UNIVERSITY OF BALLARAT

ESTIMATION OF POPULATION USING SATELLITE IMAGERY

Jack Harvey

SCHOOL OF INFORMATION TECHNOLOGY
AND MATHEMATICAL SCIENCES

March 1999

ESTIMATION OF POPULATION USING SATELLITE IMAGERY

by

Jack Harvey B.Sc. Dip.Ed.

Submitted in total fulfillment of the degree of

Doctor of Philosophy

School of Information Technology and Mathematical Sciences
University of Ballarat
P.O. Box 663
Gear Avenue, Mount Helen,
Ballarat, Victoria 3353,
Australia

March 1999

STATEMENT OF AUTHORSHIP

Except where explicit reference is made in the text of the thesis, this thesis contains no material published elsewhere or extracted in whole or in part from a thesis by which I have qualified for or been awarded another degree or diploma. No other person's work has been relied upon or used without due acknowledgment in the main text and bibliography of the thesis.

Jack Harvey
29/3/99

Interviewer: “But isn’t it …. difficult?”
Sir Arthur Streeb-Greebling: “I think your choice of the word ‘difficult’ is an extremely good one…..”

“Teaching Ravens to Fly Under Water”
Not Only… But Also…
Peter Cook & Dudley Moore

Abstract

Whilst population counts cannot be obtained directly from remote sensing imagery, human habitation is associated with recognisable physical characteristics. Visual interpretation of aerial photographs has been employed in the past for population estimation, but with the advent of digital imagery it has become possible in principle to devise mathematical models based on spectral reflectances which can directly provide small-area population estimates in a timely, flexible, and automated fashion. Calibration or training of such algorithms requires "ground truth" population data. Because this is generally available only at census enumeration district (ED) level and above, research to date in this area has generally utilised regression models for estimating ED populations on the basis of spatially aggregated reflectance data. Two extensions to this methodology have been explored, utilising 6-band Landsat Thematic Mapper (TM) imagery.

Firstly, the ED aggregate approach was extended to include a range of spectral and textural transformations which might provide improved surrogate indicators of population, and by the use of non-linear functional forms in the regression models.

Secondly, an alternative approach was developed for estimating population at the level of individual pixels. Pixels were first classified as residential or non-residential by the method of maximum likelihood. Ground truth populations for residential pixels were imputed from census counts for EDs, and these were regressed on the various remote sensing measures – both the basic bands and selected spectral and spatial transformations. An expectation-maximisation (EM) approach was incorporated to iteratively refine the imputed pixel populations.

Model identification and development was based on an image centred on the Australian provincial city of Ballarat. This was followed by three phases of refinement and validation. Firstly, a number of candidate procedures were tested on a second image centred on the neighbouring provincial city of Geelong, as a result of which the preferred procedure was chosen and a number of refinements were made, including a contextual reclassification phase. Secondly, the Ballarat-trained regression model was applied, in normalised form, to several further Australian images, including large metropolitan areas with populations of up to 3 million. The model was found to be moderately robust, for urban areas, to geographical and temporal differences in season and climate, but less robust to differences in intensity of urbanisation. Finally, methods were developed for local training of an estimation equation on a small sample of population data from within an image, and for tuning the training sample to minimise estimation bias. Coefficients of determination for population density of individual

EDs were typically in the range .80-.90 for training samples and .70-.80 for full image validation. Total urban populations for a number of images were estimated with errors in the range –2% to +4%. In urban areas of moderate density, mean absolute proportional errors for individual EDs were typically in the range 15-20%.

It was concluded on a number of grounds that the pixel-based estimation procedure is preferable to aggregate-based procedures. A generic modelling framework for population estimation from TM imagery is specified, which is proposed as the basis of a feasible operational procedure. The limitations of the methodology at both high and low extremes of population density are considered, and procedures outlined for the incorporation of ancillary information from other sources. A number of avenues for potential further investigation are outlined, together with some potential applications.

Acknowledgements

Many people have supported and assisted me during the long and interrupted life of this project. My thanks go firstly to my supervisor Professor Binh Pham for her encouragement and timely suggestions, and for helping me bring the task to fruition. Thanks to my first supervisor, Dr Joseph Leach, of the then Ballarat University College Remote Sensing Unit, for introducing me to remote sensing and image analysis, and to Professor Dennis Else for his advice and assistance during the early stages. Thanks also to Associate Professor Bruce Forster of the Centre for Remote Sensing, University of New South Wales, for his assistance in focussing my early ideas. Thanks also to my statistical and mathematical colleagues in the School of Information Technology and Mathematical Sciences and its predecessors at the University of Ballarat, and at the Ballarat and Western Victoria Regional Information Bureau; particularly Lyn Roberts, Associate Professor Gerry Anderson, Shani Clark, Dr Barney Glover, Dr Ewan Barker, David Stratton, Greg Simmons, Cilinda Atkins and especially toward the end of the project John Yearwood, for listening and for making helpful observations; to my respected mentor and ex-colleague Jim Snow for his thoughtful comments; to Professor Niels Becker and Professor Mike Titterington for advice regarding the EM algorithm; to my computing colleagues and technical staff, especially Associate Professor Paul Kelly, Geoff Boyd, Gary Walker, Stephen Bell, Colin Linehan, Peter Cowan, John Fogarty and Andrew Segrott for computing life-support; to Patrick Halewood, of the BUC Remote Sensing Unit, and to Matt Gibson and Miranda Kerr of the UB Centre for Environmental Management for advice about technical problems of mutual concern; to Dianne Elshaug, for countless resets of a troublesome first generation network router; to Peter Brushfield, Christine Silvestroni and staff of the City of Greater Geelong and the Victorian Office of Housing, for information about the demography of Geelong; to the Australian Bureau of Statistics and the Public Sector Mapping Agencies for census data and associated geographical information; to Bob Walker of Geoimage P/L and to the Australian Centre for Remote Sensing for the provision of TM images; to Maxine Kingston, Debra Walker and Kaye Lewis for their patient and unflinching response to the ubiquitous red correction pen; and especially my heartfelt gratitude to Meredith, Georgia, Tom and Sam Harvey for their loving support and their bemused acceptance of the idiosyncratic working hours, the long Archimedean showers, and the frequently preoccupied air of their husband and father.

Jack Harvey
29/3/99

Contents

Statement of Authorship ii

Abstract iv

Acknowledgments vi

List of Tables xiii

List of Figures xvi

List of Landsat TM Images xviii

List of Abbreviations xx

1. The Use of Remote Sensing for Population Estimation 1

1.1 INTRODUCTION 1

1.2 HISTORICAL OVERVIEW 2
- 1.2.1 Dwelling Counts 2
- 1.2.2 Measured urban area 3
- 1.2.3 Measured land use areas 3
- 1.2.4 Spectral radiance characteristics of individual pixels 4

1.3 CONCEPTUAL FRAMEWORK 7
- 1.3.1 Sensor characteristics 7
- 1.3.2 Approaches to spatial aggregation 7
- 1.3.3 Form and complexity of models 8

1.4 RESEARCH QUESTIONS 13

1.5 OUTLINE OF THE RESEARCH AND THE THESIS 14

2. Theoretical and Methodological Bases for the Research 16

2.1 INTRODUCTION 16

2.2 LANDSAT THEMATIC MAPPER MULTISPECTRAL RADIANCE DATA 16

2.3 SPATIAL DOMAIN AND SPECTRAL DOMAIN DATA REPRESENTATIONS 18
- 2.3.1 Colour representation 18
- 2.3.2 Spatial and spectral displays 19
- 2.3.3 Data transformations 19

2.4 SPECTRAL DOMAIN TRANSFORMATIONS 20
- 2.4.1 Single band transformations 20
- 2.4.2 Single-valued band comparisons or indices 20
- 2.4.3 Many-to-many transformations 21
- 2.4.4 The principal components (PC) transformation 21
- 2.4.5 Hue-saturation-intensity (HSI) transformations. 24

2.5 SPATIAL DOMAIN TRANSFORMATIONS 25

2.5.1 Aggregate measures 25
2.5.2 Neighbourhood measures (spatial filters) 26
2.5.3 Measures of image texture 26
2.5.4 Development of a texture index based on pairwise differences 29

2.6 SUPERVISED CLASSIFICATION 30
2.6.1 Maximum likelihood classification (MLC) 30

2.7 SELECTION OF CLASSIFIERS 33
2.7.1 Linear discriminant analysis 33
2.7.2 Stepwise discriminant analysis 34
2.7.3 Relationship between DA and MLC 35

2.8 LINEAR MODELS 35
2.8.1 Multiple linear regression 35
2.8.2 Stepwise regression 38
2.8.3 Regression through the origin 39
2.8.4 Statistical assumptions, variable transformations and alternative models 39

2.9 REGRESSION ANALYSIS WITH INCOMPLETELY DETERMINED DATA 42
2.9.1 An algorithm for iterative refinement of estimates 42
2.9.2 Models with a transformed dependent variable 45

2.10 NON-LINEAR METHODS 45
2.10.1 Interactive effects 45
2.10.2 Rule-based methods 47

2.11 MEASURES OF PERFORMANCE, VALIDITY AND ROBUSTNESS 47
2.11.1 Internal validity, external validity and robustness 47
2.11.2 Indicators of internal validity 48
2.11.3 Sample size considerations 48
2.11.4 Sample-dependence of stepwise procedures: external validation procedures 49
2.11.5 Multicollinearity 50
2.11.6 Type I error rates in stepwise analyses 51
2.11.7 Modified performance measures with transformed dependent variables 52

2.12 ISSUES OF PARAMETRISATION, PRESENTATION AND EVALUATION 53
2.12.1 Population density vs. total population 53
2.12.2 Presentation and evaluation 53

2.13 SUMMARY 56

3. Data Preparation and Integration 58

3.1 INTRODUCTION 58

3.2 THE STUDY AREAS 58
3.2.1 The primary study area 59
3.2.2 The secondary study area 61
3.2.3 Further study areas 62

3.3 POPULATION AND RELATED DATA 62
3.3.1 Population estimates 62
3.3.2 Dwelling count estimates 63
3.3.3 Census collection district boundaries 64

3.4 LANDSAT THEMATIC MAPPER DATA 64

3.5 COMPUTING METHODS 65

3.6 ESTABLISHING GROUND TRUTH POPULATION DATA 66

3.7 RADIOMETRIC AND GEOMETRIC CORRECTION OF THE IMAGE 67
3.7.1 Radiometric Correction 67
3.7.2 Approaches to rectification and registration 68
3.7.3 Sources of geometric distortion 69
3.7.4 Methods of co-registration 71
3.7.5 Explicit Parametric Transformations 72
3.7.6 Warping Polynomials 73

3.8 SUMMARY 76

4. Estimates Based On Census Collection District Aggregate Measures 77

4.1 INTRODUCTION 77

4.2 POPULATION DENSITY ESTIMATES FROM AVERAGE REFLECTANCES 78

4.3 POPULATION DENSITY ESTIMATES: MEASURES OF SPATIAL VARIABILITY 82

4.4 POPULATION DENSITY ESTIMATES: SPECTRAL TRANSFORMATIONS 84

4.5 DWELLING DENSITY ESTIMATES 86

4.6 EVALUATION OF REGRESSION MODELS BASED ON CENSUS COLLECTION DISTRICT AGGREGATES 86

4.7 SUMMARY 96

5. Estimates Based on Individual Pixels 97

5.1 INTRODUCTION 97

5.2 A BASIC PIXEL-BASED REGRESSION MODEL 98

5.3 CLASSIFICATION OF PIXELS FOLLOWED BY REGRESSION 99
5.3.1 Supervised land use classification: categories and training sets 99
5.3.2 Initial classification based on the 6 TM bands 100
5.3.4 Application to the full image: population density estimates 101
5.3.5 CD population estimates: preliminary evaluation of the model 101
5.3.6 Strategies for improving the model 101

5.4 DATA TRANSFORMATIONS 102
5.4.1 Spectral domain transformations 102
5.4.2 Preliminary screening of potential discriminators/predictors 103
5.4.3 Spatial domain transformations 105
5.4.4 Provision for non-linearity and interactions 106

5.5 CLASSIFICATION USING TRANSFORMED EXPLANATORY VARIABLES 106
5.5.1 Selection of classification variables 106
5.5.2 Classification of the image 109

5.6 REGRESSION USING TRANSFORMED EXPLANATORY VARIABLES 110
5.6.1 Selection of regression models 110
5.6.2 Application to the full image: population density estimates 111
5.6.3 CD population estimates 111

5.7 ITERATIVE REFINEMENT OF THE REGRESSION MODELS 111
5.7.1 Application of the iterative procedure 111

5.7.2 Application of iterated models to the full image: population density estimates 113
5.7.3 Some technical issues 113

5.8 CD POPULATION AND POPULATION DENSITY ESTIMATES: COMPARISON OF THE MODELS 114

5.9 TRANSFORMATIONS OF THE DEPENDENT VARIABLE 121

5.10 THE FORM OF THE MODEL 125

5.11 SUMMARY 127

6. Application of Estimation Algorithms to a Second Geographical Area 128

6.1 INTRODUCTION 128

6.2 ESTIMATION BASED ON CENSUS COLLECTION DISTRICT AGGREGATES 129

6.3 ESTIMATION BASED ON INDIVIDUAL PIXELS 140

6.4 MODIFICATIONS TO THE PIXEL-BASED ESTIMATION PROCEDURE 146
6.4.1 A proposed explanation of scaling error 146
6.4.2 Local classification of the secondary study area 148
6.4.3 Adjustments for under-estimation and over-estimation 150
6.4.4 CD-based adjustment for multiple dwelling structures 150
6.4.5 CD-based adjustment for inflated counts of residential pixels in rural areas 154
6.4.6 Pixel-based high and low density adjustments 155
6.4.7 Examination of remaining discrepancies 158

6.5 SUMMARY 164

7. The Iterative Re-Estimation Algorithm 166

7.1 INTRODUCTION 166

7.2 THE RELATIONSHIP OF THE ITERATIVE RE-ESTIMATION ALGORITHM TO THE EM ALGORITHM 166
7.2.1 The EM algorithm 166
7.2.2 The iterative re-estimation algorithm 168
7.2.3 Relationship to the EM algorithm 170
7.2.4 Conclusion 173

7.3 SAMPLING VARIATION 174

7.4 SIMULATION 176
7.4.1 Simulated populations 176
7.4.2 Monte Carlo sampling from the simulated populations 177
7.4.3 Assessment criteria 178
7.4.4 Interpretation and conclusions 180

7.5 SUMMARY 191

8. Normalised Population Estimation Models 192

8.1 INTRODUCTION 192

8.2 THE SUPPLEMENTARY TEST AREAS 193
8.2.1 Characteristics of the study areas and images 193
8.2.2 Classification of the images 193
8.2.3 Spectral characteristics of the test images 200

8.3 ADJUSTMENT OF THE ESTIMATION EQUATION 200

8.3.1 Preliminary adjustments 200
8.3.2 Normalising adjustments 202

8.4 APPLICATION OF NORMALISED MODELS 204
8.4.1 Methodology 204
8.4.2 Results 205

8.5 SUMMARY 215

9. Local Training of The Population Estimation Equation 216

9.1 INTRODUCTION 216

9.2 SELECTION OF REGRESSION TRAINING SAMPLES 216

9.3 APPLICATION OF LOCAL REGRESSION TRAINING 217
9.3.1 Analysis of the samples 217
9.3.2 Results 220
9.3.3 Adaptive adjustments to training sets 230
9.3.4 Estimates for Statistical Local Areas 234

9.4 CHARACTERISTICS OF THE ESTIMATION EQUATIONS 237

9.5 SUMMARY 238

10. Conclusions and Recommendations: Towards a Feasible Operational Methodology for Population Estimation from Landsat TM Imagery 240

10.1 SUMMARY OF THE STUDY 240

10.2 CONCLUSIONS 242

10.3 ADVANTAGES OF PIXEL-BASED ESTIMATION 243

10.4 RESULTS, OUTCOMES AND PERFORMANCE 245

10.5 THE MODEL AND ITS IMPLEMENTATION 249
10.5.1 The basic model and procedure 249
10.5.2 Enhancements: adjustment for anomalies 251

10.6 DIRECTIONS FOR FURTHER RESEARCH 252
10.6.1 Improving estimation at low population densities 252
10.6.2 Improving estimation at high population densities 253
10.6.3 Other aspects 254

10.7 APPLICATIONS 255
10.7.1 Direct use of the methodology: estimation of population 255
10.7.2 Indirect use of the methodology: hybrid methodologies 255
10.7.3 Back to Earth 256

Landsat TM Images 257

Appendices 263

A. Transformations from RGB to HSI co-ordinates 264

B. Development and Comparative Evaluation of Texture Measures by Simulation 265

C. Implementation of Procedures for Chapters 3 - 6 271

D. Collection District resident population and dwelling count estimates as at 14/2/88 276

E. Collection District Aggregate-based Methods: Primary and Secondary Study Area Descriptive Statistics and Model Diagnostics 281

F. Selected Results from Exploratory Discriminant Analysis and Regression Analysis on Samples from the Primary Image 290

G. Sampling Variation in Results of Iterative Re-estimation 292

H. Summary Statistics for the Distributions of CD Population and CD Population Density in the Supplementary Study Areas 293

I. Landuse/Landcover Classes in Supplementary Images 295

J. Normalisation Formulae 296

K. Summary of Least Accurately Estimated CDs in Adelaide Image 297

References **298**

List of Tables

2.1	Landsat TM spectral bands	16
2.2	Scope of data sets	50
2.3	Summary of performance measures	50
3.1	Study Areas	59
3.2	Statistical structure of Ballarat Statistical District	61
3.3	Statistical structure of Geelong Statistical District	61
3.4	Specifications of supplementary images	65
3.5	Comparison of warping polynomials	75
4.1	Summary of regression variables	80
4.2	Population density models based on Census Collection District means	80
4.3	Variable suffix nomenclature	82
4.4	Population density models based on Census Collection District means and spatial variation measures	83
4.5	Selected spectral transformations	84
4.6	Population density models based on Census Collection District means and spatial variation of selected spectral transformations	85
4.7	Dwelling density models based on Census Collection District means	87
4.8	Dwelling density models based on Census Collection District means and spatial variation measures	87
4.9	Dwelling density models based on Census Collection District means and spatial variation of selected spectral transformations	88
4.10	Summary of selected models for population density based on TM data aggregated over Census Collection Districts	91
4.11	Summary of selected models for dwelling density based on TM data aggregated over Census Collection Districts	92
4.12	Summary of estimated Census Collection District populations based on TM data aggregated over Census Collection Districts	93
4.13	Summary of estimated Census Collection District dwelling counts based on TM data aggregated over Census Collection Districts	94
5.1	Categories of land use and land cover	100
5.2	Summary of spectral domain transformations	102
5.3	Selected spectral transformations	104
5.4	Summary of stepwise discriminant analyses	107

5.5	Summary of the selected sequence of discriminant analyses	109
5.6	Summary of stepwise regression analysis on pixels classified as residential	110
5.7	Iterative refinement of a regression model for pixel population based on the 6 TM bands	112
5.8	Coefficients of determination for four regression models for pixel population: with and without iterative refinement	113
5.9	Summary of selected models for estimating Census Collection District population densities: based on a two-phase pixel classification and regression procedure	115
5.10	Summary of selected models for estimating Census Collection District populations: based on a two-phase pixel classification and regression procedure	118
5.11	Alternative regression models for estimating the population associated with a pixel classified as residential	122
5.12	Comparison of estimated population densities and populations for Census Collection Districts: based on a two-phase procedure of classification with various regression models	124
6.1	Application of population density models based on Census Collection District aggregates to secondary study area	130
6.2	Application of dwelling density models based on primary study area Census Collection District aggregates to secondary study area	131
6.3	Summary of estimated Census Collection District populations based on application of population density models based on primary study area CD aggregates to secondary study area	132
6.4	Summary of estimated Census Collection District dwelling counts based on application of dwelling density models based on primary study area CD aggregates to secondary study area	133
6.5	Comparison of twelve population and dwelling density estimation models for primary and secondary study areas	134
6.6	Demographic characteristics of primary and secondary study areas	137
6.7	Comparison of estimated population densities for Census Collection Districts in primary and secondary study areas: based on a two-phase pixel classification and regression procedure	142
6.8	Comparison of estimated populations for Census Collection Districts in primary and secondary study areas: based on a two-phase pixel classification and regression procedure	143
6.9	Categories of land use and land cover: comparison of classes used in the two study areas	149
6.10	Comparison of estimated population densities for Census Collection Districts in primary and secondary study areas: based on a two-phase pixel classification and regression procedure with adjustments for extreme densities	152
6.11	Comparison of estimated populations for Census Collection Districts in primary and secondary study areas: based on a two-phase pixel classification and regression procedure with adjustments for extreme densities	153

6.12	Census Collection Districts with discrepant estimates of population density and/or population	160
7.1	Underlying pixel population models used for simulation	177
7.2	Summary of simulation results: estimates of regression coefficients	181
7.3	Summary of simulation results: estimates of population of individual pixels	185
7.4	Assessment of sampling strategies	190
8.1	Characteristics of study areas and images	194
8.2	Some features misclassified as residential	199
8.3	Normalised forms of the population estimation model	203
8.4	Coefficients of the normalised models	208
8.5	Summary of selected models for estimating census collection district population densities based on local classification and normalised regression procedures	210
9.1	Regression training samples	218
9.2	Summary of selected models for estimating census collection district population densities: based on local training of both classification and regression procedures	221
9.3	Summary of selected models for estimating statistical local area population densities and populations: based on local training of both classification and regression procedures	236
9.4	Numbers of positive and negative regression coefficients in 94 estimation equations	237
10.1	Comparison of pixel-based and aggregate-based estimation methods	244
10.2	Comparison of some results of this research with comparable published results	248

List of Figures

2.1	The principal components transformation	22
2.2	Maximum likelihood classification	32
2.3	The linear discriminant function	33
2.4	The linear regression function	37
2.5	Regression with incompletely determined data	43
2.6	Modelling of interaction	46
3.1	Study Areas	59
3.2	Ballarat Statistical District: showing Census Collection District (CD) boundaries and ground control points	60
3.3	Geelong Statistical District: showing Census Collection District (CD) boundaries and ground control points	62
4.1	Population and dwelling density estimates for 138 Census Collection Districts: ground truth vs. remote sensing estimates from base and enhanced models	95
5.1	Population density estimates for 138 Census Collection Districts in the primary study area: ground truth vs. remote sensing estimates from four variants of four models	116
5.2	Population estimates for 138 Census Collection Districts in the primary study area: ground truth vs. remote sensing estimates from four variants of four models	119
5.3	Population density and population estimates for Census Collection Districts: ground truth vs. remote sensing estimates	125
5.4	Contours of differences and ratios	126
6.1	Population and dwelling density estimates for Census Collection Districts: ground truth vs. remote sensing estimates from base and enhanced models	136
6.2	Estimation error vs. population and dwelling density	139
6.3	Comparison of estimated populations for Census Collection Districts in primary and secondary study areas: based on a two-phase pixel classification and regression procedure	144
6.4	Population density and population estimates for Census Collection Districts: ground truth vs. adjusted remote sensing estimates for primary and secondary study areas	154

6.5	Estimation error in Census Collection Districts population density and population: by population density and population	163
7.1	Schematic representation of the EM algorithm	167
8.1	Census Collection District boundaries for supplementary study areas	195
8.2	Spectral characteristics of the test images	201
8.3	Population density and population estimates for census collection districts: ground truth vs. remote sensing estimates from normalised Ballarat 1988 models	211
9.1	Population density and population estimates for census collection districts: ground truth vs. remote sensing estimates from locally trained models	227
9.2	Comparison of three key indicators for training samples and whole images	231
9.3	Population density and population estimates for statistical local areas: ground truth vs. remote sensing estimates from locally trained models	235

List of Landsat TM Images

1	Ballarat study area: Quasi natural colour RGB - TM bands 3, 2, 1	258
2	Ballarat study area: Green-enhanced quasi natural colour RGB - TM bands 3, 2+4, 1	258
3	Ballarat study area: 12 class MLC based on 6 TM bands	258
4	Ballarat study area: 12 class MLC based on 25 spectral and spatial transformations of 6 TM bands	258
5	Ballarat study area: Difference to sum ratio of TM bands 1 and 5	258
6	Ballarat study area: spatial variability in difference to sum ratio of TM bands 1 and 5 Standard deviation over a 3 pixel × 3 pixel neighbourhood	258
7	Ballarat study area: Estimated population density with classification and regression based on 6 TM bands.	259
8	Ballarat study area: Estimated population density with classification and iterated regression based on 6 TM bands.	259
9	Ballarat urban area: Estimated average population density based on 6 TM bands with iterated regression, smoothed with a mean filter over a 7 pixel × 7 pixel neighbourhood.	259
10	Ballarat study area: Estimated population density based on 6 TM bands with iterated regression, low density contextual reclassification and high density enhancement.	259
11	Ballarat urban area: Quasi natural colour RGB - TM bands 3, 2, 1	259
12	Ballarat urban area: Estimated population density based on 6 TM bands with iterated regression, low density contextual reclassification and high density enhancement.	259
13	Geelong study area: Quasi natural colour RGB - TM bands 3,2,1	260
14	Geelong study area: Estimated population density based on locally trained classification and normalised Ballarat regression model	260
15	Ballarat study area (1994): Quasi natural colour RGB - TM bands 3,2,1	260
16	Ballarat study area (1994): Estimated population density based on locally trained classification and regression	260
17	Adelaide study area: Quasi natural colour RGB - TM bands 3,2,1	261
18	Adelaide study area: Estimated population density based on locally trained classification and regression	261
19	Sydney study area: Green-enhanced quasi natural colour RGB - TM bands 3, 2+4, 1	261

20	Sydney study area: Estimated population density based on locally trained classification and regression	261
21	Brisbane study area: Quasi natural colour RGB - TM bands 3,2,1	262
22	Brisbane study area: Estimated population density based on locally trained classification and normalised Ballarat regression model	262
23	Kalgoorlie study area: Quasi natural colour RGB - TM bands 3,2,1	262
24	Kalgoorlie study area: Estimated population density based on locally trained classification and normalised Ballarat regression model	262

List of Abbreviations

ABS	Australian Bureau of Statistics
BIL	Band interleaved by line (format)
BSD	Ballarat Statistical District
CD	(Census) Collection District
COV	Coefficient of variation
DV	Dependent variable
ED	(Census) Enumeration District
e.r.p.	Estimated resident population
EV	Explanatory variable
GCP	Ground control point
GSD	Geelong Statistical District
HSI	Hue-saturation-intensity
LGA	Local Government Area
MSS	(Landsat) Multi-spectral scanner
PDTI	Pairwise difference texture index
RGB	Red-green-blue
SD	Standard deviation
SLA	Statistical Local Area
SPOT	Système Probatoire d'Observation de la Terre
TM	(Landsat) Thematic Mapper

Chapter 1

The Use of Remote Sensing for Population Estimation

1.1 INTRODUCTION

Population estimation in developed countries has in the past been based on two methodologies. Regular censuses have provided comprehensive baseline data. This has been supplemented by inter-censal population estimates derived from mathematical models which utilise statutory data about such indicators as births, deaths, marriages, school enrolments, land ownership and occupancy, and housing construction.

Remote sensing techniques have in relatively recent times provided a third methodology, whose rationale and potential was expressed by Henderson (1979) thus:

> *Estimates of population cannot be obtained directly from remote sensing imagery. However, simple models employing visible physical characteristics can be designed that infer population densities by surrogate...the result is a rather precise population estimate derived without actually counting the people.*

In some developing countries, these developments have coincided with and perhaps assisted in the emergence of national demographic programs. The use of remote sensing in conjunction with official census or survey operations has been reported and/or evaluated in Afghanistan (Dayal and Khairzada, 1976), Bolivia (National Aeronautics and Space Administration, 1978), Nigeria (Morgan, 1984; Olorunfemi, 1986) and Sudan (Stern, 1984). Checks of census accuracy based on remote sensing data have been carried out in Jamaica by Eyre et al (1970) and in the USA by Clayton and Estes (1980). The Australian Bureau of Statistics routinely uses aerial reconnaissance as part of the process of regular reselection of sample areas for its monthly population surveys (Crockett, 1990).

In the USA, the potential for census-related applications were examined in reports by Durland (1975) and General Electric Corporation (1977). Brugioni (1983) went so far as to suggest that

a remote sensing program could replace the US census. Morrow-Jones and Watkins (1984) and other respondents have rightly criticised Brugioni's rather extravagant claims. Nevertheless, at a time when censuses in developed countries are under fire for their expense and perceived shortcomings, and when there is pressure for at least a reduction in their frequency, remote sensing has real potential for the provision of selective, timely, and economical inter-censal population and housing estimates, which furthermore need not be limited in geographical scope to standard census areas.

In many underdeveloped countries with high rates of population growth, remote sensing has the potential to provide methodologies for monitoring changes produced by population pressure, where no methodologies exist at present.

With regard to the precision alluded to by Henderson above, it is true that some researchers have obtained quite accurate estimates of the total populations of large urban areas. In the case of small areas, a high degree of accuracy has only been attained when the visual interpretation of large scale aerial photography has been employed, and where the study areas have been relatively homogeneous. Small area population and housing estimates obtained for more heterogeneous areas from small scale high altitude photographs or from digitally analysed satellite imagery have to date been much less accurate. Improved techniques must be developed if orbital remote sensing is to become operationally feasible for this purpose.

1.2 HISTORICAL OVERVIEW

Lo (1986a, p.53) distinguished four approaches to the use of remote sensing data for population estimation. Three were established methods involving visual interpretation of (analogue) photographic images, the population estimates being based on:

- counts of dwelling units
- measured urban land areas
- measured land-use areas

The fourth method is radically different, being based on automated analysis of digital images, and utilising the spectral radiance characteristics of the individual pixels of an image.

