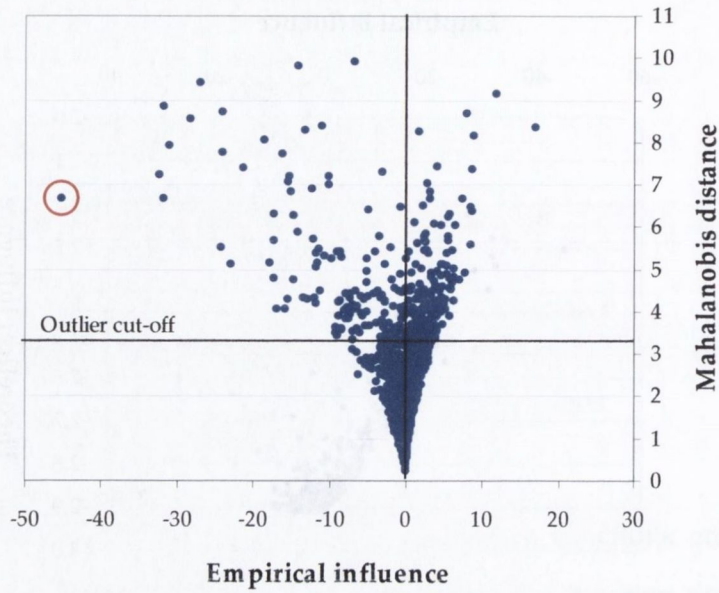
Brooks compared theoretical and empirical influence functions and found little difference in the results although clearly the empirical function entails a greater computational burden as the PCA must be repeated  $n$  times.<sup>345</sup> Both the empirical and theoretical influence functions were calculated for the five variable deprivation index being calculated. Figure 4.22 shows a plot of the empirical values against the theoretical values. Empirical influence values are lowest at 0 while theoretical influence values are lowest at -2.91 in this example. For the subsequent discussion the empirical influence values will be used as they are more simply defined and interpretable. A negative value in the empirical influence function indicates an increased eigenvalue with the inclusion of that observation.

Figure 4.22 Empirical and theoretical influence function values



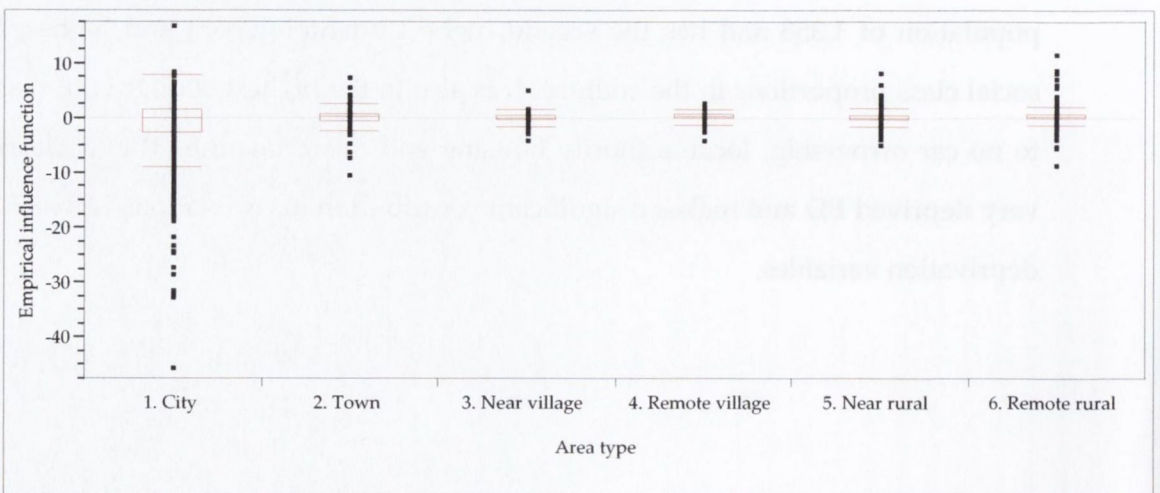
As outliers have been defined in terms of Mahalanobis distance, it is interesting to note to what extent outliers are also influential observations. The plot in Figure 4.23 shows Mahalanobis distance by empirical influence value. As can be seen from the outlier cut-off line, the most influential observations are also all outliers. However, some of the outliers are not particularly influential. An empirical influence value of zero indicates an observation with no influence on the eigenvalue. This merely confirms the statement that outliers need not be influential. With the given dataset most of the observations that are influential are also outliers. The circled observation in Figure 4.23 is John's A ED in Limerick City. This ED has a population of 1,358 and has the second highest unemployment and highest low social class proportions in the country. It is also in the highest 30 EDs with respect to no car ownership, local authority housing and overcrowding. This is clearly a very deprived ED and makes a significant contribution to correlations between the deprivation variables.

Figure 4.23 Mahalanobis distance by empirical influence



The empirical influence function values are given by area type in Figure 4.24 below. It can be seen from the many negative values for city EDs that they tend to increase the eigenvalue of the first principal component. This underlines the fact that although city EDs account for only 13.6% of EDs, they exert a lot of influence on the nature of the first principal component and the amount of variance explained by it.

Figure 4.24 Empirical influence by area type



Some area types may exert greater influence on the principal components than others. The consequence of this is that, as with outliers, they may introduce a regional or urban-rural bias into the resultant deprivation index. However, unlike specific outliers, the identification of influential observations may assist in the interpretation of a deprivation index but it is not a necessity to adjust for them.

## 4.7 Summary

A number of methods of dimension reduction have been used to combine deprivation indicators into a single, or sometimes multiple, indices. The most common techniques have been principal components analysis (PCA) and factor analysis (FA). A subjective comparison suggested that PCA was more appropriate for combining variables for a deprivation index.

A discussion of sources of error in deprivation indices included the problem of regional variation in deprivation indicators. The variation extends to the correlations between variables indicating that in different regions, different combinations of variables may be more consistent proxies for deprivation. To account for this regional variation, a new method of PCA – geographically weighted PCA – was developed and applied to Irish data. In this new method, the correlation matrix is recomputed for each small area with increased weight given to observations that are geographically close to the focus ED. A range of distance decay functions for computed weights were tested. The methodology was applied to an illustrative example using Irish data. As part of GW-PCA, the local weights for each indicator can be mapped to show the relative importance of each indicator regionally.

This new methodology represents an important advance as regional variation in the deprivation indicators is explicitly accounted for in the process of dimension reduction.

Some well-known examples of the use of the method are given in the appendix. The method is particularly useful for the analysis of data from experiments where the response is a continuous variable and the treatment is a categorical variable. The method is also useful for the analysis of data from experiments where the response is a continuous variable and the treatment is a continuous variable. The method is also useful for the analysis of data from experiments where the response is a continuous variable and the treatment is a continuous variable.

A number of authors have proposed various methods for the analysis of data from experiments where the response is a continuous variable and the treatment is a categorical variable. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1].

The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1].

The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1]. The method proposed in this paper is based on the method proposed by [1].

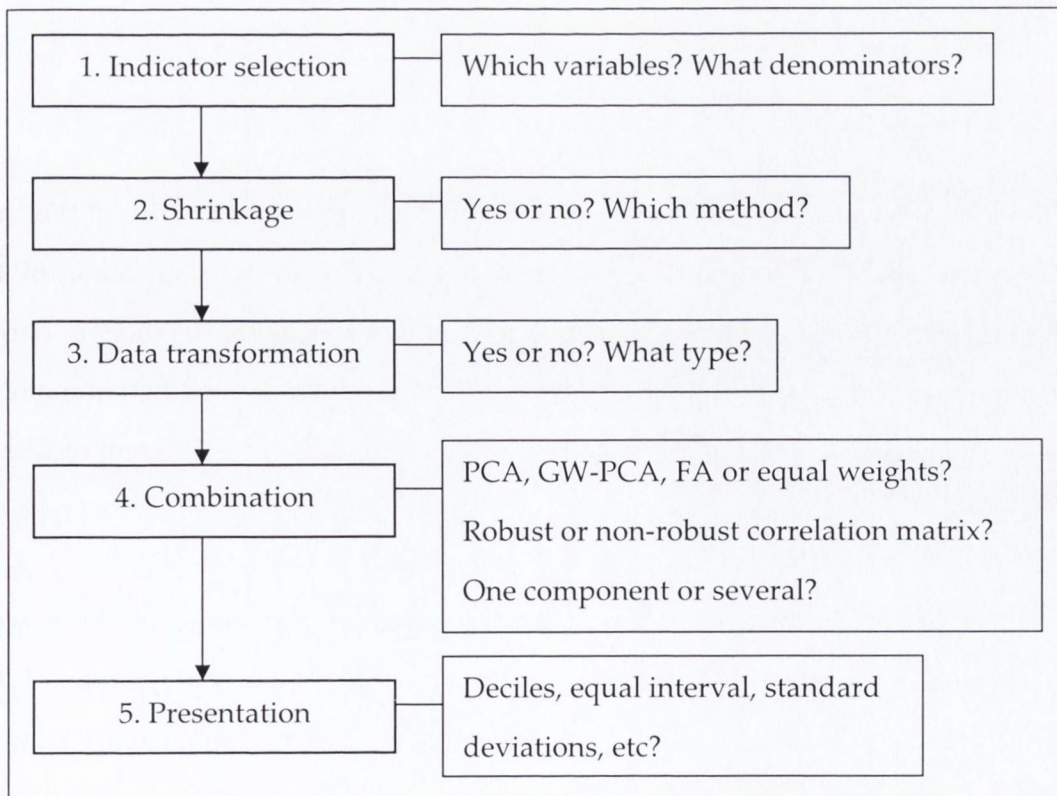
## 5 Sensitivity analysis

The development of a deprivation index involves a number of key steps where the developer must make decisions regarding the content and computation of the index. The choices made can have substantial impacts on the appearance and defensibility of the final index. A logical analysis would be the sensitivity of the calculated index to different choices of method and data. The first section of this chapter defines the key stages in the development of a deprivation index. In the final sections of the chapter there is a simple sensitivity analysis followed by a more in-depth sensitivity analysis.

### 5.1 Key steps in deprivation index development

The flow chart in Figure 5.1 outlines the key steps and some of the options that might arise in those steps.

Figure 5.1 Flow chart of deprivation index calculation





When an index is comprised of domains of deprivation, the steps to calculating the indices are the same except that there may be a second process of transformation and combination prior to presentation if a single general index is being calculated. Each of the key steps will be dealt with individually to assess what potential impacts might arise depending on the choices made.

### **5.1.1 Indicator selection**

As was shown in section 3.1.1, a large number of different indicators have been used in different deprivation indices. The choices are often governed by data availability and context. Even within the same jurisdiction, very different selections of indicators are used as can be seen from three Irish indices.<sup>181 183 185</sup> The developers of all of these indices would cite theoretical reasons and probably research evidence for the choices they have made. Indicators are chosen because they supposedly reflect some aspect of deprivation and yet a consensus cannot be reached as to which indicators achieve this successfully or at least do so without introducing unacceptable urban or rural bias. Even though the researchers are attempting to produce a measure of the same underlying problem, they do so with different variables.

### **5.1.2 Shrinkage**

Shrinkage is used for many variables in the various English,<sup>264</sup> Scottish,<sup>299</sup> Welsh,<sup>300</sup> Northern Irish<sup>301</sup> and Irish<sup>183</sup> deprivation indices. However, given some of the criticisms levelled at the use of shrinkage, application of the method is not an automatic choice.<sup>265</sup> Additionally, as was shown in chapter 3, the different methods of shrinkage, and choice of mean to shrink to, can produce quite different results.

### **5.1.3 Transformation**

A number of options exist for data transformation and some of these were illustrated in section 3.3 in terms of how they affect the distribution of the transformed variable. Depending on the method of data combination, a multi-normal distribution might be assumed in which case transformation will probably be required for some variables. If normality is not a prerequisite then

transformation is optional and its application may alter the results of the deprivation index.

#### **5.1.4 Data combination**

At this step there are a number of choices to be made, all of which can have a major impact of the deprivation index being generated. First of all, the methodology to be used for combination: PCA, FA or the simple use of arbitrary equal weights. The use of equal weights has largely fallen out of favour leaving the various forms of PCA and FA to choose from. One might choose PCA or, as proposed in section 4.4, GW-PCA. One might choose FA with any of a number of options regarding rotation. There are then possibilities regarding cut-offs for factor loadings below which a variable will effectively have its loading reduced to zero. Analysis can be performed using the covariance or the correlation matrix, both of which will produce different results. Robust estimation of the covariance or correlation matrix can be used to make the analysis more resistant to the influence of outliers. Finally, there are numerous methods for deciding on component retention, some of which have been noted to lead to over-retention. In the case of FA, it is the subjective choice of the researcher that dictates how many factors will be retained and as was discussed in 4.2.6, depending on the number of factors retained and rotation applied, quite different factor loadings can be obtained.

#### **5.1.5 Presentation**

This step takes place after index calculation and so the raw scores are unaffected by decisions made at this point. However, the choice of deciles or other classifications can impact on how areas are grouped and whether an ED will be regarded as deprived or not. Few users of an index will compute a context specific index based on the scores. Generally the index provided is the index used and so the index should not over- or under-represent deprived areas. Large groupings will generally lead to moderately deprived areas being regarded along with the highly deprived areas. Conversely, small groupings might result in quite deprived EDs being overlooked because they are not in the most deprived group.

For ease of understanding, deprivation scores are typically ordered and expressed in deciles. The advantage of this is the transparency and ease of understanding of a simple ten point scale (e.g. an ED is in the most deprived 10% of areas). There is no ambiguity about the choice of cut-offs which might occur if a method such as k-means clustering or similar classification method was used. An alternative method was originally used by Kelly and Sinclair whereby the histogram of index values would mimic the distribution of the histogram of scores.<sup>182</sup> Haase and Pratschke use equal intervals for some of their indices but provide little information on how some of their labels were derived.<sup>185</sup>

The importance of the choice of cut-offs relates primarily to perception and to the use of indices in resource allocation. A person who sees a map of deprivation automatically assumes that areas in the same class are equivalent – that they have the same level of deprivation. When there are many areas in a class, it becomes possible that the areas at the top and bottom of that class are very different in terms of their indicator values yet they have the same class. Without resorting to large numbers of index values, this is unavoidable. For resource allocation purposes, funding may be allocated to areas according to their deprivation index value rather than score. In such an instance two consecutively ranked EDs with different index values (i.e. on either side of a cut-off) could receive very different funding when, with a different cut-off definition they would receive the same funding. This perhaps points to the need for a caveat regarding the interpretation of index values so as not to place too much weight on the significance of the index value but rather to use the scores when applicable.

## **5.2 Assessing the choices**

Given the many ways in which a large set of indicators can be reduced, shrunk, transformed, combined and presented, there is substantial scope for the generation of indices that show very different spatial distributions of what is meant to be the same measure. Therefore it stands to reason that researchers would at least

perform some form of sensitivity analysis to determine how much impact their decisions have had on the ranking of EDs. Some of the decisions are based on subjective reasoning and theory relating to deprivation measurement, such as the choice of indicators or of FA over PCA. Conducting a sensitivity analysis might be seen as an admission that the proposed theory was either deficient or unreliable. On the other hand, a sensitivity analysis could be used to support a given theory if it showed other methods to produce more variable results.

Some examples of sensitivity analysis in deprivation index development are described. For example, Field carried out a sensitivity analysis on an index of relative disadvantage.<sup>346</sup> He tested the impact of omitting variables on the final index values. He found that simplification of the index was possible while maintaining a good correlation with the index computed using a full set of variables. This, however, may point towards redundancy in his original choice of variables rather than a robust index. A revision of the Scottish deprivation index included a sensitivity analysis of the insurance variable which was deemed controversial.<sup>298</sup> Analysis showed that the exclusion of the variable had little impact on the index which was seen as a vindication of its inclusion. A sensitivity analysis was conducted on the Scottish index of multiple deprivation (SIMD) to assess the impact of changing the weights applied to domain scores in the process of combination into a single index.<sup>293</sup> It was found that even large changes to the weights had little effect on the SIMD with correlations between the actual index and the simulated index being 0.987 and higher. With the large number of small areas included, the correlation coefficient might not have been the most appropriate statistic for comparison. Klasen performed a simple sensitivity analysis on some of the variables used in his South African index of deprivation although it is not a rigorous analysis.<sup>295</sup>

Where deprivation is used to determine resource allocation, the impact of a different ranking can seriously affect funding for that ED.<sup>347</sup>

## 5.3 Example of a sensitivity analysis

This section will use an example of sensitivity analysis to illustrate the issues that may arise. Two types of analysis are outlined: a simple approach of testing the application of shrinkage and transformation on the variables and a more complex analysis looking at random selection of methods at each stage of the process.

### 5.3.1 Simple sensitivity analysis

For this analysis, only the sensitivity of the deprivation index to the application of shrinkage and transformation are tested. For simplicity, only the following four scenarios are tested:

1. No shrinkage or transformation of variables
2. Only shrinkage applied to all variables
3. All variables transformed but not shrunk
4. Shrinkage and transformation applied to all variables

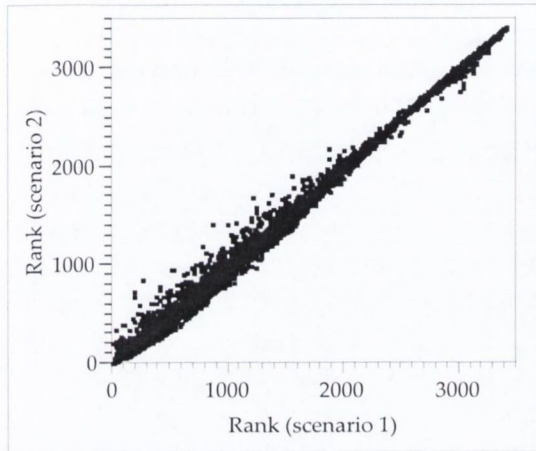
The shrinkage method used was that of Longford and the transformation method was empirical logit which can be applied when the numerator is zero. The scenarios were applied prior to combination by PCA. Four variables were used for this analysis: proportion unemployment (UE), proportion low social class (SC), proportion households with no car (NC), and proportion households in Local Authority housing (LH). The PCA and ranks were computed for each scenario and plotted against the ranks for scenario 1 in Figure 5.2 below. It can be seen that the ranks for some observations change substantially, particularly if both shrinkage and transformation are applied.

The application of transformation appears to have a greater impact than the application of shrinkage estimation. After shrinkage, 13.5% of EDs had a different decile than without shrinkage. After transformation, this figure increased to 28.6% with some EDs changing up to three deciles. It is worth noting, however, that most of the substantial shifts occurred in EDs with mid-range values (e.g. with and without shrinkage an ED with deprivation index values of 4 and 6 respectively). It would be of greater concern if EDs at the extremes were primarily affected. Even

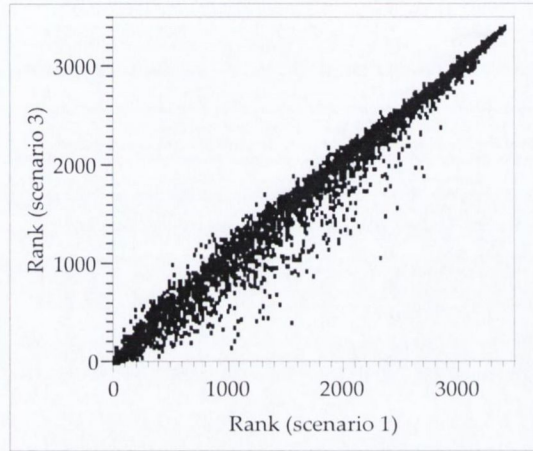
so, given the impact of shrinkage and transformation, it is important that the use of either is fully justified.

Figure 5.2 Plots of scenario 1 ranks against ranks for scenarios 2 to 4

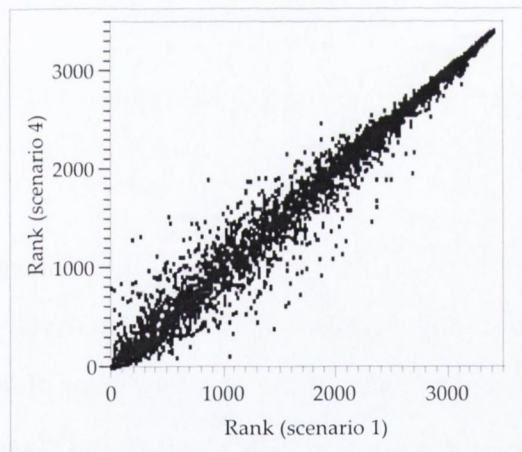
(a) Scenario 1 v. scenario 2



(b) Scenario 1 v. scenario 3



(c) Scenario 1 v. scenario 4



Scenarios:

1. No shrinkage or transformation of variables
2. Only shrinkage applied to all variables
3. All variables transformed but not shrunk
4. Shrinkage and transformation applied to all variables

The percentage EDs in each area type in the least and most deprived deciles are given in Table 5.1 below. Shrinkage alone reduces the number of rural EDs in the extreme deciles. This can be explained by the fact that these EDs tend to have small populations and thus shrinkage has a greater impact, drawing extreme values closer to the mean and consequently closer to mid-range deprivation scores. Shrinkage increases the number of urban EDs with extreme deprivation values. This is a natural consequence of rural EDs being drawn to the centre of the distribution. Transformation has a greater impact on urban areas, reducing the

number of deprived EDs and increasing the number of affluent EDs. This is partly due to urban EDs with large values for NC and LA, which are not reduced by shrinkage, being brought closer to the mean by the empirical logit transformation.

Table 5.1 Percentage EDs in the least and most deprived deciles by scenario and area type

Area type	Least deprived decile				Most deprived decile			
	Scenario*				Scenario*			
	1	2	3	4	1	2	3	4
1. City	16.7	19.9	18.2	21.0	38.1	38.5	37.0	36.8
2. Town	5.6	9.4	5.1	9.0	21.8	23.5	23.1	23.9
3. Near village	2.5	3.8	1.3	3.1	7.5	8.2	8.8	9.4
4. Remote village	1.4	1.4	1.4	1.4	15.5	16.9	14.1	15.5
5. Near rural	13.3	13.1	13.5	13.0	1.8	2.1	2.0	2.3
6. Remote rural	6.1	4.2	5.6	4.0	5.6	4.7	5.5	5.0

- \* Scenarios:
1. No shrinkage or transformation of variables
  2. Only shrinkage applied to all variables
  3. All variables transformed but not shrunk
  4. Shrinkage and transformation applied to all variables

### 5.3.2 Detailed sensitivity analysis

Having seen from the previous section that both shrinkage and transformation can impact on deprivation index values, a more detailed sensitivity analysis is attempted. For this analysis, the decisions made at each of the first four steps of the flow chart in Figure 5.1 will be dictated by random choices. The number and choice of variables can change. The decision of which variables to shrink is made randomly, as is the decision to transform variables. The choice of combining variables using equal weights or PCA is also made randomly.

The following ten deprivation indicators were chosen:

- Proportion unemployment (UE)
- Proportion low social class (SC)
- Proportion households with no car (NC)
- Proportion households in Local Authority housing (LH)

- Proportion population who left school before the Junior Cert./Inter Cert. (ES)
- Proportion lone parent families (LP)
- Proportion households with no central heating (CH)
- Proportion population with a disability (DP)
- Proportion population not in the labour force (LF)
- Proportion population unable to work (UW)

The variables were chosen on the basis that they should all be proxies for different aspects of deprivation with minimal redundancy. The correlation matrix for the ten variables is given in Table 5.2. All variables are positively correlated with each other.

The steps in the sensitivity analysis were as follows:

1. Randomly select  $m$  of the 10 variables available (with  $5 \leq m \leq 10$ )
2. Randomly select variables from  $m$  chosen and apply shrinkage
3. Randomly select variables from  $m$  chosen and transform
4. Combine data using PCA
5. Calculate scores and ranks and store
6. Complete 5,000 iterations of steps 1 to 5
7. Calculate median and inter-quartile range of ranks for each ED

As this example is for illustrative purposes, the sensitivity analysis was specified in a simplistic manner. Robust estimation was not included and only Longford shrinkage was used as it was shown previously to be the most appropriate method. The only method of transformation used was empirical logit as when the numerator is zero the log and standard logit transformations cannot be computed. GW-PCA was not included as the additional number of parameters that can be varied is substantial and may make interpretation of the results more difficult. Also, the use of GW-PCA does not generally have a large impact on results unless a severe distance decay parameter is specified which leads to large numbers of EDs requiring the retention of a second component.



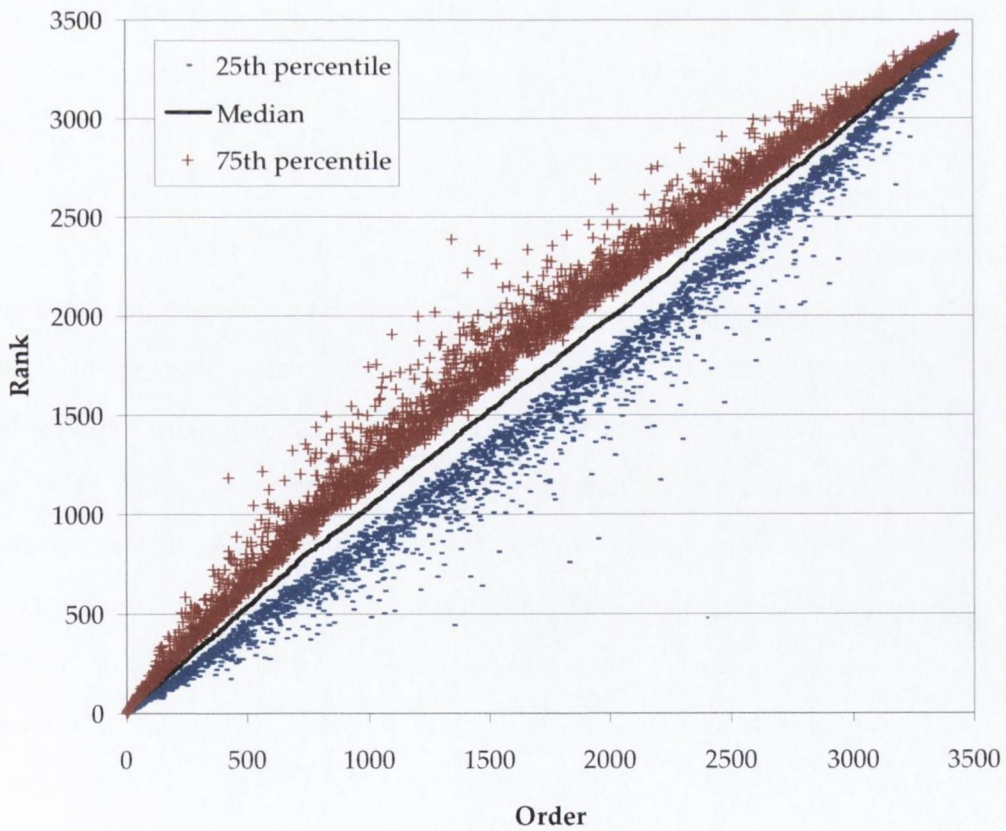
Table 5.2 Correlation matrix for the ten deprivation variables

Variable	UE	SC	NC	LH	LP	CH	DP	LF	UW	ES
UE	1.00	0.57	0.59	0.59	0.54	0.30	0.29	0.12	0.35	0.35
SC	0.57	1.00	0.40	0.49	0.42	0.43	0.32	0.36	0.40	0.64
NC	0.59	0.40	1.00	0.64	0.67	0.43	0.44	0.15	0.34	0.16
LH	0.59	0.49	0.64	1.00	0.71	0.22	0.29	0.09	0.35	0.20
LP	0.54	0.42	0.67	0.71	1.00	0.32	0.33	0.14	0.32	0.19
CH	0.30	0.43	0.43	0.22	0.32	1.00	0.25	0.38	0.20	0.46
DP	0.29	0.32	0.44	0.29	0.33	0.25	1.00	0.54	0.73	0.32
LF	0.12	0.36	0.15	0.09	0.14	0.38	0.54	1.00	0.43	0.57
UW	0.35	0.40	0.34	0.35	0.32	0.20	0.73	0.43	1.00	0.37
ES	0.35	0.64	0.16	0.20	0.19	0.46	0.32	0.57	0.37	1.00

The sensitivity analysis was initially run using the ten variables listed above. A total of 5,000 iterations were calculated and recorded. The eigenvalue of the first principal component was recorded to assess which combinations of data, shrinkage and transformation would account for most of the variance. Using the sensitivity analysis in this manner is analogous to an optimisation routine to find the best method of constructing an index given a selection of variables to choose from and the options of shrinkage and transformation.

Figure 5.3 shows the median and inter-quartile range of ranks for each ED ordered by median rank. EDs at the extremes have much narrower inter-quartile ranges than EDs closer to the mid-range or ranks.

Figure 5.3 Median and inter-quartile range of ranks for each ED



The plot in Figure 5.3 shows only the median and inter-quartile range which does not reveal the full extent of variation. For example, Farranshone ED in Limerick city can be labelled any value from decile 1 to decile 10 depending on the combination of variables, shrinkage and transformation used. This is because from the ten variables available, there are five with high values and five with low values for this particular ED. Such a variation is startling given that the input variables are all proxies for deprivation and that shrinkage and transformation should also affect all EDs. This further highlights the point that the choice of methodology is important.

### 5.3.3 Most probable deprivation index values

By recording the number of times an ED is allocated to each decile, it is possible to generate a probability of an ED being in any given decile. For example, if an ED is in the first decile in ten percent of the iterations, there is a probability of 0.1 that it is in first decile. The decile with the highest probability can be said to be the decile the ED is most likely to be in, based on the parameters applied in the sensitivity analysis. It is interesting to compare the most likely decile with the actual decile given by a fixed selection of indicators and transformation.

A baseline deprivation index is developed for the example deciles using four variables: proportion unemployment (UE), proportion low social class (SC), proportion households with no car (NC), and proportion households in Local Authority housing (LH). These variables are combined using PCA with no shrinkage or transformation. The sensitivity analysis is then run for 5,000 iterations combining any four of the full list of ten variables with random application of Longford shrinkage and/or logit transformation prior to combination by PCA. The results are used to calculate the most probable deprivation decile for each ED. Table 5.3 shows the frequency of EDs by decile for the example deprivation index against the most probable decile from the sensitivity analysis.

Table 5.3 Example versus most probable decile

Example decile	Most probable decile									
	1	2	3	4	5	6	7	8	9	10
1	204	75	37	13	2	5	4	1	1	0
2	74	129	59	33	14	15	8	4	1	5
3	39	77	86	56	30	26	18	5	3	2
4	19	40	72	63	59	38	33	8	6	4
5	17	36	40	60	61	66	36	16	9	1
6	6	22	24	23	50	84	64	41	25	4
7	7	8	20	19	30	58	78	78	38	6
8	3	8	2	5	8	27	63	109	102	15
9	5	0	0	0	0	1	4	63	209	60
10	3	0	0	0	0	0	0	0	42	298

It can be seen that a number of EDs show substantial differences between the most probable decile and the decile from the example index. The percentages of EDs that show no difference or shift no more than one decile are 38.6% and 70.0%, respectively. This indicates that for the large majority of EDs the difference between the example index and the most probable decile is small. This stands to reason as all of the indicators used are positively correlated and should lead to similar deprivation indices. However, there are still a number of EDs that show marked differences and they are of interest. Given that all of the indicators are positively correlated with each other, it is important to understand how one selection of indicators may give such markedly different results for a small number of EDs.

The EDs showing the most extreme differences between the example and most probable decile are listed in Table 5.4 below. Only EDs where there is a difference of 8 or more between the example decile and most probable decile are listed. Indicator values are also shown in the form of the observed value divided by the mean. Thus a value of greater than 1 indicates an above average value. It should be noted that all the EDs for which the most probable decile is lower than the example ED are city EDs that have particularly high values for the no car ownership variable. The EDs where the most probable decile is higher than the example decile are more mixed in terms of area type and, although having mostly low values in the four indicators that comprise the example index, have high values elsewhere.

Table 5.4 EDs with most extreme differences between example and most probable decile

ED name	Area type	Population	Example decile	Most probable decile	Indicator*									
					UE	SC	NC	LH	LP	CH	DP	LF	UW	ES
Arran Quay C	City	2375	10	1	1.76	0.72	3.91	1.81	1.13	1.56	0.71	0.32	0.40	0.40
North City	City	3942	10	1	1.38	0.66	4.62	1.61	0.88	1.66	0.64	0.32	0.40	0.24
Dock A	City	1879	10	1	1.72	1.02	3.07	0.67	1.09	1.81	0.77	0.33	0.39	0.18
Rathmines West A	City	4749	9	1	1.12	0.60	3.23	0.92	0.85	1.47	0.95	0.53	0.48	0.33
Rathmines West B	City	3526	9	1	1.68	0.56	3.33	0.85	0.92	1.43	0.90	0.48	0.61	0.27
Rathmines West D	City	3275	9	1	1.38	0.50	2.63	1.50	1.21	1.48	0.84	0.59	0.55	0.30
South Gate A	City	1431	9	1	1.37	0.86	2.93	0.66	0.91	1.75	0.64	0.39	0.59	0.24
Eyre Square	City	4066	9	1	1.34	0.93	3.00	1.02	1.07	2.00	0.63	0.56	0.56	0.28
Clogher	Remote rural	233	1	9	0.60	0.56	0.51	0.20	1.26	1.05	1.30	1.24	2.56	1.41
Newcastle	Near village	2355	2	10	0.49	0.68	0.65	0.30	0.80	0.27	1.99	1.06	2.98	1.01
Streamhill	Near rural	125	2	10	0.80	0.57	0.30	0.68	0.79	1.22	6.01	1.81	11.95	1.77
Farranshone	City	1175	2	10	0.43	0.53	1.30	0.22	0.96	0.34	3.21	1.54	5.37	0.79
Carriglea	Near rural	557	2	10	0.58	0.93	0.25	0.10	0.88	0.69	2.42	1.32	4.02	0.79
Ardrahan	Remote rural	375	2	10	0.30	0.62	0.68	0.80	0.86	0.59	1.97	1.26	3.31	1.00

\* Indicator values are expressed as the observed divided by the mean. Values greater than 1 equate to above average values (e.g. a value of 2 refers to twice the mean).

This type of analysis highlights the importance of the choice of indicators and the need to assess the effect of a different choice of indicators. It also once again brings to attention the problem of indicators that are regionally biased. It shows that a number of city EDs may be labelled as very deprived largely on the strength of a single indicator: no car ownership. Replacement of this indicator with a different, and arguably equally appropriate deprivation indicator, may have marked consequences for the labelling of these city EDs. It also suggests that even a jack-knife type approach to ascertain the sensitivity of the deprivation index values to the choice of indicators may be a useful tool in determining the appropriateness of indicators.

#### **5.3.4 Optimal index development**

As the first eigenvalue is recorded for every iteration of the above exercise, it is possible to investigate the best and worst performing combinations in terms of the amount of variance explained by the first principal component. This could be used as a basis for an optimisation procedure to determine an 'optimal' method of constructing a deprivation index. Such a method could be criticised for lacking a theoretical basis although if the input variables and methods of shrinkage and transformation are all justifiable then the results of an optimisation might be useful for comparative purposes.

From the 5,000 iterations, the combination of variables and methods that resulted in the PCA with the largest eigenvalue is outlined in Table 5.5 below. Of the five variables, shrinkage is applied to three and transformation to only one. The transformed variable is the car ownership variable (NC) which is highly skewed and tends to be higher in urban areas. For this combination, 66.0% of the variance was accounted for by the first principal component and, using parallel analysis, only one component was required.

Table 5.5 Variables, shrinkage and transformations applied for PCA result with largest eigenvalue

Variable	Shrinkage	Transformation
UE	Longford	None
SC	Longford	None
NC	Longford	Empirical logit
LH	None	None
LP	None	None

The above result assumes that the combination with the largest eigenvalue is in some way the best combination. For instance, it may be desirable to produce an index that minimised the differences between urban and rural areas in terms of proportion of EDs in each decile. A chi square statistic can be used to compare different deprivation indices in terms of the proportion of EDs of each area type in each decile. Equation 5.4 gives the chi square formula.