1.2.1 Dwelling Counts

Estimated dwelling counts based on the visual interpretation of large scale black and white photography have a long history, being reported by Porter (1956), Green (1957), Eyre et al (1970), Hsu (1971), Collins and El-Beik (1971), Dayal and Khairzada (1976), Lo and Chan (1980), Watkins (1984), and Watkins and Morrow-Jones (1985). Medium to small scale high altitude color infrared images were used as the basis of housing counts by Lindgren (1971),

Duecker and Horton (1971), Clayton and Estes (1980), and Lo (1986b). Lo (1989) overlaid a raster grid on high altitude and large format space photographs in order to produce GIS-compatible estimates of dwelling counts and population.

1.2.2 Measured urban area

Models which assume a direct mathematical relationship between the population of an urban area and its size are referred to as allometric models (Webster, 1996). The relationship between total urban population and measured urban area has been investigated in a number of studies based on small scale high altitude images, including those of Wellar (1969), Holz et al (1973), Anderson and Anderson (1973), Ogrosky (1975), and Lo and Welch (1977). Various linear, power and logarithmic functional forms have been employed.

Holz et al (1973) developed a linear regression model for estimating populations of urban centres of 2500 persons or more in the Tennessee River valley, based on measures obtained from visual interpretation of small scale aerial photographs. Using census population estimates as their dependent variable, they obtained coefficients of variation (R^2) of .902 and .774 for the two years of their study.

Ogrosky (1975) applied a similar method in a study of urban centres in the Puget Sound area of Washington state, whose populations ranged from 11,000 to 531,000. The images used, from high altitude infrared photography, had a similar ground resolution to the then imminent Skylab orbiter. The regression model obtained, which had an R^2 value of .973, predicted *log*(population) for each centre on the basis of four measures obtained visually from the images, of which by far the most important was *log*(area).

Lo and Welch (1977) produced population estimates for urban centres in China from Landsat MSS images, using linear and curvilinear models relating population to area alone.

These area-related approaches have three characteristics in common.

- The populations estimated are of whole urban centres.
- The indicators, particularly measured urban area, are obtained by visual interpretation of the image as a whole.
- Census population estimates are used as the best feasible approximation to ground truth data.

1.2.3 Measured land use areas

This is a refinement to the allometric approach, in which urban areas are broken down into sub-types. Kraus et al (1974), Thompson (1975) and Lo (1979) related population to measured

urban areas of various land use types, via characteristic population densities associated with each land use type. Olorunfemi (1984) related population to the proportions of different land use types in each of a set of test areas in Nigeria.

1.2.4 Spectral radiance characteristics of individual pixels

The regression models cited above were generally aimed at estimating the populations of relatively large regions, usually whole urban areas, and invariably involved a substantial amount of visual interpretation. But with the advent of digital imagery it has become possible in principle to devise mathematical models which relate population to the spectral reflectances of the individual pixels of a raster image.

Of course the accuracy of estimates of small area populations obtained directly and automatically from satellite data in this way is inherently limited, since the linkage between the quantity to be estimated, population, and the indicator variables, which are measures of surface reflectance, is of its nature indirect, conjectural, and potentially complex.

The first published attempt to estimate small area populations from multispectral imagery was that of Iisaka and Hegedus (1982), who used regression models based on the spectral radiance characteristics of individual pixels for the purpose of estimating the population of 88 relatively small (500m × 500m) sections of the residential areas of suburban Tokyo.

Their predictor variables were not visually interpreted characteristics, but rather measures directly obtained for each pixel by remote sensing - the radiance values of the four spectral bands of the Landsat multispectral scanner (MSS).

Strictly, Iisaka and Hegedus did not operate at the level of individual pixels. Because census-based ground truth population data was available for 500m grid squares, the Landsat data was resampled to 50m × 50m pixels co-registered with the grid. Average (mean) reflectances calculated over the 10 pixel × 10 pixel grid squares were used as the explanatory variables in the regression analysis.

Iisaka and Hegedus considered only straightforward linear models. Their final equations, obtained by stepwise linear regression, expressed estimated population as a linear function of the mean radiances of MSS bands 4, 6 and 7, with R^2 values of .70 and .59 for the two years studied.

This research demonstrated that there is a relationship between small area populations and average spectral radiances which can be modelled moderately well by a simple linear function. Refinement of this methodology might be expected to lead to an improved accuracy of estimation.

In a brief report Stern (1984) described attempts to locate villages in the Sudan and estimate their populations from Landsat MSS imagery, using a classification approach based on various spectral and textural transformations of the MSS bands. No quantitative results were reported.

The work reported herein was conceived and commenced in 1990. Subsequently, a number of contemporaneous studies have been reported.

Langford et al. (1991) used a classification-based approach to estimate the populations of 49 census wards (clusters of census enumeration districts) in northern Leicestershire. The explanatory variables were the numbers of pixels in each of five land use categories (industrial/commercial, dense residential, ordinary residential, uninhabited, agricultural), obtained by performing a supervised classification of a Landsat TM image. This was conceptually similar to the previously cited work of Kraus et al (1974), Thompson (1975), Lo (1979) and Olorunfemi (1984), but the implementation was digital. For multiple regression models with all five explanatory variables, two explanatory variables (dense and ordinary residential) and one explanatory variable (all residential) R^2 values of .85, .82 and .75 respectively were obtained. However, the last two models were forced through the origin, and the basis for calculation and interpretation of the R^2 values was not clear (see section 2.8.3).

Langford et al. subsequently diverged from remote sensing population estimation per se, focussing instead, in a series of related publications (Langford and Unwin, 1994; Fisher and Langford, 1995; Fisher and Langford, 1996), on a hybrid population estimation methodology using the technique of dasymetric mapping (Wright, 1936 cited in Fisher, 1989). Beginning with information available for a set of relatively coarse geographical aggregates (in this context the known population of census enumeration districts), data from a second source is used to produce an estimated distribution at a finer level of aggregation (in this context remote sensing imagery was used to classify pixels and hence to geographically distribute the known census populations within EDs).

Yuan et al. (1997) applied a similar dasymetric analysis to that of Langford et al. to census enumeration districts in central Arkansas.

Lo (1996) applied similar approaches to both those of Iisaka and Hegedus and Langford et al. to the estimation of the population and dwelling unit numbers in 44 tertiary planning units (TPUs) in Kowloon, Hong Kong, using SPOT multispectral imagery. Five different regression models were reported for each. In four cases, the form was linear and the dependent variable was population (or dwelling) density. The explanatory variables were: means of SPOT bands 1, 2 and 3; mean of SPOT band 3 alone; mean population per pixel in high and low density residential classes; proportion of pixels in the high density residential class. In the fifth case, the form was logarithmic, the dependent variable was population (dwelling count), and the

explanatory variable was the number of pixels in the high density residential class. In each case the models were estimated using 12 TPUs and then applied to the full set of 44 TPUs. R^2 values were reported for only the fourth and fifth of these models in the training phase, the values being .88 and .77 respectively, the latter of which must be interpreted in the context of a model that was logarithmic in form (see Section 2.11.7). Results for the full set of TPUs were summarised in terms of the relative error in the total estimated population, which ranged from –5.3% to +5.3%, and the "absolute mean relative error" for individual TPUs, which ranged from 64% to 99% after deletion of 4 extreme outliers. The corresponding results for dwelling unit estimation ranged from –10.1% to +5.0% and from 50% to 77%. Whilst the overall totals were estimated reasonably accurately, the individual TPUs were not, reflecting the difficulty of applying remote sensing methodology to an area of very high population density including many multi-level and multi-functional structures.

Webster (1996) developed models for estimating dwelling unit densities in the 47 suburbs of Harare, Zimbabwe. The explanatory variables, derived from both SPOT and TM images and based on a subsample of pixels within each suburb, were characterised as measures of tone (6 TM bands); measures of texture (3 measures derived from a classification of pixels into urban and non-urban: urban pixel density, homogeneity and entropy); and one measure of context – distance from the city centre. Results from five models were reported, one based on each of the three texture variables in turn, and two (with two and three explanatory variables) selected using stepwise regression. R^2 values were in the range .69 to .81.

In the same paper, a similar but more extensive analysis was reported for the numbers of dwelling units in 65 grid squares on a transect through Cardiff, Wales, which were co-registered with a dwelling count database. Of a reported 70 texture statistics investigated, the 7 chosen by stepwise regression were described as measures of 'edginess' and 'ripple', generated using line detection algorithms and Fourier and Laplace transform methodology. R^2 values for linear and logarithmic models were reported as .86 and .97 respectively. In addition to the issue of capitalisation on chance (see section 2.11.4), the latter value was inflated both by a forced zero intercept (as pointed out by the author) and by the logarithmic form of the model (see section 2.11.7). Tables showing absolute and relative errors were presented but no summary statistics were reported for these.

The researches of Langford et al., Lo and Webster in the period since 1991 have followed similar paths to some of those traversed by the author during the same period. Together with the work of Iisaka and Hegedus, their reports provide benchmarks against which the methodologies, issues and results reported herein can be compared and assessed.

This body of work also illustrates an important issue about scale which arises when aggregated areas are unequal in size. When the explanatory variables are aggregate measures such as average spectral characteristics or proportions of pixels in different classes, then the natural dependent variable is population (or dwelling unit) density. When the explanatory variables are pixel counts, then the natural dependent variable is the total population of the aggregate. When the aggregated areas are equal in size, as with grid squares, the distinction is immaterial.

1.3 CONCEPTUAL FRAMEWORK

The work of Iisaka and Hegedus and all of the subsequent work cited above involved in each case the development (calibration, training) and testing of a model for estimating ground truth (census) population using, as explanatory variables some function(s) of spectral characteristics derived from remote sensing imagery. A conceptual framework for encompassing and extending this work was conceived as having four key aspects: sensor characteristics; approaches to spatial aggregation; form and complexity of models; and extent of validation.

1.3.1 Sensor characteristics

The work of Iisaka and Hegedus was based on Landsat MSS imagery. *Prima facie*, it might be expected that the increased number of spectral bands and the higher spectral and spatial resolution of Landsat TM might be expected to lead to more accurate estimates of population.

1.3.2 Approaches to spatial aggregation

The data employed in remote sensing approaches to population estimation occur at two levels of spatial aggregation:

Census aggregates: standardised grid cells or enumeration districts (EDs). Ground truth population figures are available at this level

Remote sensing image pixels. Spectral characteristics are available at this level, which is typically one or more orders of magnitude smaller in size than census aggregates.

Approaches to multi-level analysis

The resulting problem of multi-level analysis (Goldstein, 1995) or areal interpolation (Goodchild and Lam, 1980; Goodchild et al., 1993) can be approached in two ways, either by:

- aggregating the remote sensing data to the level of the population data, or
- disaggregating the population data to the level of individual pixels

The former approach was used by Iisaka and Hegedus and in all the subsequent research into population estimation per se cited above. The aggregation of the remote sensing data took various forms: averages, texture statistics and counts or proportions of pixels in different classes.

In the dasymetric approach, known census aggregate populations are disaggregated on the basis of broad pixel classifications, but there is no attempt to model the populations of individual pixels directly from the remote sensing data.

There are a number of potential advantages to be gained by explicitly modelling at the level of single pixels rather than larger aggregates. On a theoretical level, it might be conjectured that the relationships between human habitation and reflectances might be better defined and expressed at the level of single pixels; both spectral characteristics and population densities can vary greatly within an extended area. On a practical level, both classification and many textural measures involve a single pixel approach. As to outcomes, pixel-based estimates would enable population density images to be produced, and would be compatible with geographical information systems (GIS).

On the negative side, the essential difficulty lies in the fact that ground truth data for population estimation cannot feasibly be obtained at individual pixel level for any but the smallest areas. Most readily available ground truth and ancillary demographic data is only available for larger areas. Disaggregation of the population of an ED into constituent populations associated with each pixel is less straightforward conceptually and computationally than the derivation of aggregate spectral characteristics for an ED.

Notwithstanding these difficulties, the latter as well as the former approach has been employed in the present work.

1.3.3 Form and complexity of models

Some lines of investigation under this heading are:

Mathematical models involving data transformations and/or non-linear functional forms

There is no *a priori* reason to believe that the simple linear models employed by Iisaka and Hegedus are the most appropriate. Spectral transformations which have been widely used in remote sensing contexts include normalised bands, band ratios, band difference to band sum ratios, hue-saturation-intensity transformations and principal components (see for example Richards, 1986; Harrison and Jupp, 1990; Langford et al., 1991).

Spatial statistics, such as between-pixel variance measures and other indicators of texture have been found useful particularly in an urban context by Hsu (1978), Forster (1981, 1983), Kivell et al (1989), Barnsley and Barr (1996), Webster (1996) and Heikkonen et al. (1997, 1998).

The incorporation of non-linear functional forms such as the logarithm of the dependent variable is an approach frequently adopted when a dependent variable exhibits wide variation. This is frequently the case when the dependent variable is the population or population density of a diverse set of geographical units. Logarithmic and other curvilinear transformations were used in a number of the studies already alluded to, including those of Anderson and Anderson (1973), Ogrosky (1975), Lo and Welch (1977), Lo (1995) and Webster (1996).

Classification of the study area into land use zones prior to regression modelling

The relationship between population and spectral reflectance is not invariant across a range of land uses. For example, a roof surface in a residential area is associated with population, whereas the same type of surface in an industrial or commercial area is not. The incorporation of a classification phase would enable the regression analysis to be focussed on a more homogeneous class of residential pixels.

Iisaka and Hegedus (1982) reported that they used a clustering procedure to assign their test sites into a number of classes, but whilst it appears that this classification was used for ancillary analyses, there is no evidence in their report that any stratification was incorporated into the regression analysis.

Nevertheless they attained encouragingly high R^2 values (.70 and .59) with a rather blunt instrument - averages over 10×10 MSS pixels and straight linear functions of band reflectances. They then omitted, post hoc, the one-third of their test sites which were fitted least well, and the R^2 values rose to .88 and .81. It appears from their report that many of these sites included obvious anomalies such as major roads or open land. If such anomalies can be removed before the regression stage, some improvement should ensue.

The resulting method would essentially combine the third and fourth of Lo's methods listed above. First, all pixels in the image would be classified into broad land use categories. For all but the most common residential categories, characteristic population densities would be assigned (zero in the case of non-residential classes). But for the predominant residential categories, a regression analysis would be performed.

As to classification techniques, the shift from photo interpretation techniques to more objective statistical methods of supervised classification such as the method of maximum likelihood and linear discriminant analysis occurred in the late 1970s in the context of digitised aerial

photography (contrast, for example, Gautam, 1976, with Hsu, 1978 or Scarpace and Quirk, 1980).

Over the next decade researchers such as Jackson et al. (1980), Congalton et al (1983), Stern (1984), Tom and Miller (1984), Haack (1984), Toll (1984), Martin et al (1988), Gong and Howarth (1990a, 1990b), explored these and conceptually related unsupervised classification (or clustering) techniques with Landsat MSS, Landsat TM, SPOT XS, SPOT HRV and other orbitally acquired data. However, as is discussed by Webster (1996) and comprehensively documented by Barnsley and Barr (1996), the improvement in urban classification which was anticipated with data from higher resolution sensors largely failed to occur. The problem is one of heterogeneity. Unlike rural land-use and land-cover classes, which are often quite homogeneous, many urban land-use classes, perhaps the residential class more than any other, are inherently mixtures of different land cover types at pixel or sub-pixel scale.

In the last decade, many lines of enquiry have been followed for improving classification in remote sensing contexts, particularly with high levels of heterogeneity. These include: analysis of fractal dimensions (De Cola, 1989; Fotheringham, 1989; Lam, 1990); fuzzy set theory (Wang, 1990; Gopal and Woodcock, 1994); mixed pixel or end member analysis (Smith et al., 1990); knowledge-based systems (Wharton, 1987; Moller-Jensen, 1990; Bolstad and Lillesand, 1992); neural networks (Chen et al., 1995; Foody et al., 1995; Foody, 1996); iterative methods based on maximum likelihood (Van Deusen, 1995); genetic programming (Rioli and Line, 1995); classification and regression trees (Heikkonen et al., 1997; Heikonnen and Varfis, 1998); and various procedures which utilise spatial patterns or other contextual information for iterative reclassification (Treitz et al., 1992; Gong and Howarth, 1992; Barnsley and Barr, 1996; Sharma and Sarkar, 1998).

Notwithstanding this range of experimentation, it was anticipated that in the context of the present project even a moderately successful classification into residential versus non-residential classes using the well-established and computationally accessible maximum likelihood method had the potential to substantially improve the performance of subsequent regression modelling for population estimation. This was the approach adopted by Langford et al. (1991), by Lo (1995), by Webster (1996) and by the author. In the present instance, it was ultimately embellished by what is effectively a subsequent phase of contextual reclassification.

Models which link population to reflectance data indirectly through intermediate dwelling-related variables

Essentially, this would involve a synthesis of Lo's first and fourth methods, with visual dwelling counts being replaced by variables such as percentage housing cover, which might be directly estimated on a per pixel basis.

Such a two-stage procedure, as well as providing dwelling-related estimates of interest in their own right, might also enable refinements which are feasible at one stage but not the other.

For example, the intermediate dwelling measures, unlike population, relate directly to ground cover, and hence the nature of their relationship to surface reflectance should be more amenable to physical analysis. Ground truth data for small areas can also be obtained photographically. For population, ground truth data can only feasibly be inferred from censuses, and then only for standard geographical aggregations.

The second stage, the link between dwellings and population, is via measures such as occupancy ratios, which may also be able to be estimated spectrally, either in relation to land use classes or in some other way.

The work of Forster (1980a, 1980b, 1981, 1983) embodied the first of these two stages, and also incorporated more complex models as advocated in 2) above.

Forster estimated various housing measures including percentage housing cover and number of dwellings per area from Landsat MSS spectral data, using regression models incorporating various data transformations.

Data was obtained from 70 ground truth sites in the metropolitan area of Sydney, Australia, each site consisting of an 8×5 pixel block. Ground co-ordinates of the ground truth sites were obtained from the Landsat line and pixel co-ordinates by a polynomial transformation based on 100 ground control points distributed across the study area. The 40 pixels of each ground truth site were then located on panchromatic aerial photographs. Within each pixel area, 20 random points were chosen, and the type of ground cover at each point was assigned after examination of the photographs. Hence, estimates were obtained of the proportion of housing and other types of cover for each pixel. These estimates provided the ground truth data against which the Landsat estimates were evaluated.

The regression equations obtained by Forster, which included as predictors various transformed variables such as brightness vectors, band ratios, and textural variables in the form of between-pixel variances, attained R^2 values in excess of .80.

Lo (1995) has demonstrated that remote sensing estimates of dwelling unit counts tend to be more accurate than those of population, though not uniformly so. In some contexts, dwelling counts may be of interest in their own right (Webster, 1996). But from the perspective of population estimation, when the additional imprecision in the relationship between dwelling counts and population is considered, it has not been established that improved estimates of population would result.

Dwelling counts were estimated in the first phase of the present work, but following similarly modest improvements in accuracy to those reported by Lo, the approach was ultimately abandoned in favour of more direct population estimation.

1.3.4 Extent of validation

Validation of models for population estimation can be considered at two levels:

1) Internal validation – Given a "training set" of entities (pixels, EDs etc.) on the basis of which some procedure / function of the spectral characteristics is chosen, how well does that function reproduce the populations of the entities in the training set i.e. can any relationship between population and remote sensing characteristics be established at all?

2) External validation – How well does the procedure / function perform when applied beyond the training set i.e. how robust and generally applicable is the procedure / function / relationship?

Whilst it might be interesting from a theoretical or conceptual perspective to demonstrate the existence of a relationship in a particular context, for an estimation procedure to have any operational use *for estimation* its external validity must be demonstrated over some extended domain.

Whilst Iisaka and Hegedus (1982) mentioned external validation, their reported results appear to pertain only to the training set, and hence they were not externally validated at all. In Langford et al. (1991), external validation was limited to two indirect demonstrations of face validity – one involving a different aggregation scheme (National Grid kilometre squares vs. census wards) for the same geographical area, and the other involving the same area at the time of an earlier census. Again, whilst Webster (1996) sounded notes of caution about the geographic robustness of his results, no external validation was undertaken. Of the recently reported population estimation procedures, only Lo (1995) undertook clear and explicit external validation, by training his models on a sample of Hong Kong TPUs, and then applying them to the remaining TPUs. In no case was robustness or validity examined beyond the particular test image.

In the present work considerable investigation of external validity has been undertaken, both within particular images (regions) and between different images (regions).

1.4 RESEARCH QUESTIONS

The basic aims of this research were twofold:

- to extend and refine statistical image analysis methodologies for directly estimating small area populations and population densities from Landsat TM images.
- to validate the procedures developed and to explore their robustness to geographical and seasonal differences within Australia, and hence to explore the potential of this methodology to provide a genuine operational alternative to existing methods of population estimation.

In pursuing these aims, a number of specific research questions and specific research hypotheses, some of which emerged during the research, were addressed. Research hypotheses regarding methodology were:

- That the capability of linear population estimation models is enhanced by the incorporation of spectral transformations of TM data.
- That the capability of linear population estimation models is enhanced by the incorporation of spatial transformations of TM data.
- That the capability of linear population estimation models is enhanced by the incorporation of mathematical transformations of the dependent population variable.
- That the capability of linear population estimation models is enhanced by modelling the population of individual pixels rather than that of larger spatial aggregates.
- That the capability of linear population estimation models based on individual pixels is enhanced by classification of the pixels into different landcover/landuse classes.
- That classification of pixels is enhanced by the incorporation of spectral transformations of TM data.
- That classification of pixels is enhanced by the incorporation of spatial transformations of TM data.
- That classification of pixels is enhanced by the incorporation of a second stage of contextual reclassification.

Regarding validity and robustness, the research questions addressed to varying degrees were the extent to which both the general procedures developed and the specific details of the models were robust, firstly beyond the immediate training data, and then more broadly to differences in:

- geographical location, land cover and climate;

- time and season;
- intensity of human settlement.

The final objective was to specify a feasible operational procedure for estimating population from TM imagery. More specific aims related to this objective were:

- to identify the nature and extent of non-remote sensing inputs required;
- to identify the nature and extent of human intervention and interpretation required;
- to estimate the accuracy of population estimates attainable at the macro (major metropolitan centre), intermediate (Statistical Local Area, provincial city) and micro (Census Collection District) levels.

1.5 OUTLINE OF THE RESEARCH AND THE THESIS

The theoretical bases for each stage of the analysis are outlined in Chapter 2.

The procedures were developed and evaluated using Landsat TM images of the mixed urban/rural areas surrounding and including the provincial cities of Ballarat and Geelong, the state capital cities of Adelaide, Sydney and Brisbane, and the remote mining centre Kalgoorlie. The sources of both remote sensing data and population data, the preparation of the remote sensing images, the co-registration of images with census boundaries and the computational methods used to link the remote sensing data and population data and to perform statistical analyses are described in Chapter 3.

The first approach, in which populations of 138 census collection districts in Ballarat were estimated using aggregated remote sensing indicators, is reported in Chapter 4. A range of spectral and spatial transformations was examined and a number of preferred models selected by stepwise regression analysis.

The second approach, based on individual pixels, is developed in Chapters 5-9.

Chapter 5 traces the initial development of this methodology on the Ballarat image. An initial maximum likelihood classification of pixels was followed by regression modelling on a sample of those pixels classified as residential. A variety of spectral and spatial transformations were investigated, as well as nonlinear functional forms and an iterative refinement of the estimated ground truth populations assigned to individual pixels.

External validation of the procedures developed in Chapters 4 and 5 was conducted on a second image. A number of candidate estimation procedures arising from both approaches were applied to an image of the nearby Geelong region. As a result of this testing, one of the pixel-based approaches was selected as the preferred methodology, a number of refinements were

made to it, and a detailed evaluation of its performance undertaken. These steps are reported in Chapter 6.

Central to the chosen procedure was an algorithm for iteratively refining the estimated ground truth populations assigned to individual pixels. It was decided, prior to applying the methodology on a larger scale, to place this algorithm in a broader theoretical context, and to undertake a thorough investigation of its sampling variability and other characteristics by Monte Carlo simulation. This work is reported in Chapter 7.

Attempts were then made to find a normalising transformation which would render the procedure robust to changes in climate, season, and to the limited extent possible within the Australian context, culture. Various normalising transformations were tested on a second image of Ballarat, and on images of the other cities and regions. This work is reported in Chapter 8.

Since only a modest degree of robustness was achieved, the alternative of local training on a small subset of each region was investigated, with rather more success. These explorations and the outcomes are reported in Chapter 9.

Chapter 10 includes a comparative summary of outcomes, a generic specification for application of the recommended methodology, a consideration of its limitations, a review of further avenues for refinement and improvement, and an outline of some potential applications.

Sections of Chapters 4 and 5 have been published in preliminary form in Harvey (1996).

Chapter 2

Theoretical and Methodological Bases for the Research

2.1 INTRODUCTION

This chapter includes some general background material about remote sensing imagery and methodologies, and introduces the theoretical bases and the technical methods which are used throughout the study.

Sections 2.2 and 2.3 introduce Landsat TM data and the various ways it can be represented and displayed. Sections 2.4 and 2.5 are concerned with various mathematical transformations that can be applied to remote sensing data, in both the multivariate spectral domain and the two dimensional spatial domain.

Both the aggregate-based and pixel-based approaches to population estimation utilise some form of linear modelling, and the pixel-based approach also includes a classification phase. Sections 2.6 and 2.7 deal with discrimination and classification, and sections 2.8 and 2.9 discuss various aspects of linear regression models. Two non-linear aspects of the methodology are considered in Section 2.10.

Assessment of performance, validity and robustness are addressed in Section 2.11 The final section 2.12 considers some related issues of parameterisation and presentation.

2.2 LANDSAT THEMATIC MAPPER MULTISPECTRAL RADIANCE DATA

Landsat 5 follows a near polar, sun synchronous orbit at an altitude of 705km, and with a period of 98.9 min. Image data for each particular area is acquired on the north-south traverse, at around 9.30 am local time.

The Thematic Mapper (TM) is a mechanical scanning device which sweeps 16 transverse scan lines simultaneously across a swath of width 185 km. The orientation of the 16 TM sensors is

such that the spacing of the scan lines on the ground is 30m. The sensors produce a continuous output which is sampled at a rate which corresponds to a 30m spacing of samples along the scan line also. The optical characteristics of the TM are such that the instantaneous field of view (IFOV) also corresponds to this spacing, so that each sampled pixel ideally represents the integrated response of the TM sensors to the radiation reflected from a 30m square on the ground, with the pixels (or strictly the squares that they represent) being contiguous in both across-track (along-scan) and along-track directions. In practice, atmospheric attenuation and scattering degrade the signal to some degree both radiometrically and spatially (see Section 3.7).

The TM senses radiation in seven spectral bands, whose characteristics are summarised in Table 2.1.

The IFOV of band 6 is 120m×120m, corresponding to a 4 pixel×4 pixel square in the other bands. For uniformity, data for this band is also formatted as if for 30m pixels, with each data value being repeated for the 16 appropriate pixels. However, the underlying resolution incompatibility renders band 6 generally unsuitable for incorporation in multispectral analyses. This study is based on the 6 spatially compatible bands. Henceforth, reference will be made to these bands only.

The pixel brightness in each spectral band is proportional to the incident radiance, measured in watt per steradian per square metre. The dynamic range of the TM is 8 bits. The output is expressed as an integer in the range 0-255.

Table 2.1 Landsat TM Spectral Bands

Band	Wavelegth (μm)	Description
1	0.45 - 0.52	blue
2	0.52 - 0.60	green
3	0.63 - 0.69	red
4	0.76 - 0.90	near infrared
5	1.55 - 1.75	mid infrared
7*	2.08 - 2.35	mid infrared
6	10.4 - 12.5	thermal infrared

* Band 7 is numbered out of sequence as a result of being added late in the design period when the numerical designations of the other bands were well established.

2.3 SPATIAL DOMAIN AND SPECTRAL DOMAIN DATA REPRESENTATIONS

The full raw dataset for a TM scene[1] or subscene can be thought of as being 8-dimensional, consisting of a 2-dimensional $r \times p$ spatial array, where

r = number of rows

p = number of pixels per row,

and where each point in the spatial array has an associated 6-dimensional vector of spectral brightnesses.

The term "image" is used in three ways:

(i) (occasionally) the totality of the 8-dimensional raw data i.e. synonymous with "scene" or "subscene".

(ii) (more commonly) a derived data set with 2 spatial dimensions and either: 1 spectral dimension e.g. a single spectral band or the ratio of 2 spectral bands; or 3 such spectral dimensions.

(iii) (usually) a visual representation or realisation of (ii) e.g. a colour-coded video display or print.

2.3.1 Colour representation

Harrison and Jupp (1990) distinguish two types of colour representation of images.

In the case of a single spectral dimension, a colour look-up table (*lut*) is used to assign a graded range of colours (typically up to 2^8 =256 in number) to the range of numerical values which occur, producing a *pseudocolour* image. A monochrome or *greyscale* image is a special case in which only black, white and the intermediate shades of grey are used.

Three spectral dimensions may be displayed as follows. The human eye perceives colours as the proportions of red, green and blue wavelengths that it detects. These colours are known as the additive primaries. Colour monitors use three colour guns to excite red, green and blue phosphors, and hence additively create the full range of colours on a screen. Three variables may be displayed by assigning each to one of the guns, whose intensities then represent the levels of the corresponding variables. The resulting *colour composite* or RGB image has a much larger range of available colour shadings than a pseudocolour image (typically $(2^8)^3 \approx$

[1] Technically, the term “scene” refers to the standard geographical units into which TM output is divided for purposes of identification and distribution.

17,000,000). If the red, green and blue TM bands are displayed in this way, the resulting image is a 'true' or 'natural' colour composite. Any other mapping is called a false colour composite.

In the present study, both natural and false colour composite RGB images have been used for general geographical orientation, for co-registration with overlaid census boundaries, for visual selection of classification training sets, and for identifying causes of anomalous results. Pseudocolour images have been used for displaying single TM bands, various derived measures, classifications and population densities.

2.3.2 Spatial and spectral displays

When a pseudocolour or colour composite image is displayed, the spatial domain provides the explicit structural basis of the display, whilst the quantity displayed represents a 1- or 3-dimensional aspect of the spectral domain characteristics.

Alternatively, the spectral domain may be used as the explicit basis of the representation. The response of each pixel can be represented by a point in 6-dimensional Euclidean space, of which any 1, 2 or 3 dimensions may be displayed graphically at one time. It is conventional to use histograms for displaying the values of a single spectral variable, and 2- or 3-dimensional crossplots (or scatter-plots) in the other cases. Because the number of points (pixels) is usually large, spatial information is usually omitted from histograms and crossplots, although some information about spatial areas may be incorporated, for example by colour coding of categories.

In the present study, histograms have been used extensively for purposes of distributional analysis and as an aid to image enhancement for the purpose of detecting features and facilitating visual judgements.

2.3.3 Data transformations

There is no *a priori* reason to believe that population will be best indicated by a simple linear combination of TM bands. Potentially useful data transformations which have been widely used in many remote sensing contexts are now reviewed and considered.

The mathematical transformations which are applied to multispectral data fall into two categories - those which are applied to the data one pixel at a time, and those which involve neighbouring pixels also.

The former type operate on one point in space at a time. They take as input one or more spectral bands and produce as output one or more new spectral variables. These are referred to as *spectral domain transformations* or *point transformations.*

The latter type operate on one spectral band only, but involve more than one point in space. These are referred to as *spatial domain transformations*.

Both types[1] are considered further in the following sections.

2.4 SPECTRAL DOMAIN TRANSFORMATIONS

2.4.1 Single band transformations

The simplest spectral domain or point transformations are those applied to a single spectral band.

These include linear and distribution-based transformations routinely applied for radiometric correction or contrast enhancement of images (See Section 3.7).

Transformations which might be considered for analytic purposes include raising to a power, or taking a root, logarithm, or exponent. However, most common analytic point transformations are multivariate in nature, with multiple inputs and either single or multiple outputs. Some standard procedures are briefly outlined.

2.4.2 Single-valued band comparisons or indices

The pixel brightnesses in two or more spectral bands may be added, subtracted, multiplied or divided. Multiplication has not proved useful and is little used. Band differences and ratios are most common (Richards, 1986, p. 146).

Differencing is used, for example, to detect temporal changes by comparing the same band in two co-registered images from different dates.

Three generic ratio-based vegetation indices of increasing complexity are:

$$I1 = \frac{\text{infrared}}{\text{red}}$$

$$I2 = \frac{\text{infrared} - \text{red}}{\text{infrared} + \text{red}} \qquad \text{normalised difference vegetation index (NDVI)}$$

$$I3 = \sqrt{I2 + 1.0} \qquad \text{transformed vegetation index (TVI)}$$

The first of these is a simple band to band ratio, which is effectively a transformation to a single polar co-ordinate. The second is a band-difference to band-sum ratio, division by the sum

---

1 The terms “tone” and “texture” are sometimes used respectively to refer to the spectral and spatial dimensions of an image (see for example Wang, L. and He, D.C. (1990)).

having the effect of "normalising" or limiting the values of the index to the range (-1,1). I2 carries the same information as I1, and is related to it by the formula

$$I2 = \frac{I1 - 1}{I1 + 1}$$

In the third case, a constant is added to remove negative values, and then the square root is taken.

A related TM-based index used as an indicator of urban density (Kawamara et al., 1996) is the band-difference to band-sum ratio:

$$UI = \frac{B7 - B4}{B7 + B4}$$

Another type of common index is the basic band normalising ratio:

$$I4 = \frac{band}{\Sigma\, bands}$$

The purpose of normalisation is to reduce the effect of variations in overall light intensity.

In the present study, extensive investigation was made of band to band ratios, band-difference to band-sum ratios and band normalising ratios.

2.4.3 Many-to-many transformations

In the three transformations which follow, the rectangular co-ordinate framework based on the original spectral bands is replaced by a new co-ordinate framework based on a new set of derived variables.

2.4.4 The principal components (PC) transformation

Essentially, the principal components transformation is a *linear transformation* in which the vector **y** representing a pixel point in the new co-ordinate system is related to the original co-ordinates **x** by the equation

$$\mathbf{y} = \mathbf{T}\mathbf{x}$$

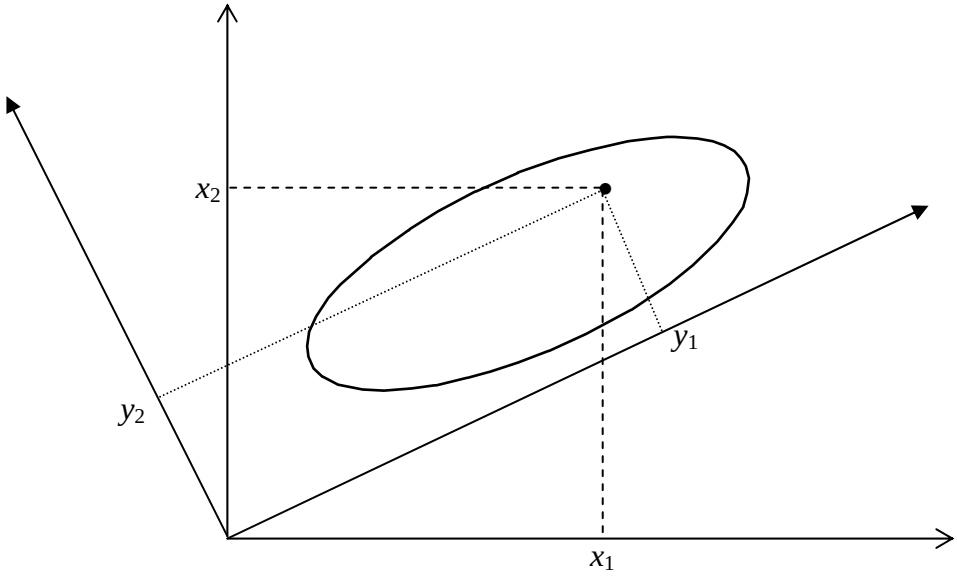
Geometrically, this represents a rotation of the reference axes. The matrix **T** is chosen so as to align the new axes with the spatial distribution of the data points in such a way that the new *y* variables (the principal components) are, unlike the original *x* variables, uncorrelated with one another.

This is illustrated schematically in Figure 2.1, where the ellipse represents the region of 2-dimensional spectral space occupied by a set of data points, of which a typical one is plotted.

To accomplish this particular rotation, it can be shown that **T** must be the transposed matrix of eigenvectors of the covariance matrix $\mathbf{\Sigma}_x$.

It can be further shown that the covariance matrix of the *y*s, $\mathbf{\Sigma}_y$, is diagonal, with the diagonal entries (the variances of the *y*s) being equal to the eigenvalues of $\mathbf{\Sigma}_x$, which are always non-negative. It also follows that since the off-diagonal elements of $\mathbf{\Sigma}_y$ (the covariances of the *y*s) are zero, the *y*s are uncorrelated.

Figure 2.1 The principal components transformation



The rows of **T** can always be ordered in such a way that the eigenvalues occur in descending order of magnitude. Thus, as can be seen in Figure 2.1, y_1, the first principal component (PC1), has the largest variance, y_2 the next largest variance, and so on. Geometrically, PC1 is aligned in the direction of greatest spread of the pixel points in the original co-ordinate space. PC2 is orthogonal to PC1, and subject to this constraint, is in the direction of the next greatest spread, and so on.

These directions are determined by the coefficients of the linear transformation, i.e. the elements of the matrix **T**, also known as weights or loadings.

Commonly, all spectral bands are positively correlated, in which case PC1 will load positively on all bands, and will be a measure of overall brightness. The second and subsequent components will represent orthogonal contrasts between various band combinations, with each further component exhibiting less pixel-to-pixel variation.

The higher the correlations between the original bands, the more rapid is the decrease in component variances, and the greater is the proportion of the total information which is contained in the first few components.

The following figures, taken from Richards (1986, p.137), illustrate a typical principal components analysis (PCA) for a 4 band Landsat MSS image. The covariance matrix is

$$\boldsymbol{\Sigma}_x = \begin{bmatrix} 34.89 & 55.62 & 52.87 & 22.71 \\ 55.62 & 105.95 & 99.58 & 43.33 \\ 52.87 & 99.58 & 104.02 & 45.80 \\ 22.71 & 43.33 & 45.80 & 21.35 \end{bmatrix}$$

Its eigenvalues are:

eigenvalues	253.44	7.91	3.96	0.89
% of total	95.20	3.00	1.50	0.30

The transformation matrix, whose columns are the eigenvectors, is

$$\mathbf{T} = \begin{bmatrix} 0.34 & -0.61 & 0.71 & -0.06 \\ 0.64 & -0.40 & -0.65 & -0.06 \\ 0.63 & 0.57 & 0.22 & 0.48 \\ 0.28 & 0.38 & 0.11 & -0.88 \end{bmatrix}$$

The pattern of the loadings indicates that the 1st PC is a positively weighted combination of all four bands (i.e. overall brightness); the 2nd PC is essentially the difference between the visible and infrared bands; the 3rd is essentially the difference between visible red and green; and the 4th is essentially the difference between the two infrared bands.

However, the first component encompasses 95% of the variance in the data, and hence 95% of the information for distinguishing between pixels. At the other extreme, the last component has negligible variance and hence negligible information content. It is to be expected that, when displayed, the first component will exhibit strong contrast and structure, whilst the last will appear almost totally as noise of low amplitude.

In a similar way, most of the information contained in a 6 band TM image can often be compressed into a 1-, 2- or 3-dimensional PC image for more convenient and effective display or further analysis.

As has been pointed out above, the rotation of the co-ordinate axes results in at least some components which are bipolar, for which data co-ordinates may be positive or negative. For convenience, in the image analysis context, negative values may be avoided by a translation of the reference origin by an appropriate amount. Thus in practice the affine transformation

$$\mathbf{y} = \mathbf{Tx} + \mathbf{c}$$

where **c** is a constant vector, may be used.

The Kauth-Thomas transformation

The Kauth-Thomas or "tasselled cap" transformation (Kauth and Thomas, 1976) originally defined in terms of four band MSS data, is not so much a type of transformation as a realisation of a principal components transformation in an agricultural context. Kauth and Thomas found four orthogonal directions, the first three of which broadly corresponded to soil brightness, greenness and yellowness, with the remaining dimension being essentially random noise. These four components, like principal components, have decreasing variances.

Whilst the context-driven approach of Kauth and Thomas was very different to the purely statistical basis of principal components analysis, the resulting transformation matrix (after Richards, 1986, p.145)

$$\begin{bmatrix} 0.43 & -0.29 & -0.83 & 0.22 \\ 0.63 & -0.56 & 0.52 & 0.01 \\ 0.59 & 0.60 & -0.04 & -0.54 \\ 0.26 & 0.49 & 0.19 & 0.81 \end{bmatrix}$$

is, apart from the arbitrary reversal of sign in the last component, very similar to the PC transformation matrix above.

In this study, the principal components of the 6 TM bands were investigated as possible indicators of population.

2.4.5 Hue-saturation-intensity (HSI) transformations.

The red, green and blue *additive primary* dimensions of a colour composite image (see Section 2.2) can be re-parameterised in terms of the three variables hue, saturation and intensity. For an illustration of the representation of HSI in the RGB colour cube see for example Harrison and Jupp, 1990, p18. The *intensity* diagonal represents the shades of grey from black to white. On the level surfaces of intensity, *hue* is an angular measure representing what is commonly referred to as colour, whilst *saturation*, measured radially from the intensity axis, represents the strength of the colour. The *subtractive primaries* yellow, cyan and magenta occur at the remaining corners of the cube.

Harrison and Jupp report a variety of mathematical implementations of the HSI concept, based on triangular, conical, cylindrical and spherical co-ordinates, among others. The HSI transformation supplied with ERMAPPER software is essentially rectangular in nature. The level surfaces of intensity are the faces of the RGB cube, the saturation contours on each face

are L shaped, and the hue contours are 'stripes' parallel to the R-W, G-W and B-W diagonals of the faces.

In this study rectangular, triangular and cylindrical HSI transformations were examined. The formulae are given in Appendix A.

2.5 SPATIAL DOMAIN TRANSFORMATIONS

As has been discussed above, spectral domain or point transformations are applied to each pixel one at a time, and usually involve more than one spectral band.

In contrast to this, spatial transformations are applied in the spatial domain. The input is a single spectral band (which may be either a primary data band or a derived band resulting from a univariate or multivariate spectral domain transformation), and the output value is a function of the input values of a number of pixels. A distinction can be drawn between the structural or morphological approaches of pattern recognition which seek to distinguish shapes and objects, and statistical measures of spatial variation (though the two approaches are not mutually exclusive). Within the statistical domain, two approaches can be distinguished, based on either aggregates over an extended area, or on the immediate neighbourhood of each pixel.

2.5.1 Aggregate measures

If an image can be partitioned into a number of contiguous areas, then it may be useful to calculate statistical measures for each such area.

Iisaka and Hegedus (1982) used means of each spectral band to estimate populations within grid squares. In the present study, this approach was applied to census collection districts (CDs), with means, variances, standard deviations and coefficients of variation (standard deviation/mean) of both raw TM bands and a number of spectrally transformed variables being evaluated as predictors.

These bulk variability measures indicate the magnitude of the variation over an extended spatial area, in contrast to the measures of local variability or texture discussed below.

Chavez (1992) has investigated the scale-dependence of spatial standard deviations calculated at particular spacings or spatial frequencies, using high pass Laplacian filters and variogram techniques. Whilst it is possible that scale-specific variability measures may have different properties than overall regional measures, they were not utilised in the present study.

2.5.2 Neighbourhood measures (spatial filters)

Neighbourhood operations are generally defined in terms of a square window or template with an odd number of pixels per side (often just 3), which is passed over the image and centred on each pixel in turn. The output value for each pixel is some function of the input values of that pixel and of its neighbouring pixels within the window.

Convolution operations

In a convolution operation, the central pixel value is replaced by a linear combination (i.e. a weighted sum) of the pixel values within the window. The array of weights is referred to as the *kernel* of the convolution.

In image analysis and image processing, convolution operations are routinely employed for both *smoothing* (low pass filters) and *sharpening* (high pass filters) of images. Niblack (1986) gives an extensive summary.

In the case of smoothing, averaging procedures with all non-negative weights are used. Sharpening techniques, such as those for detecting and enhancing lines and edges, utilise various patterns of positive and negative weights.

Non-convolution template operations

Other smoothing transformations, which are based either on structural analysis or on statistical alternatives to averaging, include the *median filter*, in which the central pixel value is replaced by the median of the pixel values within the window.

Various measures of image *texture* have been proposed, which are based on statistical measures of the variability of the pixel values within the window.

2.5.3 Measures of image texture

Three obvious related measures of inter-pixel variability are *spatial variance*, *spatial standard deviation* and *spatial coefficient of variation*, which have been investigated in the context of urban applications of remote sensing by Forster (1981), Woodcock and Strahler (1987), Forster and Jones (1988), Takeuchi and Tomita (1988), Kivell et al (1989), Ng (1990), Forster and Xing (1992) and Forster (1993), and in a more rural village-oriented study by Stern (1984).

Haralick (1978, 1986) and Rosenfeld and Kak (1982) have reviewed a wide range of other statistical and structural approaches to the analysis of texture in remote sensing and other image analysis contexts. The methods and measures include auto-correlation and autoregressive time series techniques, digital transforms, mathematical morphology, gray-tone co-occurrence, edge density, relative extrema density, run lengths, Markov random fields and random mosaic

models. Wang and He (1990) have developed a statistical approach which involves a scheme for coding the texture of each pixel neighbourhood into a single numerical value called a "texture unit number", and then analysing the distribution or spectrum of these numbers over a whole image or a subset thereof.

Gong and Howarth (1990a) have reported the application of an edge density measure to the classification of land-use in a mixed urban/rural setting. However, the method is computationally expensive, requiring multiple passes over the image with different window sizes, and it involves visual examination and subjective decision-making by the user. More recently, these authors have proposed (Gong and Howarth, 1992) a multidimensional classification procedure in which the dimensions are the frequencies of occurrence of each possible grey level within a relatively large (9×9) window. For multivariate data, principal components analysis followed by requantisation are used to reduce the number of possible grey levels. This method too is computationally complex, and requires the selection of an optimum window size.

Webster (1996) reported the use of a battery of texture measures in the context of estimation of dwelling counts in urban areas. Some were measures of spatial variation derived from the class membership of neighbouring pixels, and others, designed to detect the repetitive patterns of street grids and described as measures of ‘edginess’ and ‘ripple’, were generated using line detection algorithms and Fourier and Laplace transform methods.

Heikkonen and others (Heikkonen et al., 1997; Heikkonen and Varfis, 1998) defined a number of co-occurrence measures computed from spatial gray-level dependence matrices (conceptually similar to Webster’s neighbour-based measures). They also used Gabor filters from the domain of signal processing. These are spatial sinusoids localised by a Gaussian window which provide measures of self correlation or morphological self similarity (again conceptually related to Webster’s measures of repetitiveness).

Hsu (1978) carried out a very comprehensive study of relatively straightforward approaches. The 23 different texture measures he applied to the classification of digitised black-and-white aerial photographs can be categorised into four types, as follows.

(i) Deviation measures

These are measures of the average deviation of pixels in the window from some central value, such as the central pixel value or the mean of all the pixels in the window. The types of average used by Hsu were mean absolute (MA) deviation, mean squared (MS) deviation and root mean squared (RMS) deviation. In the case of deviations from the mean, the resulting statistics are the familiar mean deviation, variance and standard

deviation. Other averages such as the median could be used. Hsu also included higher order moments which measure skew and kurtosis.

(ii) Pairwise-difference measures

In these measures, which Rosenfeld (1984) termed measures of "local busyness", an average is calculated for the differences between chosen pairs of pixels within the window e.g. all pairs, nearest neighbours, second nearest neighbours. Again, the averaging procedure may be MA, MS or RMS.

(iii) Proportion measures

In this case, the calculated statistic is the proportion of pixels in the window with values above (or below) some chosen threshold or datum.

(iv) Measures based on spatial wave-form characteristics

These are based on the frequency and amplitude of peaks and troughs in both x and y directions. They include average number of peaks and troughs, average difference in magnitude of peaks and troughs, and average distance of peaks and troughs from the centre of the window.

Hsu applied these measures to the classification of eight general land-use types, using stepwise discriminant analysis. He reported "hit rates" (correct classifications) above 95% for the training sets, and 85-90% for the full data set, with all but 4 of the 23 variables contributing significantly to the classification.

In the present study, a central task was to distinguish between residential and other types of land use.

Woodcock and Strahler (1987), Forster and Jones (1988), Takeuchi and Tomita (1988), Ng (1990), Forster and Xing (1992) and Forster (1993) have shown that inter-pixel variation is maximised when the individual land cover elements are similar in size to the pixels. The key elements of the residential environment are dwellings and other structures, yards, lawns, gardens, trees and paved areas including streets. The dimensions of these elements are of the same order of magnitude as a 30m square Landsat TM pixel. As a consequence, residential areas display a high degree of inter-pixel variation in all TM bands.

Whilst other land cover classes such as open forest also exhibit inter-pixel homogeneity, it may be possible to find a combination of texture and other spectral characteristics which clearly indicates residential land use.

2.5.4 Development of a texture index based on pairwise differences

In classifying pixels on the basis of their spectral response alone (See Section 5.5), it was observed that many rural pixels associated with roads and shorelines in particular were incorrectly classified as residential.

The aim was to discriminate between, on the one hand both lines and edges, regardless of their direction, and on the other hand, a more amorphous texture. Whilst many standard line detecting and edge detecting filters exist, the particular combination of requirements appeared to be sufficiently unusual to necessitate the development of a more specific filter.

Nine standard variants of a 3×3 window were set up with cells set to high or low values. Four represented lines in different directions, four represented lineal boundaries in different directions, and one, with all cells at a constant intermediate level, represented no lineal pattern. With superimposed random error, the ninth window had the amorphous texture characteristic of mixed residential pixels.

After extensive experimentation with simulated data (see Appendix B) a composite measure was developed, based on two of Hsu's pairwise difference measures, which can be specified as follows.

The value assigned to the central pixel is the lesser of:

- the average of the absolute values of the differences between 2nd nearest neighbours in the diagonal direction (which in the case of a 3×3 window is just the average absolute difference between pixels in diagonally opposite corners), and
- the average of all but the two largest of the absolute values of the differences between nearest neighbours around the perimeter of the window.

This measure is referred to in Chapter 5 as the *pairwise difference texture index*.

This measure achieved strong discrimination between the random pattern and all the geometrical patterns tested even in the presence of a substantial degree of random noise, regardless of the of the particular averaging procedure used. For simplicity and speed of computation it was decided to use the root mean square averaging procedure.

Application of the same set of measures to simulated data based on a 5×5 window resulted in a greatly increased computational burden, and produced no improvement in discrimination. Accordingly, it was decided to use measures based on a 3×3 window only.

The index was coded in C as a user-defined ERMapper filter.

2.6 SUPERVISED CLASSIFICATION

Supervised classification is the commonly used procedure of assigning each pixel in an image into one of several predefined land use or land cover categories. (The less common exploratory approach in which the categories are not predefined, but are suggested by the analysis, is referred to as unsupervised classification or, in statistical parlance, cluster analysis.) In the present study, the key land use, "residential", was predetermined, along with various other categories.

Having defined the classes, the next step is to select "training sets" of typical representative pixels from each class. In the present study this was done in part by visual examination of the image, and in part from known statutory land use zones.

Next, a particular classification algorithm, which may be deterministic or probabilistic, is chosen, and the statistical characteristics of the training set data are used to estimate the parameters which characterise each class in multispectral space, also known as the signature of the class. This empirical estimation of parameters constitutes the "training" of the classifier.

The trained algorithm is then applied to the whole image, with each pixel being assigned to a class. The output of the classification essentially takes the form of a new band of categorical data, which may be used in its own right as the basis of thematic displays, or as in the present study, it may provide a basis for further analysis.

The effectiveness of the classification procedure may be gauged by examining the "confusion matrix", which shows the number of pixels in each training set assigned to each class. The leading diagonal gives the numbers of correct assignments, and the off-diagonal elements give the numbers of incorrect assignments. The proportion that are correctly classified provides a measure of internal validity.

2.6.1 Maximum likelihood classification (MLC)

The simplest classification algorithms are based on a deterministic partitioning of the multivariate spectral space. More sophisticated methods involve the probabilistic assignment of pixels to classes with distributions which overlap in the multivariate space. The maximum likelihood classifier is the most commonly used algorithm of the latter type.

As the name suggests, this algorithm assigns each pixel to the class which is most likely (in a technical sense which is explained below) to contain it, given the empirical information about the classes obtained from the training sets.

Symbolically, given a set of n spectral classes represented by

ω_i, i=1,...,n

and a particular pixel at location **x** in multivariate space, the pixel is assigned to the class for which the conditional probability (known as the likelihood)

$p(\omega_i/\mathbf{x})$

is a maximum.

Now, by Bayes' rule (Mendenhall, 1979)

$p(\omega_i/\mathbf{x}) = p(\mathbf{x}/\omega_i)\ p(\omega_i) / p(\mathbf{x})$

For a particular pixel, $p(\mathbf{x})$ is constant with respect to changes in *i* (it is the overall probability of finding a pixel from any class at location **x**). Thus the required class is the one for which the product

$p(\mathbf{x}/\omega_i)\ p(\omega_i)$

is a maximum.

The second term in this product is called a *prior probability*. It is the *a priori* probability associated with class *i* - that is, the overall probability with which any pixel would be assigned to this class by guesswork before the classification is carried out. The prior probabilities may be assigned on the basis of prior knowledge about the relative preponderance of the classes. If nothing is known, they may be assumed equal, in which case the term is simply omitted.

The first term is the (conditional) probability that a pixel from class *i* would occur at location **x**. To compute this term, the distribution of each class must be specified. The standard procedure is to assume that each class has a multivariate normal or Gaussian distribution in multispectral space. The training set data are used to estimate the two parameters of this distribution for each class: the mean vector $\mathbf{m}_i$, which geometrically represents the centroid, and the covariance matrix $\mathbf{\Sigma}_i$, which determines the shape, spread and orientation of the distribution, i.e. the region of multispectral space occupied by pixels of the particular class.

By way of illustration, Figure 2.2 depicts four classes in 2-dimensional space, each with a different bivariate normal distribution. The third axis (*p*) represents the probability of a pixel belonging to each of the classes. (Strictly *p* is the probability density, the probabilities being represented by the volume under each "hill".) A low probability contour is shown for each class. The distributions differ in location, shape, spread and orientation (i.e. they have different mean vectors and different covariance structures). Three of the classes overlap each other to various degrees, whilst the fourth is clearly separated from the other three.

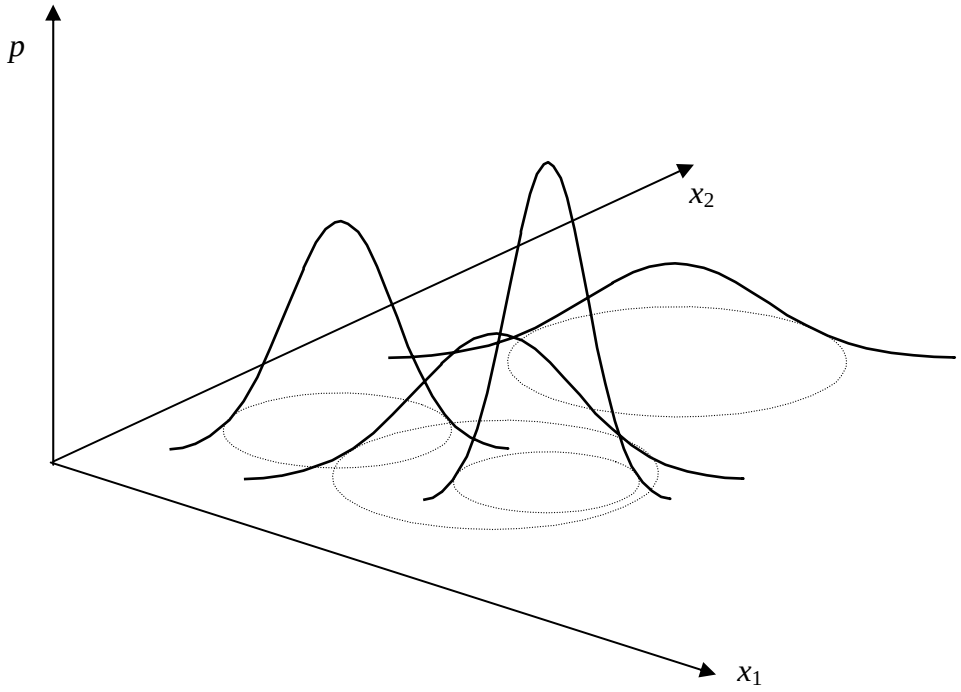
In practice, because of the exponential form of the Gaussian probability function, it is convenient to work with the logarithm of the likelihood product given above. It can be shown (Johnson and Wichern, 1982, p 497) that the so called log-likelihood criterion which results and which is to be maximised has the form

$$L(\omega_i : \mathbf{x}) = 2 \ln p(\omega_i) - \ln|\Sigma_i| - (\mathbf{x} - \mathbf{m}_i)'\Sigma_i^{-1}(\mathbf{x} - \mathbf{m}_i)$$

with the first term being omitted if the prior probabilities are assumed to be equal.

Finally, a threshold level T_i may be set for each class, with the pixels whose log-likelihoods fall below this threshold remaining unclassified.

Figure 2.2 Maximum likelihood classification



The final decision rule is then

$\mathbf{x} \in \omega_i$ if $L_i(\mathbf{x}) > L_j(\mathbf{x})$ for all $j \neq i$

and $L_i(\mathbf{x}) > T_i$

In Figure 2.2 the probability contour shown might correspond to the threshold level. These and the intersections between the distributions define the decision boundaries.

A more extensive treatment of the above is given in Richards (1986).

If the covariance matrices in all classes are equal, the inter-class boundaries are linear, and MLC is equivalent to linear discriminant analysis (see Section 2.7). In the case of unequal covariance

matrices, the boundaries are quadratic in form (hyperbolae & ellipses), and the term quadratic discriminant analysis is sometimes used.