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^m \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (5.4)$$

Where:  $O_{ij}$  = observed table cell value

$E_{ij}$  = expected table cell value

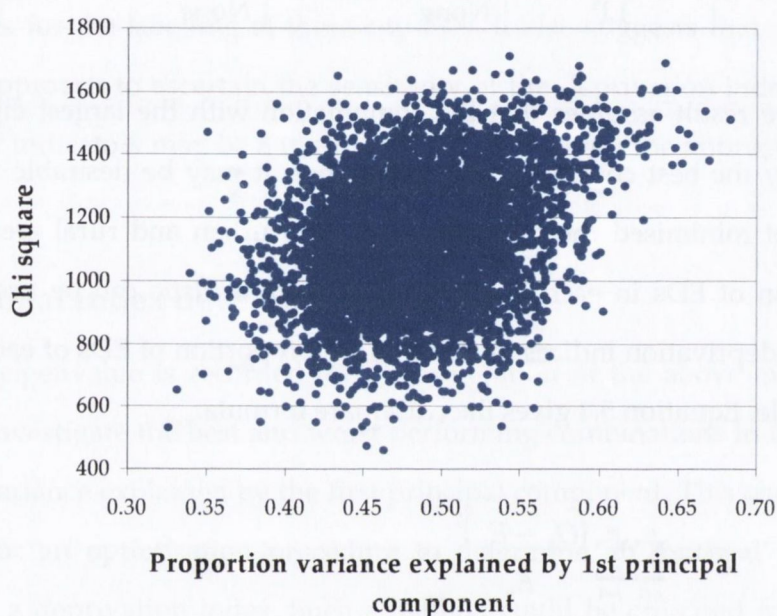
$n$  columns and  $m$  rows

A low chi square value would indicate that the observed values are close to the expected so an optimisation might seek to minimise the chi square value. There is, however, a potential trade-off between low chi-square and a high eigenvalue. For the previous sensitivity analysis, the chi square values were recorded where the expected was for 10% of EDs in each area type to be in each decile. Figure 5.4 shows the plot of chi square against the percentage variance explained by the first principal component. In order to select a combination with a large eigenvalue and a low chi square, the values were standardised and summed. The optimal



combination had eigenvalue and chi square values of 2.805 (explaining 56.12% of the variance) and 596.42, respectively. Only one principal component was required.

Figure 5.4 Plot of chi square against the percentage variance explained by the first principal component



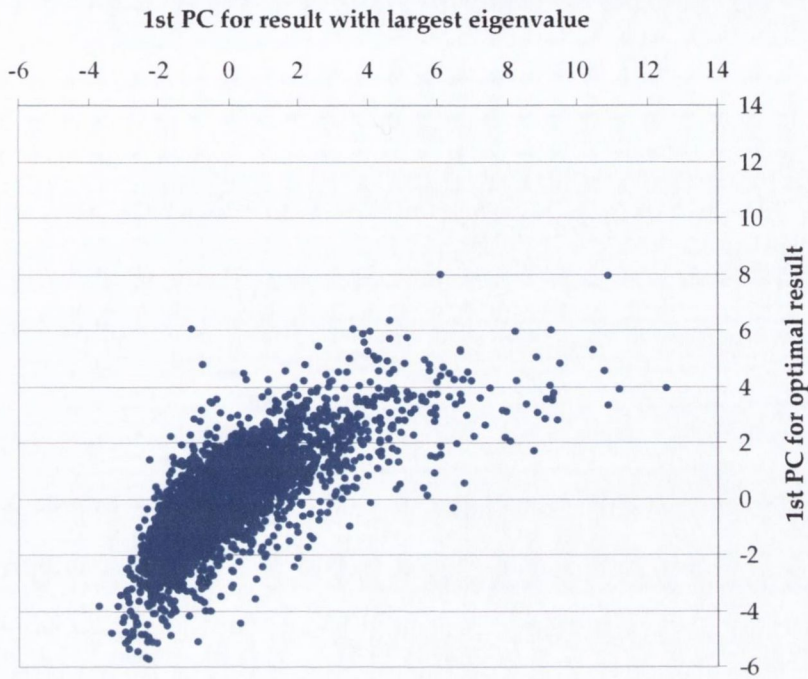
The variables, shrinkage and transformation combination for this optimal result are given in Table 5.6 below. Only two of the five variables are the same as that for the result with the largest eigenvalue. The chi square for the result with the largest eigenvalue is 1176.61, which indicates a much larger departure from expected.

Table 5.6 Variables, shrinkage and transformations applied for optimal PCA result

Variable	Shrinkage	Transformation
Unemployment (UE)	None	None
Low social class (SC)	None	Empirical logit
Disabled persons (DP)	None	Empirical logit
Unable to work (UW)	None	Empirical logit
Early school leavers (ES)	Longford	None

The first principal components were computed and Figure 5.5 gives the plot of the two PCs.

Figure 5.5 Plot of two 1st principal components for 'optimal' results



Although the principal components scores have an  $R^2$  of 0.54, when reduced to deciles there is a large amount of difference between the two deprivation indices. It has been assumed that it is desirable to have near equal proportions of EDs in each area type in each decile. Such an assumption ignores the fact that different area types might legitimately be more typically affluent or deprived than other area types. As an exercise, however, it is interesting to analyse what set of variables might lead to such an index.

As the algorithm used to compute the sensitivity analysis is not optimised to search for answers with high eigenvalues or low chi square values, it is likely that the above combinations are not optimal. Although it does represent a good combination it is probable that better combinations could be found. The routine only performed 5,000 iterations. To find the best 5 variables from a selection of 10, where any number of variables may have shrinkage and/or transformation applied, there are 242,172 possible combinations to test. An efficient search routine would have to be developed to find the optimal or near optimal combination without

testing all of the possible combinations. Other factors such as sensible application of shrinkage and transformation based on the nature and distribution of the variables would also have to be considered.

The application of a sensitivity analysis highlights the fact that the large number of choices facing the developer of a deprivation index need to be carefully considered. Two seemingly appropriate sets of decisions may yield two quite different indices with some areas labelled deprived in one and affluent in the other.

## 5.4 Summary

The application of sensitivity analysis makes it possible to assess the relative importance of different decisions in the process of developing a deprivation index. The initial analysis in section 5.3.1 varied only the application of shrinkage and transformation to a fixed set of indicators prior to combination by PCA. It is interesting to note that the application of shrinkage alone, when compared to no shrinkage, leads to an increase in the number of city EDs in the least and most deprived deciles and a decrease in the number of rural EDs in the least and most deprived deciles. This is due to the tendency for rural EDs to have smaller populations which in turn leads to greater shrinkage. As a consequence, rural EDs are brought closer to the mean values resulting in middle of the range deprivation values. The reverse is seen in more heavily populated city EDs where little shrinkage occurs and more extreme values are preserved.

The plots in Figure 5.2 highlight the fact that shrinkage alone has only a small impact on deprivation scores compared to using the raw values. The use of log transformation has a much more noticeable effect. However, it is the choice of variables that is perhaps the most critical aspect in the development of a deprivation index. It is evident from Figure 5.3, for example, that some small areas can be assigned very different ranks depending on the indicators selected. This is despite the fact that all of the variables used can be considered as reasonable proxies for deprivation. The differences arise in EDs where some of the indicators

have very low values and others high values. This suggests that these EDs are deprived in some respects and not deprived in others. Depending on the selection of indicators used, these EDs may be classified at any point along the spectrum from deprived to not deprived. It is arguable that this issue could be overcome by including all of these variables in the course of constructing a deprivation index. However, inclusion of all appropriate variables can lead to a significant averaging effect.

In section 5.3.4 the prospect of developing an 'optimal index' was discussed with an example. This methodology might provide a route for deciding on a suitable set of indicators, shrinkage and transformation for a given context. It may be desirable, for example, to develop a deprivation index that had the maximum possible correlation with small area lung cancer mortality. Rather than using a standard index, it is possible to find the combination of indicators that best predicts a particular outcome. This approach could be applied in health service resource allocation problems with respect to specific services where the deprivation index used has been tailored to the outcome or measure of interest.



## 6 Discussion

The associations between health, socioeconomic status and rurality were discussed in chapter 1. At an individual level it is clear that low socioeconomic status is closely linked to poor health due to personal behaviour, living conditions and the wider neighbourhood social context. Disadvantaged people are more likely to smoke and be physically inactive, are more likely to live in poor quality housing, are more likely to have a poor diet and are more likely to live in a neighbourhood with higher levels of crime and violence. Numerous studies have found that when individual level characteristics are aggregated to an area level, the socioeconomic status or deprivation of the neighbourhood is also well correlated with the health status and outcomes of the people who live in that area. Furthermore, aside from the socioeconomic status of an area, the degree to which it is urban or rural also has associations with health. Area measures of deprivation and rurality are widespread although the definitions of such measures may have a large impact on any analyses linking deprivation and rurality to health. This study has sought to explore the issues relating to the development of such measures.

### 6.1 Urban-rural issues

It was shown in section 1.4 how both health and poverty show urban-rural variations. While area level health differences may be attributable to a combination of both environmental factors and selective migration, the findings across studies are inconsistent. Some find improved health in rural areas whilst others find that the health differences can be accounted for by socioeconomic differences. The variations in poverty across the urban-rural divide may be largely affected by measurement issues. Shaw's model (page 26) shows how rural deprivation is comprised of three categories: household, opportunity and mobility deprivation.<sup>245</sup> Area level poverty measures generally focus on household aspects, ignoring the problems associated with the lack of opportunities or mobility that may arise from geographic isolation. While opportunity and mobility can be problematic in urban environments, they are not typically due to constraints of physical accessibility. As

such, the rural experience of poverty may arise from a different set of circumstances than urban poverty. Indeed, some of the criticisms levelled at existing deprivation indices stem from possible urban bias propagated by the choice of variables.

The construction of an urban-rural classification is a fundamental step in either assessing the importance of urban-rural differences or to integrate the distinction into research. Rather than a simplistic urban-rural dichotomy, it was shown that there is an urban-rural continuum with different settlement types in between the clearly urban and clearly rural areas. A review of methods of urban-rural classification found that most seek to define a simple urban-rural divide with no account of grades of urbanicity or rurality. The existing Irish method classes a town of 1,500 or more persons as urban with all other population living in rural areas. While it is apparent that the class cut-off points for any nominal classification will have to be chosen, basing the decision on a single variable may lead to substantial misclassification compared to using a different variable. It was shown, for example, that an otherwise rural ED may display an attribute that would classify it as urban if the classification was based on that attribute alone. Therefore a multivariate approach is justified.

A method for producing an urban-rural classification for Ireland was developed and applied in chapter 2. A set of suitable area level attributes were chosen, each of which had been used previously to produce an urban-rural classification. These attributes included settlement size, population density, access to settlements and land use. The consequence of choosing a multivariate approach is that the subsequent classification method is more complex than for a single variable approach. Several clustering techniques were tested along with a multi-criteria classification (MCC) method. Ultimately the MCC option, combined with discriminant analysis, was used. Although k-means clustering resulted in slightly better goodness of fit, it would lead to a classification that was difficult to interpret and label.

The resultant classification is hierarchical where each class can be further subdivided according to the level of detail required by the user. For 2002 it was found that 31.2% of the population live in rural EDs, 7.7% in village EDs, 25.5% in town EDs and the remaining 35.6% in city EDs. The aggregate city and town population is 61.1% compared to the CSO urban population of 59.6% in 2002.<sup>260</sup> The CSO, however, only offer a binary classification.

A cross-temporal analysis was also produced showing the changes in area types from 1986 to 2002. The proportion city population has remained stable although the number of city EDs has increased slightly. The biggest changes have been the increase in town EDs and town population. As the distinction between near and remote village and rural EDs is dictated by proximity to population centres, the number of remote EDs have decreased due to the increasing town populations outside of the main cities and commuting belts.

The classification method presented does have a few weaknesses. As it is an ED based classification, labels are attached based on the living environment of the majority of people in a given ED. This does mean that a substantial minority in, for example, a rural ED might actually be living in the suburbs of a town. This weakness is inherent in the use of areal boundaries. An alternative may be to develop the classification for grid squares although relatively few Irish datasets are available at such a spatial level. A further weakness is that no sensitivity analysis was carried out to determine the impact of changes to the methodology. For instance, k-means clustering was used to determine cut-offs for a number of the variables. If a different method was used, such as hierarchical clustering, the cut-offs may have been different resulting in a different classification. No small area health data were available to assess the extent of urban-rural health differences.

The use of a number of complementary variables enabled the development of an index that captures different aspects of the urban-rural divide. It also made it



possible to develop an index that retains more detail and distinction between area types than a simple single variable index. The index has been widely used throughout this study with systematic variation being observed across the area types for a range of variables. Furthermore, the use of several classes, as opposed to a simple dichotomy, makes it possible to develop a greater understanding of the spatial variation of health and socioeconomic data.

The main findings relating to urban-rural classifications are as follows:

- There is an urban-rural continuum whereby there are a variety of area types that exist that are between the urban and rural extremes
- Representing urban-rural differences as a simple dichotomy is misleading and ignores the urban-rural continuum
- Single variable classifications do not adequately capture the variety of areas that exist
- A detailed small area urban-rural classification has been presented for Ireland using a range of techniques and variables

## **6.2 Deprivation indices**

As was discussed in chapter 1, Townsend introduced the notion of deprivation in terms of exclusion from activities and resources that members of society would normally expect to have access to. Deprived individuals are the 'have-nots' of society. Measurement of area deprivation has evolved over the years but the basic construct tends to be the same: a set of indicators are selected and then combined in some manner into a smaller set of scores which are typically presented grouped into deciles. Different developers have justified very different decisions to produce indices, although sensitivity analyses are rarely conducted to assess the impact of those decisions. The basic steps for producing a deprivation index were outlined in Figure 5.1 (page 187) with a range of options available to the developer at each step.

### 6.2.1 Indicator selection

The choice of indicators is very often constrained by data availability. In most instances, census data form the basis for a deprivation index although in the UK there has been a move to more routinely collected data. This switch has been due to the relatively infrequent UK census which is only once every ten years as opposed to every five years in Ireland. The use of routinely collected data brings with it a range of problems. Some data sets are collected at higher levels of aggregations and small area rates need to be inferred. Some data sets are based on sample information which may have substantial measurement errors associated with it.

The indicators are generally intended to act as proxies for poverty or deprivation, typically in the absence of detailed income information. Common indicators include: unemployment, overcrowding, low social class, lone parents, low education, car ownership and rented accommodation. These indicators represent portions of the populations that are known to be at increased risk of poverty. Not all unemployed people are in poverty, but they are more likely to be in poverty than employed people.

A criticism of deprivation indices is that they frequently incorporate indicators that introduce an urban bias. Car ownership tends to be lower in urban areas, partly due to the greater availability of public transport. Urban inhabitants can choose not to own a car when for many rural inhabitants it is a necessity. Rural areas appear less deprived as a greater proportion of people own a car even though the necessity of owning a car and the associated running costs may only increase disadvantage. Similarly for rented accommodation, there is greater availability of rental houses in the urban areas which consequently increases the proportion of the population living in rented accommodation in urban areas. These two variables tend to be concentrated in urban areas and, if both are included in a deprivation index, can lead to a disproportionate number of urban EDs being considered as very deprived. It is arguable that rural deprivation should be assessed independently of

urban deprivation on the grounds that the suitable indicators are sufficiently different to warrant separate indices.

It was shown in section 4.4.2 that the impact of correlations of urban deprivation indicators in mostly rural EDs can increase the number of urban EDs classified as most deprived. The Irish context is interesting in that although over 60% of the population live in city and town EDs, only 20.5% of EDs are classed as city or town using the classification developed in this study. With 79.5% of observations being of a rural nature, this has implications for a national deprivation index. As was shown, car ownership has higher correlations with other deprivation indicators in rural areas, which in turn increases the weight associated with car ownership. As car ownership rates experience more extreme values in urban areas, this variable contributes substantially to high deprivation values found in urban areas. A pragmatic solution may be to exclude variables that display a very overt urban-rural gradient on the grounds that their inclusion creates an urban or rural bias in the deprivation index.

As a final consideration, Gordon suggested the use of weighting indicators according to the probability of poverty for the population represented by that indicator.<sup>262</sup> For example, if unemployed persons have a 50% probability of being in poverty, then a weight of 0.5 should be given to the unemployment variable. With this suggestion in mind, it is perhaps sensible to restrict the choice of deprivation indicators to those that represent populations with a substantial probability of being in poverty.

### **6.2.2 Shrinkage and transformation**

Prior to combination, deprivation indicators are frequently transformed using several techniques. Principal among these is the method of shrinkage. When dealing with small area data, the denominator population for any given indicator may be quite small and as a consequence, the indicator may be susceptible to large changes by chance. Shrinkage is used to bring indicators closer to a mean value

with the amount of shrinkage being proportional to the standard error associated with the observation. An area with a large population should have a small standard error and the amount of shrinkage applied should be negligible. On the other hand, an area with a very small population will likely have a large standard error and the indicator will be brought closer to the mean.

In chapter 3, three different shrinkage methods were outlined and assessed. The first has been used by Noble et al. in the English indices of deprivation<sup>264</sup>, the second is a univariate method outlined by Longford<sup>291</sup> while the third is an empirical Bayes method.<sup>275</sup> A simulation exercise was performed to analyse the characteristics of each shrinkage technique for a range of datasets with different means and standard deviations. The empirical Bayes method was shown to apply more shrinkage for datasets with a higher mean. A more serious problem was highlighted in section 3.2.2.2 when it was shown that shrinkage with the Noble method is strongly influenced by the numerator rather than just the denominator. As a consequence, several areas with the same denominator but slightly different numerators may have quite different shrinkage weights applied. A further albeit less significant problem with the Noble method is that it does not shrink to the observed mean. Due to the application of an empirical logit transformation, there is a shift in the mean which can impact on the relative positions of observations. Figure 3.3 (page 104) illustrates the potential problem that this might entail with two observations being shrunk towards each other when in fact they are both to the same side of the mean.

The Longford and empirical Bayes methods were shown to fail under certain circumstances. The failure of these techniques appears to be associated with indicators with very low standard deviations. It is argued that for indicators with a very low standard deviation, the need for shrinkage is questionable as most observations are very close to the mean to start with. A quirk of the empirical Bayes method is that the standard deviation at which failure might occur increases with the mean of the indicator. Given the characteristics of the three methods

tested, the Longford method appears to be the most appropriate for applying shrinkage.

Aside from the choice of which method of shrinkage to use, there is also the question of what mean to shrink to. The mean effectively acts as an expected value that the observed local value is shrunk towards. Typically this may be the national mean but a local mean has also been used in some indices. The use of a national mean creates the problem of shrinkage to an urban mean by virtue of the fact that the majority of the population live in urban areas. In Ireland it could be said that the average person is urban while the average ED is rural. In the event of an indicator exhibiting an urban-rural difference, shrinkage to the national mean brings rural EDs closer to the urban mean. Furthermore, the more populace urban EDs tend to have less shrinkage applied than the less populated rural counterparts.

A local mean might be a local authority or administrative district, the justification being that due to regional variation a local mean may be a more appropriate expected value. It was shown in section 3.2.3 that the choice of mean can lead to increased spatial autocorrelation if the chosen localities are small. Administrative boundaries are arbitrarily defined, and an ED on the border of a district may be quite different from the average ED in that district. Indeed, the ED may be more like its neighbouring EDs in a neighbouring district.

To overcome the problem of local shrinkage being constrained by arbitrary boundaries, a Monte Carlo approach was proposed which enabled shrinkage to a local mean for randomly defined districts of varying size. One thousand shrunken values are calculated for each ED based on these randomly defined districts. This method results in a smaller increase in spatial autocorrelation than the use of fixed district boundaries. The drawback to the Monte Carlo approach is that it is not transparent and different choices of minimum and maximum allowable district size affect the results.

Section 3.2.5 outlined an analysis of the impact of shrinkage to different means. The difference between using the national mean and the Monte Carlo method were small but use of a district mean led to more substantial differences. With shrinkage to a district mean, some EDs shifted by up to four deciles from an index calculated with no shrinkage. The number of EDs that were subject to such large shifts was relatively small – using district shrinkage only 3.2% of EDs moved more than one decile. Given the arbitrary nature of district or administrative boundaries, it would seem more logical to use a Monte Carlo approach, or even to select a district of contiguous EDs with a certain minimum population around the ED for the purpose of applying shrinkage.

At a Royal Statistical Society meeting in 2001, Longford put forward arguments against the use of shrinkage in the UK deprivation index.<sup>265</sup> The criticism was in part directed at shrinkage to a district mean alone, rather than national or some combination of the national and district means. It is also suggested that further transformations of the shrunken values may give rise to additional problems. In his commentary, Longford also criticises the use of shrunken values in the subsequent factor analysis used to combine indicators. This stems from the variance-covariance matrix being based on shrunken values rather than the raw values. The application of district level shrinkage means that the ward values are no longer independent of each other. Ideally the ward values should be independent prior to factor analysis.

In response to the above criticisms, the team of Noble et al. who were responsible for the UK deprivation index outlined their rationales for shrinkage and factor analysis.<sup>348</sup> They defend district shrinkage on the grounds that it is sensitive to regional variation in underlying means, the effects are small and, in most cases, the index will ultimately be used at a district level. As regards the use of factor analysis, the small numbers of indicators used in some domains was a function of data availability and it was reasoned that any index is better than no index.

Both the criticisms and the responses have merit. Shrinkage undoubtedly leads to increased spatial autocorrelation but it is arguable that the effects are sufficiently small to be ignored. The presupposition of a single factor model could be overcome by the application of parallel analysis (PA) as a test of how many components to retain. The purpose of factor analysis in this context is not to identify the 'correct' set of indicators; it is merely a method to identify suitable weights which can be readily achieved by PCA.

All of these arguments assume the necessity for shrinkage in the first place. Due to the small populations of some rural EDs it is likely that the proportion unemployed or of low social class may fluctuate substantially by chance. As an extreme example, the denominator for unemployment for Branchfield ED in Sligo is only 44 compared to 18,271 in Blanchardstown-Blakestown ED in Dublin. Both EDs have similar unemployment levels which are below the national average. If the number of unemployed in the Sligo ED increased by 1, the proportion unemployed would increase from 0.045 to 0.068 – above the national average. To achieve a similar increase in the Dublin ED would require an increase of 382 unemployed persons. When an indicator for an ED is sensitive to such small changes it would seem to be appropriate to use shrinkage.

### **6.2.3 Combining indicators**

Chapter 4 dealt with the problem of combining indicators using dimension reduction techniques such as clustering, multidimensional scaling (MDS), factor analysis (FA) and principal components analysis (PCA). The latter two methods produce a continuous score and are commonly used in the development of deprivation indices. The benefit of a score over the groupings produced by clustering or MDS is that areas can be ranked and relative positions compared.

The choice between PCA and FA is more difficult to resolve. The existence of a set of one or more independent hidden variables is intrinsic to FA. The user will generally select a set of variables, hypothesise how many underlying variables exist

and what they represent, and then determine that many factors. The factor loadings are then inspected to attach the predefined labels to the factors based on the initial hypothesis. A different choice of the number of factors to determine can result in a set of factors with very different loadings. A further complication is that rotation can be applied to render the factors more easily interpretable. A range of rotation methods are available and, depending on which is used, can lead to different results. In favour of FA is that it incorporates an error component, acknowledging that some variance is unexplained by the common factors.

The use of FA is often coupled with a sense that some underlying variable, namely deprivation, is being accurately measured. This completely ignores the fact that the analysis is being conducted at an area level. The overlap between indicators is entirely unknown. That an area contains people who are unemployed and people who do not own a car does not necessarily entail that the unemployed do not own cars. There is an increased probability on the basis of income restraints but it is not a certainty. Chatfield and Collins conclude that FA "allows the experimenter to impose his preconceived ideas on the raw data".<sup>308</sup> They recommend that FA is not used in most practical situations.

At the 2001 Royal Statistical Society meeting mentioned in the previous section, Chalmers discussed issues surrounding the use of factor analysis to combine indicators in the UK deprivation index.<sup>349</sup> Chalmers' criticisms are that factor analysis was not a suitable method for combining indicators. He states that in each domain, only single-factor models were considered when it was possible that a second factor existed. Furthermore, he states that the small numbers of indicators used in each domain may not lead to an accurate ranking of the wards. Noble et al. stated that a single-factor model was used because each set of indicators was selected so that only one factor would exist.<sup>348</sup>

Alternatively, there is the option of PCA which is purely an arithmetic combination of the variables based in the covariances or correlations of the variables. There are



no assumptions made about distribution or how many components might be required. There is no error component. A PCA based on covariances is problematic when the variables show large differences in variances – the variables with the largest variances will then receive the largest weights. This can be avoided by using a correlation matrix rather than the covariance matrix. Rather than depending on the judgement of the researcher, a statistical method such as parallel analysis (PA) can be used to determine how many components to retain.

It has already been mentioned that some variables have an urban-rural gradient, and it was noted in section 4.3.3 that the correlations between some variables display regional variation. It was illustrated that the overcrowding variable, which appears to be a reasonable proxy for deprivation at a national level, is a very poor proxy in many counties. A possible solution of geographically weighted PCA (GW-PCA) was proposed in section 4.4 whereby a distance weighted PCA is performed for each small area in turn. This is achieved through a weighted correlation matrix with the weights calculated using one of a number of distance decay functions. The weights for each variable are calculated for each ED before being applied to the variables which have been standardised in the normal manner.

A case study was presented in section 4.5.5 of a four variable deprivation index calculated using GW-PCA. A combination of distance decay function and cut-off distance was selected that resulted in only one principal component being required for all areas. Correlations between variables were shown to be subject to substantial regional variation, as were the consequent weights. The percentage variance explained by the first component varied from 46.1% to 82.1%, with the highest values being found in the Dublin and Limerick city areas. The main advantage over a global PCA could be seen in the weights associated with the variables, particularly those with a distinct urban-rural difference. The car ownership variable had higher weights in rural areas than in the main city areas of Dublin and Cork compensating for the lower rates found in rural areas. Unfortunately, the car ownership values observed in Dublin are so low in some

city centre areas that even very low weights do not greatly diminish the contribution of the car ownership variable to the high deprivation scores observed in parts of Dublin.

The major benefit of this new methodology is that variable weights are determined locally. This goes some way to addressing the problem of variables that bias an index towards urban or rural areas. It also provides substantial information that enables the researcher to assess the appropriateness of the chosen indicators. Typically the indicators are selected on theoretical grounds and then combined for all areas using PCA or FA even though locally the choice of indicators may be a rather poor reflection of deprivation. GW-PCA allows the researcher to study local correlations and eigenvalues to assess the regional performance of the PCA.

#### **6.2.4 Domains**

It is understood that a single index might not capture the many aspects of deprivation, some of them poorly correlated. It has already been pointed out that poverty and deprivation may be expressed differently in urban and rural areas. It is now recognised that some forms of deprivation may be present independently of others and for targeted intervention it is more appropriate to consider those different forms of deprivation separately.

Prior to the calculation of deprivation in distinct domains, it was not unusual to calculate a number of indices from a single set of variables by retaining a number of factors in a FA approach. An example of this would be the Irish index of deprivation developed by Haase.<sup>184</sup> In the adoption of such an approach there is a prior hypothesis as to what domains exist and as to what variables contribute to each domain. Once FA has been applied to the data and the factors have been appropriately labelled, it is said that the factors confirm the hypothesis. Little mention is given to the sometimes counter-intuitive loadings that appear and a cut-off loading value is often applied to prevent the inclusion of variables with little

influence on a factor. Although this is done to make the factors easier to interpret, it might also be used to remove loadings that are difficult to explain.

The definition of distinct domains is appealing not least because all of the variables are included with a view to reflecting that domain. A single factor or component model is used and good correlations between the variables can be anticipated. Instead of attempting to produce a single catch-all index with a sometimes disparate collection of variables, a range of domains are represented by appropriate and pertinent indicators. A domain of access to services can be expected to highlight remote rural areas. A domain of crime will probably highlight more urban areas with a high level of social disorder. Indicators of access and crime might consequently correlate poorly with each other but due to the separation of domains, that is not an issue. An area can be deprived in one domain and not deprived in another.

Ultimately it is still desired to combine domains into a single index of multiple deprivation. This is problematic as some domains will almost certainly have poor if not negative correlations. Thus far in the UK an approach of arbitrarily chosen weights has been used and, with no obvious alternative, this is probably the most pragmatic approach.

For resource allocation and targeting interventions, the use of domains has much greater utility. If, for example, it was intended to alleviate educational disadvantage on the basis of a general deprivation index, it is entirely possible that some or even many of the people in most need of assistance might not receive it as the most deprived areas might not all be the most educationally disadvantaged.

A final point is that identifying income and social inequalities is only one aspect of identifying where health inequalities are likely to occur. Other determinants of health, such as the physical environment (e.g. air pollution) or access to amenities (e.g. recreational space for physical exercise), are also important and the separate

identification of where inequalities in these determinants might be evident is equally important in addressing health inequalities. Indeed, the broad groups of health determinants may form a basis for the selection of domains of deprivation.

### **6.2.5 Robust analysis and influence functions**

The GW-PCA method results in data output that can assist in determining the relative importance of indicators in different geographic areas. It does not, however, identify observations that might be considered outlying or influential. When using data combination techniques such as PCA and FA, the derived weights are quite susceptible to bias because of a small number of outliers or influential observations. As was discussed, an observation can be outlying, influential or both – it is not a given that an outlier is influential or vice versa.

In terms of PCA or FA, a single or small group of outlying observations can influence the correlation or covariance matrix disproportionately. In relation to a deprivation index, this may take the form of a few observations with counter-intuitive values (e.g. they may be deprived in relation to some variables but affluent in relation to others). These observations may dramatically alter the correlation matrix which in turn impacts on the weights associated with the variables. A group of methods for dealing with outliers in PCA and FA is referred to as robust estimation. These methods generally identify the outlying observations based on distance from the multidimensional mean and then compute the correlation matrix with the identified outliers either down-weighted or omitted altogether.

In section 4.6.1.1, robust estimation methods were applied in the construction of a simple deprivation index. The frequency of outliers was identified by area type and it was found that 33.4% of city EDs were considered outliers, compared with 5.6% of remote rural EDs. This is probably due to the extreme indicator values, particularly for car ownership, that can be observed for some city EDs. The application of robust estimation raises an important point: having applied robust

methods to estimate the correlation matrix, the eigenvalues and eigenvectors of that robust matrix are then calculated and applied to unadjusted indicator values. Thus observations that were outliers at the start of computation might be even more outlying in terms of deprivation score values. This is because the weights derived from a robust matrix may be higher for the indicators that are largely responsible for the outlying observations being considered outliers. It is a similar problem to weights from a predominantly rural set of observations leading to an increase in the deprivation values of urban EDs.

The application of robust methodology is, perhaps, a double edged sword. On the one hand it can eliminate the undue influence of outliers, but on the other hand, they are still present as outliers in the resultant deprivation index. It may be that the best option is to apply robust estimation and flag the outliers. The characteristics of the outliers, in terms of the indicator values, should at least be inspected as they may point towards an indicator or spatial feature that needs to be investigated further.

Closely associated with outlier detection is the area of influence functions. There are two components for which influence may be of interest: variables and observations. The influence of variables is quantified by the eigenvector values, or weights. A variable with a large weight is more influential than one with a small weight. Some of the analysis relating to GW-PCA looked at ways to determine the contribution of a variable locally. The influence of individual observations is a more complex problem. Two influence functions were illustrated and the results showed that most influential observations were outliers but not all outliers were influential. The most influential observations were found to be city EDs.

### **6.2.6 Sensitivity analysis**

To illustrate the utility and feasibility of sensitivity analysis an example was given in Chapter 5. It was shown that even a simple sensitivity analysis to test the impact of shrinkage and transformation highlighted how such operations can affect the

resultant deprivation scores. A more complex analysis was also conducted to investigate the impact on deprivation scores of choice of indicators as well as shrinkage and transformation. Depending on the indicators selected, an area could be classified from most to least deprived even though all of the indicators were positively correlated with each other. This illustrated that for some areas the deprivation score can be very sensitive to the choice of indicators.

It was possible, using the results of the sensitivity analysis, to determine the most likely deprivation decile for each area based on a pool of indicators, shrinkage, transformation and combination methods. Using a fixed set of four indicators to generate a baseline index, it could be seen that the most likely decile for a number of EDs was markedly different from the baseline decile.

It was also shown that the method of sensitivity analysis could be used to develop an 'optimal' index. For example, the difference between area types (e.g. urban and rural) could be minimised in terms of how many EDs were classified into each deprivation decile. Such a method could be used to select variables that minimise urban-rural deprivation differences. The primary benefit of this technique would be the development of an index for which the bias towards a particular area type has been minimised.

It is evident that sensitivity analysis is a useful tool in deprivation index development, primarily to assess the impact of the various choices made regarding indicator selection, shrinkage, transformation and data combination. It is particularly useful for highlighting areas that are subject to substantial variation in terms of what deprivation decile they are in.

The main findings pertaining to deprivation index development are as follows:

- Indicator selection is subjective and often dependent on data availability
- Shrinkage is a method for improving the reliability of indicators by adjusting the indicator to a central value

- The importance of numerators in the Noble method and the characteristics of the empirical Bayes method at larger mean-standard deviation combinations make the Longford method most appropriate methodology
- The use of small fixed districts in shrinkage can lead to increased shrinkage and spatial autocorrelation and shrinkage to a misrepresentative mean
- As a district or national mean might misrepresent an ED, a Monte Carlo method of district delineation was presented
- PCA is the most appropriate method for combining variables for deprivation index development
- However, PCA is susceptible to regional bias when estimating variable weights
- A new method of PCA – Geographically Weighted PCA – was presented which enables local calculation of weights to account for regional variation
- The relationship between area types, indicator correlations and indicator weights is complex and should be analysed to determine if an index is biased towards a particular area type
- Robust methods of weights determination can be used to diminish the impact of outliers, although this may inadvertently increase the outlying status of these observations
- Sensitivity analysis is recommended as a technique for assessing the choices made in developing a deprivation index

### **6.3 The health context**

The purpose of developing deprivation indices has generally been either explicitly or implicitly driven by an attempt to identify areas that are likely to a higher need for health services. In the case of Jarman, the index was developed specifically to predict general practice workload.<sup>217</sup> As deprivation indices gained acceptance they have become useful tools in generally identifying areas with adverse social conditions.

Given the numerous examples cited in section 1.3.1, it is clear that there are many instances of significant positive correlations between area deprivation and morbidity. An appropriate validation of a deprivation index might therefore be an analysis of correlations with a range of health status measures.

In the course of this study it has not been possible to validate an index using correlations with health measures. This is due to the fact that none of the main health datasets in Ireland are available at a small area level. The hospital in-patient records are collected at county level, as are perinatal statistics. Mortality data are collected at a sub-county level although still at a high level of aggregation. Cancer morbidity is now being coded to a small area level although this is not widely available and the quality of the small area coding in rural areas is questionable. That is not to say that the development of deprivation indices without suitable validation data is futile. As was discussed in section 1.3.1, numerous studies in the UK have shown strong positive correlations between area level deprivation and poor health. By using similar indicators to the UK it can be expected that positive correlations between deprivation and poor health exist in Ireland. Furthermore, positive correlations between deprivation and poor health have been shown to exist at more aggregated area levels in Ireland.<sup>137</sup>

In the Health Information Strategy report, it is proposed that for key health information systems the data will be geocoded by small area.<sup>350</sup> It is proposed that this would occur during the second phase of actions which is envisaged as being in years 3 to 5 of the strategy implementation. This suggests that routine geocoding will be implemented over the next three years as the report was published in 2004. Such a proposal would be greatly aided by the introduction of a national post code system. It is hoped that these initiatives will be implemented and that within the next three years small area coding of records will be routine in all of the major health datasets.



## 6.4 General remarks

The typical intervention for people living in poverty has been to augment their income by way of benefits. This is a direct attempt to improve financial resources which should in turn reduce deprivation by providing the means to acquire necessities. However, the nature of a number of environmental factors that influence health, such as green space, air pollution and crime, is such that the only way an individual can diminish the impact of these factors is to migrate out of the area. In other words, an individual has little power to affect the environment other than to move to a different healthier environment. This suggests that governments and local authorities should acknowledge the need to promote healthy environments and to intervene directly to improve conditions.