2.7 SELECTION OF CLASSIFIERS

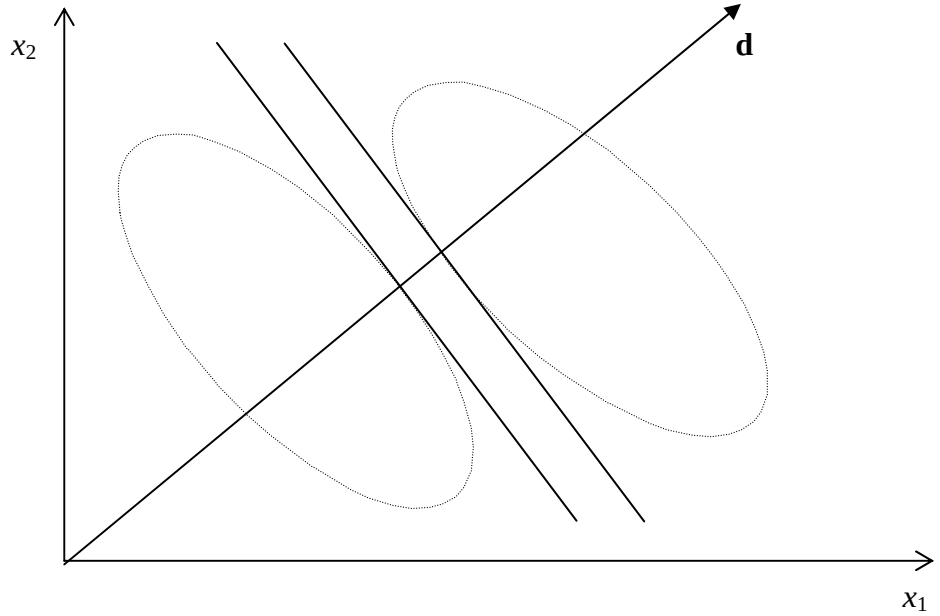
Maximum likelihood classification is computationally expensive if the number of spectral dimensions is large. Furthermore, it is usually the case that the classes will be more separable in some dimensions than others. As an alternative to multiple runs of the MLC algorithm, stepwise discriminant analysis (Cliff, 1987) or canonical analysis (an almost synonymous term used by some authors such as Richards, 1986) provides a systematic method of selecting an appropriate subset of the most strongly discriminating variables for use in the MLC procedure.

2.7.1 Linear discriminant analysis

Leaving aside the "stepwise" aspect for the moment, the aim of a linear discriminant analysis with a fixed set of variables is to find those directions in multivariate space in which the separation between a number of predefined classes with the same covariance structures is maximised. This is a mathematicallty equivalent problem to MLC though viewed from a different perspective.

The ellipses in Figure 2.3 represent two distinct classes in 2-dimensional spectral space. (They may be thought of as near-zero probability contours.) The two classes are almost identical with respect to variable x_2. Whilst x_1 provides some discrimination between the classes, they have a degree of overlap in this direction also. However, in the direction of the discriminant function vector **d**, the classes are maximally separated.

Figure 2.3 The linear discriminant function



The separation criterion used is $\frac{\text{variance between classes}}{\text{variance within classes}}$.

It can be shown, analogously to principal components analysis (see Section 2.2), that the required directions and the values of the variance ratio in those directions, are given by the eigenvectors and eigenvalues of the matrix

$$\Sigma_B \Sigma_W^{-1}$$

where Σ_B = between classes covariance matrix, esimated from the class means

Σ_W = within classes covariance matrix, estimated by pooling or averaging the covariance matrices for each class.

Analogously to principal components analysis, the direction associated with the largest eigenvector, called the first discriminant function, is the direction of greatest separation of the classes. The second discriminant function gives the next preferred direction for discrimination, and so on (though, unlike principal components, successive discriminant functions are not generally orthogonal). The number of potential discriminant functions is one less than the lesser of the number of variables (spectral dimensions) and the number of classes. However, as with PCA, the first few functions may account for most of the class separation.

2.7.2 Stepwise discriminant analysis

When a substantial number of potential discriminating variables are available for consideration, stepwise discriminant analysis provides a systematic method for selecting an appropriate parsimonious subset of the variables. In principle, the stepwise discriminant algorithm proceeds as follows.

Firstly, each variable is examined in turn, and the one on which the classes are most separated, in terms of the above criterion, is taken as the starting point of the analysis.

Each of the remaining variables is then paired with the first chosen variable in turn, and the discriminant functions calculated. The pair of variables which discriminate best are retained.

The procedure continues with one more variable being included at each step until no significant improvement in discrimination is achieved, at which point both the selected set of variables and the degree of discrimination they achieve is known.

It is also possible for a variable included in the model at some stage to become redundant at a later stage, in the sense that its removal would not cause a statistically significant reduction in discriminating power. To avoid iterative cycling, it is usual to set the significance probability

for such removals somewhat higher than that for entry, typical values being .10 and .05 respectively.

In practice the matrix computations are performed incrementally, without the need for full inversion and eigenvalue calculations at each trial step.

Stepwise discriminant analysis is an example of a general sub-optimal iterative search technique called the method of steepest ascent. It does not necessarily lead to the global optimum - this can only be guaranteed by examining all possible combinations of variables. However, the loss in discriminating power relative to the optimum is generally small and the computational savings are such that the trade-off is generally regarded as acceptable.

2.7.3 Relationship between DA and MLC

Stepwise discriminant analysis has two aspects: an initial model identification and estimation phase, in which the class covarince matricesd are assumed to be equal; and a classification phase, in which the covariance matrices may or may not be assumed equal. Maximum likelihood classification is equivalent to the second phase, and is generally based on the individual class covariance matrices.

There is no conflict in this. At the exploratory stage, averaging over the classes leads to a single criterion for assessing overall discriminating power of the available variables. But having chosen the variables to be used, it is important for achieving high accuracy at the classification stage, to use of all the available information about the particular spread, shape and orientation of each class.

In the present study, maximum likelihood classification was available in the image processing software, but not discriminant analysis. Using statistical software, stepwise discriminant analysis was applied “offline” to a suite of 80 variables derived from the training set data, of which sets of 6, 10, 15 and 25 were chosen for use with the maximum likelihood classification algorithm of the image processing software. (See Section 3.5 for details of the software used.)

2.8 LINEAR MODELS

2.8.1 Multiple linear regression

Given for each of a set of entities or cases, the measurements on a number of variables x_1, x_2,...,x_p, and a variable y, the aim of a multiple linear regression analysis is to find the linear combination of the x variables which best estimates the value of y for each entity.

The *x* variables are referred to as predictors or explanatory variables (and often, less inappropriately, as independent variables, though they are frequently not independent of one another, either in the colloquial sense nor technically in a statistical sense), whilst *y* is called the response variable or dependent variable. In the context of the main modelling phase of the present study, the entities are either CDs (Ch 4) or pixels (Ch 5), the predictors are spectral values and other measures derived from them, and the dependent variable is a demographic characteristic such as a population density. In the following discussion, the entities are assumed to be pixels.

For a particular pixel, say the *i*th, the estimated value of *y* is given by the regression equation

$$\hat{y}_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} \qquad (1)$$

where b_0 is called the constant term

and b_1 to b_n are called the regression coefficients.

The minimisation criterion which is generally used is the so-called "least squares criterion". The statistic which is minimised is the sum of the squares of the deviations of each $\hat{y}$ from the corresponding *y* value

$$SSD = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2)$$

where *n* is the number of pixels.

The data for a set of *n* pixels may be represented as an *n*-vector **y** and an $n \times (p+1)$ matrix **X**, whose first column entries are all equal to 1 (this corresponds to the constant term) and whose subsequent columns correspond to the *p* variables x_1 to x_p. The vector of estimates $\hat{y}$ is given by the regression equation

$$\hat{\mathbf{y}} = \mathbf{Xb} \qquad (3)$$

where **b** is the $(p+1)$-vector of regression coefficients.

Equation (1) corresponds to one row of the matrix equation (3).

Also, in matrix terms,

$$SSD = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

$$= (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \qquad (4)$$

It can be shown (Myers, 1990) that the vector **b** which minimises *SSD* is given by

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

This vector of regression coefficients characterises the regression model of equation (1) or (3).

Geometrically, the data can be represented as a "cloud" of points in (p+1) dimensional space, one dimension representing y and the other p dimensions representing x_1 to x_p.

For p=1, the data points can be plotted in 2 dimensions and the equations (1) or (3) represent the 1-dimensional "line of best fit" through the 2-dimensional "data cloud".

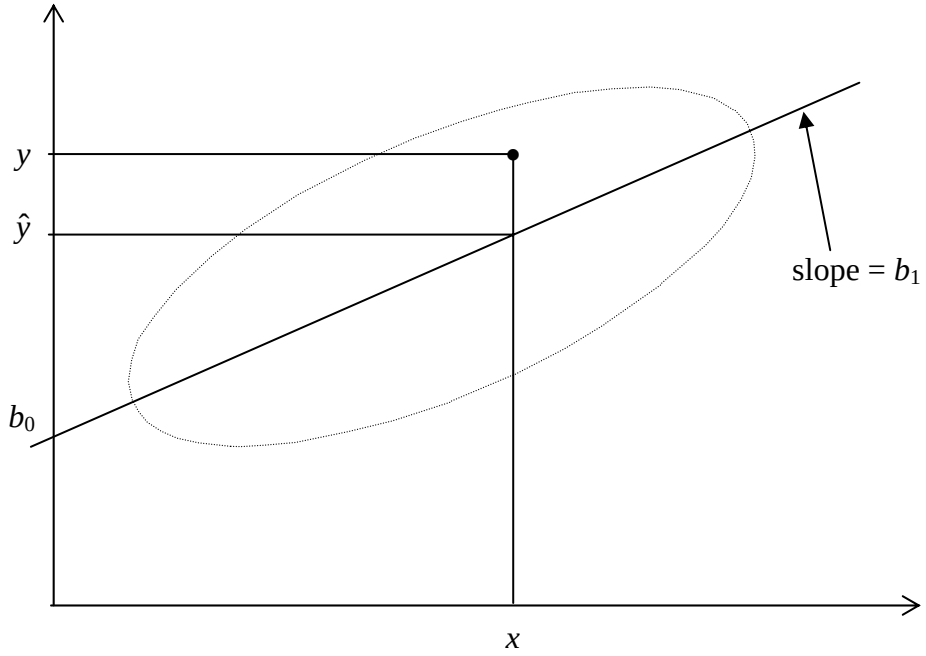
In Figure 2.4 the ellipse represents the region occupied by the data points. The regression line has the equation

$$\hat{y} = b_0 + b_1 x$$

A typical data point (x,y) is shown, as is the regression estimate $\hat{y}$ corresponding to it.

For p=2, we have instead a 2-dimensional "plane of best fit" in 3-dimensional space. In general, we have a p-dimensional hyperplane of best fit in (p+1)-dimensional space.

Figure 2.4 The linear regression function



The above calculations can be performed for any set of data **y** and **X**. Whilst the best estimates of y (in the least squares sense) are thereby obtained, it does not follow that the estimates are sufficiently accurate for the x variables to be regarded as practically useful predictors of y.

The strength of the linear relationship between y and the x variables determines the accuracy of the predicted values. This is usually indicated by using the test data set to calculate the coefficient of determination, R^2. This is the square of the correlation between the $\hat{y}$ values and

the *y* values, which is conventionally expressed as a percentage, and interpreted as "the proportion of the variation in *y* which can be attributed to its relationship to the *x* variables".

The estimates $\hat{y}$ define points on the fitted line, plane or hyperplane. Geometrically, R^2 is a measure of the proximity of the actual data points to this line, plane or hyperplane.

If R^2 is sufficiently large that a useful relationship is regarded as having been established, then the model can be used to estimate or predict *y* for data points (pixels) for which the *x* values are known but *y* is not. This was the central aim of the present study.

Other measures of predictive performance which are based on the *residuals* $(y_i - \hat{y}_i)$ include the standard deviation of the residuals, or *root mean square error* (RMSE), the *mean absolute deviation of the residuals*, and the standard deviation or mean absolute deviation or of the proportional errors (or relative errors)

$$\frac{(y_i - \hat{y}_i)}{y_i}$$

i.e. the *standard deviation of the proportional errors*, and the *mean absolute proportional error* (MAPE). These are discussed further in Section 2.12.

2.8.2 Stepwise regression

The rationale and methodology of stepwise regression are similar to those of stepwise discriminant analysis discussed in Section 2.6.

When a substantial number of potential predictor variables is available for consideration, some will be better predictors of the dependent variable than others. Also, the predictors are likely to be interrelated amongst themselves to some degree. Hence there will be both superfluity and redundancy in the full set of variables. Stepwise regression analysis provides a systematic method for selecting an appropriate, near optimal, parsimonious subset of the variables. In principle, the stepwise regression algorithm proceeds by fitting a sequence of models to the data as follows.

Firstly, *y* is regressed on each *x* variable in turn, and the one which is the best predictor of *y* (using a criterion such as the highest R^2 value), is taken as the starting point of the analysis.

Each of the remaining variables is then paired with the first chosen variable in turn, and the regression equations calculated. The variable which produces the largest increase in R^2 is selected as the second predictor.

The procedure continues with one more variable being included at each step until no significant improvement in R^2 is achieved, at which point the selected set of variables, the final regression

equation and the predictive performance of the model as indicated by the final value of R^2, are known.

It is also possible for a variable included in the model at some stage to become redundant at a later stage, in the sense that its removal would not cause a statistically significant reduction in predictive power. To avoid iterative cycling, it is usual to set the significance probability for such removals somewhat higher than that for entry, typical values being .10 and .05 respectively. These levels were used throughout this study, unless otherwise specified.

In practice the matrix computations are performed incrementally, without the need for full inversion calculations at each trial step.

For a more detailed discussion of the above, see Cliff (1987).

2.8.3 Regression through the origin

In contexts where it is known that the zero points of a dependent variable *y* and a predictor *x* should coincide, a regression line can be fitted using the least squares principle but subject to the constraint that the constant term must be zero i.e. the line must pass through the origin. Whilst such a line may be more appropriate in particular circumstances, it does not usually fit the data as well as the unconstrained line. However direct comparison is complicated by the fact that since the regression through the origin does not in general pass through the mean of the data, R^2 can not be meaningfully calculated in the usual way. A conceptually similar indirect measure can be calculated, but it tends to exaggerate the extent of the reduction in goodness of fit (Myers, 1990). In the present study, regression through the origin was used in the context of comparing CD population density estimates produced by two different methods.

2.8.4 Statistical assumptions, variable transformations and alternative models

Ordinary least squares (OLS) regression is predicated on an assumed additive linear model of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

where the random errors ε_i are independent and identically normally distributed with constant variance. Under these assumptions, the OLS estimation procedure is equivalent to maximum likelihood estimation, and the standard inference based on *t* and *F* tests is valid.

If, either on theoretical grounds or on the basis of empirical evidence, these assumptions appear to be violated - non-linear relationships, non-normal errors or non-constant variance - this

framework can be extended to encompass transformations of both the *y* and or *x* variables, but the fundamental assumptions about functional forms and error structures remain. In particular, the *y* variable is assumed to be measured on a continuous scale.

An alternative approach is the broader class of generalised linear models, in which transformations are treated rather more integrally via the concept of link functions, and which explicitly accommodate non-constant error variances and non-normal error distributions, including discrete distributions like the Poisson distribution for independent count data. Generalised linear models are usually estimated by rather more computationally intensive iterative numerical algorithms for maximum likelihood estimation.

In the context of modelling population, some related conceptual issues arise. Population density is a time dependent and scale dependent concept. As Langford and Unwin (1995) point out, when the size of the areas for which it is calculated are reduced, population density becomes more grainy and variable. The population density of a tower block of apartments is much higher than that of the suburb in which it is located.

With regard to time dependence, instantaneous population density maps of a major city at midday and midnight would look very different. For most demographic and planning purposes (though not for example for the positioning of display advertising), population is assigned by place of residence. Raw census counts are a little anomalous in this regard, being based on the place of census-night accommodation.

In spatial terms, the concept of residential population density changes in character as one reaches the quantum scale of the individual residence. On the scale of CDs, an individual's residence is unambiguously and discretely located within a particular CD. The same cannot be said at the level of pixels, whose boundaries intersect property lines, structures and even rooms.

Thus one can conceptualise either an instantaneous time dependent discrete pixel population, or a notional residential pixel population which is time invariant (for most pixels, on a scale of days or weeks), but which is not discrete.

Poisson regression with an identity link function has been suggested for modelling population dependence on a binary variable such as a classification (Flowerdew and Green, 1989). Poisson regression with a log link has been used for modelling migration between Canadian census divisions with origin and destination populations and distance as explanatory variables (Amrhein and Flowerdew, 1989).

In the context of modelling ED aggregate populations the distinction between Poisson and OLS regressions with a logarithmic transform might be expected to diminish, since for large counts the Poisson distribution is well approximated by the normal. In fact, Langford et al. (1991)

reported very little difference in the coefficients obtained using a linear OLS and Poisson model with identity link.

In the case of modelling the relationship between population and the spectral characteristics of individual pixels, for the Poisson parameterisation to be appropriate then not only must the population of a single pixel be regarded as a discrete count, but also individuals in the population must be assumed to be acting independently in their choice of place of residence, which is not realistic. As Flowerdew and Green point out, the clumping of population suggests a compound or generalised Poisson model.

Alternatively, since a place of residence is not a single point but an extended area which might contribute fractionally to a number of pixels, the "population" of a single pixel can conversely be conceptualised as a continuous variable, in which case a normal error distribution perhaps has more face validity than a Poisson. However, in this case too the independence assumption does not hold for adjacent pixels.

It is certainly the case that in high density areas, the error distribution is positively skewed which *prima facie* is more characteristic of a Poisson distribution. However, this is predominantly a systematic effect associated with multi-level accommodation, which is arguably more appropriately dealt with using ancillary parameters (see Chapter 10).

Leaving aside issues such as independence of the locational behaviour of individuals, the Poisson regression models for population modelling suggested by Flowerdew and Green (1989) and used by Langford et al. (1991), as discussed in Section 2.8.4, would seem to be more congruent with the instantaneous discrete conceptualisation of population. The ordinary least squares models with normal errors, used in by Iisaka and Hegedus (1982), Forster((1980b, 1981, 1983), Langford et al. (1991), Fisher and Langford (1994), Lo (1995), Webster (1996) and Yuan et al. (1997) would seem to be more congruent with the alternative concept of a notional time invariant residential population assigned to each pixel, with individuals in some cases at least contributing fractionally to the notional populations of adjacent pixels.

Throughout the pixel-based phase of this study, discrete ground truth CD populations are distributed amongst constituent pixels in various ways. These imputed populations are not discrete, and are consistent with the latter conceptualisation. Hence, OLS models, sometimes including transformed dependent variables, have been used throughout.

Another aspect of generalised linear models in a spatial context is the capacity to relax the assumption of independence of the error terms for different pixels, and to make explicit provision in the linear model for estimating the nature and degree of spatial dependence between the populations of neighbouring areas. This approach was not adopted in the present study, whose methodology was tailored to the capablities of the available statistical and image

processing software (see Section 3.5 for details of the software used). Instead, the spatial dependence aspect was dealt with through the use of the texture measures described in Section 2.5 and an approach to contextual reclassification (see Section 6.4).

2.9 REGRESSION ANALYSIS WITH INCOMPLETELY DETERMINED DATA

2.9.1 An algorithm for iterative refinement of estimates

The inherent difficulty with the pixel-based approach is that we have no ground truth data for the dependent variable, the population of each pixel. We only have population figures for aggregates of pixels, in this instance for each CD.

In this section a heuristic argument is advanced for a methodology for making initial estimates of the population of each pixel, and then iteratively refining those estimates. This approach could be applied in any multi-level situation, spatial or otherwise, where the dependent variable is incompletely determined in this way (i.e. only constrained to the extent of fixed aggregate subtotals). Essentially, it is a least squares approximation to an EM (expectation-maximisation) algorithm (Dempster, Laird and Rubin, 1977; Lee, 1997), a generic 2-stage iterative approach the use of which has been reported in a number of multi-level analysis and image analysis contexts (Titterington, 1990; Goldstein, 1995). A detailed consideration of the relationship between this iterative refinement algorithm and the EM algorithm can be found in Section 7.2. Whilst this algorithm was developed independently, the author has since become aware that the EM approach was suggested (though not implemented) by Flowerdew and Green (1989) in the closely related context of areal interpolation for combining data from two incompatible sets of spatial zones.

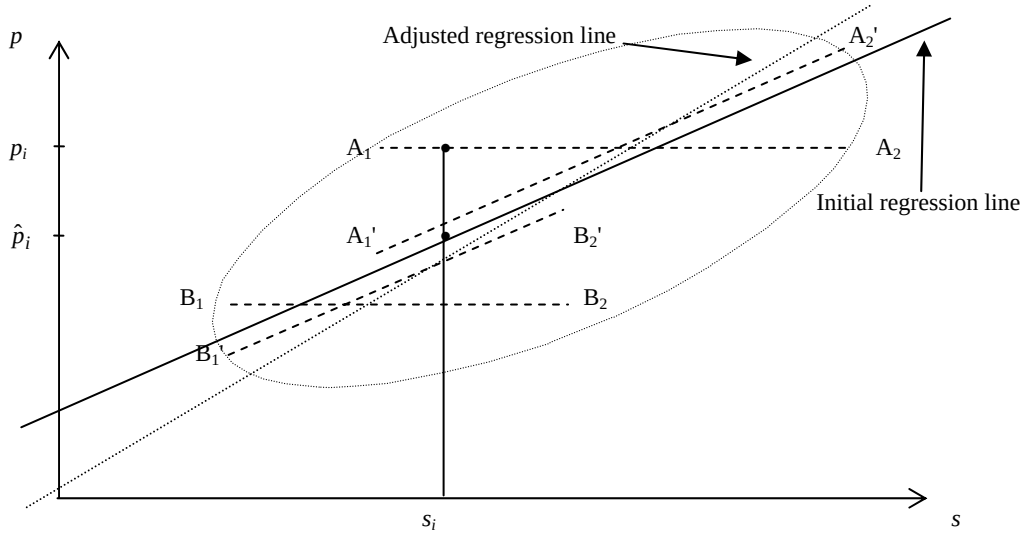
Consider the relationship portrayed in Figure 2.5 between pixel population p and a multivariate vector of remote sensing predictors **s**, represented schematically by a single dimension s. Suppose a particular CD has a ground truth population P, and includes n relevant pixels (those classified as residential). Initially we make the simplest assumption, of constant population density, and assign to each pixel an equal share of the CD population., i.e. all pixels in the CD are assigned the same population

$$p_i = \frac{P}{n} \qquad \text{i} = 1,\ldots, n$$

Since these pixels will in general have different spectral characteristics i.e. different s values, they will be represented in Figure 2.5 by data points along a line such as A_1A_2, parallel to the s axis. For points near A_1, the regression estimate $\hat{p}_i$ will be less than the assigned population p_i (as illustrated in Figure 2.5), and the converse will be true for points near A_2.

We argue as follows. We do not believe that the populations of each pixel are in fact equal. Some of our assigned populations are no doubt too high and others are too low. Assuming that there is in fact an underlying linear relationship between the dependent variable *p* and the explanatory variable *s*, the regression equation represents our best estimate of this relationship based on our partial knowledge about the actual values of *p*. *Prima facie*, it seems that the populations assigned to pixels near A_1 were too high, and those near A_2 too low. If we redistribute population away from those pixels near A_1 and towards those near A_2, whilst maintaining the constant CD total, we can produce a new set of assigned values which are consistent with the known CD totals, but which might be expected to lie closer to the true values. A regression line fitted to this revised set of data might be expected to better represent the true relationship between population and spectral response.

Figure 2.5 Regression with incompletely determined data



Intuition suggests, and it can be shown (see Section 7.2), that the optimal such redistribution in a least squares sense, which minimises the sum of squared residuals about the regression line while holding the sum of the *p*-values constant, is to adjust as follows:

$$p_{i(adj)} = \hat{p}_i + \bar{r}$$

where $\hat{p}_i$ is the regression estimate

$$\bar{r} = \sum_{i=1}^{n} (p_i - \hat{p}_i) \Big/ n$$

This has the effect of making all the residuals equal i.e. mapping A_1A_2 onto $A_1'A_2'$, parallel to the regression line. Similarly the data for the pixels from another CD might be reassigned from the line B_1B_2 onto $B_1'B_2'$, and so on.

We make this adjustment for the pixels within each CD, then re-estimate the regression model. Since the second model should fit the adjusted data better than the first model fitted the initial data, the value of R^2 will be expected to increase. Furthermore, the geometry of the horizontal distributions of points make it likely that positive residuals will predominate for higher values of *s*, and negative residuals for lower values of *s*. As a consequence, the adjusted regression line will be likely to be steeper, as illustrated. This also suggests that the model initially fitted to the averaged data is likely to underestimate the sensitivity of *p* to changes in *s*.

When the process is iterated, R^2 increases monotonically towards some limiting value $R^2_L<1$. A value of $R^2_L=1$ would imply that it is possible to distribute the fixed CD populations amongst their component pixels in such a way that some linear combination of the *s* variables will reproduce them exactly. Whilst this could and would eventually happen if only one CD were involved, with more than one CD any non-linearity in the relationship between the CD totals and the *s* variables precludes it.

When this procedure was first applied in the present study (see Chapter 5), a monotonic increase was observed in R^2. The value of R^2 increased, initially at a rapid rate, but with a steadily decreasing rate of increase, appearing to converge towards a limiting value around .85.

A model thus obtained provides an upper limit to the accuracy of prediction that could be attained were the individual pixel populations known.

As to validity and efficacy, there is no direct way of internally validating the procedure - no way of knowing whether the adjusted pixel populations are more or less accurate than the equal proportions assigned originally. It must be stressed that the increase in "accuracy" as indicated by the increase in R^2 means nothing of itself and may be quite spurious.

So why do it? With each iteration, the regression coefficients change. The hope is that by better representing the populations of individual pixels, we will better estimate the relationships between pixel population and the predictor variables. Ultimately, the validity and efficacy of this procedure, as with the rest of the estimation procedures, can only be assessed by external validation, in terms of the accuracy of the CD aggregate estimates produced for the whole training set, the whole image, and for other images. In the present study, substantial improvements were achieved using this procedure.

A detailed investigation into the properties and characteristics of this procedure and the resulting regression coefficients is reported in context in Chapter 8.

2.9.2 Models with a transformed dependent variable

The iterative refinement procedure of Section 2.9.1 requires some modification before it can be applied in cases where the dependent variable is transformed. If, as in this study, the lowest values of the dependent variable are close to zero (relative to the scale of the data), the redistribution of residuals can lead to adjusted values which are negative. This may be conceptually problematical, as in the case of population, but even so it may still lead to improved (non-negative) estimates of larger aggregates (see Chapter 5). However with transformed data, more immediate practical problems arise. In the present study, square root and logarithmic transformations were used. In each of these cases the range of permissible variable values is restricted. In the case of the logarithmic transformation the untransformed variable must be positive. In the case of the square root transformation the transformed variable must be non-negative, and whilst negative values of the untransformed variable are technically permissible, the one-to-one relationship between transformed and untransformed values (which is essential for backtransformation) is lost. So in both cases, negative values of the untransformed variable are a problem. In the present study, because the constraint on pixel populations is an additive one, it is in the domain of the untransformed (or backtransformed) variable where adjustments are made, and where negative values may consequentially arise.

The problem can be overcome by a relatively minor ad hoc adjustment to the iterative process, the only cost of which is to inject a small arbitrary perturbation into the least squares process. The procedure and the rationale are as follows. At each iteration, the fitted values for each pixel are backtransformed and the backtransformed values are adjusted in the usual fashion. In the context of the present study, a negative population estimate is regarded as being below an indicative threshold, and is therefore readjusted to zero. A compensating readjustment must be made to the positive pixel values, which in the absence of the negative values will sum to a greater figure than the CD total. The sum of the negative contributions is obtained for each CD, and the remaining positive pixel values are reduced proportionately, so that the correct CD total is maintained.

2.10 NON-LINEAR METHODS

2.10.1 Interactive effects

If the relationship between a dependent variable y and a predictor x_1 depends on the value of another predictor x_2 , the two variables x_1 and x_2 are said to interact.

Interactive effects are inherently non-linear, but in linear analyses such as discriminant analysis and multiple regression they can be modelled without departing from the linear form of the model, by the inclusion of product terms such as x_1x_2.

In the more flexible regime of knowledge-based or rule-based expert systems, interactions are the norm. They are explicitly expressed in boolean terms such as

IF x_1 ..... AND x_2 ..... THEN y = ...

The difference between these approaches is illustrated in Figure 2.6.

The inequality

$x_1x_2>1$

defines a region bounded by a hyperbola (dashed line), whereas the boolean statement

$x_1>1$ AND $x_2>1$

defines a related but different region bounded by two lines.

Similarly, the inequality

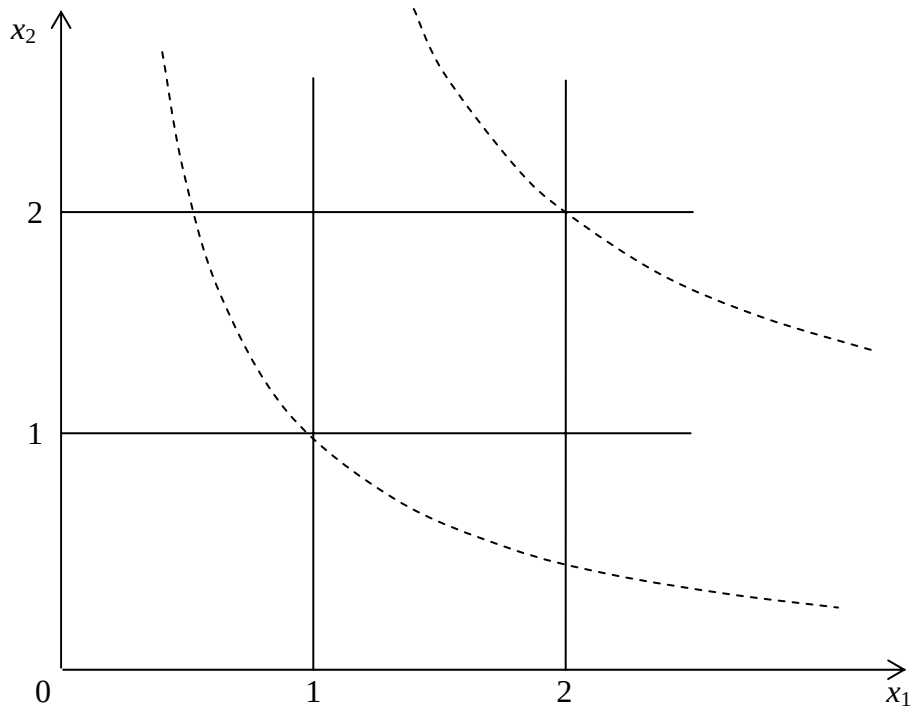
$1<x_1x_2<4$

defines a region bounded by two hyperbolae, whilst

$1<x_1<2$ AND $1<x_2<2$

defines a square subset of that region.

Figure 2.6 Modelling of interaction



In the present study, multiplicative representations of interactions have been included at the classification and regression modelling stages.

2.10.2 Rule-based methods

Researchers such as Wharton (1987), Moller-Jensen (1990) and Bolstad and Lillesand (1992) have investigated the application of rule-based artificial intelligence or expert system techniques to land cover classification. These methods usually involve the boolean combination of remotely sensed data with other available thematic data.

In this study, the thresholding built into the final model implicitly represents a set of rule-based adjustments overlaid on a core relationship which is linear.

2.11 MEASURES OF PERFORMANCE, VALIDITY AND ROBUSTNESS

The discussion in this section is framed in terms of the pixel-based analyses of the primary and secondary images of Ballarat and Geelong described in Chapters 5 and 6, though the regression aspects apply also to the CD-based procedures of Chapter 4, and both aspects apply to the further scope of external validation undertaken in Chapters 8 and 9.

Essentially, the pixel-based estimation algorithms developed in this study consist of two phases:

(i) classification of each pixel as residential or non-residential

(ii) estimation of the population (or number of dwellings) attributable to each pixel, by a multiple regression model.

2.11.1 Internal validity, external validity and robustness

Validation of both classification and regression procedures can be considered at two levels, which are characterised in this report as follows:

1) ***Internal validation*** – Given a "training set" of entities (pixels CDs etc.) on the basis of which some procedure/function of the spectral characteristics is chosen, how well does that procedure/function work for the training set itself?

2) ***External validation*** – How well does the procedure/function perform when applied beyond the training set i.e. how generally applicable is the procedure/function?

Robustness is almost synonymous with external validity, but it carries the connotation of validity over a broader domain. In the remote sensing context a procedure trained on a sample or subset of a particular image may be externally validated by applying it to other samples or subsets from the same image or a similar image. If it also works for other rather different

images it is more likely to be described as robust. Validity and robustness are not absolute terms nor absolutely distinguishable – it is a matter of degrees of generalisation.

2.11.2 Indicators of internal validity

The internal validity of a regression model is indicated by measures such as the coefficient of determination R^2, the standard deviation of the residuals, and various relative error statistics (see Section 2.8).

The internal validity of a classification algorithm can also be measured by R^2 type measures such as Wilks Λ, and also in terms of the percentages of correct classifications in the training set data (the "hit rates") derived from the confusion matrix (see Section 2.6).

Whilst these are often reported at face value, there are a number of statistical issues which should be addressed.

2.11.3 Sample size considerations

In any multivariate analysis, the sample size is an important determinant of both the validity of the models obtained, and also of the sensitivity with which relationships are detected. On the one hand, small sample sizes can lead to “overfitting” or capitalising on chance. On the other hand, automatic model selection procedures used with samples which are very large can result in over-sensitivity and the development of models which are unnecessarily complex. Authors such as Tabachnick and Fidell (1996) have given rules of thumb for determining appropriate sample sizes for multiple regression and multiple discriminant analyses, in terms of the number of predictors and the number of groups. A sample size at least an order of magnitude larger than the number of predictors or groups is desirable if reliable variable selections are to be made.

One of the perennial problems of exploratory statistical analysis is that variables are easier to generate than are cases to test them on. In the present context this is certainly so with the aggregate-based models of Chapter 4, where the number of variables considered is almost as large as the number of CDs in the primary study area. One should be aware that models identified in this way are not likely to generalise robustly.

Pixel-based models do not suffer from such paucity of training observations. The sample sizes in Chapter 5 and beyond were typically in the order of thousands, in line with the principles enunciated by Tabachnick and Fidell.

2.11.4 Sample-dependence of stepwise procedures: external validation procedures

Stepwise algorithms in general can be said to "capitalise on chance". In both discriminant analysis and regression, the final form and the specific detail of the model are chosen to maximise performance for the particular sample of data under consideration. The chosen model is thus biased by the characteristics of the particular data used, and the calculated values of the performance measures are misleading; whilst they provide a measure of internal validity, they are likely to over-estimate the performance of the algorithm when it is applied to other data.

If the statistical population about which inferences are to be made is well defined and homogeneous, and the sample data is broadly representative, then less biased algorithms and more accurate measures of performance can be obtained by cross-validation methods (Myers, 1990). The simplest such approach is to split the test data, usually into two or three sets - one set for selecting the explanatory variables, one for estimating the parameters of the model, and one for assessing its performance.

A more sophisticated regression cross-validation technique involves omitting one data point at a time, fitting the model to the reduced data set, and obtaining an estimate for the omitted data point. This procedure is repeated for each point in turn, leading to a set of so-called "deleted residuals"

$$(y_i - \hat{y}_{i,-1})$$

where $\hat{y}_{i,-1}$ is the estimate of $\hat{y}_i$ from the model determined by all the data except point *i*. Whilst R^2 is based on the ordinary residuals of equation (2) above, the deleted residuals provide the basis for alternative measures which are less subject to the bias discussed above.

However, in the present context, there is no single well defined homogeneous population. Rather, there is a hierarchy of possible generalisations beyond the starting point of the training data. We can consider the applicability of the algorithm to other sections of this particular image of Ballarat, to other "similar" images (e.g. Ballarat on other occasions, other Victorian provincial cities), other "less similar" images (e.g. an Australian capital city metropolitan area), quite dissimilar images (e.g. other parts of regional Australia, other countries) and so on.

In this study, the performance of the algorithm has been evaluated at two levels beyond that of the immediate training set data.

Firstly, performance statistics based on aggregate estimates for the 138 census collection districts (CDs) in the Ballarat image were calculated. These are in effect performance measures based on all pixels in the image.

Secondly, the algorithm was applied, without retraining or modification, to an image of a neighbouring provincial city (Geelong) from the same date, and similar aggregate statistics computed.

These aggregate measures provide the most useful indication of the effectiveness and generality of the estimation procedure. Nevertheless, the various available intermediate performance measures have also been calculated and reported. These measures can be summarised, and the scope of the data on which they were based can be approximately quantified, as in Tables 2.2 and 2.3.

Table 2.2 Scope of Data Sets

Data set	Description	Approximate number of pixels
1	Entire test area of Ballarat image	700,000
2	Residential pixels	70,000
3	Maximum likelihood classification (MLC) training sets	70,000
4	Stepwise discriminant analysis (SDA) test data (1 in 10 sample of MLC training sets)	7,000
5	Stepwise regression analysis (SRA) test data (1 in 50 sample of residential pixels)	1,400
6	Test area of Geelong image	500,000

Table 2.3 Summary of Performance Measures

Stage	Measure	Data set	Scope
SDA	Hit rate	4	7,000 pixels
MLC	Hit rate	3	70,000 pixels
SRA	R^2	5	1,400 pixels
Aggregates	R^2, slopes, relative errors	2,1	138 CDs - effectively the full set of residential pixels, or the full image
Geelong	R^2, slopes, relative errors	6	225 CDs - the full image

2.11.5 Multicollinearity

The outcome of a multiple regression analysis, both with regard to which explanatory variables are selected and with regard to the estimated regression coefficients of the selected variables, is most sample-dependent when there is a substantial degree of correlation among the explanatory variables. Geometrically, the spread of data in multivariate space is not sufficiently broad to support a consistent orientation of the hyperplane of best fit[1]. Regression equations resulting from different training samples may have very different patterns of coefficients and even utilise

[1] Myer (1996) uses the analogy of a planar object balanced on a picket fence. Bob Dylan's image is more graphic: "It balances on your head just like a mattress balances on a bottle of wine – your brand new leopard skin pillbox hat" (in this case, a line of wine bottles!).

different subsets of predictors. Under these circumstances individual regression coefficients must be interpreted with even more caution than usual, since the notional varying of one explanatory variable whilst holding all others constant cannot occur in practice.

Individual correlation coefficients very close to ±1 are a sufficient but not necessary condition for multicollinearity to be present. A more reliable indication is a measure called the variance inflation factor (VIF), which is based on the multiple correlation coefficient between each explanatory variable and the rest. VIFs greater than 10 are generally regarded as indicating some cause for concern about multicollinearity (Myers, 1990).

When, as in the present study, the main motivation for the regression analysis is prediction, multicollinearity may or may not be problematical to this aim. If different samples have individual multicollinearity structures which are not reflected in the population, then estimates for points in the population whose combination of explanatory variable values are not represented in the sample (away from the wine bottles in a perpendicular direction) will have large variances. But if multicollinearity exists in the population and is well represented in the sample data, accuracy of predictions will not be affected. Regression equations resulting from different training samples may have very different patterns of coefficients and even utilise different subsets of predictors, and yet produce quite consistent estimates.

The simplest remedy for multicollinearity is to omit the affected variable(s). However, there is a trade-off. Deletion of selected variables will inevitably result in some reduction in overall predictive power of the model.

In the present study, there was evidence of multicollinearity in all of the regression models for pixel population, particularly involving the visible bands. TM band 2 was most consistently affected with VIFs typically in the range 15-25, whilst bands 1 and 3 each had VIFs around 10 in one or more images. Whilst the general pattern was consistent across all images, the Sydney and Brisbane images exhibited higher levels of multicollinearity than the more southerly images, with bands 5 and 7 also having VIFs around 10. However, since for each image the multicollinearity was consistent across samples, and the sample covariance matrices were similar to those of the whole image, it was considered that in each case the samples reflected the underlying multicollinearity in the population. For this reason, and because omission of TM band 2 resulted in considerable reduction in R^2, it was decided to retain this variable in the suite of predictors.

2.11.6 Type I error rates in stepwise analyses

Another related problem with stepwise regression is that, whilst the *p*-to-enter value may be set at say .05 for each step, as the number of available candidate variables increases, so too the

probability that at least some of the selected variables will be chosen in error due to chance relationships, becomes much greater than .05.

This is analogous to the well established concept of an experimentwise type I error rate, associated with multiple comparisons in the analysis of statistical experiments.

In the regression context, a correction factor is routinely applied to R^2 to make adjustment for the number of terms included in the model (adjustment for degrees of freedom). In the pixel-based models in the present study, the sample sizes (and total degrees of freedom) are generally so large that such adjustments are of little consequence.

However, the very fact that *df* are large allows for the testing of many variables and many transformations. In this situation, the chance selection of spurious variables is increasingly likely. External validation of models is again the safeguard.

2.11.7 Modified performance measures with transformed dependent variables

Transformation of the dependent variable in a regression analysis is a common response to nonlinearity of relationships or violation of statistical assumptions by the data. The following example illustrates the fact that when the principal aim is estimation rather than investigation of structural relationships, the resulting improvements in model fit indicated by the standard measures may be illusory.

Consider a logarithmic model of the form

$$\log\hat{(}y_i) = b_0 + \sum_{j=1}^{p} b_j x_{ij}$$

for which R^2 =.901. (See Table 4.2, Section 4.1)

R = $\sqrt{.901}$ = .95 is the correlation between the observed and fitted values of log(*y*).

However it is *y*, not log(*y*), that we wish to estimate. Of more concern is the correlation between the observed values *y* and the estimates $\hat{y}$ obtained from $\log\hat{(}y)$ by an exponential back-transformation, which in this instance is .84. The square of this correlation (.704), which can also be interpreted as the R^2 value obtained from the regression of *y* on $\hat{y}$, gives a more meaningful indication of the estimation performance of such a model than does the R^2 from the transformed model.

Similar calculations can be applied to RMSEs, mean absolute proportional errors, etc.

Measures based on back-transformed estimates have been quoted wherever appropriate in this study.

2.12 ISSUES OF PARAMETRISATION, PRESENTATION AND EVALUATION

2.12.1 Population density vs. total population

As was pointed out in Section 1.2.4, when building regression models based on geographical aggregates of unequal area, the question arises as to what is the most appropriate dependent variable. If the explanatory variables are aggregate measures such as average spectral characteristics or proportions of pixels in different classes, then the natural dependent variable is population density. If the explanatory variables are pixel counts, then the natural dependent variable is the total population of the aggregate. If the aggregated areas are equal, as with grid squares, the distinction is immaterial.

In the first phase of this study (see Chapter 4), the geographical basis was census collection districts (CDs) and the explanatory variables were average spectral characteristics. Hence population density was used as the dependent variable.

An alternative would be to regress total CD population on the aggregate measures weighted by CD area, but it was considered that in a mixed urban-rural context, this might have the undesirable result of producing very positively skewed marginal distributions for the explanatory variables, with a few very large low density rural CDs unduly influencing the outcome of the regression analysis.

2.12.2 Presentation and evaluation

In the second phase (see Chapter 5 et seq.), regression models were fitted at the scale of individual pixels. Since all pixels in an image are (nominally) the same size, the total vs. density issue does not arise. However, for purposes of validation, the estimated pixel populations were aggregated to CD level and compared with the ground truth CD totals. This raises a number of issues which are now considered.

Orientation of plots

To maintain consistency throughout, ground truth values were plotted on the vertical (Y) axis and remote sensing estimates on the horizontal (X) axis.

Measures of accuracy, consistency and bias

First, a word about the usage in this report of two terms which have quite precise technical meanings in the context of mathematical statistics, but which can also be used in less technical discourse to refer to the same broad concepts.

The estimated values of some quantity, such as the population densities of a number of CDs, may be inaccurate in two ways. They may consistently underestimate or consistently

overestimate the true values, in which case they are said to be *biased*. If not, the estimates and the procedure that generated them can be said to be unbiased.

Whether or not there is bias present, estimates for individual CDs may vary above or below the true values, to a degree which may be large or small in comparison to any bias present. This is referred to as *variability* or conversely *consistency*.

The portmanteau term *accuracy* covers both aspects. Accuracy implies both consistency and lack of bias. Inaccuracy may be due to either bias or variability, or both.

CD aggregate measures for pixel-based models

Consider a plot of some ground truth data for a set of CDs, *g*, (which may either be a total or a density), vs. a remote sensing estimate of it, *r*.

Suppose that *r* is the set of fitted values from a linear regression model fitted to *g*, as is the case with the many of the models of Chapter 4. Then if *g* were to be regressed on *r*, the OLS line of best fit would necessarily have zero intercept and unit slope, and measures such as R^2 would have the same values as in the original regression. Furthermore, the residuals from the second regression are the same as those from the first, i.e. they represent the errors of estimation.

Such an analysis would be redundant, uninformative and pointless. But if *r* is the result of some other less direct estimation process, as in the models with transformed dependent variables in Chapter 4, the pixel-based procedures of Chapter 5, or the external validation of a regression equation on second set of data, then regressing *g* on *r* provides new information about how well the estimation algorithm can recover the CD data. Goodness of fit criteria include a high R^2 value (in the case of the models with transformed dependent variables these are the backtransformed values discussed in Section 2.11.7), an intercept near zero (in relative terms) and a slope near unity, regardless of whether or not the line is forced through the origin (see Section 2.8.3). A slope other than unity is an indication of *bias* in the estimation procedure, whilst R^2 is a measure of *consistency*. In particular, it is quite possible to obtain a high R^2 with CD estimates which are consistently off target.

Note however, that such a secondary regression analysis is indicative only and is part of the validation process – not the estimation process. The remote sensing estimates are the *r* values which have already been obtained – not the fitted values which result when the ground truth *g* values are regressed on the *r* values. A corollary to this is that the estimation errors are the differences *r*-*g*, not the residuals from the regression of *g* on *r*.

(Note that in the linear CD aggregate based models, these quantities are identical. Furthermore, in this case the issue of bias does not arise, since the residuals are constrained to sum to zero. In

this context, R^2 is a reasonable measure of overall accuracy. This is not so when reporting CD results for pixel-based models, or results for other validation sets.)

Population density vs. total population

The two types of plot throughout this report, based on CD population densities and CD populations (see for example Figure 6.3, Section 6.3), each have a characteristic feature, which is to some degree an artifact of the nature of census CDs. On both types of plot, the main body of points is scattered about a positively sloping line indicative of positive correlation between the ground truth figures and the remote sensing estimates. On the population density plots, the outlying points are generally above and to the left of the linear "main sequence". These represent CDs whose ground truth population density is substantially underestimated. These are generally CDs with high population densities. Because CDs are designed to have roughly similar populations, such CDs are usually small in area. Whilst the relative error of underestimation may be large, so long as estimates are constrained to be non-negative it cannot exceed 100%, and so the absolute error in population cannot exceed the CD population.

On the population plots, the most extreme points are generally below and to the right of the linear "main sequence". These represent CDs whose ground truth population is substantially overestimated. These are generally large CDs with low population densities. Whilst an overestimated population density may still be low in absolute terms, when it is leveraged by the large area involved, the resulting error in population may be very large. There is no upper limit to overestimation.

Underestimation is not so noticeable on population plots, because it is relatively small in magnitude and generally occurs near the origin. For the same reason, overestimation is not so noticeable on population density plots.

Regression models fitted to CD population densities have different characteristics from those fitted to CD populations. In particular they are less sensitive to the effects of overestimation in large low density CDs. Conversely, regression models fitted to CD populations are less sensitive to the effects of underestimation in small high density urban CDs.

Because of the urban focus of most of the interest in population estimation, most models in this report are fitted to CD population densities.

It is arguable that in the context of population estimation, the proportional errors in the estimates may often be of more interest and importance than their absolute magnitudes. Relative error measures are the same for both population density and population. Therefore, such measures are sensitive to both types of discrepancy.

Relative errors

The relative or proportional error of estimation

$$RE = \frac{r - g}{g} \times 100\%$$

is a measure of performance which is identical for both population and population density.

Since the errors of estimation for individual CDs may be positive or negative, averages of the absolute errors or the relative errors give indications of any consistent bias in the estimation process. But even in the absence of bias, estimates for individual CDs may vary above and below the true values. Averages of the absolute (unsigned) values, either of the absolute errors or the relative errors (one inevitably runs into some terminological inconsistency here), are indicators of the overall accuracy of estimates, which as well as variability or consistency between CDs, may or may not also include bias effects. Two such measures have been used in this report: the *mean absolute proportional error* (MAPE – referred to by Lo, 1995, as "absolute mean relative error"); and because it is less susceptible to the inflationary influence of a few extreme outliers, the *median absolute proportional error*, i.e. the proportional error which is exceeded in half of the CDs. For brevity, these statistics are referred to as *mean relative error* and *median relative error* respectively.

Bias is addressed by calculating (signed) relative errors for the total populations of whole regions and the urban sections of regions. Accuracy in estimating these totals is of course an important objective in its own right.

2.13 SUMMARY

This chapter has provided a general orientation to multispectral remote sensing imagery and an outline of the theoretical bases and technical methods which are used throughout the study.

Some key issues included:

- the tone-texture or spectral-spatial dichotomy of multispectral remote sensing imagery;
- mathematical transformations that can be applied to remote sensing data in both the multivariate spectral domain and the two dimensional spatial domain;
- statistical methods for classification of the pixels of an image;
- statistical methods for modelling the putative relationship between population and remote sensing indicators;
- the assessment of performance, validity and robustness of models, and in particular the use and interpretation of R^2 and relative error measures in various contexts;

- associated issues of parameterisation and presentation.

In the next chapter, the specific data sets used in the study and the preliminary preparation of the data are described, together with an outline of the computational methods used to implement the analyses.

Chapter 3

Data Preparation and Integration

3.1 INTRODUCTION

This study involved the integration of two types of data from different sources: ground-based demographic data and satellite-based reflectance data. In this chapter, the sources and specific details of both sets of data are given, and the methods used to prepare and integrate them are described.

Section 3.2 introduces the six areas and seven images used in the study. Full details of the first two study areas are also given here; more detail of the remaining five areas is given in context in Chapter 8. Sections 3.3 and 3.4 give details of the ground-based demographic data and the satellite-based reflectance data respectively. Section 3.5 outlines the computing methods used to implement the integration of the two sets of data and the subsequent analyses.

Sections 3.6 and 3.7 are concerned with the temporal and spatial aspects of data integration. Section 3.6 outlines the methods used to estimate ground truth populations at the dates the images were acquired. Section 3.7 concerns the spatial alignment of the remote sensing imagery with the ground-based census geography.

A number of variations in methodology alluded to in Section 3.7 came about because the work was carried out over a number of years. Advances in hardware and software capability and changes in data formats and availability meant that different methods were employed for similar tasks at different stages of the work. Some tasks such as initial data acquisition and co-registration which had to be performed in a painstaking, time-consuming first-principles fashion for the primary image became routine or trivial at later stages.

3.2 THE STUDY AREAS

Six mixed urban/rural areas of Australia were employed in the study, surrounding and including the provincial cities of Ballarat and Geelong, the state capital cities of Sydney, Brisbane and

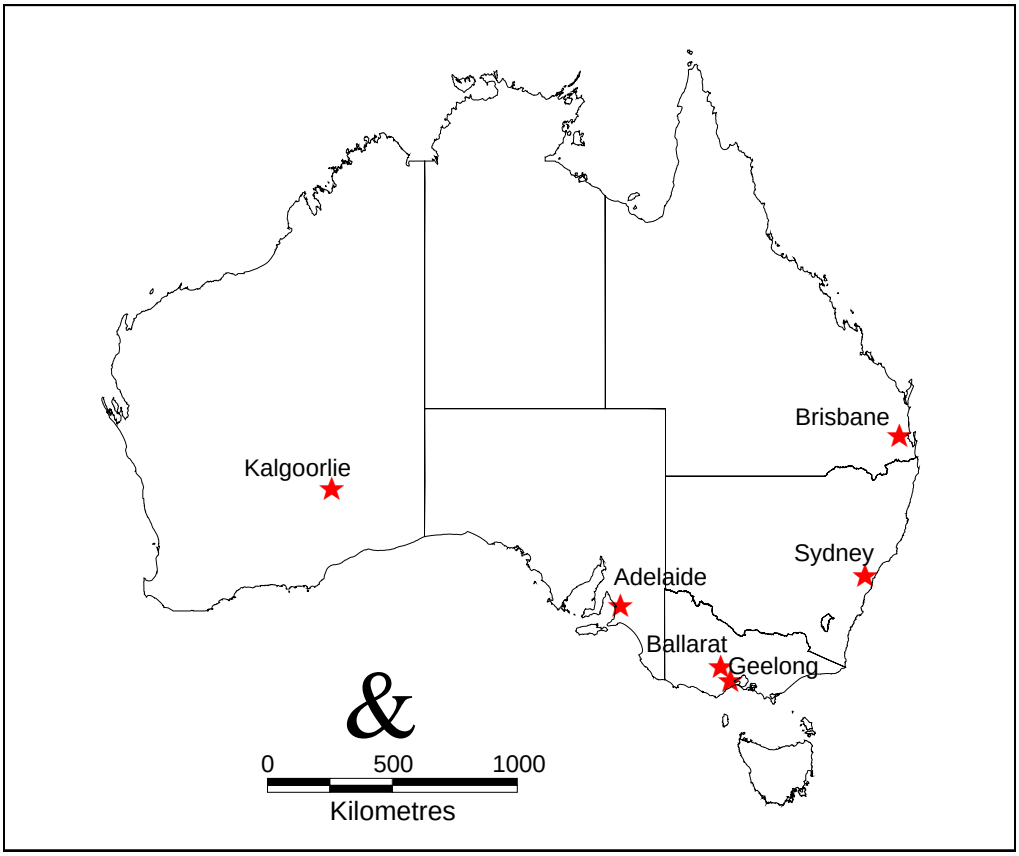
Adelaide (population rankings 1, 3 and 4 amongst Australian cities), and the remote mining centre Kalgoorlie. Figure 3.1 and Table 3.1 show the locations and some basic characteristics of the study areas.

Table 3.1 Study Areas

Name	State	Area (sq. km.)	Population	Year
Ballarat	Victoria	634	79,179	1988
Ballarat*	Victoria	199	35,711	1994
Geelong	Victoria	352	147,910	1988
Adelaide	South Australia	10735	1,158,625	1997
Sydney	New South Wales	3524	3,283,889	1996
Brisbane	Queensland	4623	1,488,880	1989
Kalgoorlie	Western Australia	62	30246	1989

*The 1994 image of Ballarat included only part of the original 1988 study area

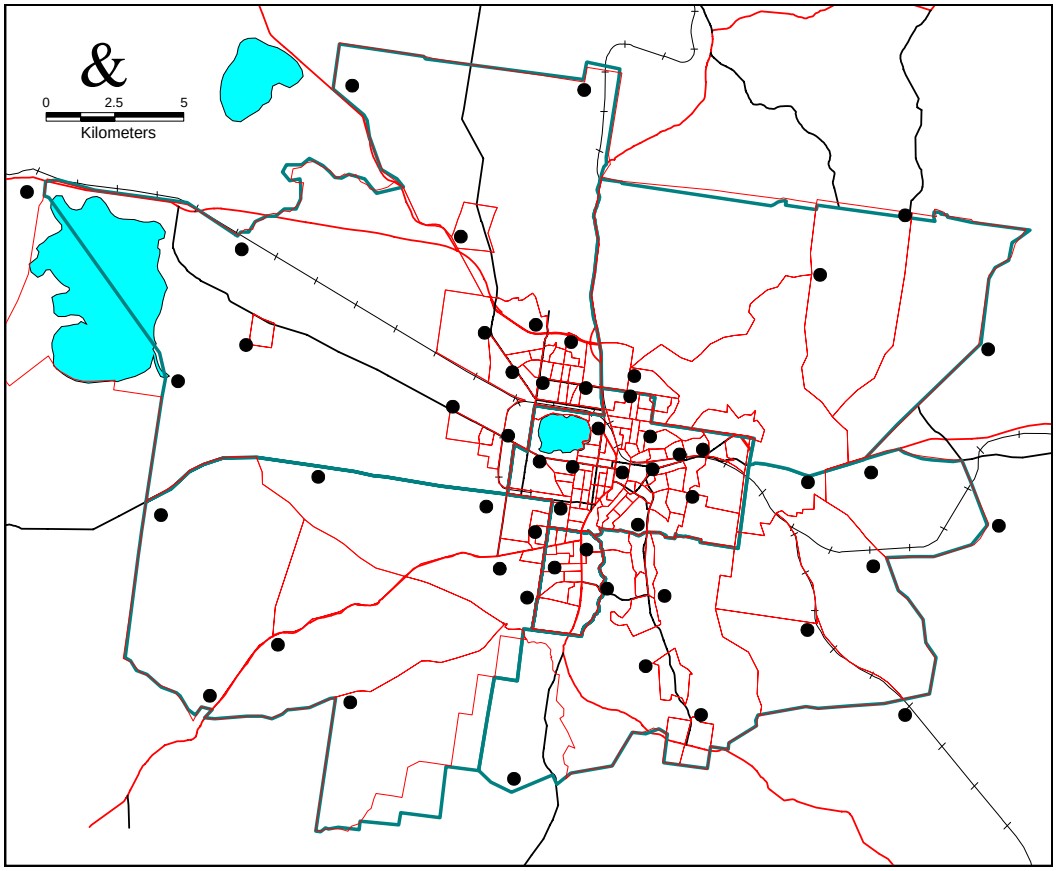
Figure 3.1 Study Areas



3.2.1 The primary study area

The primary area selected for study was Ballarat Statistical District (BSD), an inland region of some 634 sq. km. in extent, centred on the provincial city of Ballarat, 110 km west of Melbourne, Victoria, Australia.

Figure 3.2 Ballarat Statistical District
Showing major roads, SLA and CD boundaries and ground control points



Source: ABS

BSD, as defined by the Australian Bureau of Statistics (ABS), comprises the Ballarat urban area, which is delineated using criteria based on population density, together with a surrounding rural area which is expected to encompass urban expansion over an extended period.

BSD encompasses six Statistical Local Areas (SLAs), which correspond to legal Local Government Areas (cities, shires, boroughs, etc.), or parts thereof. The six SLAs which made up BSD included two central urban LGAs in their entirety, and the urban and near-urban sections of the four surrounding, predominantly rural shires.

For census purposes, each SLA is further subdivided into Census Collection Districts (CDs). For the 1986 Census of Population and Housing, BSD comprised 138 CDs, of which 122 were classified by the ABS as urban. The main criterion for classifying a CD as urban is an average population density of 200 persons per sq. km. or more, supplemented by contextual rules aimed at reducing fragmentation (ABS, 1998). The urban CDs comprise the Ballarat urban area (118 CDs) and four outlying satellite suburbs or townships, each of which consists of a single CD.

The estimated population of BSD in 1988 was 79179, of which the urban area contributed 70222.

The statistical structure of BSD at the time of the 1986 census is summarised in Table 3.2. Figure 3.2 shows 1986 SLA boundaries and 1996 CD boundaries.

Table 3.2 Statistical structure of Ballarat Statistical District in 1986

Statistical Local Area (SLA)	Number of 1986 Census Collection Districts (CDs)		
	Urban	*Rural*	*Total*
City of Ballaarat	72	-	72
Borough of Sebastopol	11	-	11
Shire of Ballarat (Part A)	27	3	30
Shire of Bungaree (Part A)	2	4	6
Shire of Buninyong (Part A)	7	6	13
Shire of Grenville (Part A)	3	3	6
Total BSD	122	16	138

3.2.2 The secondary study area

The secondary study area was Geelong Statistical District (GSD), a similarly mixed urban/rural area of some 352 sq. km. in extent, centred on the port city of Geelong, 90 km south east of Ballarat.