Haynes and Gale compared the association between deprivation and health in rural and urban parts of England and Wales.<sup>232</sup> They found that in rural areas and in inner London, people's health was better than predicted by their deprivation scores. They concluded that if health resources were to be allocated based on social deprivation it would put rural and inner London wards at an advantage. Saul and Payne looked at the prevalence of specific morbidities in relation to socioeconomic measures and found the strength of the relationship depended on the illness being investigated.<sup>201</sup> Kawachi and Kennedy also found that the correlation between income and mortality varied depending on the choice of income measure.<sup>351</sup> The significance of these studies is that they highlight the sensitivity of the correlation between health and deprivation to the choice of deprivation measure and the consequences that might have in terms of resource allocation. It was shown in the sensitivity analysis in section 5.3 that the choice of indicators, shrinkage, transformation and combination can have significant effects on the resulting deprivation index. To assume that the relationship between an index and health is uniform across areas is also incorrect and basing any resource allocation formula on deprivation alone could lead to a poor allocation of resources.

A British Medical Journal editorial by Cox discusses the problem of poverty in rural areas and, more specifically, how it is more hidden in rural areas.<sup>352</sup> He points to the fact that extensive and persistent poverty exists in rural areas but that it is not often highlighted. It is interesting to note that a number of the studies highlighted in chapter 1 were unable to concur on whether the health of rural dwellers was better or worse than that of their urban counterparts. It is possible that due to the sparser population distribution the issue of relative deprivation is less pervasive in rural areas. Relative deprivation refers to the awareness of socioeconomic inequalities between oneself and other members of society. Relative deprivation has been shown to have an impact on health independent of personal deprivation.<sup>353</sup> In urban areas, where there is greater mixing and arguably greater socioeconomic inequalities, people in poverty might be more aware of affluence which can lead to a greater sense of relative deprivation. This heightened relative deprivation may then add to health inequalities in urban areas.

The importance of the urban-rural divide has been stressed in much of the literature and has been the reason for developing an urban-rural classification in this study. It is, of course, rather simplistic to think that all rural areas are similar. The map in Figure 3.7b, for example, shows the local means for unemployment. Unemployment levels in rural Donegal are much higher than those in rural Kerry. Both are remote rural areas but there are clear socioeconomic differences nonetheless. A study of mortality in England and Wales found that mortality was higher in more northern areas, largely due to higher smoking rates in the Northern region.<sup>354</sup> It suggests that rather than just an urban-rural differential, a basic geographic differential might also be significant. Not only is it important whether an area is urban or rural, but also the latitude and longitude. This may come back to the first law of geography: "everything is related to everything else, but near things are more related than distant things."<sup>355</sup> The fact that two EDs are both rural farming areas is not enough to presume that they are similar - the localities of the EDs must also be considered.

What is perhaps most surprising about the many deprivation indices that have been produced is that they typically include no assessment of spatial autocorrelation or urban-rural variation. To the best of my knowledge, no deprivation index has been developed using robust estimation during the data combination stage. The use of PCA or FA also provides a statistic on the amount of variance explained by each component but these values are not generally provided. It would appear that deprivation index developers favour a degree of secrecy about their discipline. The bulk of the literature revolves around the issue of selecting indicators and the finer points of PCA or FA without ever addressing the myriad of other issues that can have a significant impact on the actual deprivation index. Rarely is there any reference to validation or sensitivity analysis, as though the worst is feared about what might be revealed. Given the wide usage of deprivation indices in various forms of resource allocation and planning, it should be a minimum requirement that some form of sensitivity analysis is performed to ascertain how misleading the index might be.

A final point relates to equivalence. A deprivation index is presented in deciles with the assumption that two EDs in the same decile are somehow equivalent in terms of deprivation. This assumption rests on the fact that if two EDs have similar levels in each indicator then they have a similar level of deprivation. Coming back to the car ownership variable, let us say that 20% of the population in an ED does not own a car. If that ED is in an urban area, it is arguable that some of that 20% have chosen not to own a car perhaps because public transport is more convenient. If the ED is in a rural area, such an argument is less defensible. The lack of car in a rural area might also represent a greater degree of exclusion and diminished access than in an urban area. Therefore, depending on whether the ED is rural or urban, the value of 20% not owning a car may be considered more or less deprived. Context can be important for some variables and the manner in which this affects equivalence should at least be discussed in deprivation index development. In the absence of regional information on individual perceptions of what constitutes deprivation, this problem can only be accounted for by caveats.

## **7 Conclusions & recommendations**

This study set out to develop an urban-rural classification for Ireland, assess the characteristics of shrinkage and methods of combining deprivation indicators, and to identify the key problems and possible solutions associated with area-level deprivation measure methodology. These issues have been addressed in the preceding chapters. On the basis of the work completed in this study, the following conclusions and recommendations are presented.

### **7.1 Conclusions**

An individual's health status is affected by a range of factors including occupation, social support, stress and living environment. Health status also varies by socioeconomic status and geographic location. While it is understood that persons with low socioeconomic status will generally have poorer health, the links between geographic location and health are less clear.

The positive correlation between low socioeconomic status and poor health holds at a small area level. Area level deprivation indices are useful aggregate measures of socioeconomic status and can be calculated using census data and other routinely collected data. Positive correlations between increased area level deprivation and poor health have been shown in a number of studies.

A detailed small area classification system was produced using a range of indicators combined using multi-criteria classification and discriminant analysis. It is argued here that acknowledging the urban-rural continuum is a more appropriate approach to urban-rural classification than assuming a simple dichotomy. Future studies investigating differences between urban and rural areas should distinguish between areas using such a multi-level classification as opposed to a simple dichotomous division of areas.

Deprivation indicators frequently vary regionally such that local means can differ substantially. This may partly reflect regional differences in deprivation but it can also point towards a lack of equivalence and comparability across area types. When selecting indicators it is important to assess the extent of regional variation and whether it is systematic by area type (e.g. are values much higher in city and town EDs than in village or rural EDs). The manner in which indicator choice and area type may affect a deprivation index through the correlation matrix is complex. In the Irish context the large proportion of rural EDs can lead to an increased number of urban EDs being considered as very deprived, depending on the set of indicators chosen to form the index.

Shrinkage has become a popular technique to improve the reliability of small area deprivation indicators prior to dimension reduction by Factor Analysis (FA) or Principal Components Analysis (PCA). Three methods have been assessed in this study and each was shown to have limitations. The method of shrinkage described by Longford was found to be the most appropriate method for deprivation index development. The use of a district mean can lead to increased shrinkage and increased spatial autocorrelation. It can also give rise to small areas on the edge of a district being shrunk to the district mean when the neighbouring district might be more representative in terms of socioeconomic characteristics. To overcome this problem, a Monte Carlo approach to district delineation was developed in this thesis. In light of the importance of indicator selection, shrinkage is relatively less critical to the rankings of areas in a deprivation index.

Of the number of methods available for combining indicators into a single area level deprivation index, PCA was found to be the most appropriate based on a subjective analysis. The main alternative, FA, was shown to be less transparent and the results too susceptible to arbitrary choices made by the user. It was shown, however, that there is frequently a regional variation in the correlations between deprivation indicators which can lead to inappropriate weights being applied in some regions. To overcome this problem, a new method of PCA – Geographically

Weighted PCA (GW-PCA) – was developed in this study. GW-PCA facilitates the localised estimation of weights. This is an important new methodology as it can be seen how much variation is explained locally by the first principal component as well as giving the regional variation in weights.

PCA can be affected by outlying observations. Robust correlation matrix estimation methods are available and can be used to obtain robust PCA estimates of weights. While this will diminish or remove the effects of outliers on the correlation matrix, the outliers are retained in the resultant deprivation index and may have even more extreme values after robust analysis.

There are a number of defined steps in the development of a deprivation index: indicator selection, shrinkage, transformation, dimension reduction and presentation. At each stage a number of methods are available to the developer and the choices made have implications for the resultant deprivation index. Sensitivity analysis is a useful tool for assessing the implications of those choices and should be used when developing a deprivation index.

It is envisaged that a greater number of Irish health related datasets will be routinely small area coded in the coming years. Researchers will be able to investigate issues relating to health inequalities and health geography in greater detail as these data become available. It is therefore important that appropriate techniques are used for both area level deprivation measurement and for distinguishing between different area types.

The sensitivity analyses used in this thesis highlighted the impacts of different methodologies and choices on the ranks of small areas in a deprivation index. The use of small area health data will make it possible to investigate whether those impacts have positive or negative effects on the correlation between a deprivation index and health status measures. Such analyses may make it possible to develop a 'best practice' approach for the development of a deprivation index.

It is intended to apply the methodological lessons and advances of this thesis to the development of a new national deprivation index for Ireland as and when the 2006 census data become available.

## 7.2 Recommendations

On the basis of the findings of this study, a number of recommendations are made.

The following are general recommendations:

- When developing an urban-rural classification the urban-rural continuum must be recognised. A simple dichotomous classification based on a single variable should not be used.
- Mortality and morbidity data should be collected at a small area level. This would enable both the validation of deprivation indices and analyses to better understand the links between socio-economic indicators and health status. These data are collected by a range of bodies but primarily in Ireland under the auspices of the Health Services Executive and Central Statistics Office. It is imperative that small area coding of residences, if not exact latitude and longitude, must become a routine element of patient records.

The following recommendations are intended for those developing deprivation indices:

- To avoid the introduction of bias to a specific area type, deprivation indicators should be tested for systematic variation by urban-rural area type. Consideration should be given to excluding indicators that show marked systematic variation by area type.
- The method described by Longford should be used for the shrinkage of deprivation indicators.
- Shrinkage to a district mean can lead to excessive shrinkage and so should only be applied if districts are not comprised of small numbers of areas. Ideally, a Monte Carlo method of district delineation should be used for the purposes of shrinkage.

- To account for regional variation, deprivation indicators should be combined using GW-PCA.
- Robust correlation matrix estimation should be employed to adjust for the impact of outliers. However it must be used with caution as under some circumstances outliers can be made even more extreme.
- Sensitivity analysis should always be used to assess the impact on the resultant deprivation index of choices regarding indicator selection, shrinkage, data transformation and combination of indicators.
- Consideration should be given to developing optimal deprivation indices tailored to a specific context to maximise the correlation between the index and a health status or outcome measure of interest.



The actual for-against-variant information indicated should be considered...  
...the actual for-against-variant information indicated should be considered...  
...the actual for-against-variant information indicated should be considered...

7.2. Recommendations  
Sensitivity analysis should always be used to assess the impact of the...  
...Sensitivity analysis should always be used to assess the impact of the...  
...Sensitivity analysis should always be used to assess the impact of the...

...The following recommendations are included for...  
...The following recommendations are included for...  
...The following recommendations are included for...

...to avoid the inclusion of bias in a meta-analysis...  
...to avoid the inclusion of bias in a meta-analysis...  
...to avoid the inclusion of bias in a meta-analysis...

...Consideration should be given to including...  
...Consideration should be given to including...  
...Consideration should be given to including...

## 8 References

1. Marmot M. Income inequality, social environment, and inequalities in health. *Journal of Policy Analysis and Management* 2001;20(1):156-159.
2. de Hollander AEM, Staatsen BAM. Health, environment and quality of life: an epidemiological perspective on urban development. *Landscape and Urban Planning* 2003;65(1-2):53-62.
3. Marmot M. Social determinants of health inequalities. *The Lancet* 2005;365(9464):1099-1104.
4. Fleming DW. More evidence, more action: Addressing the social determinants of health. *American Journal of Preventive Medicine* 2003;24(3, Supplement 1):1-1.
5. Dahlgren G, Whitehead M. Policies and strategies to promote social equity in health. Stockholm: Institute of Futures Studies, 1991.
6. San Francisco Department of Public Health. 2002 Overview of health. San Francisco: San Francisco Department of Public Health, 2002.
7. Denton M, Walters V. Gender differences in structural and behavioral determinants of health: an analysis of the social production of health. *Social Science & Medicine* 1999;48(9):1221-1235.
8. Denton M, Prus S, Walters V. Gender differences in health: a Canadian study of the psychosocial, structural and behavioural determinants of health. *Social Science & Medicine* 2004;58(12):2585-2600.
9. Walters V, McDonough P, Strohschein L. The influence of work, household structure, and social, personal and material resources on gender differences in health: an analysis of the 1994 Canadian National Population Health Survey. *Social Science & Medicine* 2002;54(5):677-692.
10. Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJL. Selected major risk factors and global and regional burden of disease. *The Lancet* 2002;360(9343):1347-1360.
11. Bartley M, Blane D, Brunner E, Dorling D, Ferrie J, Jarvis M, et al. Social determinants of health - the solid facts. In: Wilkinson R, Marmot M, editors. Copenhagen: World Health Organization, 2003.
12. Gisselmann MD. The influence of maternal childhood and adulthood social class on the health of the infant. *Social Science & Medicine* 2006;63:1023-1033.
13. Latif AHA, Green DA, Li WCW. The effect of deprivation on child health in Bro Taf. *Public Health* 1999;113:211-214.
14. Case A, Fertig A, Paxson C. The lasting impact of childhood health and circumstance. *Journal of Health Economics* 2005;24:365-389.
15. Davey Smith G, Hart C, Blane D, Hole D. Adverse socioeconomic conditions in childhood and cause specific adult mortality: prospective observational study. *British Medical Journal* 1998;316:1631-1635.
16. Hack M. Young adult outcomes of very-low-birth-weight children. *Seminars in Fetal and Neonatal Medicine* 2006;11(2):127-137.
17. Doyle LW, Faber B, Callanan C, Morley R. Blood Pressure in Late Adolescence and Very Low Birth Weight. *Pediatrics* 2003;111(2):252-257.

18. Irving RJ, Belton NR, Elton RA, Walker BR. Adult cardiovascular risk factors in premature babies. *The Lancet* 2000;355(9221):2135-2136.
19. Donma MM, Donma O. Low birth weight: a possible risk factor also for liver diseases in adult life? *Medical Hypotheses* 2003;61(4):435-438.
20. Matheson FI, Moineddin R, Dunn JR, Creatore MI, Gozdyra P, Glazier RH. Urban neighborhoods, chronic stress, gender and depression. *Social Science & Medicine*;In Press, Corrected Proof.
21. Latkin CA, Curry AD. Stressful neighbourhoods and depression: a prospective study of the impact of neighbourhood disorder. *Journal Of Health And Social Behavior* 2003;44(1):34-44.
22. Lin N, Ensel WM. Life stress and health: stressors and resources. *American Sociological Review* 1989;54(3):382-399.
23. Elliott M. The stress process in neighborhood context. *Health & Place* 2000;6(4):287-299.
24. Carroll D, Harrison LK, Johnston DW, Ford G, Hunt K, Der G, et al. Cardiovascular reactions to psychological stress: the influence of demographic variables. *Journal of Epidemiology and Community Health* 2000;54(11):876-877.
25. Kivimaki M, Leino-Arjas P, Luukkonen R, Riihimaki H, Vahtera J, Kirjonen J. Work stress and risk of cardiovascular mortality: prospective cohort study of industrial employees. *BMJ* 2002;325(7369):857.
26. Bruner EJ, Kivimaki M, Siegrist J, Theorell T, Luukkonen R, Riihimaki H, et al. Is the effect of work stress on cardiovascular mortality confounded by socioeconomic factors in the Valmet study? *Journal of Epidemiology and Community Health* 2004;58(12):1019-1020.
27. Lochner K, Kawachi I, Kennedy BP. Social capital: a guide to its measurement. *Health & Place* 1999;5:259-270.
28. Veenstra G, Luginaah I, Wakefield S, Birch S, Eyles J, Elliott S. Who you know, where you live: social capital, neighbourhood and health. *Social Science & Medicine* 2005;60:2799-2818.
29. Poortinga W. Social capital: An individual or collective resource for health? *Social Science & Medicine* 2006;62(2):292-302.
30. Skrabski Á, Kopp M, Kawachi I. Social capital in a changing society: cross sectional associations with middle aged female and male mortality rates. *Journal of Epidemiology and Community Health* 2003;57:114-119.
31. Drukker M, Buka SL, Kaplan C, McKenzie K, Van Os J. Social capital and young adolescents' perceived health in different sociocultural settings. *Social Science & Medicine* 2005;61:185-198.
32. Ziersch AM. Health implications of access to social capital: findings from an Australian study. *Social Science & Medicine* 2005;61(10):2119-2131.
33. Buchanan A. Children aged 0-13 at risk of social exclusion: Impact of government policy in England and Wales. *Children and Youth Services Review* 2006;28(10):1135-1151.
34. Robila M. Economic pressure and social exclusion in Europe. *The Social Science Journal* 2006;43(1):85-97.
35. Schonfelder S, Axhausen KW. Activity spaces: measures of social exclusion? *Transport Policy* 2003;10(4):273-286.