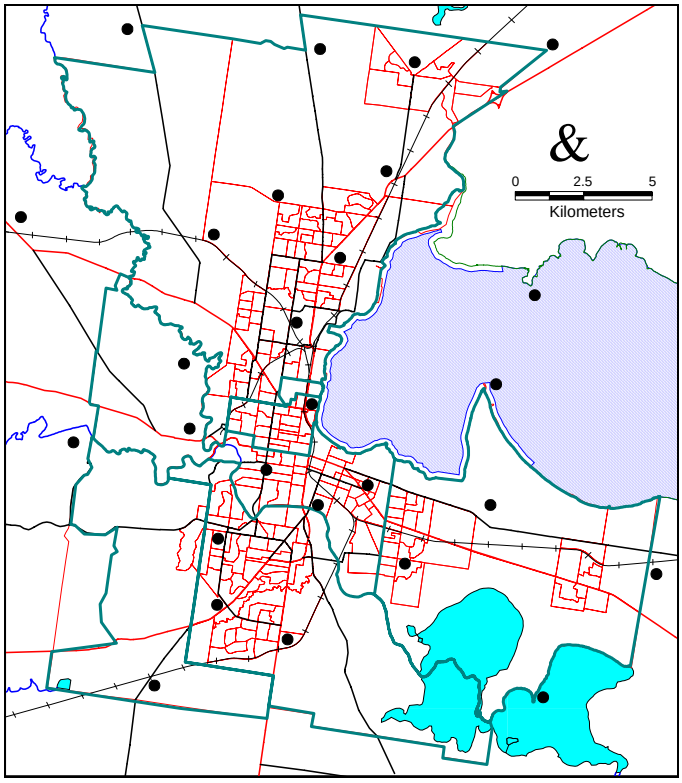
The estimated population of GSD in 1988 was 147,910 to which the urban area contributes some 142,250. The GSD comprised eight SLAs, which for the 1986 census were divided into 225 CDs, of which 218 were classified as urban. These include the Geelong urban area and two outlying satellite suburbs or townships, each of which consists of several CDs.

The statistical structure of GSD is summarised in Table 3.3. Figure 3.3 shows 1986 SLA boundaries and 1996 CD boundaries.

Table 3.3 Statistical structure of Geelong Statistical District

Statistical Local Area (SLA)	Number of 1986 Census Collection Districts (CDs)		
	Urban	*Non-urban*	*Total*
City of Geelong	25	-	25
City of Geelong West	27	-	27
City of Newtown	19	-	19
City of South Barwon	48	1	49
Shire of Bannockburn (Part A)	-	2	2
Shire of Barrabool (Part A)	-	1	1
Shire of Bellarine (Part A)	24	4	28
Shire of Corio (Part A)	71	3	74
Total GSD	214	11	225

Figure 3.3 Geelong Statistical District
Showing major roads, SLA and CD boundaries and ground control points



Source: ABS

3.2.3 Further study areas

In the latter phase of the study, four further areas were involved, three of them being much more extensive regions centred on large capital cities, and also part of the Ballarat study area was re-examined on a different date. CD boundary maps and a comparative summary of all study areas can be found in context in Section 8.2. See also TM images 15, 17, 19, 21, and 23.

3.3 POPULATION AND RELATED DATA

3.3.1 Population estimates

Three types of population estimate are published by ABS.

Census counts by place of enumeration

ABS conducts a five-yearly Census of Population and Housing, the most recent for which detailed data was available being that of June 30, 1986. Population counts based on people's actual location on census night are published for all levels of geographic aggregation down to the lowest level, the CD.

Census counts by place of usual residence

These are census counts adjusted to take account of people who are not at their place of usual residence on census night. These people are omitted from the count in the SLA in which they were enumerated and added to count for the SLA corresponding to their home address. These adjusted counts provide a better estimate of resident population, but are not available for CDs - only for SLAs and larger geographical aggregates.

Estimated resident population

These estimates are produced annually for SLAs and larger geographical aggregates but not for CDs. In census years, the census counts by place of usual residence are adjusted upwards for the effects of temporary absences overseas. Adjustment is also made for census under-enumeration, the magnitude of which is estimated by an intensive post-census sample survey. In the intervening non-census years, mathematical models are used to estimate population changes from the census baseline, employing a range of statutory data such as births, deaths, school enrolments, building approvals and commencements, and so on. After each census, the estimates for the previous intercensal period are revised.

3.3.2 Dwelling count estimates

Census counts

Dwellings are enumerated in the five-yearly ABS Census within four categories:

* Private dwellings

* Unoccupied private dwellings

* Caravans etc. in caravan parks

* Non-private dwellings.

In all but the first category, enumerated units correspond to distinct physical structures. However, the private dwelling category generally predominates. In this category, a dwelling is regarded as the space occupied by a household (see Appendix D for detailed definitions). There may be one or more dwellings contained within a physical structure, be it a house, a block of flats, or whatever. Each private dwelling is categorised according to the type of structure which contains it. Counts of each dwelling category, and of each structure category for private dwellings, are published for all levels of geographical aggregation.

Other potential sources

Databases pertaining to property ownership, valuations, land use zoning, building regulations, utilities and services, etc., contain relatively current and comprehensive information about the

existence, location and status of dwelling structures. However, such data is neither integrated nor readily accessible at this time.

3.3.3 Census collection district boundaries

Digitised 1986 census boundaries of the CDs in the two study areas were obtained from the Australian Survey and Land Information Group of the Department of Administrative Services (AUSLIG). The data was in vector polygon form, consisting of an ordered sequence of vertex co-ordinates for each CD. The co-ordinates for each point were latitude and longitude expressed to 5 decimal places (in the order of 1m).

For the supplementary study areas, 1996 census boundaries expressed in AMG co-ordinates were obtained from CDATA96 (ABS, 1997).

3.4 LANDSAT THEMATIC MAPPER DATA

The primary and secondary images

The satellite data used in the earlier phases of the study consists of two subsets of the Landsat Thematic Mapper (TM) scene path 93 row 86F, of February 14, 1988. Each Landsat TM pixel corresponds to a 30m square on the earth's surface. Since the satellite travels in an approximate N-S direction and the TM sensor makes sweeps perpendicular to the satellite's path (Lo, 1986, p.31), images are oriented approximately N-S/E-W.

The raw Ballarat subscene comprised a rectangle of 1350 pixels (40.5 km) east-west by 1008 pixels or rows (30.2 km) north-south which was skewed as part of the rectification process into a parallelogram within a 1412 pixel × 1008 pixel rectangle .

The raw Geelong subscene comprised a rectangle of 900 pixels (27 km) east-west by 1008 pixels or rows (30.2 km) north-south, which has been skewed and rotated as part of the rectification process into a parallelogram within a 1119 pixel × 1174 pixel rectangle.

Images 1 and 13 are quasi-natural colour images of the two areas. Image 2 is a green-enhanced quasi-natural colour image of the primary study area.

Supplementary images

The images used in the later phases of the study were rectangular subsets of larger images which in all but one case had already been rectified to Australian Map Grid (AMG) co-ordinates.

The details of these subscenes are listed in Table 3.4.

Table 3.4 Specifications of Supplementary Images

Image	Size (pixels/line ×lines)	Pixel size (m)	Date of acquisition	TM path/row
Ballarat	616 × 697	30	15/12/94	93 86
Adelaide	5010 × 6187	25	2/2/97	97 84
Sydney	2740 × 3678	25	8/12/96	89 83
Brisbane	2965 × 3616	30	16/9/89	89 79
Kalgoorlie	1201 × 1078	30	27/9/89	109 81

3.5 COMPUTING METHODS

Most image analysis was performed using ER Mapper Versions 2.0, 3.0 (1991), 3.1 (1992), 5.0 (1995) and 5.2 (1997) using an X-terminal linked to a Sun SPARCstation 2. Additional user-defined code for ER Mapper was written in C. Some preliminary data preparation was done using MicroBRIAN (1988).

Maps of the study areas were produced using CDATA96 (ABS, 1997) census mapping software and Mapinfo 4.1 (1997) GIS software. Mapinfo was also used to locate ground control points for the secondary study area.

Notwithstanding the rapid advances in hardware and software capability, it remains the case, as has been recently discussed by Mesev (1998), that proprietary remote sensing and GIS software packages are quite limited in their capacity to undertake anything but the most straightforward statistical analysis. In this study, most statistical analysis was performed offline using Excel (1990-97), Minitab (1989-97), SPSS-X (SPSS Inc, 1988) and SPSS for Windows (SPSS Inc, 1993-98).

A range of incompatible native data formats for both raster and vector data was involved, and handshaking did not always proceed seamlessly, necessitating the development of both systematic interchange routines and ad hoc patches. Pascal programs were written to enable data sampling and data interchange between the image analysis programs, which use band-interleaved-by-line (BIL) binary format for raster data, and the spreadsheet and statistical programs, which most conveniently import data in ASCII format. Pascal programs were also written for: geometric correction of the primary image; pre-processing and co-registration of the CD boundary data for the primary study area; the simulation study of texture measures; systematic manipulation of vector data files; the initial implementation of the iterative regression algorithm.

The final implementation of the iterative regression algorithm and the associated simulation study was via Minitab macros, which were also used extensively during model testing and evaluation. Excel was also used extensively for data interchange, and Excel macros were used

in the post-processing and summarising of the results of regression simulations. ERMAPPER vector and raster dataset header files, vector data files and algorithm files were also manipulated extensively to achieve objectives beyond the standard capabilities of the graphical user interfaces.

A detailed breakdown of the computational steps involved in the project, and the technical details of their implementation, are summarised in tabular and schematic form in Appendix C.

3.6 ESTABLISHING GROUND TRUTH POPULATION DATA

Because the date of acquisition of the satellite data for the primary and secondary study areas, 14/2/88, was some 19 months after the census of 1986, it was decided to bring together all available information in order to estimate the population of each CD as of that date.

Published population and dwelling data as described in Chapter 3 were obtained for the CDs and SLAs in the primary study area using the Supermap (1988) census data retrieval system, and a number of printed sources (Ballarat and Western Victoria Regional Information Bureau, 1989, Australian Bureau of Statistics 1987, 1989, 1990). These were used as the basis for calculating population and dwelling estimates as at 14/2/88, the date of acquisition of the satellite data, for each of the 138 CDs in the primary study area.

CD population data is only available as raw counts by place of enumeration from 5-yearly censuses - in this case data from 1981 and 1986 was used. SLA data is available from the censuses and from the ABS estimated resident population (e.r.p.) series.

The procedure essentially involved three phases.

Firstly, the 1986 CD e.r.p.s were estimated by comparing the 1986 SLA e.r.p.s with the corresponding 1986 SLA census counts, and applying the resulting SLA differentials to the counts for the CDs within each SLA.

Secondly, the intercensal rate of population change for each CD was compared with that of the SLA it lies within. This differential was then used, together with the annual SLA e.r.p. figures, as a basis for extrapolating the e.r.p. of each CD beyond 1986.

It was observed that when the resulting CD estimates were summed for each SLA, the total exceeded the estimate for the SLA as a whole. It was then proved that the estimation procedure has a small systematic bias towards over-estimation. Consequently, a final adjustment was made to ensure that the CD estimates for each SLA summed to the SLA estimates. See Appendix D for a fuller explanation of the methodology and details of the calculations.

The same approach was used in the secondary study area.

As for the supplementary areas, because detailed 1986 and 1991 census information at CD level was not available to the author for regions outside Victoria, and because the scale and complexity of such adjustments would render the task very much more difficult for large and unfamiliar urban areas, attempts were made to obtain images acquired close to the census of August 1996, with mixed success. Of the large supplementary study areas, two images (Adelaide and Sydney) were acquired soon after the 1996 census (within 4 and 6 months respectively). Another image (Ballarat) was acquired 18 months before the 1996 census but included no areas of rapid change. It was decided that these images would serve the purpose of the study quite adequately without detailed adjustments of the sort carried out on the primary and secondary study areas. The two remaining images obtained, those of Brisbane and Kalgoorlie, were acquired in 1989, seven years before the only census data available to the author. Some analysis was nevertheless carried out on these images, but their usefulness was limited because of the temporal mismatch.

3.7 RADIOMETRIC AND GEOMETRIC CORRECTION OF THE IMAGE

Remote sensing images in their raw form are not located relative to any standard frame of reference, such as latitude and longitude or Australian Map Grid (AMG) co-ordinates. They are also subject to various distortions in both spectral and spatial domains. Corrections for radiometric distortion may or may not be made, but geometric correction or rectification and registration of an image is always necessary if it is to be integrated with ground-based data, as was the case in this study.

3.7.1 Radiometric Correction

Radiometric distortion refers to any lack of correspondence between the measured brightness values of the pixels in a scene and the true brightness. Richards (1986, p.33) distinguishes two broad categories. Firstly, within a particular band the relativity between image brightness and scene brightness can vary from pixel to pixel over the image. Secondly, within a particular pixel the relativity between image brightness and scene brightness can vary from band to band. Both effects can arise either as a result of atmospheric conditions or of sensor characteristics.

Of the instrumentation effects, the most potentially serious is within-band variation caused by differences in response between parallel detectors, of which Landsat TM has 16 per band. This problem is indicated by visible horizontal striping of an image. In the present study, no striping was evident in any band of the data, and so no corrective steps were taken.

The effect of the atmosphere is to scatter radiation, and hence obscure some of the detail in the image. Different atmospheric conditions in different parts of a scene will result in within-band variations, the most extreme being localised total obscurity due to clouds, smoke etc.

Two broad mechanisms of atmospheric scattering are distinguished: Rayleigh scattering, due to the air molecules themselves, and Mie scattering, which involves larger particles such as those of dust, smoke, fog and cloud. Both mechanisms are wavelength dependent, the shorter wavelengths being the more scattered in each case. This dependence leads to a lack of calibration between bands. One important way in which this is manifested is a brightening of the darkest pixels (such as deep water), but to a different degree in each band, leading to a non-zero threshold brightness in each band, whose magnitude is inversely related to wavelength.

Explicit correction for these effects requires detailed information about atmospheric conditions (temperature, relative humidity, pressure, visibility) which were not available in the present study.

However, implicit corrections for the threshold or haze effect may be made based on statistical analysis of the image. Whilst more elaborate algorithms have been employed (see for example Lavreau, 1991), a first order approach is to assume a constant haze noise level within each band and to correct for it by subtracting the dark threshold value from all data values within each band. This procedure was adopted in the present study, the thresholds for the six bands of the primary image being 45,14,10,4, 1 and 1. It is worth noting that whilst such an affine transformation has an effect on non-linear derived measures such as band ratios, it has no substantive effect on linear functions of the bands.

This "haze removal" procedure is distinct from the "histogram equalisation" or "stretching" procedures employed for enhancing the visual contrast of images. In this study, "95th percentile equalisation", in which a linear transformation is applied such that the data value at the 5th percentile is set to zero and that at the 95th percentile is set to the maximum value of 255, was frequently employed for contrast enhancement of displayed images.

An important difference is that stretching produces the same standardised brightness range in all bands, whereas haze removal retains the information about the relative brightness ranges in the different bands.

3.7.2 Approaches to rectification and registration

Changes in software capability and available data formats during the life of the study impacted greatly on the tasks of image rectification and registration.

At the time of the initial processing of the primary image, the image data was available only in raw form with no georeferencing, the 1986 CD boundaries were available only in (latitude, longitude) form, no GIS software was available and the available remote sensing software had no warping (rectification, registration, resampling) capability. It was decided to leave the raster image unchanged and co-register the CD vectors to it using parametric adjustments followed by a polynomial warp fitted from first principles using offline statistical processing.

Some time later, when the secondary image was prepared, the remote sensing software had a warping capability, but the CD boundaries were still in the same form. In this case, the CD vectors were transformed from first principles into approximate AMG co-ordinates, and the raster image warped in the more usual manner to co-register with the CD boundaries.

Later again, when the supplementary images were processed, most of the images were supplied already approximately registered to the AMG grid, and the 1996 CD boundaries were available in a GIS in AMG co-ordinates, so that final adjustment to the co-registration was a trivial task. The one image which was in raw form was registered to the CD vectors in the routine manner.

The remaining sections of this chapter pertain mainly to the procedures used for the primary image.

3.7.3 Sources of geometric distortion

Richards (1986, pp.43-50) discusses the following seven sources of geometric distortion of remotely sensed images:

- aspect ratio distortion
- earth curvature
- panoramic distortion
- variations in platform altitude, velocity and attitude
- sensor scan nonlinearities
- earth rotation effects
- scan time skew

The first of these refers to the fact that some sensors, such as Landsat MSS, produce pixels which do not correspond to a square region on the ground. However, a Landsat TM image has an aspect ratio of 1, since the along-scan sampling interval and the scan line spacing are equal, nominally both 30m.

Earth curvature and panoramic effects cause the instantaneous field of view (IFOV) represented by a pixel to vary in both size and shape across the image. The narrow swathe of Landsat TM ensures that both of these effects are negligible. Since the study area was west of the satellite nadir, the maximum panoramic distortion would occur at the western edge of the image, where it was estimated using the methods of Richards (1986) at less than 1%.

Variations in platform motion are not considered to be a major problem, since corrections for these effects are applied to Landsat data before distribution (Richards, 1986, p.48).

Sensor scan nonlinearities may arise from non-uniform velocity of the scanning mechanism. This is not considered to be a major problem in the case of the TM's oscillating mirror arrangement, which is designed so that the acceleration associated with the endpoints of the oscillation takes place beyond the range of data acquisition.

Whilst none of the above is regarded as serious in its own right, together they will always cause some departure from ideal image geometry. The method of warping polynomials, which was used to register the CD boundaries to the image (see Section 3.7.6), also implicitly provides an ad hoc correction for these effects.

The most serious geometric distortion in TM images, which cannot be corrected in this way, is a skewing effect caused by the finite scan rate of the sensor and the rotation of the earth during each scan.

The TM has 16 parallel sensors for each band, so that on each scan it images a strip of 16 rows of pixels. During the time taken to scan one such strip and position for the next, the satellite moves forward along its orbit a distance corresponding to 16 pixels or 480m on the earth's surface, and also the earth's surface rotates some distance from west to east. These motions have three consequences. The first is a slight skewing of each pixel and each row of pixels due to the forward motion of the satellite. The second is a slight compression of each pixel in the east-west direction relative to the nominal dimension of 30m. These effects are minute, and can be disregarded. The third, far more substantial effect is that each strip of 16 rows in the image is incrementally displaced towards the west on the ground. Using the methods of Richards (1986), the magnitude of this displacement at Ballarat was calculated as approximately 26m, or just under one pixel width.

When a raw TM image containing any regular geometric features is displayed as a rectangle, then each strip of 16 rows is quite visible. The successive displacement of each strip to the right produces a skew towards the east as the image is traversed from top to bottom.

Because of its discrete stepwise nature, this distortion cannot be corrected by a continuous mathematical transformation such as a warping polynomial. However, an approximate correction method is readily available, which does not require pixel resampling but which

slightly over-corrects. This is to displace each successive strip of 16 rows in the image by one pixel to the left. In this study, such a correction was made programmatically to the raw TM data files, by inserting appropriate blocks of null values.

An eighth potential source of geometric error discussed by Forster (1980a) is a topographic effect - a displacement of pixels in the scan direction due to differences in surface elevation within the scene. In the present study, this effect would be most pronounced in the western section of the image, which is furthest from the satellite nadir. This area happens to be a lava plain with little topographic variation. Calculations based on the methods of Richards (1986) confirmed that topographic distortion was not a major problem in this study, having a maximum in the order of 6m in the urban area and perhaps 10m around a few high points in the eastern half of the image. Some correction for this effect was of course also implicitly incorporated in the polynomial transformation.

3.7.4 Methods of co-registration

An adjunct to rectification is co-registration, in which the remote sensed image is aligned with a map projection or with another raster or vector dataset (in this instance the CD boundaries).

In the absence of a developed co-registration capability in the software available to the author at the time, co-registration of the primary image and CD boundaries was carried out from first principles. There are three possible approaches. The parametric approach involves mathematically modelling the motion of the satellite and the formation of the image. The method of warping polynomials is a statistical approach which produces an optimal polynomial transformation of a specified order without explicitly modelling the relationships or mechanisms. The third approach is to combine aspects of both methods. Whilst Trinder (reported in Forster, 1980a) suggests that all three methods lead to similar results, others (microBRIAN Version 2.2, 1988) have recommended the third approach, by first explicitly modelling those aspects of the transformation which are well defined, then applying a warping polynomial to complete the task, the rationale being to simplify the required polynomial as far as possible. It was decided to adopt such a hybrid approach.

When an image is registered to a standard co-ordinate system, the original pixels no longer align with the co-ordinate grid. This necessitates resampling, in which the original pixel values are processed using one of a number of standard algorithms, to produce estimated values for a new set of aligned pixels. The new pixel values are derived either by averaging or interpolating over a neighbourhood, or by selecting the nearest neighbour and using its data. In the case of the primary image, rather than resample, it was decided to register the vector CD boundaries to the row and pixel co-ordinates of the raster image, rather than the more conventional registration of

the image to the map co-ordinates. The following sections describe the steps in co-registration of the CD boundaries to the skew-corrected primary image.

The secondary image was co-registered to the CD boundaries at a later time with the benefit of substantial software enhancements. It was initially skew-corrected, then co-registered to the CD boundaries in the conventional manner, using a cubic warp and nearest neighbour resampling, which has the advantage for the purposes of the present study of preserving the band-to-band relativities (ERMapper 5.0, 1995, p 390).

3.7.5 Explicit Parametric Transformations

Transformations were applied to the CD boundary co-ordinates to shift the reference origin, to correct the aspect ratio, to align the co-ordinate frame with that of the TM image, and to rescale.

Reference Origin

The origin for latitude and longitude is the intersection of the equator and the Greenwich meridian. This was shifted to the point 37.5° S 143.8° E, within the study area, by the transformation

$$L' = L - 143.8$$

$$l' = l - 37.5$$

where L = longitude (east)

l = latitude (south)

Aspect Ratio

Latitude and longitude are directly proportional to distances N-S and E-W respectively. However, whilst the constant of proportionality for latitude does not depend on longitude, the proportionality constant for longitude does vary, being itself proportional to the cosine of the latitude. Hence anywhere but on the equator, a particular difference in longitude represents a smaller distance than a numerically equal difference in latitude. This aspect ratio was corrected by the transformation

$$L'' = L' \cos l$$

Alignment

Landsat 5 follows an approximately circular orbit whose plane intersects the longitudinal plane at an angle of 8.2°. (In fact the sun-synchronous orbit precesses by 360° per year, or approximately 1° per day, or 0.07° per revolution). From this it was calculated that the nominal path of Landsat 5 at Ballarat (latitude 37.5° S) is inclined at 10.36° to N-S. The origin for row

and pixel counts is at the top left corner of the image, which is consistent with the east-south sense of the longitude and latitude co-ordinates. Alignment with this frame of reference required only a clockwise rotation of 10.36°, by means of the transformation

$$X = l' \sin\theta + L'' \cos\theta$$

$$Y = l' \cos\theta - L'' \sin\theta \quad \text{where } \theta = 10.36°$$

Scale

Since 1° of latitude corresponds to about 110 km, the scale change

$$x = 100X$$

$$y = 100Y$$

results in a scale on which 1 unit represented about 1.1 km, or 37 pixel widths. This was done to reduce the scale mismatch with the pixels, thereby avoiding numerical problems at the next stage.

The overall transformation to this point was

$$x = 100((l\text{-}37.5)\sin 10.36° + (L\text{-}143.8)\cos l \cos 10.36°)$$

$$y = 100((l\text{-}37.5)\cos 10.36° - (L\text{-}143.8)\cos l \sin 10.36°)$$

3.7.6 Warping Polynomials

To complete the transformation from latitude and longitude to row (R) and pixel (P) co-ordinates, polynomials of fifth degree in x and y were fitted to the data from 59 ground control points, using standard least squares techniques.

Ground Control Points

Ground control points (GCPs) were selected using an enhanced quasi-natural colour RGB image (R = band 3, G = band 2 + band 4, B = band 1) of the study area. The 59 points selected were predominantly road/road or road/rail intersections, with a few involving creeks or fence lines. They were chosen on the basis of even distribution, good image definition, and location on CD boundaries, for which accurate co-ordinates were known. A higher concentration of GCPs was selected in the more densely populated urban areas where CDs are smaller and accurate registration was most critical. The distribution of GCPs is shown in Figure 3.2.

Of the 28 points outside the main urban area of Ballarat, 18 lie in a ring on or close to the boundary of the study area, and 10 lie in a band midway between the urban area and the outer ring. The spacing of these rural GCPs is typically 5-10 km.

There are 31 points in the main urban area. About half are on or near the periphery and the rest on a rough grid with a spacing of approximately 1.5 km, or 50 pixels.

Choice of warping polynomials

The method of least squares, or multiple linear regression, was used to determine an appropriate pair of polynomials which fitted the data as closely as possible i.e. which mapped the (x,y) co-ordinates of the 59 GCPs close to their observed line and pixel (L,P) locations in the image.

In the standard regression terminology, there were two dependent variables, L and P. The candidate predictor variables in each case were a constant term and all powers and cross-products of x and y up to and including 5th order - 21 terms in all.

If it were reasonable to assume that the underlying relationships between the two sets of co-ordinates are functionally simple, but that there may be a substantial random measurement error component in the L and P values of the GCPs, then it would be appropriate to select a minimum set of predictors by a procedure such as stepwise regression, the general principles of which have been outlined in Section 2.8.

If, on the other hand, the pattern of image distortion is smooth but more complex, and the measurement errors relatively small, then it is appropriate to proceed to more complex models regardless of the statistical significance of the extra terms. The tests of significance will err conservatively under these conditions, because the residuals of the simpler models will contain a substantial "lack of fit" component as well as random error. The aim here is not to be parsimonious, but to model the contortions of the complex response surface as closely as possible.

The logical conclusion of this line of reasoning is to just fit a model containing all 21 terms. In practice, with high order polynomial models, the degree of correlation between the terms is such that one or more terms are likely to be almost exactly linearly dependent on the others. Including all terms in the regression model can lead to numerical instability and erroneous results, particularly at the extremities of the image.

In this study, because the CD boundaries generally followed recognisable features in the image, it was possible to visually evaluate the performance of both types of model for the boundary data as a whole rather than just at the GCPs, and hence decide the most appropriate strategy.

Firstly, the standard stepwise procedure was used, with F-to-enter set at 4.0 (approximately corresponding to p=.05), to derive a minimal polynomial for each of L and P. This is referred to as Model 1.

Secondly, the stepwise procedure was rerun with the statistical inclusion criterion greatly relaxed (F-to-enter = 0.1, p=.75), with the result that the only terms excluded were those which

were extremely highly correlated with terms already in the model, and which might lead to numerical instability. The result was Model 2.

Finally, as a check, forced multiple regression was used to fit polynomials which contained all 21 terms. This was Model 3.

The standard errors of the three models are compared in Table 3.5.

Table 3.5 Comparison of Warping Polynomials

Dependent variable	Number of terms in model	Standard Error			
		All GCPs (n = 58)		*Urban GCPs (n=31)*	
		Pixels	*(metres)*	*Pixels*	*(metres)*
Model 1					
Line	9	1.02	(30.6)	0.74	(22.2)
Pixel	5	1.25	(37.5)	0.67	(20.1)
Model 2					
Line	16	0.70	(21.0)	0.45	(13.5)
Pixel	17	0.76	(22.8)	0.47	(14.1)
Model 3					
Line	21	0.71	(21.3)	not calculated	
Pixel	21	0.83	(24.9)	not calculated	

Clearly, the more complex polynomials of Model 2 produced a better fit to the GCPs. Models 1 and 2 both performed better in the urban area, where the density of GCPs was higher. As expected, Model 3 provided no improvement over Model 2. Indeed, numerical instability actually led to an increase in the calculated standard errors.

Models 1 and 2 were then applied to the CD boundary co-ordinates and the results overlaid on the Landsat image and examined visually.[1]

If the residuals in Model 1 had been predominantly due to random measurement error, then the improvement in fit attained by Model 2 at the GCPs would not have been maintained across the image as a whole.

In fact, with the exception of two sparsely populated areas at the north-east and south-west extremities of the study area, the fit of all the CD boundaries was uniformly improved by Model

[1] During the visual examination, some localised discrepancies were discovered which were clearly due to errors in the CD boundary co-ordinates as supplied. The positions of the small rural townships of Miners Rest and Cardigan Village (CDs 80 and 81) had been translated some 2 rows up and 3.5 pixels to the left, presumably as the result of inadvertent movement during the digitising process. A compensating adjustment was made to these co-ordinates before final co-registration.

2. This indicates that the discrepancies in Model 1 were largely due, not to random noise, but to a smooth pattern of distortion which was better modelled by the more complex polynomials of Model 2. Accordingly, Model 2 was adopted.

The final equations for the warping transformation were:

$$L = 296.057 - 0.05858x + 36.0084y + 0.062483x^2 + 0.030666y^2$$
$$+ 0.03483xy - 0.004692x^3 - 0.0044253xy^2 + 0.005852x^2y - 0.000199x^4$$
$$- 0.000811x^2y^2 - 0.0001203x^3y + 0.00001818x^5 + 0.00003441x^2y^3$$
$$+ 0.00006346x^3y^2 - 0.00004086x^4y$$

$$P = 616.896 + 37.0608x - 0.5989y + 0.019426x^2 - 0.00717y^2 - 0.07719xy$$
$$- 0.002171x^3 + 0.010388y^3 - 0.003914xy^2 + 0.004066x^2y$$
$$- 0.00011504x^4 - 0.0012794y^4 + 0.0004581xy^3 - 0.00042159x^2y^2$$
$$+ 0.00032904x^3y + 0.00000274x^5 + 0.00004608y^5$$

For purposes of display, the transformed boundaries were overlaid in vector form on the TM image. For purposes of analysis, they were also used to define 138 regions on the image, and hence, using the ER Mapper IF INREGION( ) function, a data band containing the CD identification of each pixel was also defined. This band formed the essential link between the remote sensing data based on pixels and the ground truth data based on CDs.

3.8 SUMMARY

In this chapter, the methods used to acquire, prepare and integrate ground-based demographic data and satellite-based reflectance data have been described.