36. Cattell V. Poor people, poor places, and poor health: the mediating role of social networks and social capital. *Social Science & Medicine* 2001;52(10):1501-1516.
37. Baumeister RF, Twenge JM, Nuss CK. Effects of Social Exclusion on Cognitive Processes: Anticipated Aloneness Reduces Intelligent Thought. *Journal of Personality and Social Psychology* 2002;83(4):817-827.
38. Siegrist J. Place, social exchange and health: proposed sociological framework. *Social Science & Medicine* 2000;51(9):1283-1293.
39. Twenge JM, Catanese KR, Baumeister RF. Social Exclusion Causes Self-Defeating Behavior. *Journal of Personality and Social Psychology* 2002;83(3):606-615.
40. Ferrie JE, Shipley MJ, Newman K, Stansfeld SA, Marmot M. Self-reported job insecurity and health in the Whitehall II study: potential explanations of the relationship. *Social Science & Medicine* 2005;60(7):1593-1602.
41. Virtanen P, Vahtera J, Kivimaki M, Pentti J, Ferrie J. Employment security and health. *Journal of Epidemiology and Community Health* 2002;56(8):569-574.
42. Broom DH, D'Souza RM, Strazdins L, Butterworth P, Parslow R, Rodgers B. The lesser evil: Bad jobs or unemployment? A survey of mid-aged Australians. *Social Science & Medicine* 2006;63:575-586.
43. Sacker A, Clarke P, Wiggins RD, Bartley M. Social dynamics of health inequalities: a growth curve analysis of aging and self assessed health in the British household panel survey 1991-2001. *Journal of Epidemiology and Community Health* 2005;59(6):495-501.
44. Kivimaki M, Virtanen M, Vartia M, Elovainio M, Vahtera J, Keltikangas-Jarvinen L. Workplace bullying and the risk of cardiovascular disease and depression. *Occupational and Environmental Medicine* 2003;60(10):779-783.
45. Robert A. Baron JHN. Workplace violence and workplace aggression: Evidence on their relative frequency and potential causes. *Aggressive Behavior* 1996;22(3):161-173.
46. Niedhammer I, David S, Degioanni S. Association between workplace bullying and depressive symptoms in the French working population. *Journal of Psychosomatic Research* 2006;61(2):251-259.
47. Roberts RE, Lee ES. Occupation and the Prevalence of Major Depression, Alcohol, and Drug Abuse in the United States. *Environmental Research* 1993;61(2):266-278.
48. HSA. Fatality statistics: Health & Safety Authority, 2006.
49. Bena A, Mamo C, Marinacci C, Pasqualini O, Tomaino A, Campo G, et al. Risk of repeat accidents by economic activity in Italy. *Safety Science* 2006;44(4):297-312.
50. Beland F, Birch S, Stoddart G. Unemployment and health: contextual-level influences on the production of health in populations. *Social Science & Medicine* 2002;55(11):2033-2052.
51. Turner JB. Economic context and the health effects of unemployment. *Journal Of Health And Social Behavior* 1995;36(3):213-229.
52. Bartley M. Unemployment and ill health: understanding the relationship. *Journal of Epidemiology and Community Health* 1994;48(4):333-337.
53. Jin RL, Shah CP, Svoboda TJ. The impact of unemployment on health: a review of the evidence. *Canadian Medical Association Journal* 1995;153(5):529-540.

54. Mathers CD, Schofield DJ. The health consequences of unemployment: the evidence. *Medical Journal of Australia* 1998;168:178-182.
55. Hammarström A, Janlert U. Early unemployment can contribute to adult health problems: results from a longitudinal study of school leavers. *Journal of Epidemiology and Community Health* 2002;56:624-630.
56. Wadsworth MEJ, Montgomery SM, Bartley MJ. The persisting effect of unemployment on health and social well-being in men early in working life. *Social Science & Medicine* 1999;48(10):1491-1499.
57. Martikainen PT, Valkonen T. Excess mortality of unemployed men and women during a period of rapidly increasing unemployment. *The Lancet* 1996;348(9032):909-912.
58. Novo M, Hammarström A, Janlert U. Do high levels of unemployment influence the health of those who are not unemployed? A gendered comparison of young men and women during boom and recession. *Social Science & Medicine* 2001;53:293-303.
59. Ruhm CJ. Healthy living in hard times. *Journal of Health Economics* 2005;24:341-363.
60. Bellaby P, Bellaby F. Unemployment and ill-health: local labour markets and ill health in Britain 1984-1991. *Work, Employment & Society* 1999;13(3):461-482.
61. Williams CC. Does work pay? Spatial variations in the benefits of unemployment and coping abilities of the unemployed. *Geoforum* 2001;32:199-214.
62. Goldman D, Lakdawalla D. Understanding health disparities across education groups. *NBER Working Paper Series No. 8328*. Cambridge, MA: National Bureau of Economic Research, 2001.
63. Combat Poverty Agency. Educational disadvantage in Ireland. Dublin: CPA, 2003.
64. Huisman M, Kunst AE, Bopp M, Borgan J-K, Borrell C, Costa G, et al. Educational inequalities in cause-specific mortality in middle-aged and older men and women in eight western European populations. *The Lancet* 2005;365(9458):493-500.
65. van Lenthe FJ, Gevers E, Joung IMA, Bosma H, Mackenbach JP. Material and behavioral factors in the explanation of educational differences in incidence of acute myocardial infarction: the Globe Study. *Annals of Epidemiology* 2002;12:535-542.
66. Rasmussen JN, Rasmussen S, Gislason GH, Buch P, Abildstrom SZ, Køber L, et al. Mortality after acute myocardial infarction according to income and education. *Journal of Epidemiology and Community Health* 2006;60:351-356.
67. Yarnell J, Yu S, McCrum E, Arveiler D, Hass B, Dallongeville J, et al. Education, socioeconomic and lifestyle factors, and risk of coronary heart disease: the PRIME Study. *Int. J. Epidemiol.* 2005;34(2):268-275.
68. Cavelaars AE, Kunst AE, Geurts JJ, Crialesi R, Grotvedt L, Helmert U, et al. Differences in self reported morbidity by educational level: a comparison of 11 western European countries. *Journal of Epidemiology and Community Health* 1998;52(4):219-227.
69. Elo IT, Preston SH. Educational differentials in mortality: United States, 1979-1985. *Social Science & Medicine* 1996;42(1):47-57.

70. Bopp M, Minder CE. Mortality by education in German speaking Switzerland, 1990-1997: results from the Swiss National Cohort. *International Journal of Epidemiology* 2003;32(3):346-354.
71. van Oort FVA, van Lenthe FJ, Mackenbach JP. Cooccurrence of lifestyle risk factors and the explanation of education inequalities in mortality: results from the GLOBE study. *Preventive Medicine* 2004;39(6):1126-1134.
72. Howden-Chapman P. Housing and inequalities in health. *Journal of Epidemiology and Community Health* 2002;56:645-646.
73. Dunn JR. Housing and inequalities in health: a study of socioeconomic dimensions of housing and self reported health from a survey of Vancouver residents. *Journal of Epidemiology and Community Health* 2002;56:671-681.
74. Ellaway A, Macintyre S. Does housing tenure predict health in the UK because it exposes people to different levels of housing related hazards in the home or its surroundings? *Health & Place* 1998;4(2):141-150.
75. Macintyre S, Ellaway A, Hiscock R, Kearns A, Der G, McKay L. What features of the home and the area might help to explain observed relationships between housing tenure and health? Evidence from the west of Scotland. *Health & Place* 2003;9(3):207-218.
76. Clinch JP, Healy JD. Housing standards and excess winter mortality. *Journal of Epidemiology and Community Health* 2000;54(9):719-720.
77. Aylin P, Morris S, Wakefield J, Grossinho A, Jarup L, Elliott P. Temperature, housing, deprivation and their relationship to excess winter mortality in Great Britain, 1986-1996. *Int. J. Epidemiol.* 2001;30(5):1100-1108.
78. Evans J, Hyndman S, Stewart-Brown S, Smith D, Petersen S. An epidemiological study of the relative importance of damp housing in relation to adult health. *Journal of Epidemiology and Community Health* 2000;54(9):677-686.
79. Platts-Mills TAE, Vervloet D, Thomas WR, Aalberse RC, Chapman MD. Indoor allergens and asthma: Report of the Third International Workshop. *Journal of Allergy and Clinical Immunology* 1997;100(6):S2-S24.
80. Holmes P, Tuckett C. Airborne particles: exposure in the home and health effects. Leicester, UK: MRC Institute for Environment and Health, 2000.
81. BeruBe KA, Sexton KJ, Jones TP, Moreno T, Anderson S, Richards RJ. The spatial and temporal variations in PM10 mass from six UK homes. *Science of The Total Environment* 2004;324(1-3):41-53.
82. Evans GW, Lepore SJ, Shejwal BR, Palsane MN. Chronic residential crowding and children's well-being: an ecological perspective. *Child Development* 1998;69(6):1514-1523.
83. Evans GW, Lercher P, Kofler WW. Crowding and children's mental health: the role of house type. *Journal of Environmental Psychology* 2002;22(3):221-231.
84. Heukelbach J, Feldmeier H. Scabies. *The Lancet* 2006;367(9524):1767-1774.
85. Wanyeki I, Olson S, Brassard P, Menzies D, Ross N, Behr M, et al. Dwellings, crowding, and tuberculosis in Montreal. *Social Science & Medicine* 2006;63(2):501-511.
86. Maas J, Verheij RA, Groenewegen PP, de Vries S, Spreeuwenberg P. Green space, urbanity, and health: how strong is the relation? *Journal of Epidemiology and Community Health* 2006;60:587-592.

87. Ashton JR. Health and greening the city. *Journal of Epidemiology and Community Health* 2002;56:896.
88. Takano T, Nakamura K, Watanabe M. Urban residential environments and senior citizens' longevity in megacity areas: the importance of walkable green spaces. *Journal of Epidemiology and Community Health* 2002;56:913-918.
89. Ross CE. Neighbourhood disadvantage and adult depression. *Journal Of Health And Social Behavior* 2000;41(2):177-187.
90. Chaix B, Leyland AH, Sabel CE, Chauvin P, Råstam L, Kristersson H, et al. Spatial clustering of mental disorders and associated characteristics of the neighbourhood context in Malmö, Sweden, in 2001. *Journal of Epidemiology and Community Health* 2006;60:427-435.
91. Sundquist K, Theobalda H, Yang M, Lia X, Johansson S-E, Sundquist J. Neighborhood violent crime and unemployment increase the risk of coronary heart disease: A multilevel study in an urban setting. *Social Science & Medicine* 2006;62:2061-2071.
92. Storr CL, Chen C-Y, Anthony JC. "Unequal opportunity": neighbourhood disadvantage and the chance to buy illegal drugs. *Journal of Epidemiology and Community Health* 2004;58(3):231 - 237.
93. Bolland JM. Hopelessness and risk behaviour among adolescents living in high-poverty inner-city neighbourhoods. *Journal of Adolescence* 2003;26:145-158.
94. Crowder K, South SJ. Neighborhood distress and school dropout: the variable significance of community context. *Social Science Research* 2003;32:659-698.
95. Ross NA, Tremblay S, Graham K. Neighbourhood influences on health in Montreal, Canada. *Social Science & Medicine* 2004;59:1485-1494.
96. Pearce J, Witten K, Bartie P. Neighbourhoods and health: a GIS approach to measuring community resource accessibility. *Journal of Epidemiology and Community Health* 2006;60:389-395.
97. Ross CE, Mirowsky J. Neighborhood disadvantage, disorder, and health. *Journal Of Health And Social Behavior* 2001;42(3):258-276.
98. Janghorbani M, Stenhouse EA, Jones RB, Millward BA. Is neighbourhood deprivation a risk factor for gestational diabetes mellitus? *Diabetic Medicine* 2006;23:313-317.
99. Diez Roux AV. Investigating neighborhood and area effects on health. *American Journal of Public Health* 2001;91(11):1783-1789.
100. Kappos AD, Bruckmann P, Eikmann T, Englert N, Heinrich U, Hoppe P, et al. Health effects of particles in ambient air. *International Journal of Hygiene and Environmental Health* 2004;207(4):399-407.
101. Samet JM, Dominici F, Curriero FC, Coursac I, Zeger SL. Fine Particulate Air Pollution and Mortality in 20 U.S. Cities, 1987-1994. *The New England Journal of Medicine* 2000;343(24):1742-1749.
102. Ware JH. Particulate Air Pollution and Mortality -- Clearing the Air. *The New England Journal of Medicine* 2000;343(24):1798-1799.
103. Zmirou D, Gauvin S, Pin I, Momas I, Sahraoui F, Just J, et al. Traffic related air pollution and incidence of childhood asthma: results of the Vesta case-control study. *Journal of Epidemiology and Community Health* 2004;58(1):18-23.
104. Information Management Unit. Health Statistics 2002. Dublin: The Stationery Office, 2003.

105. CSO. Deaths from principal causes registered in the years 1998 to 2005, <http://www.cso.ie/statistics/principalcausesofdeath.htm>.
106. Elvik R. How much do road accidents cost the national economy? *Accident Analysis & Prevention* 2000;32:849-851.
107. Künzli N, Kaiser R, Medina S, Studnicka M, Chanel O, Filliger P, et al. Public-health impact of outdoor and traffic-related air pollution: a European assessment. *The Lancet* 2000;356(9232):795-801.
108. Chambers JA, Swanson V. A health assessment tool for multiple risk factors for obesity: Results from a pilot study with UK adults. *Patient Education and Counseling* 2006;62(1):79-88.
109. Marmot M. Smoking and inequalities. *The Lancet* 2006;368(9533):341-342.
110. Peto R, Lopez AD, Boreham J, Thun M. Mortality from smoking in developed countries: 1950 - 2000 (2nd edition). Oxford: Oxford University, 2006.
111. Room R, Babor T, Rehm J. Alcohol and public health. *The Lancet* 2005;365(9458):519-530.
112. Britton A, McPherson K. Mortality in England and Wales attributable to current alcohol consumption. *Journal of Epidemiology and Community Health* 2001;55(6):383-388.
113. Arndt V, Rothenbacher D, Krauledat R, Daniel U, Brenner H. Age, alcohol consumption, and all-cause mortality. *Annals of Epidemiology* 2004;14(10):750-753.
114. Hart CL, Smith GD, Hole DJ, Hawthorne VM. Alcohol consumption and mortality from all causes, coronary heart disease, and stroke: results from a prospective cohort study of Scottish men with 21 years of follow up. *BMJ* 1999;318(7200):1725-1729.
115. Frischer M, Goldberg D, Rahman M, Berney LEE. Mortality and survival among a cohort of drug injectors in Glasgow, 1982-1994. *Addiction* 1997;92(4):419-427.
116. Darke S, Ross J. Suicide among heroin users: rates, risk factors and methods. *Addiction* 2002;97(11):1383-1394.
117. Bartu A, Freeman NC, Gawthorne GS, Codde JP, Holman CDAJ. Mortality in a cohort of opiate and amphetamine users in Perth, Western Australia. *Addiction* 2004;99(1):53-60.
118. Bryant WK, Galea S, Tracy M, Markham Piper T, Tardiff KJ, Vlahov D. Overdose deaths attributed to methadone and heroin in New York City, 1990-1998. *Addiction* 2004;99(7):846-854.
119. Fridell M, Hesse M. Psychiatric severity and mortality in substance abusers: A 15-year follow-up of drug users. *Addictive Behaviors* 2006;31(4):559-565.
120. Palepu A, Tyndall MW, Leon H, Muller J, O'Shaughnessy MV, Schechter MT, et al. Hospital utilization and costs in a cohort of injection drug users. *Canadian Medical Association Journal* 2001;165(4):415-420.
121. Laine C, Hauck WW, Gourevitch MN, Rothman J, Cohen A, Turner BJ. Regular Outpatient Medical and Drug Abuse Care and Subsequent Hospitalization of Persons Who Use Illicit Drugs. *JAMA: The Journal of the American Medical Association* 2001;285(18):2355-2362.
122. Krantz MJ, Mehler PS. Treating Opioid Dependence: Growing Implications for Primary Care. *Archives of Internal Medicine* 2004;164(3):277-288.



123. Pallapies D. Trends in childhood disease. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis* 2006;608(2):100-111.
124. James WPT, Nelson M, Ralph A, Leather S. Socioeconomic determinants of health: The contribution of nutrition to inequalities in health. *BMJ* 1997;314(7093):1545.
125. Must A, Spadano J, Coakley EH, Field AE, Colditz G, Dietz WH. The Disease Burden Associated With Overweight and Obesity. *JAMA: The Journal of the American Medical Association* 1999;282(16):1523-1529.
126. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, et al. Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001. *JAMA: The Journal of the American Medical Association* 2003;289(1):76-79.
127. Hu FB, Manson JE, Stampfer MJ, Colditz G, Liu S, Solomon CG, et al. Diet, Lifestyle, and the Risk of Type 2 Diabetes Mellitus in Women. *The New England Journal of Medicine* 2001;345(11):790-797.
128. Mark DH. Deaths Attributable to Obesity. *JAMA: The Journal of the American Medical Association* 2005;293(15):1918-1919.
129. Flegal KM, Graubard BI, Williamson DF, Gail MH. Excess Deaths Associated With Underweight, Overweight, and Obesity. *JAMA: The Journal of the American Medical Association* 2005;293(15):1861-1867.
130. Bender R, Zeeb H, Schwarz M, Jockel K-H, Berger M. Causes of death in obesity: Relevant increase in cardiovascular but not in all-cancer mortality. *Journal of Clinical Epidemiology* 2006;59(10):1064-1071.
131. Simon GE, Von Korff M, Saunders K, Miglioretti DL, Crane PK, van Belle G, et al. Association Between Obesity and Psychiatric Disorders in the US Adult Population. *Archives of General Psychiatry* 2006;63(7):824-830.
132. Bauman A, Craig CL. The place of physical activity in the WHO Global Strategy on Diet and Physical Activity. *International Journal of Behavioral Nutrition and Physical Activity* 2005;2(10).
133. WHO. Global strategy on diet, physical activity & health. *Final Resolution WHA57.17*, 2004.
134. Smith GD, Shipley MJ, Batty GD, Morris JN, Marmot M. Physical activity and cause-specific mortality in the Whitehall study. *Public Health* 2000;114(5):308-315.
135. Erikssen G, Liestol K, Bjornholt J, Thaulow E, Sandvik L, Erikssen J. Changes in physical fitness and changes in mortality. *The Lancet* 1998;352(9130):759-762.
136. Carnethon MR, Gidding SS, Nehgme R, Sidney S, Jacobs DR, Jr., Liu K. Cardiorespiratory Fitness in Young Adulthood and the Development of Cardiovascular Disease Risk Factors. *JAMA: The Journal of the American Medical Association* 2003;290(23):3092-3100.
137. Barry J, Sinclair H, Kelly A, O'Loughlin R, Handy D, O'Dowd T. Inequalities in health in Ireland - hard facts. Dublin: Department of Community Health & General Practice, Trinity College, 2001.
138. Balanda KP, Wilde J. Inequalities In Mortality 1989-1998: A Report On All-Ireland Mortality Data. Dublin: The Institute of Public Health In Ireland, 2001.