The six study areas and seven study images were introduced, as was the variety of software used to implement the integration of the two types of data and to carry out the analyses.

A substantial proportion of the chapter was devoted to what now might seem rather tortuous and primitive approaches to image rectification and co-registration. Changes in available data formats and software capabilities throughout the period of the study impinged particularly in this area, which became progressively more straightforward as the study progressed.

This concludes the introductory and preparatory phases of the thesis. The work proper begins in Chapter 4 with an account of investigations into extensions and enhancements to the aggregate-based methodology of Iisaka and Hegedus (1982).

Chapter 4

Estimates Based On Census Collection District Aggregate Measures

4.1 INTRODUCTION

The geographical basis of the investigations reported in this chapter is the Census Collection District (CD). Regression analysis was performed on data for the 132 CDs in the primary study area.

Whilst Iisaka and Hegedus (1982) had used linear combinations of MSS band averages across 500m grid squares, it was decided to examine a broader range of possible relationships by incorporating a variety of standard spectral and spatial transformations of the TM data, including a number used by Forster (1980b, 1981, 1983) in the context of urban land use.

Section 4.2 is concerned with models based on band averages, together with squares, cross products and ratios of those averages. In Section 4.3, measures of spatial variation in TM reflectances across the CD are incorporated. In Section 4.4, a selected set of spectral transformations made at the individual pixel level are introduced (though the basis of their selection is reported in context in Chapter 5); means and spatial variation measures of each of these measures are incorporated into the regression models. In Section 4.5, the same sequence of models is applied to the estimation of dwelling densities and counts.

Throughout this section of the work, for reasons which are explained in Section 4.2, two variants of the basic form of model were also considered, with logarithmic and square root transformations being applied to the dependent variable (the population or dwelling density).

In Section 4.6, all of the models tested are comparatively evaluated. Six models were chosen for external validation on the secondary image. (The results of this are reported in Chapter 7.)

4.2 POPULATION DENSITY ESTIMATES FROM AVERAGE REFLECTANCES

Initially, collection district population density was regressed on average reflectances in the 6 TM bands across each CD.

This was done for two reasons: firstly, to establish whether results similar to those reported by Iisaka and Hegedus (see Section 1.2) could be obtained by a similarly basic procedure, and secondly, to establish a benchmark against which the performance of more refined procedures could be evaluated.

Using the CD identification data band (see Section 3.7), the mean reflectance in each TM band was calculated for each of the 138 CDs. Population density *D* (persons/sq.km) was calculated by dividing the ground truth population estimate *P* by the CD area.

Population densities ranged from 4.5 to 5142 persons/sq.km., with a mean value of 1556.9 persons/sq.km. (see Appendix E). The lowest densities were generally found in the large rural CDs. With regard to the CD average reflectances in the 6 TM bands, bands 1, 2, and 3 were all positively correlated to a substantial degree. Among the other bands there was a chain of moderate to high correlations but no clear clustering. The strongest correlations with population density were band 5 (-.52), band 4 (-.40) and band 1 (+.34). This suggests that the presence of people is weakly associated with a tendency for high reflectance in the short visible (blue) wavelengths, perhaps associated with paved surfaces and some roofing materials, and relatively low reflectance in the near infrared, associated with vegetation (or relative lack of it).

A multiple regression analysis was performed using the 6 TM band reflectances as predictors. The saturated 6-variable model had $R^2 = .539$. Stepwise selection resulted, with very little reduction in R^2, in the following 4-variable model (with variables in order of entry):

$$\hat{D} = 72.3 - 135.6\ b5 + 332.0\ b7 - 151.0\ b3 + 61.6\ b4$$

with $R^2 = .537$ $\quad s = 763.8$

The same procedure was also applied to the 122 urban CDs only, yielding

$$\hat{D} = -48.9 - 203.5\ b5 + 365.5\ b7 + 114.7\ b4 - 99.7\ b1$$

with $R^2 = .459$ $\quad s = 760.5$

As is often the case, the best multivariate set of predictors is not entirely made up of variables which are individually most correlated with the dependent variable. The residual standard deviations are similar for both models. The lower value of R^2 in the second case can be attributed to the reduced range of population densities when the rural areas are excluded.

Comparison with Iisaka and Hegedus' reported R^2 values (.59 and .70 for two different years) confirmed the expectation that population estimation might be more difficult in a mixed urban/rural area in regional Australia. There is perhaps a greater degree of local heterogeneity within CDs even in many urban sections of a provincial city, than would be expected in the suburban areas of a metropolis such as Tokyo.

An obvious problem was that the values of the dependent variable ranged over 3 orders of magnitude. An analysis of the residuals ($D - \hat{D}$) revealed pronounced positive skew, variance increasing with increasing *D*, and some evidence of a concave upward curvilinear trend (see Appendix E), all of which are commonly associated with a large range in the dependent variable. A standard corrective approach in such circumstances is to transform the dependent variable by taking the logarithm or the square root. Both of these transformations were investigated.

A second approach to the representation of non-linear relationships within a linear framework is to transform the explanatory variables. The pool of potential predictors was enlarged by applying to the CD means a range of transformations: firstly the squares of the 6 basic band means, then the 15 band mean to band mean cross-product terms, the 15 pairwise band-to-band ratios, and finally the 15 pairwise difference-to-sum ratios.

Models involving these variables were fitted to the whole data set and also to the urban subset. As the aim was to encompass both rural and urban areas, and as it seemed that the models performed no better when applied to the urban areas alone (see Table 4.2), analysis of the urban subset was discontinued, and all subsequent analyses were done on the full data set.

In the context of additive linear modelling, difference-to-sum ratios have an inherent additive symmetry, whilst band ratios do not. Because of this, the 15 reciprocal band ratios were also investigated, but they were found to have little effect on the results reported below. This is probably because the range of variation in each ratio was quite limited, so that any reciprocal relationships were adequately represented to first order accuracy by negative linear terms.

The full set of variables, and the models selected by stepwise regression, are summarised in Tables 4.1 and 4.2. It is important to realise that the variables *pij*, *rij* and *dij* are functions of means, not means of functions; each variable is the product or ratio of mean values derived from CD aggregate figures, rather than the mean of a product or ratio calculated for each individual pixel. The latter approach is considered in Chapter 5.

Table 4.2 shows that, beginning from a base R^2 value of around .5, the incorporation of squares, cross-products or ratios increased this to around .7. Application of either the square root or the logarithmic transformation to D resulted in a further increase in R^2 to within the .8 to .9 range.

Table 4.1 Summary of Regression Variables

Generic name	Number of variables	Description
bi	6	mean of TM band i
si	6	square of bi
pij	15	cross-product bi×bj
rij	15	ratio bi/bj
dij	15	difference-to-sum ratio (bi-bj)/(bi+bj)
Total	42	i, j = 1,2,3,4,5,7

Many of the potential predictors were quite highly correlated, so that the number of variables retained by the stepwise procedure was reasonably small, ranging from 1 to 8. Also in most cases the specificity of the chosen set of variables is not high; using "best subsets" regression it is possible to select alternative sets which perform almost as well. For example in the basic 4-band model, band 1 or band 2 can replace band 3 without much loss, since all three are highly correlated.

Table 4.2 Population Density Models based on CD Means
Summary of Stepwise Regression Results

Potential Predictors (number)	Urban area only (*n*=122)		Urban and rural areas (*n*=138)					
	Dependent variable *D*		Dependent variable *D*		Dependent variable $\sqrt{D}$		Dependent variable ln*D*	
	Selected Predictors	R^2	Selected Predictors	R^2	Selected Predictors	R^2 (R^2_b)	Selected Predictors	R^2 (R^2_b)
bi (6)	b5 b7 b4 b1	.459	**b5 b7 b4 b3**	**.537**	**b5 b7 b4 b3**	**.652 (.557)**	b5 b7 b3 b2	.684 (.343)
bi si (12)	s1 b1 s7 b4 b5	.618	s5 b7 s1 b5 b4 s4 s7	.735	**s1 b1 s7 b4 b5 s4**	**.847 (.755)**	s5 b7 s1 b1 s7 b5	.861 (.687)
bi si pij (27)	s5 p47 b5	.501	s5 p47 p57 p45	.730	s2 b4 p24 s7 b5	.845 (.757)	s5 p14 b4 s1 p47	.855 (.656)
bi si pij rij (42)	r23	.351	r57 r14 r37	.696	r57 r15 r17 r47	.844 (.757)	p45 s1 r17 p35 p37 p13 s3 r37	.901 (.704)
bi si rij dij (42)	r23	.351	r57 r14 d35 d14 d37 r25	.753	**r57 r14 d14 r15**	**.846 (.762)**	r13 r14 r15 s4 s1 s7	.910 (.719)

Note:
- Models selected for further investigation are shown in boldface.
- Predictors in each model are listed in the order of selection.
- Parenthesised R^2_b values are based on back-transformation (see Section 2.9). They are the R^2 values obtained when the dependent variable *D* is regressed on the estimate of *D* (obtained by inverse transformation of the regression estimates for the transformed *D* values), and hence provide a more realistic indication than R^2 of the predictive accuracy of the model.

Two points are noteworthy. Firstly, except in the case of the logarithmic models, when band ratios (varibles whose names begin with "r" or "d") were included they completely displaced the other predictors (see last two rows of Table 4.2). Secondly, band 2 only appears relatively infrequently in Table 4.2. It seems that in the multispectral context, visible green may be the least discriminating spectral signature of human habitation.

The residuals from these models were examined in the light of the demographic characteristics of the individual CDs (Ballarat and Western Victoria Regional Information Bureau, 1989). The residual distributions tended to be positively skewed, with the extreme positive values being consistently associated with the same few CDs. The CD whose population was consistently underestimated by the greatest amount contains a large multi-storeyed geriatric institution. The populations of 6 CDs consisting predominantly of high density public housing were also substantially underestimated throughout. Conversely, whilst overestimation was not so extreme in most models, populations did tend to be overestimated in older established areas where there are relatively high proportions of small households in disproportionately large houses.

Whilst the R^2 values obtained for these refined models were encouraging, there are a number of reasons for circumspection. Firstly, in models with a transformed dependent variable, the R^2 value exaggerates the precision of estimation (see Section 2.11.7). In this situation, R^2_b values based on back-transformation give a more meaningful indication of comparative predictive performance. Values of this statistic are included in parentheses in Table 4.2 and subsequent tables. In the case of the square root transformation they are in general marginally higher than the R^2 values for the corresponding untransformed model, and in the case of the logarithmic transformation they are substantially lower. This indicates that whilst transformation may lead to a more appropriate from of model, the resulting increase in R^2 may be illusory, in the sense that the population estimates produced by the transformed models are not substantially more accurate, and in some cases less so.

Secondly, even though increases in R^2 have been obtained, with values around .75 (i.e. with correlations in the order of .87), the uncertainty range associated with population estimates for individual CDs is still quite substantial.

Finally, as with any stepwise regression procedure, there is the aspect of capitalisation on chance (see Section 2.11.4) to consider. These models were generated using data from the entire primary test image, with no external validation. The results obtained when selected regression models were applied to another image are reported in Chapter 7.

The fitted values and residuals from the three types of model were examined graphically (see Appendix E). On the basis of goodness of fit, parsimony (number of predictors), and the statistical characteristics of the residuals (normality, homogeneity of variance, non-random

patterning) it appeared that the square root transformation was consistently more appropriate than the logarithmic transformation. This conclusion was borne out in the more complex models which followed.

Further sets of explanatory variables were then examined. At each stage, one or more models were selected as the best of each class. From Table 4.2, four models were retained for further examination. The first square root transformation model was selected as representing a more statistically appropriate base model than the untransformed base model. The regression equation (with terms in order of selection) was:

$$Est(\sqrt{D}) = \text{ - 1.4 - 2.33 b5 + 5.37 b7 + 1.22 b4 - 2.04 b3}$$

with $R^2 = .652$ and $R^2_b = .557$

The other four square root transformation models represented a considerable improvement over the base model. There was little to separate them. Two were chosen for further consideration: the second, based on band means and their squares, and the fifth, based on ratios of band means. The regression equations (with terms in order of selection) were:

$$Est(\sqrt{D}) = \text{ - 171.34 - 0.140 s1 + 9.220 b1 + 0.0344 s7+ 4.874 b4 - 1.952 b5 - 0.0359 s4}$$

with $R^2 = .847$ and $R^2_b = .755$

and

$$Est(\sqrt{D}) = \text{ 345.35 - 68.41 r57 - 275.56 r14 + 226.89 d14 + 120.30 r15}$$

with $R^2 = .846$ and $R^2_b = .762$

4.3 POPULATION DENSITY ESTIMATES: MEASURES OF SPATIAL VARIABILITY

Because it was conjectured that inter-pixel variability may provide key indicators of population density, the variability throughout each CD was calculated for each TM band. The measures used were variance, standard deviation and coefficient of variation (standard deviation/mean).

Table 4.3 Variable Suffix Nomenclature

Suffix	Meaning	Example
none	mean	b5
s	standard deviation	b5s
c	coefficient of variation	b5c
v	variance	b5v

The results of stepwise regression analyses using the 6 band means and the 18 (3×6) variation measures, are shown in Table 4.4. In this and subsequent tables, the suffix notation shown in Table 4.3 has been used.

Table 4.4 shows that a linear model based on the means and standard deviations of the basic TM bands within each CD produced somewhat better population density estimates (R^2 =.751) than the models based on means and functions of means. The addition of variance or coefficient of variation terms produced little further improvement. A similar pattern was evident with the square root models. Again the logarithmic models tended to incorporate more variables but did not perform as well.

On balance, the preferred prediction model from Table 4.4 was the final square root transformation model, for which the regression equation (with terms in order of selection) was:

$$Est(\sqrt{D}) = 75.20 - 2.19\ b5 - 245.20\ b7c - 70.36\ b4c + 0.171\ b7v + 0.851\ b4 + 2.88\ b7$$

$$- 0.124\ b1v + 69.57\ b1c + 70.37\ b5c$$

with $R^2 = .840$ and $R^2_b = .780$

Table 4.4 Population Density Models based on CD Means and Spatial Variation Measures: Summary of Stepwise Regression Results

Potential predictors (number)	Dependent variable *D*		Dependent variable $\sqrt{D}$		Dependent variable ln*D*	
	Selected predictors	R^2	Selected predictors	R^2 (R^2_b)	Selected predictors	R^2 (R^2_b)
mean (6)	b5 b7 b3 b4	.537	b5 b7 b4 b3	.652 (.557)	b5 b7 b3 b2	.684 (.343)
mean, std dev (12)	b5 b1s b7 b4 b4s	.751	b5 b7 b1s b4 b4s b7s	.802 (.766)	b5 b7 b4 b7s b3 b4s b2	.763 (.497)
mean, std dev, coeff of var (18)	b5 b1c b4c b7 b4s b5c	.759	b5s b5 b7c b4c b7 b3 b4s	.821 (.769)	b5 b7 b4 b7c b3 b4s b5c b3s	.791 (.533)
mean, std dev, coeff of var, variance (24)	b5 b1c b4c b7 b4v b1	.768	**b5 b7c b4c b7v b4 b7 b1v b1c b5c**	**.840 (.780)**	b5 b7 b3 b4 b4v b7c b5c b3s	.794 (.524)

Note:

- Models selected for further investigation are shown in boldface.
- Predictors in each model are listed in the order of selection.
- Parenthesised R^2_b values are based on back-transformation (see Section 2.9). They are the R^2 values obtained when the dependent variable *D* is regressed on the estimate of *D* (obtained by inverse transformation of the regression estimates for the transformed *D* values), and hence provide a more realistic indication than R^2 of the predictive accuracy of the model.

4.4 POPULATION DENSITY ESTIMATES: SPECTRAL TRANSFORMATIONS

The band-to-band relationships discussed in Section 4.1 were implemented on CD aggregates (average values) for each band. Of more interest in the later phases of this study were the localised relationships between band values for each pixel. It was decided to try out something of a hybrid methodology at this stage, by incorporating a selection of such transformations, calculated at pixel level then averaged across CDs rather than vice versa. Of the vast number of spectral transformations that could be applied to the 6 TM bands at the individual pixel level (see Section 2.3), 14 were identified as having some capacity to discriminate between residential and other land uses. These are listed in Table 4.5. Details of how they were selected are discussed in context in Chapter 5.

Values of these 14 variables were calculated for each pixel in the study area, and CD aggregate means, standard deviations, coefficients of variation and variances were derived. These 56 (14×4) variables were combined with the 24 variables based on the individual TM bands, and a number of stepwise regression analyses were then applied to various subsets of the 80 variables. The resulting models are summarised in Table 4.6, using the same suffix notation as in Table 4.3.

Table 4.5 Selected spectral transformations

Variable	Description
nb1	normalised band 1
nb2	normalised band 2
rl4	ratio band 1 to band 4
rl5	ratio band 1 to band 5
r25	ratio band 2 to band 5
r57	ratio band 5 to band 7
ds15	difference/sum ratio bands 1, 5
ds25	difference/sum ratio bands 2, 5
ds35	difference/sum ratio bands 3, 5
ds57	difference/sum ratio bands 5, 7
ch123	cylindrical hue bands 1, 2, 3
ch125	cylindrical hue bands 1, 2, 5
rh123	rectangular hue bands 1, 2, 3
rh125	rectangular hue bands 1, 2, 5

The models with untransformed population density as the dependent variable fall into two groups. For the first model and the last four models (which utilise only the spectrally transformed predictors), R^2 values in the range .779 to .825 were obtained. In the remaining 3 models, as happens from time to time with any incremental sub-optimal search, the stepwise procedure “stopped short” with relatively few variables selected and relatively low R^2 values reached.

Table 4.6 Population Density Models based on CD Means and Spatial Variation of Selected Spectral Transformations: Summary of Stepwise Regression Results

Potential predictors (number)	Dependent variable *D*		Dependent variable $\sqrt{D}$		Dependent variable ln*D*	
	Selected predictors	R^2	Selected predictors	R^2 (R^2_b)	Selected predictors	R^2 (R^2_b)
TM bands + transformations: means (20)	rh123 r14 ds35 r57 rh125 ch125	.794	rh123 r14 r57 ds35 rh125 ch125	.898 (.832)	r14 r57 ds35 ch123 ds15 b5 b7	.880 (.774)
TM bands + transformations: means, std devs (40)	rh123 b1s	.729	**rh123 r14 r57 ds35 rh125 rh125s**	**.904 (.843)**	rh123 rh123s b3 b4s b1 b7s b1s b5 b7	.897 (.706)
TM bands + transformations: means, std devs, coeffs of var (60)	rh123 b1s	.729	rh123 r14 r57 ds35 rh125 rh125s	.904 (.843)	rh123 rh123s b3 b4s b1 b7s b1s b5 b7	.897 (.706)
TM bands + transformations: means, std devs, coeffs of var, variances (80)	rh123 b3v b2s	.758	rh123 r14 r57 ds35 rh125 rh125s	.904 (.843)	rh123 rh123s b3 b4s b1 b5v b7c b3v ds15 rh125s b5s rh123c	.916 (.694)
Transformations only: means (14)	rh123 r14 ds35 r57 rh125 ch125	.794	rh123 r14 r57 ds35 rh125 ch125	.898 (.832)	r14 r57 ds35 ch123 ds15	.867 (.741)
Transformations only: means, std devs (28)	r14s r25 r57 r14 rh125s rh123s	.825	rh123 r14 r57 ds35 rh125 rh125s	.904 (.843)	rh123 r14 rh123s r57 ds35 ds25 rh125s	.894 (.764)
Transformations only: means, std devs, coeffs of var (42)	rh123 r14s r25 rh125 r57c	.779	rh123 r14 r57 ds35 rh125 rh125s	.904 (.843)	rh123 r14 rh123s r57 ds35 ds25 rh125s rh123c	.898 (.746)
Transformations only: means, std devs, coeffs of var, variances (56)	rh123 r14s r25 rh125 r57c	.779	rh123 r14 r57 ds35 rh125 rh125s	.904 (.843)	r14 rh123s ch123v r57 ds35 rh125s ch123s rh123c rh123v ch123	.928 (.771)

Note:

- Models selected for further investigation are shown in boldface.
- Predictors in each model are listed in the order of selection.
- Parenthesised R^2_b values are based on back-transformation (see Section 2.9). They are the R^2 values obtained when the dependent variable *D* is regressed on the estimate of *D* (obtained by inverse transformation of the regression estimates for the transformed *D* values), and hence provide a more realistic indication than R^2 of the predictive accuracy of the model.

Once again the logarithmic transformation produced apparent improvement in R^2 at the expense of parsimony, but the improvement was not apparent after back-transformation.

However the models utilising the square root of population density as the dependent variable again performed very well. A very stable 6 variable solution emerged which is the preferred model from Table 4.6. The regression equation (with terms in order of selection) is:

$$Est(\sqrt{D}) = 530.10 + 0.278\ rh123 - 92.34\ r14 - 60.81\ r57 + 165.91\ ds35 - 1.308\ rh125 - 0.370\ rh125s$$

with R^2=.904 and R^2_b=.843.

4.5 DWELLING DENSITY ESTIMATES

The same sequence of analyses as above were applied to the estimation of housing density (dwellings/sq.km). The resulting models are summarised in Tables 4.7 to 4.9.

4.6 EVALUATION OF REGRESSION MODELS BASED ON CENSUS COLLECTION DISTRICT AGGREGATES

The two sets of six models chosen to represent the range of options tested for estimating population and dwelling densities on the basis of averages of substantial aggregates of pixels are summarised in Tables 4.10 and 4.11, and Figure 3.4. Further plots produced for the purpose of exploring the patterns of residuals in the final models in Tables 4.10 and 4.11 can be found in Appendix E.

All but the first base model in each table involved a regression equation for predicting the square root of the ground truth population or dwelling density. Remote sensing density estimates were found by squaring the predicted value for each CD.

Tables 4.10 and 4.11 are each comprised of 2 parts (separated by double vertical borders). The first section pertains to the regression equation employed on the transformed dependent variable, and the second part to the direct relationship between the dependent variable and the back-transformed estimates. Figure 4.1 consists of 4 plots pertaining to the second section of these tables.

The first section of Table 4.10 shows that by progressive enhancement of the set of predictors, the value of the coefficient of determination (R^2) was increased from around 55% to just over 90%.

Table 4.7 Dwelling Density Models based on CD Means
Summary of Stepwise Regression Results

Potential predictors (number)	Dependent variable *D*		Dependent variable $\sqrt{D}$		Dependent variable ln*D*	
	Selected predictors	R^2	Selected predictors	R^2 (R^2_b)	Selected predictors	R^2 (R^2_b)
bi (6)	**b5 b7 b4 b3**	**.560**	**b5 b7 b4 b3**	**.667 (.584)**	b5 b7 b4 b3	.707 (.583)
bi si (12)	s5 b7 s3 b3 b5 b4 s4	.798	**s5 b2 s2 s7 b4 s4 b5**	**.877 (.817)**	s5 s3 s1 b1 s7 b5 b3	.883 (.781)
bi si pij (27)	s5 p14 p15 p37 p45 b7	.802	s2 b4 p24 s7 p25	.872 (.826)	s5 p14 b4 s1 p47	.867 (.801)
bi si pij rij (42)	r24 r37 r47 r15 r17 p25	.822	r57 r14 r17 r47 r15	.881 (.832)	p47 p45 s1 r17 r25 r24 b7	.887 (.760)
bi si rij dij (42)	r24 d35 r47 r15	.801	**r57 r14 d17 r47 r15**	**.881 (.832)**	d13 r13 r57 s1 b1 r47	.910 (.793)

Table 4.8 Dwelling Density Models based on CD Means and Spatial Variation Measures:
Summary of Stepwise Regression Results

Potential predictors (number)	Dependent variable *D*		Dependent variable $\sqrt{D}$		Dependent variable ln*D*	
	Selected predictors	R^2	Selected predictors	R^2 (R^2_b)	Selected predictors	R^2 (R^2_b)
mean (6)	b5 b7 b4 b3	.560	b5 b7 b4 b3	.667 (.584)	b5 b7 b4 b3	.707 (.583)
mean, std dev (12)	b5s b5 b1s b7 b4 b4s	.818	b5 b7 b1s b4 b4s b7s	.840 (.843)	b5 b7 b4 b7s b3 b4s b2	.779 (.515)
mean, std dev, coeff of var (18)	b5s b5 b5c b4c b7 b1s b4	.845	b5s b5 b7c b7 b4 b4c b7s b1s b1c	.871 (.876)	b5s b5 b7 b4 b7c b3 b4s b5c b7s	.817 (.585)
mean, std dev, coeff of var, variance (24)	b5 b1c b7 b4 b4c b7c b7v b1s	.875	**b5 b7c b7v b4c b4 b5c b7 b1v b1c**	**.888 (.878)**	b5 b7 b4 b4v b7c b5c b7s b3 b5s	.819 (.604)

Note:
- Models selected for further investigation are shown in boldface.
- Predictors in each model are listed in the order of selection.
- Parenthesised R^2_b values are based on back-transformation (see Section 2.9). They are the R^2 values obtained when the dependent variable *D* is regressed on the estimate of *D* (obtained by inverse transformation of the regression estimates for the transformed *D* values), and hence provide a more realistic indication than R^2 of the predictive accuracy of the model.

Table 4.9 Dwelling Density Models based on CD Means and Spatial Variation of Selected Spectral Transformations: Summary of Stepwise Regression Results

Potential predictors (number)	Dependent variable D		Dependent variable $\sqrt{D}$		Dependent variable $\ln D$	
	Selected predictors	R^2	Selected predictors	R^2 (R^2_b)	Selected predictors	R^2 (R^2_b)
TM bands + transformations: means (20)	r14 ds35 r57 r15 r25 b5 b1	.876	r14 r57 rh125 b2 ds15 ch123 ds25 b1	.945 (.918)	r14 r57 ds35 ch123 ds15 b5 b7	.896 (.810)
TM bands + transformations: means, std devs (40)	rh123 b1s rh125 r57 r14 ds15 ds25 rh123s	.905	**r14 b3 b2s ds15 r57 ds25 rh125 rh125s**	**.945 (.924)**	rh123 rh123s b3 b1 b4s b7s b5 b1s	.907 (.766)
TM bands + transformations: means, std devs, coeffs of var (60)	rh123 b1s ch123s b4c ch123c	.834	r14 b5 b4 b2c ds15 r57 ds35c rh125s	.946 (.922)	rh123 rh123s b3 b1 b4c b7s b5 b1s	.908 (.769)
TM bands + transformations: means, std devs, coeffs of var, variances (80)	b5s b5v rh123 b3v b2s r14 r57 ds35	.916	r14 b5 b2s b3v b4 ds15 r57 ds35c rh125s	.948 (.924)	rh123 rh123s b3 b1 b5v b4s ds25c ch125	.913 (.733)
Transformations only: means (14)	r14 ds35 r57 r15 r25 rh125 nb2	.872	r14 r57 ch125 ds57 nb2 nb1 ch123	.933 (.924)	r14 r57 ds35 ch123 ds15 ds25	.884 (.799)
Transformations only: means, std devs (28)	rh123 r14s rh125 ds15 r57 r14 ds25 rh123s	.898	r14 r14s ds15 r57 rh125 rh125s ds25	.941 (.912)	rh123 rh123s ch123 r14s	.875 (.717)
Transformations only: means, std devs, coeffs of var (42)	r14s rh125 ds15 r57 r14 ds25 ch123c rh123c	.903	r14 r14s ds15 r57 rh125 rh125s ds25	.941 (.912)	rh123 rh123s ch123 r14s	.875 (.717)
Transformations only: means, std devs, coeffs of var, variances (56)	r14s rh125 ds15 r57 r14 ds25 ch123c rh123c	.903	r14 r14v ds15 r57 rh125 rh125v ds25 ch123v nb1s	.944 (.918)	rh123 r14 rh123s ch123v ch123s rh123c r57 ds15 rh125s ch123 ch125 ch125c	.945 (.854)

Note:

- Models selected for further investigation are shown in boldface.
- Predictors in each model are listed in the order of selection.
- Parenthesised R^2_b values are based on back-transformation (see Section 2.9). They are the R^2 values obtained when the dependent variable D is regressed on the estimate of D (obtained by inverse transformation of the regression estimates for the transformed D values), and hence provide a more realistic indication than R^2 of the predictive accuracy of the model.

The second section shows that in terms of predictive accuracy, the effective coefficient of variation (R^2_b) was increased from around 55% to around 85%. This corresponds to an increase in the correlation between the remote sensing and ground truth figures from around 0.75 to around 0.92. The standard deviation of the residual (unexplained) variation (*s*) was progressively reduced from 739 to 441 persons/sq.km.

However, correlation is not the only criterion to be considered: a high correlation only implies a linear relationship, not necessarily a direct correspondence. In directly estimated linear models, the estimates obtained are unbiased so that if the dependent variable is regressed on the estimates, the line of best fit has intercept=0 and slope=1. But when the data is transformed or when a regression equation is applied to another set of data (as in Chapter 7), biases can occur in the form of both systematic offsets (zero errors), and systematic scale errors. Table 4.10 shows that when the regression is unforced, the zero errors for all models are small relative to the scale of the data, and the slope coefficients of both forced and unforced regressions are close to unity, indicating a lack of substantial bias of either type. This is confirmed by the plots in Figure 4.1.

Notwithstanding all of the above considerations regarding population densities, the ultimate criterion for population estimation is the accuracy with which actual population counts can be estimated. Table 4.12 shows the results of multiplying the density estimate for each CD by its area to generate an estimated CD population count. As well as regression results, the mean and median of the absolute values of the relative errors (see Section 2.12.2) are reported, as well as estimates for the population of the whole study area and the urban section and their relative errors. The more complex models had median proportional errors within the range 17-24% overall, and 14-21% for the urban area. Mean proportional errors were somewhat higher, indicating a positive skew characteristic of absolute value distributions[1], which may be exacerbated by the presence of a few outlying values (see for example Figure 4.1B).