139. Aveyard P, Manaseki S, Chambers J. The relationship between mean birth weight and poverty using the Townsend deprivation score and the Super Profile classification system. *Public Health* 2002;116:308-314.
140. Kivimaki M, Smith GD, Elovainio M, Pulkki L, Keltikangas-Jarvinen L, Talttonen L, et al. Socioeconomic Circumstances in Childhood and Blood Pressure in Adulthood: The Cardiovascular Risk in Young Finns Study. *Annals of Epidemiology* 2006;16(10):737-742.
141. King K, Stedman J. Analysis of air pollution and social deprivation. Abingdon, UK: AEA Technology, 2000.
142. Graham D, Glaister S, Anderson R. The effects of area deprivation on the incidence of child and adult pedestrian casualties in England. *Accident Analysis & Prevention* 2005;37:125-135.
143. Boardman JD, Finch BK, Ellison CG, Williams DR, Jackson JS. Neighbourhood disadvantage, stress and drug use among adults. *Journal Of Health And Social Behavior* 2001;42(2):151-165.
144. Janssen I, Boyce WF, Simpson K, Pickett W. Influence of individual- and area-level measures of socioeconomic status on obesity, unhealthy eating, and physical inactivity in Canadian adolescents. *American Journal of Clinical Nutrition* 2006;83(1):139-145.
145. Kivimaki M, Kinnunen M-L, Pitkanen T, Vahtera J, Elovainio M, Pulkkinen L. Contribution of Early and Adult Factors to Socioeconomic Variation in Blood Pressure: Thirty-Four-Year Follow-up Study of School Children. *Psychosomatic Medicine* 2004;66(2):184-189.
146. McGrath JJ, Matthews KA, Brady SS. Individual versus neighborhood socioeconomic status and race as predictors of adolescent ambulatory blood pressure and heart rate. *Social Science & Medicine* 2006;63(6):1442-1453.
147. Blakely T, Hunt D, Woodward A. Confounding by socioeconomic position remains after adjusting for neighbourhood deprivation: an example using smoking and mortality. *Journal of Epidemiology and Community Health* 2004;58:1030-1031.
148. Bancroft A, Wiltshire S, Parry O, Amos A. "It's like an addiction first thing... afterwards it's like a habit": daily smoking behaviour among people living in areas of deprivation. *Social Science & Medicine* 2003;56:1261-1267.
149. Thomas F, Bean K, Pannier B, Oppert J-M, Guize L, Benetos A. Cardiovascular Mortality in Overweight Subjects: The Key Role of Associated Risk Factors. *Hypertension* 2005;46(4):654-659.
150. Addor Ve, Wietlisbach V, Narring F, Michaud P-A. Cardiovascular risk factor profiles and their social gradient from adolescence to age 74 in a Swiss region. *Preventive Medicine* 2003;36(2):217-228.
151. Drewnowski A, Specter SE. Poverty and obesity: the role of energy density and energy costs. *American Journal of Clinical Nutrition* 2004;79(1):6-16.
152. Regidor E, Banegas JR, Gutierrez-Fisac JL, Dominguez V, Rodriguez-Artalejo F. Socioeconomic position in childhood and cardiovascular risk factors in older Spanish people. *Int. J. Epidemiol.* 2004;33(4):723-730.
153. Elgar FJ, Roberts C, Moore L, Tudor-Smith C. Sedentary behaviour, physical activity and weight problems in adolescents in Wales. *Public Health* 2005;119(6):518-524.

154. Harrison RA, McElduff P, Edwards R. Planning to win: Health and lifestyles associated with physical activity amongst 15,423 adults. *Public Health* 2006;120(3):206-212.
155. Lindstrom M, Hanson BS, Wirfalt E, Ostergren P-O. Socioeconomic differences in the consumption of vegetables, fruit and fruit juices: The influence of psychosocial factors. *The European Journal of Public Health* 2001;11(1):51-59.
156. Giskes K, Turrell G, Patterson C, Newman B. Socioeconomic differences among Australian adults in consumption of fruit and vegetables and intakes of vitamins A, C and folate. *Journal of Human Nutrition and Dietetics* 2002;15(5):375-385.
157. Hulshof KFAM, Brussaard JH, Kruizinga AG, Telman J, Lowik MRH. Socioeconomic status, dietary intake and 10 y trends: the Dutch National Food Consumption Survey. *European Journal of Clinical Nutrition* 2003;57(1):128-137.
158. Rivara FP, Mueller BA, Somes G, Mendoza CT, Rushforth NB, Kellermann AL. Alcohol and illicit drug abuse and the risk of violent death in the home. *JAMA: The Journal of the American Medical Association* 1997;278(7):569-575.
159. Park H, Sprince NL, Jensen C, Whitten P, Zwerling C. Health risk factors and occupation among Iowa workers. *American Journal of Preventive Medicine* 2001;21(3):203-208.
160. Robbins JM, Vaccarino V, Zhang H, Kasl SV. Socioeconomic status and diagnosed diabetes incidence. *Diabetes Research and Clinical Practice* 2005;68:230-236.
161. Smits J, Westert GP, van den Bos GAM. Socioeconomic status of very small areas and stroke incidence in the Netherlands. *Journal of Epidemiology and Community Health* 2002;56:637-640.
162. Schrijvers CTM, Mackenbach JP, Lutz J-M, Quinn MJ, Coleman MP. Deprivation, stage at diagnosis and cancer survival. *International Journal of Cancer* 1995;63(3):324-329.
163. Schrijvers CTM, Coebergh J-WW, van der Heijden LH, Mackenbach JP. Socioeconomic variation in cancer survival in the Southeastern Netherlands, 1980-1989. *Cancer* 1995;75(12):2946-2953.
164. Petrellia A, Gnavia R, Marinaccia C, Costa G. Socioeconomic inequalities in coronary heart disease in Italy: a multilevel population-based study. *Social Science & Medicine* 2006;63:446-456.
165. Lahelma E, Laaksonen M, Martikainen P, Rahkonen O, Sarlio-Lähteenkorvaa S. Multiple measures of socioeconomic circumstances and common mental disorders. *Social Science & Medicine* 2006;63:1383-1399.
166. Reisine ST, Psoter W. Socioeconomic status and selected behavioural determinants as risk factors for dental caries. *Journal of Dental Education* 2001;65(10):1009-1016.
167. Mielck A, Reitmeir P, Wjst M. Severity of Childhood Asthma by Socioeconomic Status. *Int. J. Epidemiol.* 1996;25(2):388-393.
168. Hasselberg M, Laflamme L, Ringback Weitoft G. Socioeconomic differences in road traffic injuries during childhood and youth: a closer look at different kinds of road user. *Journal of Epidemiology and Community Health* 2001;55(12):858-862.

169. Ancona C, Agabiti N, Forastiere F, Arca M, Fusco D, Ferro S, et al. Coronary artery bypass graft surgery: socioeconomic inequalities in access and in 30 day mortality. A population-based study in Rome, Italy. *Journal of Epidemiology and Community Health* 2000;54(12):930-935.
170. Bowling A. Socioeconomic differentials in mortality among older people. *Journal of Epidemiology and Community Health* 2004;58(6):438-440.
171. Mackenbach JP, Kunst AE, Cavelaars AEJM, Groenhouf F, Geurts JJM. Socioeconomic inequalities in morbidity and mortality in western Europe. *The Lancet* 1997;349(9066):1655-1659.
172. Lahelma E, Martikainen P, Laaksonen M, Aittomäki A. Pathways between socioeconomic determinants of health. *Journal of Epidemiology and Community Health* 2004;58:327-332.
173. Townsend P. *Poverty in the United Kingdom. A survey of household resources and standards of living*. Harmondsworth, England: Penguin Books Ltd., 1979.
174. Maitre B, Nolan B, Whelan CT. Reconfiguring the measurement of deprivation and consistent poverty in Ireland. *Policy Research Series, No. 58*. Dublin: The Economic and Social Research Institute, 2006.
175. Craig J, Driver A. The identification and comparison of small areas of adverse social conditions. *Applied Statistics* 1972;21(1):25-35.
176. Jarman B. Underprivileged areas: validation and distribution of scores. *British Medical Journal* 1984;289:1587-1592.
177. Townsend P, Phillimore P, Beattie A. *Health and deprivation: inequality and the North*. London: Croom Helm, 1988.
178. Carstairs V, Morris R. *Deprivation and health in Scotland*. Aberdeen: Aberdeen University Press, 1991.
179. Thunhurst C. The analysis of small area statistics and planning for health. *Statistics in Health* 1985;34(1):93 - 105.
180. Department of Environment. 1991 Deprivation index - a review of approaches and a matrix of results. London: HMSO, 1995.
181. Howell F, O'Mahony M, Devlin J, O'Reilly O, Buttanshaw C. A geographical distribution of mortality and deprivation. *Irish Medical Journal* 1993;86(3):96-99.
182. Kelly A, Sinclair H. A national deprivation index for health and health service research. Dublin: SAHRU, 1997.
183. Kelly A, Teljeur C. A new national deprivation index for health and health services research. Dublin: SAHRU, 2004.
184. Haase T. Affluence and deprivation: a spatial analysis based on the 1991 census of population. In: Pringle DG, Walsh J, Hennessy M, editors. *Poor people, poor places - a geography of poverty and deprivation in Ireland*. Dublin: Oak Tree Press, 1999.
185. Haase T, Pratschke J. Deprivation and its spatial articulation in the Republic of Ireland: new measures of deprivation based on the Census of Population, 1991, 1996 and 2002. Dublin: ADM, 2005.
186. Testi A, Ivaldi E, Busi A. An index of material deprivation for geographical areas. *Discussion Papers*. Genoa: Università degli Studi di Genova - Facoltà di Economia, 2004.

187. Salmond C, Crampton P. NZDep2001 index of deprivation user's manual. Wellington: Wellington School of Medicine and Health Sciences, 2002.
188. Pampalon R, Raymond G. A deprivation index for health and welfare planning in Quebec. *Chronic Diseases in Canada* 2000;21(3).
189. Benach J, Yasui Y. Geographical patterns of excess mortality in Spain explained by two indices of deprivation. *Journal of Epidemiology and Community Health* 1999;53:423-431.
190. Noble M, Penhale B, Smith G, Wright G, Owen T. Index of deprivation 1999 review: report for formal consultation. Stage 1: domains and indicators. Oxford: University of Oxford, 1999.
191. Noble M, Smith G, Wright G, Dibben C, Lloyd M, Penhale B. Welsh index of multiple deprivation: 2000 edition. Cardiff: The National Assembly for Wales, 2000.
192. Kearns A, Gibb K, Mackay D. Area deprivation in Scotland: a new assessment. *Urban Studies* 2000;37(9):1535-1559.
193. Noble M, Smith G, Wright G, Dibben C, Lloyd M, Shuttleworth I. Measures of deprivation in Northern Ireland. Oxford: Social Disadvantage Research Centre, 2001.
194. Adams J, Ryan V, White M. How accurate are Townsend Deprivation Scores as predictors of self-reported health? A comparison with individual level data. *Journal of Public Health* 2005;27(1):101-106.
195. Lyratzopoulos G, Heller RF, McElduff P, Hanily M, Lewis P. Deprivation and trends in blood pressure, cholesterol, body mass index and smoking among participants of a UK primary care-based cardiovascular risk factor screening programme: both narrowing and widening in cardiovascular risk factor inequalities. *Heart* 2006;92(9):1198-1206.
196. Shohaimi S, Luben R, Wareham N, Day N, Bingham S, Welch A, et al. Residential area deprivation predicts smoking habit independently of individual educational level and occupational social class. A cross sectional study in the Norfolk cohort of the European Investigation into Cancer (EPIC-Norfolk). *Journal of Epidemiology and Community Health* 2003;57(4):270-276.
197. Diez-Roux AV, Nieto FJ, Muntaner C, Tyroler HA, Comstock GW, Shahar E, et al. Neighborhood Environments and Coronary Heart Disease: A Multilevel Analysis. *American Journal of Epidemiology* 1997;146(1):48-63.
198. Payne JN, Coy J, Milner PC, Patterson S. Are deprivation indicators a proxy for morbidity? A comparison of the prevalence of arthritis, depression, dyspepsia, obesity and respiratory symptoms with unemployment rates and Jarman scores. *J Public Health* 1993;15(2):161-170.
199. Shohaimi S, Welch A, Bingham S, Luben R, Day N, Wareham N, et al. Residential area deprivation predicts fruit and vegetable consumption independently of individual educational level and occupational social class: a cross sectional population study in the Norfolk cohort of the European Prospective Investigation into Cancer (EPIC-Norfolk). *Journal of Epidemiology and Community Health* 2004;58:686-691.

200. Ostler K, Thompson C, Kinmonth A-LK, Peveler RC, Stevens L, Stevens A. Influence of socio-economic deprivation on the prevalence and outcome of depression in primary care. *British Journal of Psychiatry* 2001;178:12-17.
201. Saul C, Payne N. How does the prevalence of specific morbidities compare with measures of socioeconomic status at small area level? *J Public Health* 1999;21(3):340-347.
202. Paterson ICM, John G, Adams Jones D. Effect of deprivation on survival of patients with head and neck cancer: a study of 20,131 cases. *Clinical Oncology* 2002;14:455-458.
203. Begum G, Dunn JA, Bryan RT, Bathers S, Wallace DMA. Socio-economic deprivation and survival in bladder cancer. *British Journal of Urology* 2004;94:539-543.
204. Munro AJ, Bentley AHM. Deprivation, comorbidity and survival in a cohort of patients with colorectal cancer. *European Journal of Cancer Care* 2004;13:254-262.
205. Roper NA, Bilous RW, Kelly WF, Unwin NC, Connolly VM. Excess mortality in a population with diabetes and the impact of material deprivation: longitudinal, population based study. *British Medical Journal* 2001;322:1389-1393.
206. Watson JP, Cowen P, Lewis RA. The relationship between asthma admission rates, routes of admission, and socioeconomic deprivation. *Eur Respir J* 1996;9(10):2087-2093.
207. Eachus J, Williams M, Chan P, Davey Smith G, Grainge M, Donovan J, et al. Deprivation and cause specific morbidity: evidence from the Somerset and Avon survey of health. *British Medical Journal* 1996;312:287-292.
208. Uren Z, Fitzpatrick J. Analysis of mortality by deprivation and cause of death. In: Griffiths C, Fitzpatrick J, editors. *Geographic variations in health*. London: The Stationery Office, 2001.
209. Benach J, Yasui Y, Borrell C, Saez M, Pasarín MI. Material deprivation and leading causes of death by gender: evidence from a nationwide small area study. *Journal of Epidemiology and Community Health* 2001;55(4):239-245.
210. Cummins SCJ, McKay L, MacIntyre S. McDonald's Restaurants and Neighborhood Deprivation in Scotland and England. *American Journal of Preventive Medicine* 2005;29(4):308-310.
211. Hart JT. The inverse care law. *The Lancet* 1971;297(7696):405-412.
212. Hyndman JCG, Holman CDAJ. Accessibility and spatial distribution of general practice services in an Australian city by levels of social disadvantage. *Social Science & Medicine* 2001;53(12):1599-1609.
213. Griffiths S, Fone D, Borg A. Bone densitometry: the influence of deprivation on access to care. *Public Health* 2005;119:870-874.
214. Pell JP, Pell ACH, Norrie J, Ford I, Cobbe SM. Effect of socioeconomic deprivation on waiting time for cardiac surgery: retrospective cohort study. *British Medical Journal* 2000;320:15-19.
215. Maheswaran R, Pearson T, Jordan H, Black D. Socioeconomic deprivation, travel distance, location of service, and uptake of breast cancer screening in North Derbyshire, UK. *Journal of Epidemiology and Community Health* 2006;60:208-212.

216. Goddard M, Smith P. Equity of access to health care services: : Theory and evidence from the UK. *Social Science & Medicine* 2001;53(9):1149-1162.
217. Jarman B. Identification of underprivileged areas. *British Medical Journal* 1983;286(6379):1705-1709.
218. Carlisle R, Avery AJ, Marsh P. Primary care teams work harder in deprived areas. *J Public Health* 2002;24(1):43-48.
219. Moore AJ. Deprivation payments in general practice: some spatial issues in resource allocation in the UK. *Health & Place* 1995;1(2):121-125.
220. Connolly C, Chisholm M. The use of indicators for targeting public expenditure: the Index of Local Deprivation. *Environment and Planning C: Government and Policy* 1999;17:463-482.
221. Tunstall R, Lupton R. Is targeting deprived areas an effective means to reach poor people? An assessment of one rationale for area based funding programmes. London: Centre for Analysis of Social Exclusion, 2003.
222. Moore M, Gould P, Keary BS. Global urbanization and impact on health. *International Journal of Hygiene and Environmental Health* 2003;206(4-5):269-278.
223. Maheswaran R, Haining RP, Brindley P, Law J, Pearson T, Fryers PR, et al. Outdoor air pollution and stroke in Sheffield, United Kingdom: a small-area level geographic study. *Stroke* 2005;36:239-243.
224. Scoggins A, Kjellstrom T, Fisher G, Connor J, Gimson N. Spatial analysis of annual air pollution exposure and mortality. *Science of The Total Environment* 2004;321:71-85.
225. Pearce J, Boyle P. Is the urban excess in lung cancer in Scotland explained by patterns of smoking? *Social Science & Medicine* 2005;60(12):2833-2843.
226. Matson U. Indoor and outdoor concentrations of ultrafine particles in some Scandinavian rural and urban areas. *Science of The Total Environment* 2005;343(1-3):169-176.
227. Stewart Fahs PS, Smith BE, Serdar Atav A, Britten MX, Collins MS, Lake Morgan LC, et al. Integrative research review of risk behaviors among adolescents in rural, suburban, and urban areas. *Journal of Adolescent Health* 1999;24(4):230-243.
228. Steams P, Vickers M, Dukinfield W. Cottaging and its associated risks - an urban problem? *Perspectives in Training* 2003;226(1):972-977.
229. Spoth R, Goldberg C, Neppl T, Trudeau L, Ramisetty-Mikler S. Rural-urban differences in the distribution of parent-reported risk factors for substance use among young adolescents. *Journal of Substance Abuse* 2001;13(4):609-623.
230. Levine SB, Coupey SM. Adolescent substance use, sexual behavior, and metropolitan status: is "urban" a risk factor? *Journal of Adolescent Health* 2003;32(5):350-355.
231. Pearce J, Boyle P, Flowerdew R. Predicting smoking behaviour in census output areas across Scotland. *Health & Place* 2003;9(2):139-149.
232. Haynes R, Gale S. Mortality, long-term illness and deprivation in rural and metropolitan wards of England and Wales. *Health & Place* 1999;5:301-312.
233. Levin KA. Urban-rural differences in self-reported limiting long-term illness in Scotland. *Journal of Public Health Medicine* 2003;24(4):295-302.

234. Phillimore P, Reading R. A rural advantage? Urban-rural health differences in Northern England. *Journal of Public Health Medicine* 1992;14(3):290-299.
235. Senior ML, Williams H, Higgs G. Urban-rural mortality differentials: controlling for material deprivation. *Social Science & Medicine* 2000;51:289-305.
236. Judd FK, Jackson HJ, Komiti A, Murray G, Hodgins G, Fraser C. High prevalence disorders in urban and rural communities. *Australian and New Zealand Journal of Psychiatry* 2002;36(1):104-113.
237. Stiernstrom E-L, Holmberg S, Thelin A, Svardsudd K. A prospective study of morbidity and mortality rates among farmers and rural and urban nonfarmers. *Journal of Clinical Epidemiology* 2001;54(2):121-126.
238. Boland M, Staines A, Fitzpatrick P, Scallan E. Urban-rural variation in mortality and hospital admission rates for unintentional injury in Ireland. *Inj Prev* 2005;11(1):38-42.
239. Verheij RA. Explaining urban-rural variations in health: A review of interactions between individual and environment. *Social Science & Medicine* 1996;42(6):923-935.
240. Boyle P, Norman P, Rees P. Does migration exaggerate the relationship between deprivation and limiting long-term illness? A Scottish analysis. *Social Science & Medicine* 2002;55(1):21-31.
241. Verheij RA, van de Mheen HD, de Bakker DH, Groenewegen PP, Mackenbach JP. Urban-rural variations in health in The Netherlands: does selective migration play a part? *J Epidemiol Community Health* 1998;52(8):487-493.
242. Pacione M. The geography of deprivation in Scotland. *Transactions of the Institute of British Geographers* 1995;20(2):173-192.
243. Woodward R. "Deprivation" and "the Rural": an investigation into contradictory discourses. *Journal of Rural Studies* 1996;12(1):55-67.
244. Milbourne P. The local geographies of poverty: a rural case-study. *Geoforum* 2004;35(5):559-575.
245. Cloke P, Goodwin M, Milbourne P, Thomas C. Deprivation, poverty and marginalization in rural lifestyles in England and Wales. *Journal of Rural Studies* 1995;11(4):351-365.
246. The Countryside Agency. Indicators of rural disadvantage: guidance note. Wetherby, UK: The Countryside Agency, 2003.
247. Noble M, Wright G. Identifying poverty in rural England. *Policy & Politics* 2000;28(3):293-308.
248. Nolan B, Whelan CT, Williams J. *Where are poor households? - the spatial distribution of poverty and deprivation in Ireland*. Dublin: Oak Tree Press & Combat Poverty Agency, 1998.
249. Commins P. Poverty and social exclusion in rural areas: characteristics, processes and research issues. *Sociologica Ruralis* 2004;44(1):60-75.
250. McDonagh J. Transport policy instruments and transport-related social exclusion in rural Republic of Ireland. *Journal of Transport Geography* 2006;14:355-366.
251. Yantzi N, Rosenberg MW, Burke SO, Harrison MB. The impacts of distance to hospital on families with a child with a chronic condition. *Social Science & Medicine* 2001;52(12):1777-1791.



252. Jones AP, Bentham G, Horwell C. Health service accessibility and deaths from asthma. *Int. J. Epidemiol.* 1999;28(1):101-105.
253. Panelli R, Gallagher L, Kearns R. Access to rural health services: research as community action and policy critique. *Social Science & Medicine* 2006;62:1103-1114.
254. Lovett A, Haynes R, Sunnenberg G, Gale S. Car travel time and accessibility by bus to general practitioner services: a study using patient registers and GIS. *Social Science & Medicine* 2002;55(1):97-111.
255. Hartshorn T. *Interpreting the city: an urban geography*. 2nd ed. Canada: John Wiley & Sons, 1992.
256. Long JF, Rain DR, Ratcliffe MR. Population density vs. urban population: comparative GIS studies in China, India, and the United States. "Population Applications of Spatial Analysis Systems (SIS)", *IUSSP Conference*. Salvador, Brazil, 2001.
257. Goodall CR, Kafadar K, Tukey JW. Computing and using rural versus urban measures in statistical applications. *The American Statistician* 1998;52(2):101-111.
258. Cloke P, Edwards G. Rurality in England and Wales 1981: a replication of the 1971 index. *Regional Studies* 1986;20(4):289-306.
259. Pitblado JR, Pong RW. Geographic distribution of physicians in Canada. Ontario, Canada: Centre for Rural and Northern Health Research, Laurentian University, 1999.
260. Central Statistics Office. *Census 2002 - Vol. 1: population classified by area*. Dublin: Stationery Office, 2003.
261. Bartley M, Blane D. Commentary: appropriateness of deprivation indices must be ensured. *British Medical Journal* 1994;309:1479.
262. Gordon D. Census based deprivation indices: their weighting and validation. *Journal of Epidemiology and Community Health* 1995;49(Suppl. 2):S39-S44.
263. Osborne J. Notes on the use of data transformations. *Practical Assessment, Research & Evaluation* 2002;8(6).
264. Noble M, Wright G, Dibben C, Smith G, McLennan D, Anttila C, et al. The English indices of deprivation 2004 (revised). London: Office of the Deputy Prime Minister, 2004.
265. Longford NT. Comments on shrinkage. *Royal Statistical Society Meeting*. London, 2001.
266. O'Sullivan D, Unwin DJ. *Geographic Information Analysis*. New Jersey: John Wiley & Sons, Inc., 2003.
267. Christie SML, Fone D. Does car ownership reflect socio-economic disadvantage in rural areas? A cross-sectional geographical study in Wales, UK. *Public Health* 2003;117:112-116.
268. Pacione M. The geography of disadvantage in rural Scotland. *Tijdschrift voor Economische en Sociale Geografie* 2004;95:375-391.
269. Neylon M, Kirby B. The people of Clare 1991 - 2002: a community in transition. Ennis, Ireland: Clare County Development Board, 2006.
270. CSO. *Census 2002 - Vol. 1: population classified by area*. Dublin: Stationery Office, 2003.

271. Bibby P, Shepherd J. Developing a new classification of urban and rural areas - the methodology. London: Office for National Statistics, 2004.
272. Harris B. Accessibility: concepts and applications. *Journal of Transportation and Statistics* 2001;4(2/3):15-30.
273. Cloke PJ. An index of rurality for England and Wales. *Regional Studies* 1977;11:31-46.
274. McDade TW, Adair LS. Defining the "urban" in urbanization and health: a factor analysis approach. *Social Science & Medicine* 2001;53(1):55-70.
275. Bailey TC, Gatrell AC. *Interactive spatial data analysis*. Essex, England: Longman Scientific Ltd., 1995.
276. Greco S, Matarazzo B, Slowinski R. Multicriteria classification. In: Klösgen W, Zytkow J, editors. *Handbook of data mining and knowledge discovery*. Oxford: Oxford University Press, 2002:318-328.
277. Feinstein AR. *Multivariable analysis*. London: Yale University Press, 1996.
278. Gnanadesikan R. *Methods for statistical data analysis of multivariate observations*. New York: John Wiley & Sons, Inc., 1997.
279. Hand DJ. *Construction and assessment of classification rules*. Chichester, England: John Wiley & Sons Ltd., 1997.
280. JMP 5.0.1a [program]: SAS Institute Inc., 2002.
281. S-Plus Professional Edition 6.2 for Windows [program]: Insightful Corp., 2003.
282. Kohonen T. *Self-organizing maps*. Berlin: Springer, 2001.
283. Borgatti SP. How to explain hierarchical clustering. *Connections* 1994;17(2):78-80.
284. Armstrong MP, Xiao N, Bennett DA. Using genetic algorithms to create multicriteria class intervals for choropleth maps. *Annals of the Association of American Geographers* 2003;93(3):595-623.
285. Osaragi T. Classification methods for spatial data representation. *CASA Working Paper Series*. London: University College London, 2002.
286. Billah MB, Hyndman RJ, Koehler AB. Empirical information criteria for time series forecasting model selection. *Monash Econometrics and Business Statistics Working Papers*: Monash University, 2003.
287. European Environment Agency. Corine land cover 2000 (CLC2000) 100 m - version 8/2005: EEA, 2005.
288. Townsend P. Deprivation. *Journal of Social Policy* 1987;16:125-146.
289. Haynes R, Gale S, Lovett A, Bentham G. Unemployment rate as an updatable health needs indicator for small areas. *J Public Health* 1996;18(1):27-32.
290. Noble M, Penhale B, Smith G, Wright G, Dibben C, Lloyd M. Response to the formal consultations on the Indices of Deprivation 2000 (ID 2000). London: Department of Environment, Transport and the Regions, 2000.
291. Longford NT. Multivariate shrinkage estimation of small area means and proportions. *Journal of Royal Statistical Society, Series A* 1999;162(2):227-245.
292. Openshaw S. *The modifiable areal unit problem*. Norwich: Geo, 1984.
293. McConnachie A, Weir C. Evaluation of statistical techniques in the Scottish index of multiple deprivation. Glasgow: University of Glasgow, 2005.
294. Trewin D. Census of Population and Housing: Socio-Economic Indexes for Area's (SEIFA). Canberra: Australian Bureau of Statistics, 2004.

295. Klasen S. Measuring poverty and deprivation in South Africa. *Review of Income and Wealth* 2000;46(1):33-58.
296. Eibner C, Sturm R. US-based indices of area-level deprivation: Results from HealthCare for Communities. *Social Science & Medicine* 2006;62:348-359.
297. Jolliffe IT. *Principal Components Analysis*. 2nd ed. New York: Springer-Verlag, 2002.
298. Gibb K, Kearns A, Keoghan M, Mackay D, Turock I. Revising the Scottish area deprivation index: Volume 1. Edinburgh: The Scottish Office Central Research Unit, 1998.
299. Scottish Executive. Scottish index of multiple deprivation 2004: summary technical report. Edinburgh: Scottish Executive, 2005.
300. Local Government Data Unit. Welsh index of multiple deprivation 2005: technical report. Cardiff: Welsh Assembly Government, 2005.
301. Northern Ireland Statistics & Research Agency. Northern Ireland multiple deprivation measure 2005. Belfast: The Stationery Office, 2005.
302. McIntyre D, Muirhead D, Gilson L, Govender V, Mbatsha S, Goudge J, et al. Geographic patterns of deprivation and health inequities in South Africa: informing public resource allocation strategies. Cape Town: SADC EQUINET/IDRC & TDR/ICHSRI, 2000.
303. Noble M, Babita M, Barnes H, Dibben C, Magasela W, Noble S, et al. The provincial indices of multiple deprivation for South Africa 2001. Oxford: University of Oxford, 2006.
304. Messer L, Vinikoor L, Kaufman JS, Laraia BA. Sociodemographic deprivation domains and preterm birth. *Population Association of America 2006 Annual Meeting Program*. Los Angeles, 2006.
305. Odoi A, Wray R, Emo M, Birch S, Hutchinson B, Eyles J, et al. Inequalities in neighbourhood socioeconomic characteristics: potential evidence-base for neighbourhood health planning. *International Journal of Health Geographics* 2005;4(20).
306. Velicer WF, Jackson DN. Component analysis versus common factor analysis: some issues in selecting an appropriate procedure. *Multivariate Behavioral Research* 1990;25(1):1-28.
307. Bentler PM, Kano Y. On the equivalence of factors and components. *Multivariate Behavioral Research* 1990;25(1):67-74.
308. Chatfield C, Collins AJ. *Introduction to multivariate analysis*. London: Chapman and Hall Ltd., 1980.
309. Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* 2001;29(2):295-327.
310. De la Torre F, Black MJ. Robust principal component analysis for computer vision. *International Conference on Computer Vision*. Vancouver, Canada, 2001.
311. Warne K, Prasad G, Rezvani S, Maguire L. Statistical and computational intelligence techniques for inferential model development: a comparative evaluation and a novel proposition for fusion. *Engineering Applications of Artificial Intelligence* 2004;17:871-885.
312. Zwick WR, Velicer WF. Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research* 1982;17(2):253-269.

313. Velicer WF. Determining the number of components from the matrix of partial correlations. *Psychometrika* 1976;41:321-327.
314. Hayton JC, Allen DG, Scarpello V. Factor retention decisions in exploratory factor analysis: a tutorial on parallel analysis. *Organizational Research Methods* 2004;7(2):191-205.
315. Longman RS, Cota AA, Holden RR, Fekken GC. A regression equation for the parallel analysis criterion in principal components analysis: mean and 95th percentile eigenvalues. *Multivariate Behavioral Research* 1989;24(1):59-69.
316. Lautenschlager GJ. A comparison of alternatives to conducting Monte Carlo analyses for determining parallel analysis criteria. *Multivariate Behavioral Research* 1989;24(3):365-395.
317. Salmond C, Crampton P. Heterogeneity of deprivation within very small areas. *Journal of Epidemiology and Community Health* 2002;56:669-670.
318. Pringle DG. The geographical distribution of poverty in Ireland. *Rural Development Conference*. Tullamore, 2002.
319. Anselin L. Local Indicators of Spatial Autocorrelation - LISA. *Geographical Analysis* 1995;27:93-115.
320. Sammon M. dublincrime.com, <http://www.dublincrime.com>.
321. Pearce N. The ecological fallacy strikes back. *Journal of Epidemiology and Community Health* 2000;54:326-327.
322. Lancaster G, Green M. Deprivation, ill-health and the ecological fallacy. *Journal of Royal Statistical Society, Series A* 2002;165(2):263-278.
323. Hewson P. Deprived children or deprived neighbourhoods? A public health approach to the investigation of links between deprivation and injury risk with specific reference to child road safety in Devon County, UK. *BMC Public Health* 2004;4(15).
324. Salway R, Wakefield J. Sources of bias in ecological studies of non-rare events. *Environmental and Ecological Statistics* 2005;12:321-347.
325. Waldron G. Accidents at home are no more likely in deprived areas. *British Medical Journal* 2000;320:1276.
326. Ben-Shlomo Y, Davey Smith G. Commentary: socioeconomic position should be measured accurately. *British Medical Journal* 1999;318:844-845.
327. MacRae K. Commentary: socioeconomic deprivation and health and the ecological fallacy. *British Medical Journal* 1994;309:1478-1479.
328. Messner SF, Anselin L. Spatial analysis of homicide with areal data. In: Goodchild MF, Janelle DG, editors. *CSISS Best Practice Publications: Spatially Integrated Social Science*. Oxford: Oxford University Press, 2004.
329. Fiedler R, Hyndman J, Schuurman N. Locating spatially concentrated risk of homelessness amongst recent immigrants in Greater Vancouver: a GIS-based approach. *Research on Immigration and Integration in the Metropolis: Working Paper Series No. 06-10*. Vancouver: Vancouver Centre of Excellence, 2006.
330. Jones M, Ramsay J, Feder G, Crook AM, Hemingway H. Influence of practices' ethnicity and deprivation on access to angiography: an ecological study. *British Journal of General Practice* 2004;54:423-428.

331. Lyratzopoulos G, Havely D, Gemmell I, Cook GA. Factors influencing emergency medical readmission risk in a UK district general hospital: a prospective study. *BMC Public Health* 2005;5(1).
332. Central Statistics Office. Census 96 - Volume 7: occupations. Dublin: Stationery Office, 1998.
333. Luke BT. Simulated annealing cooling schedules, <http://members.aol.com/btluke/simanf1.htm>.
334. Goovaerts P, Jacquez G. Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics* 2004;3(1):14.
335. Microsoft Visual Basic .NET [program]: Microsoft Corporation, 2001.
336. Croux C, Haesbroeck G. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* 2000;87(3):603-618.
337. Devlin SJ, Gnanadesikan R, Kettenring JR. Robust estimation and outlier detection with correlation coefficients. *Biometrika* 1975;62(3):531-545.
338. Rousseeuw PJ, Vanden Branden K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999;41(3):212-223.
339. Engelen S, Hubert M, Vanden Branden K. A comparison of three procedure for robust PCA in high dimensions. *Austrian Journal of Statistics* 2005;34(2):117-126.
340. Li G, Chen Z. Projection-pursuit approach to robust dispersion matrices and principal components analysis: primary theory and Monte Carlo. *Journal of the American Statistical Association* 1985;80(391):759-766.
341. Yang T-N, Wang S-D. Robust algorithms for principal component analysis. *Pattern Recognition Letters* 1999;20:927-933.
342. Skočaj D, Bischof H, Leonardis A. A robust PCA algorithm for building representations from panoramic pictures. *European Conference on Computer Vision*. Copenhagen, 2002.
343. Critchley F. Influence in principal components analysis. *Biometrika* 1985;72(3):627-636.
344. Croux C, Haesbroeck G. Empirical influence functions for robust principal components. *American Statistical Association Proceedings of the Statistical Computing Section*, 1999:201-206.
345. Brooks SP. Diagnostics for principal components: influence functions as diagnostic tools. *The Statistician* 1994;43(4):483-494.
346. Field K. Measuring the need for primary health care: an index of relative disadvantage. *Applied Geography* 2000;20:305-332.
347. Mackenzie IF, Nelder R, Maconachie M, Radford G. 'My ward is more deprived than yours'. *J Public Health* 1998;20(2):186-190.
348. Noble M, Firth D, Dibben C, Lloyd M, Smith G, Wright G. Meeting on Indices of Deprivation 2000 organised by Jane Galbraith and Colin Chalmers: Statement from the Oxford team, 2001.
349. Chalmers C. ID2000: can factor analysis solve the problem of combining indicators. *Royal Statistical Society Meeting*. London, 2001.

350. Department of Health and Children. Health Information - A National Strategy. Dublin: The Stationery Office, 2004.
351. Kawachi I, Kennedy BP. The relationship of income inequality to mortality: Does the choice of indicator matter? *Social Science & Medicine* 1997;45(7):1121-1127.
352. Cox J. Poverty in rural areas. *British Medical Journal* 1998;316:722-730.
353. Åberg Yngwe M, Fritzell J, Lundberg O, Diderichsen F, Burstrom B. Exploring relative deprivation: Is social comparison a mechanism in the relation between income and health? *Social Science & Medicine* 2003;57(8):1463-1473.
354. Law MR, Morris JK. Why is mortality higher in poorer areas and in more northern areas of England and Wales? *Journal of Epidemiology and Community Health* 1998;52(6):344-352.
355. Tobler W. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 1970;46(Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods.):234-240.