Similar results for estimated total dwelling counts are shown in Table 4.13.

The inherent limitations of the CD aggregate method start to become apparent in Tables 4.12 and 4.13. In most of the models, for reasons which will be discussed in later chapters, for the 16 rural CDs which have very low densities, the densities tend to be over-estimated by amounts which though small in absolute terms, are large in relative terms. When these over-estimates are weighted or leveraged by the large areas involved, they have a substantial effect on estimates of population or dwelling counts. As a result, whilst the more complex models produce estimates

---

[1] For normally distributed errors, the so-called half normal distribution has $\frac{\mu - M}{\sigma} = 0.20$ (Kendall et al., 1987, p 117).

of the urban totals which are accurate to within a few percent, the total for the whole study area is in every case overestimated to a much greater degree.

The opposite happens in the first (untransformed) model where, because of the concave-up curvilinearity of the relationship, the densities at the low end are underestimated by the linear model to such an extent that a number of the estimates are substantially negative. When combined with the large areal weightings of these low density CDs, this results in an estimated total population which is negative.

Negative estimates are always a potential problem when the lowest densities are small in comparison to the range of densities being estimated. Langford et al. (1991) describe models which can lead to such estimates as "logically flawed", whilst Lo (1995) takes the more pragmatic view that a negative estimate can be interpreted as evidence of zero population. Webster's results (1996) also contain negative estimates but they are not discussed. Negative estimates can only be avoided by restricting models to functional forms such as logarithmic, where such results are not possible. However, other forms of model may be preferred on the basis of other performance criteria. In the present instance the square root form, which generally produces better results than the logarithmic form, is potentially quite logically flawed in that negative results automatically backtransformed by squaring lead to positive estimates. In these circumstances, the author is inclined to share Lo's view and reset any negative estimates to zero.

Since dwellings are more directly evidenced in satellite imagery than human population, it is to be expected that the accuracy of estimation would be somewhat higher for dwelling counts than for population, as indeed it generally is, though not invariably so and not by very large margins. Furthermore, it seems from an examination of Figures 4.1B and 4.1D that much of the difference in the best fitting models is associated with specific anomalous CDs (outliers) such as those in which major institutions are located (the two CDs which include a geriatric institution and two hospitals are clearly visible at the top right on Figure 4.1B), rather than with a more diffuse phenomenon across all CDs. For the larger aggregate areas (urban and total), the population estimates are just as accurate as the estimates of dwelling counts. For this reason, it was decided that estimating dwelling counts held no advantage for the estimation of population, and it was decided henceforth to focus exclusively on models for direct population estimation. Of course the same methodologies could be applied throughout to the estimation dwelling counts if that was the required outcome.

Table 4.10 Summary of Selected Models for Population Density Based on TM Data Aggregated over Census Collection Districts

Model type	Class of predictors	Dependent variable	Number of predictors	Regression equation	R^2	D vs. $\hat{D}$ Regression coeffts.* (unforced & forced)	R^2_b	s
1	Band mean	D	4	72.3 – 135.6 b5 + 332.0 b7 -151.0 b3 + 61.6 b4	.537	0+1.00 ; 1.00 (Fig 13A)	.537	739
2	Band mean	$\sqrt{D}$	4	-1.4 – 2.33 b5 + 5.37 b7 + 1.22 b4 – 2.04 b3	.652	81+1.01 ; 1.05	.557	739
3	Mean, $(\text{mean})^2$	$\sqrt{D}$	6	-171.34 – 0.140 s1 + 9.220 b1 + 0.0344 s7 + 4.874 b4 – 1.952 b5 – 0.0359 s4	.847	-58.1+1.07 ; 1.04	.755	550
4	Ratios & difference to sum ratios	$\sqrt{D}$	4	345.35 – 68.41 r57 – 275.56 r14 + 226.89 d14 + 120.30 r15	.846	1.7+1.03 ; 1.03	.762	541
5	Mean, std dev, variance, coefft of variation	$\sqrt{D}$	9	75.20 – 2.19 b5 – 245.20 b7c – 70.36 b4c + 0.171 b7v + 0.851 b4 + 2.88 b7 – 0.124 b1v + 69.57 b1c + 70.37 b5c	.840	116+.95 ; 1.01	.780	521
6	Mean & std dev of spectral transformations at pixel level	$\sqrt{D}$	6	530.10 + 0.278 rh123 – 92.34 r14 – 60.81 r57 +165.91 ds35 – 1.308 rh125 – 0.370 rh125s	.904	9.5+1.01 ; 1.02 (Fig 13B)	.843	441

* Intercept + slope; slope when forced through origin

Table 4.11 Summary of Selected Models for Dwelling Density Based on TM Data Aggregated over Census Collection Districts

Model type	Class of predictors	Dependent variable	Number of predictors	Regression equation	R^2	D vs. $\hat{D}$ Regression coeffts.* (unforced & forced)	R^2_b	s
1	Band mean	D	4	89.7 - 49.77 b5 + 110.49 b7 + 23.81 b4 - 44.18 b3	.560	0+1.00 ; 1.00 (Fig 13C)	.560	265
2	Band mean	$\sqrt{D}$	4	-0.879 - 1.451 b5 + 3.119 b7 + 0.784 b4 - 1.051 b3	.667	19.6+1.03 ; 1.06	.584	258
3	Mean, $(\text{mean})^2$	$\sqrt{D}$	7	-176.63 + 0.0101 s5 + 12.147 b2 - 0.266 s2 + 0.019 s7 + 6.540 b4 - 0.050 s4 - 3.035 b5	.877	3.5+1.02 ; 1.02	.817	171
4	Ratios & difference to sum ratios	$\sqrt{D}$	5	160.0 - 27.9 r57 - 153.6 r14 + 58.8 d17 - 33.3 r47 + 146.6 r15	.881	2.8+1.02 ; 1.02	.832	164
5	Mean, std dev, variance, coefft of variation	$\sqrt{D}$	9	59.21 - 1.29 b5 - 169.32 b7c + 0.130 b7v -34.37 b4c + 0.508 b4 + 35.46 b5c + 1.42 b7 - 0.0715 b1v + 37.68 b1c	.888	33.9+0.96 ; 1.00	.878	139
6	Mean & std dev of spectral transformations at pixel level	$\sqrt{D}$	8	312.05 - 72.85 r14 + 0.260 b3 - 0.520 b2s + 244.92 ds15 - 30.45 r57 - 165.42 ds25 -0.716 rh125 - 0.158 rh125s	.945	5.4+1.00 ; 1.01 (Fig 13D)	.924	111

* Intercept + slope; slope when forced through origin

Table 4.12 Summary of Estimated Census Collection District Populations Based on TM Data Aggregated over Census Collection Districts

Model Type*	Class of predictors	Ballarat Statistical District (*n*=138)						Ballarat urban area (*n*=122)					
		Slope (forced)	R^2	*s*	Mean % error	Median % error	Est. tot. (% error**)	Slope (forced)	R^2	*s*	Mean % error	Median % error	Est. tot. (% error**)
1	Band mean	-.002	.02	259	553.6	31.0	-33519 (-142.3%)	.57	.26	225	83.2	28.5	88178 (+26%)
2	Band mean	.12	.00	259	185.9	32.9	151019 (+91%)	.77	.29	221	72.4	30.5	74832 (+7%)
3	Mean, $(\text{mean})^2$	.66	.10	246	47.6	24.1	89094 (+13%)	.78	.32	216	33.5	20.0	74811 (+7%)
4	Ratios & difference to sum ratios	.58	.15	238	57.3	20.9	91992 (+16%)	.85	.34	213	37.7	19.4	71992 (+2%)
5	Mean, std dev, variance, coefft of variation	.28	.01	257	93.0	23.0	111574 (+41%)	.97	.49	186	36.5	21.0	68805 (-2%)
6	Mean & std dev of spectral transformations at pixel level	.58	.18	235	39.4	17.4	89849 (+14%)	.94	.56	175	28.2	13.6	70615 (+0.6%)

* As in Table 4.10, the dependent variable for model type 1 was *D*, and for all other model types $\sqrt{D}$

** Ground truth populations are: BSD 79179; Urban 70222

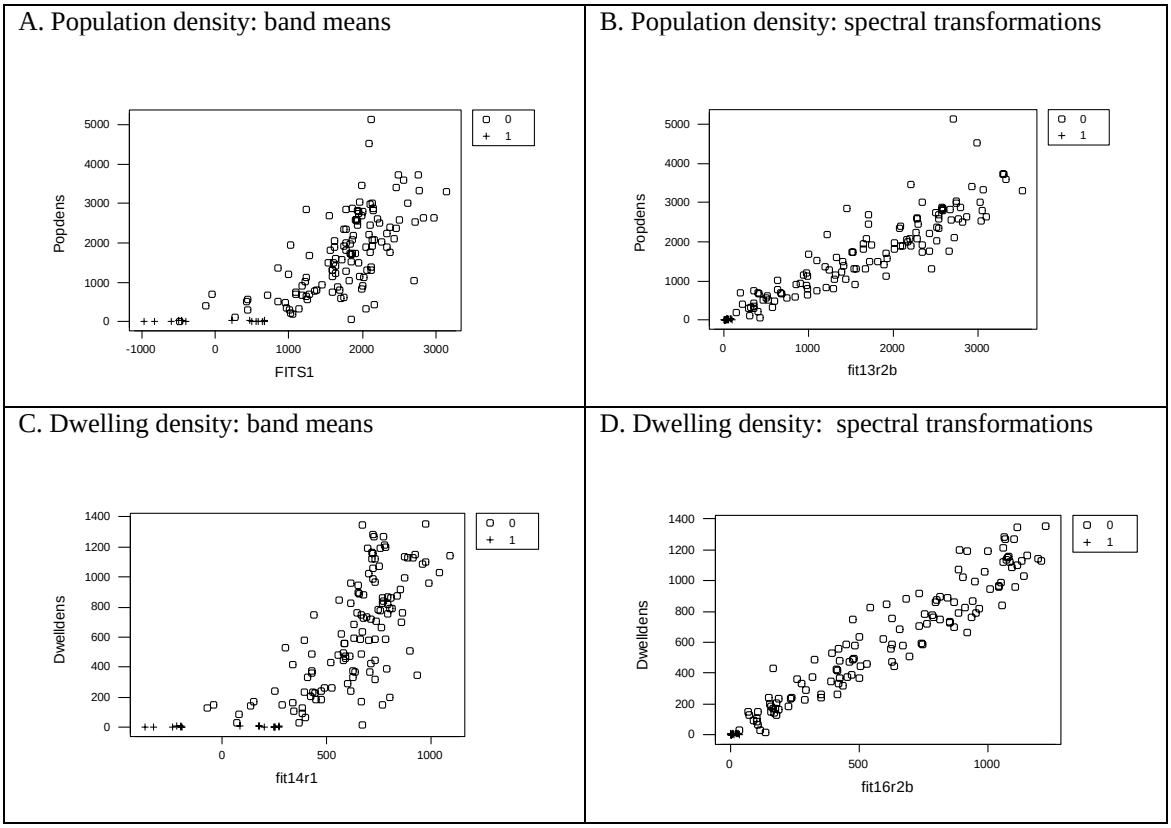
Table 4.13 Summary of Estimated Census Collection District Dwelling Counts Based on TM Data Aggregated over Census Collection Districts

Model Type*	Class of predictors	Ballarat Statistical District (*n*=138)			Ballarat urban area (*n*=122)		
		Mean % error	Median % error	Est. tot. (% error**)	Mean % error	Median % error	Est. tot. (% error**)
1	Band mean	720.9	32.7	-18649 (-169.6%)	91.0	28.0	31569 (+30%)
2	Band mean	231.6	37.5	55070 (+49%)	77.7	29.5	26645 (+9%)
3	Mean, $(\text{mean})^2$	63.2	22.9	32811 (+22%)	33.9	19.6	25237 (+4%)
4	Ratios & difference to sum ratios	57.9	18.7	31507 (+17%)	37.5	16.0	25215 (+3%)
5	Mean, std dev, variance, coefft of variation	103.3	21.6	36911 (+37%)	34.3	18.5	24104 (-0.1%)
6	Mean & std dev of spectral transformations at pixel level	36.9	16.0	29300 (+9%)	23.7	15.4	24503 (+0.6%)

* As in Table 4.11, the dependent variable for model type 1 was D, and for all other model types $\sqrt{D}$

** Ground truth dwelling numbers are: BSD 26971; Urban 24368

Figure 4.1 Population and Dwelling Density Estimates for 138 Census Collection Districts: Ground Truth vs. Remote Sensing Estimates from Base and Enhanced Models



0 Urban + Rural

The level of accuracy achieved with the most complex models is rather better than the best of the models fitted to 47 suburbs of Harare by Webster (1996) using a similar methodology but a different suite of predictors. It is comparable to the accuracy of a 7-variable model Webster fitted to 65 Cardiff grid squares using yet another different suite of 70 texture variables. A second model reported by Webster for the Cardiff data cannot be assessed comparatively, because no adjustment to R^2 was reported for its logarithmic form nor for the fact that its intercept was forced to zero (see Sections 2.8.3, 2.11.7).

The "mechanism" of the models derived in this chapter, why these particular linear combinations of these particular spatially aggregated and averaged spectral characteristics should correlate highly with populations, is conjectural. It may be possible to relate the structure of these equations to reflectance properties of materials and combinations thereof, but one must always exercise extreme caution in placing interpretations on individual regression coefficients in a multivariate context. As has already been noted, there may be many other alternative combinations of variables which would work almost as well on this set of data, and perhaps better on a slightly different set of data.

In models based on CD aggregates, this problem of capitalisation on chance is exacerbated by the relative paucity of aggregated observations. In the present instance, a total of 116 explanatory variables has been considered and assessed on the basis of a mere 138 observations. There is also the issue of the ecological fallacy; the presence of a particular relationship between population and spatially aggregated and averaged spectral characteristics does not imply that a similar relationship holds at the level of individual pixels or individual dwellings. Considering these caveats, a question that can be more readily addressed than "what is in the black box" is "how well does the black box work in other contexts". The beginnings of an answer to that question will be found in Chapter 7.

4.7 SUMMARY

In this chapter, beginning from the most basic prediction model, substantial improvements were achieved in the estimation of CD population and dwelling densities and counts on the basis of CD aggregates of remote sensing indicators. From the many models tested, six representative models were chosen (for each of population and dwelling density) for further testing on the secondary image. The first was the simplest baseline model, with an untransformed dependent variable and the 6 TM band means as predictors. The other 5 all utilised the square root of the density, and involved progressively more complex CD aggregate functions of the TM bands: basic means; squares of means, ratios of means, variation measures; and means and variation measures of selected pixel-level spectral transformations. The effective R^2 values of these models ranged from .54 to .84 for population density, and from .56 to .92 for dwelling density. The three most complex models produced quite accurate estimates of total population and total dwellings for the Ballarat urban area (all correct to within 3%), but total population and total dwellings for the low density rural sections and hence for the whole study area were substantially overestimated by all models.

The general shortcomings of the aggregate-based approach which were outlined in Section 1.3.2 were brought into sharper focus by the foregoing analysis. Firstly, because of the small effective sample size associated with the aggregate approach, the improvements in predictive accuracy bought at the price of wide net-casting and lack of parsimony were likely to prove illusory when subjected to external validation. Secondly, the sacrifice of detailed spatial information particularly within the more extensive low density CDs left no way to respond to the problem of over-estimation in these areas. The conclusion that the aggregate-based approach was likely to prove a blunt and limited instrument strongly motivated a change to the pixel-based approach of Chapter 5.

Chapter 5

Estimates Based on Individual Pixels

5.1 INTRODUCTION

As an alternative to the use of spatially aggregated data discussed in Chapter 4, regression analyses were investigated for estimating the population of each pixel, based on the spectral responses of individual pixels. Some conceptual issues regarding the modelling of the notional residential population associated with a single pixel have been addressed in Section 2.8.4.

Section 5.2 reports a straightforward regression analysis on a random sample of all pixels in the image, with the basic 6 TM bands as predictors. This approach was found to be quite inadequate unless terms were included to distinguish *a priori* between urban and non-urban pixels, and so was not taken further. In Section 5.3, a classification step is introduced, and a two phase classification and regression procedure based on just the 6 TM bands is reported. The obtained population estimates were aggregated to CD level and compared to ground truth CD figures. Section 5.4 describes the selection of a suite of spectral transformations which seemed to have the potential for improving population estimates (the use of these was also reported on in Chapter 4), and also a number of spatial transformations designed to provide information about spatial variation in the immediate neighbourhood of each pixel (i.e. image texture). Sections 5.5 and 5.6 describe the incorporation of these transformations at the classification stage and regression stage respectively. Section 5.7 describes the implementation of an ancillary improvement strategy - iterative refinement of the initial ground truth populations assigned to individual pixels. In Section 5.8, the models resulting from Sections 5.5-5.7 are comparatively evaluated by aggregating population estimates to the CD level and comparing the results with ground truth values. In Section 5.9, logarithmic and square root transformations of the dependent variable, which had been used with some success in the CD aggregate modes of Chapter 4, were applied at the pixel level. Four pixel-based models were ultimately selected for further testing on the secondary image. In Section 5.11, the reasons why the chosen models were all of simple linear form are considered.

5.2 A BASIC PIXEL-BASED REGRESSION MODEL

As a preliminary, a pixel-based regression model was fitted using only 6 predictors - the 6 untransformed TM bands.

This served the twofold purpose of developing the procedures and of providing some preliminary indication of whether this methodology was inherently superior to the CD average reflectance approach described in Chapter 4.

A 1 in 500 random sample[1] of 1398 pixels in the primary study area was selected for regression analysis. At the same time the number of pixels in each CD was counted. An imputed ground truth population p_g was calculated for each pixel by dividing the CD population P_g (estimated using the methods discussed in Chapter 4) by the number of pixels. This is based on an assumption of constant population density across all the pixels in a particular CD, and as such is quite unrealistic, especially for the large rural CDs. Subsequently these estimates were refined (see Sections 5.3 and 5.7).

A stepwise regression analysis (see Section 2.8) of estimated pixel population p_g on the 6 TM bands resulted in all bands except band 2 being selected. The final prediction equation (with variables in order of entry) was:

$$\hat{p}_g = 0.0237 + 0.0298\ b1 - 0.0212\ b3 + 0.00336\ b4 - 0.0120\ b5 + 0.0222\ b7$$

with $s = 0.3461$ $R^2 = .258$

The fit of this model was extremely poor. A plot of ground truth values vs. estimated values and a normality plot of residuals indicated pronounced skew in the imputed population distribution and so a logarithmic transformation was applied to clarify the nature of the underlying problems. Stepwise analysis resulted in the following prediction equation (with variables in order of entry):

$$\log \hat{p}_g = - 1.867 - 0.0220\ b5 + 0.0434\ b7 + 0.0107\ b4 + 0.0404\ b1 - 0.0401\ b3$$

with $s = 0.5485$ $R^2 = .264$

Whilst the overall fit was no better, a plot of ground truth values vs. estimated values for this model clarified the central problem: the very different population densities in the rural and urban areas. To explore the extent to which explicit modelling of this factor could improve the fit, a dummy or indicator variable for rural/urban difference was added, together with multiplicative interaction terms, and stepwise analyses performed again for both untransformed and logarithmic models. In the resulting prediction equations the coefficient of the indicator

---

1 See Section 2.11.3 for a discussion of sample sizes.

variable “rururb” represents the average difference in the dependent variable between rural and urban areas, and the variables suffixed with ”ru” represent the corresponding differences in the coefficients of the various TM bands; in effect separate regression relationships for urban and rural areas are represented in a single equation.

The prediction equations (with variables in order of entry) are:

$\hat{p}_g$ = 1.258 - 1.237 rururb + 0.00867 b1 - 0.0212 b5 + 0.0212 b5ru - 0.00908 b1ru + 0.0289 b7 - 0.0288 b7ru

with s = 0.2513 R^2 = .610

and

$\log \hat{p}_g$ = - 0.162 - 1.679 rururb - 0.0251 b1ru + 0.00532 b4ru + 0.00768 b1 - 0.00312 b5 + 0.00866 b7ru

with s = 0.2696 R^2 = .822

The R^2 values showed a substantial improvement. As well as the overall difference in population density, the significant interaction terms indicated that the relationship between imputed population and TM reflectances is different in the urban and rural areas, which is also apparent in the plot of ground truth values vs. estimated values for the logarithmic model. It was also apparent from the plots that most of the improvement in the fit was associated with the rural/urban differences; within each of these groups the relationships were not strong. Clearly the crude broad-brush manner in which CD ground truth populations had been equally distributed amongst all pixels was likely to be a major contributing factor, particularly with CDs with low population density. Accordingly, it was decided to persevere no further with this broad-brush approach, but rather to move immediately to a two-stage methodology with an initial classification stage.

5.3 CLASSIFICATION OF PIXELS FOLLOWED BY REGRESSION

5.3.1 Supervised land use classification: categories and training sets

The essential aim of the classification phase of the study was to distinguish residential land use from all other categories. Nevertheless, it was (correctly) anticipated that better discrimination might be achieved if other land uses/ land covers were separated into relatively homogeneous categories.

Accordingly, the twelve broad categories listed in Table 5.1 were defined. Five were land use categories pertaining to the (generally urban) built environment. The other seven were land cover categories pertaining to rural areas and urban open space.

Table 5.1 Categories of Land Use and Land Cover

Residential
Industrial
Commercial
Public use
Road
Bare ground: dark coloured soils
Bare ground: light coloured soils
Dry grass, pasture, crops
Green grass, pasture, crops
Native eucalypt forest and scrub
Pine plantation
Water

For all categories except residential, training sets were selected visually using local knowledge and a quasi-natural colour RGB image of the study area. In the case of the residential category, the training set consisted of the 25 urban CDs for which more than 90% of the total CD area was statutorily zoned as residential in 1984 (Harvey and Taylor, 1984).

5.3.2 Initial classification based on the 6 TM bands

On the basis of the selected training sets, each pixel in the image was classified into one of the 12 land use categories, using a maximum likelihood classification based on all 6 TM bands. There was no attempt at this stage to select the most discriminating variables.

The resulting classification was displayed using a 12 colour pseudocolour display (see Image 3). Whilst some classes were not well separated (e.g. industrial and commercial), the residential class appeared to be reasonably well delineated in the urban areas. However, this was much less so in the rural areas, where a number of features including many roads, some paddocks, and even a swamp were classified as residential.

It appeared that residential areas are characterised not only by the reflectance levels (i.e. colour) of individual pixels, but also by the mottled spatial pattern of neighbouring pixels.

As a result, an investigation was begun into texture measures. (See Section 2.5 for theoretical development.)

Notwithstanding the indifferent quality of this preliminary classification, the regression modelling phase was proceeded with.

The pixel classifications were saved as a new data band, and a 1 in 50 random sample of pixels classified as residential (1402 pixels) was selected for regression analysis. At the same time, a count was made of the number of residential pixels in each CD. An imputed ground truth population p_g was calculated for each pixel by dividing the estimated CD population P_g by the

number of residential pixels. This is based on an assumption of constant population density throughout the residential pixels of each CD.

A stepwise regression analysis of estimated pixel population p_g on the 6 TM bands resulted in all bands except band 4 contributing significantly to the regression model. The final prediction equation (with variables in order of entry) was

$$\hat{p}_g = 0.928 + 0.0890\ b1 - 0.0405\ b5 - 0.134\ b3 + 0.0656\ b7 + 0.109\ b2$$

with $s = 0.862$ $R^2 = 0.444$

5.3.4 Application to the full image: population density estimates

The regression equation obtained was used to calculate a population estimate for every pixel classified as residential in the full image. All pixels classified as non-residential were assigned zero population. The resulting data band was displayed in a pseudocolour display (see Image 7) which, since the pixels are of constant size, can be interpreted as population density. This display has similar visual characteristics to the residential classification. It conforms quite well with expectations in the urban areas, but population is clearly over-estimated in many parts of the rural area.

5.3.5 CD population estimates: preliminary evaluation of the model

Using the CD identification band, the pixel population estimates were aggregated to produce satellite-based population estimates P_S for each CD. These were divided by CD areas to produce CD population density estimates D_S. The two sets of estimates were compared with the corresponding ground truth figures P_G and D_G using the graphical and analytic methods of Section 4.5. The R^2 values obtained were .802 for population density and .163 for population. The result for population density was quite promising, but as was found in Chapter 4, the overestimation of low densities in the large rural CDs led to greatly inflated population estimates and the low value of R^2

A more detailed examination of this basic model is delayed until Section 5.6. We now consider a number of strategies which were investigated for improving upon it.

5.3.6 Strategies for improving the model

It was considered that the basic classification/regression approach was limited by at least four factors:

- inaccuracy of residential/non-residential classification
- weakness of the linear relationship between population and the untransformed TM bands.
- inaccuracy of imputed pixel populations.
- scale effects, especially the discontinuity between rural and urban densities.

These factors are addressed in the following sections by:

Extending the set of variables used for both classification and prediction by incorporating spectral and spatial transformations at the pixel level.

Improving the regression modelling by iterative refinement of the imputed ground truth pixel populations.

5.4 DATA TRANSFORMATIONS

5.4.1 Spectral domain transformations

There was no *a priori* reason to believe that the relationship between population and spectral response should be linear (see Section 1.3.2). A number of standard spectral domain transformations (see Section 2.4) were applied to the 6 TM bands, giving rise to 61 derived variables. These are summarised in Table 5.2.

A band difference to band sum ratio is a monotonic mapping of a band-to-band ratio from $(0,\infty)$ onto (-1,1). A hue transformation can be regarded as an extension of the 2 dimensional linear scale of the difference-to-sum ratio to a 3 dimensional circular scale representing the relative weightings of three bands.

Table 5.2 Summary of Spectral Domain Transformations

Generic name	Number of variables	Description
nbi	6	normalised band
rij	15	band to band ratio
dsij	15	band difference to band sum ratio
PCi	6	principal component
rHSI123	3	rectangular hue/saturation/intensity: bands 1,2,3
rHSI125	3	rectangular hue/saturation/intensity: bands 1,2,5
tHSI123	3	triangular hue/saturation/intensity: bands 1,2,3
tHSI125	3	triangular hue/saturation/intensity: bands 1,2,5
cHSI123	3	cylindrical hue/saturation/intensity: bands 1,2,3
cHSI125	3	cylindrical hue/saturation/intensity: bands 1,2,5
TVI34	1	transformed vegetation index: bands 3,4
Total	61	

5.4.2 Preliminary screening of potential discriminators/predictors

Because of disk storage constraints it was decided to visually screen the derived variables before proceeding to the regression analysis. This was done for the practical reason that the 1412×1008 pixel primary test image required almost 1.5 Mb per band (1 byte integer) for the 6 basic data channels, and 6 Mb per band (4 byte floating point) for the derived variables. There would be a similar further storage requirement for each spatial transformation at the next stage (see Section 5.5.3).

The screening criterion used was the extent to which each variable could discriminate between known residential and non-residential areas. This is related to the classification phase rather than the regression phase, reflecting the fact that visual qualitative relationships are more easily assessed visually than quantitative ones. It was judged that any variable which was unable to discriminate at a gross qualitative level would be unlikely to contribute more finely detailed quantitative information.

Another limitation of such a screening is that it is univariate, whereas the following analyses would be multivariate. It is conceivable that a variable which does not contribute a lot of information in isolation may make an important incremental contribution in conjunction with other variables. Nevertheless, resource limitations dictated that such a screening be undertaken.

Each variable was displayed in turn in a pseudocolour image, with a variety of standard colour enhancement transformations applied to the image histogram. These included:

- 100% histogram stretch: actual input data range is linearly mapped onto the full 1-255 display range
- 99% trim: the bottom 0.5% of values are set to 1, the top 0.5% to 255, and the remainder linearly stretched
- 95% trim: the bottom 2.5% of values are set to 1, the top 2.5% to 255, and the remainder linearly stretched
- histogram equalisation: a piecewise linear transformation resulting in an approximately uniform distribution of displayed colours
- Gaussian equalisation: a piecewise linear transformation resulting in an approximately Gaussian distribution of displayed colours
- interactive user-defined piecewise linear transformation: enables the user to selectively enhance the colour sensitivity in selected data ranges
- interactive thresholding: an extreme case of the previous method: a user-specified stepped transformation which results in a classification image based on ranges of values of a single variable.

Of the raw TM bands, only band 5 showed any substantial discrimination capacity. Nevertheless all 6 raw bands were retained in the analyses which followed, in recognition of the widely accepted hierarchical principle that models containing complex terms should also include the simpler terms from which they have been generated.

Of 55 derived variables, 37 were assessed as having little discriminating power and were discarded.

Two further variables, the fourth and fifth principal components, discriminated moderately well, but it was decided to exclude these variables on the grounds that principal component structure is largely determined by the content of a particular image, and so a procedure based on a particular principal component is unlikely to be robust when applied to other images.

The remaining 14 transformed variables are listed in Table 5.3. As the third column of the table indicates, all 14 variables exhibited characteristic levels in residential areas, though the strength of the discrimination varied. The particular characteristic level within the distribution of variable values was different for different variables, but in each case, residential areas were characterised by intermediate rather than extreme values. This has implications for regression modelling, since it suggests that any relationship between these variables and population is likely to be non-linear.

Table 5.3 Selected Spectral Transformations

Variable	Description	Visually assessed degree of discrimination between residential and non-residential areas			
		Level	Texture		
			Spatial SD	Spatial COV	PDTI
nb1	normalised band 1	high	discernable	-	discernable
nb2	normalised band 2	moderate	-	-	-
r14	ratio band 1 to band 2	moderate	discernable	-	discernable
r15	ratio band 1 to band 5	high	moderate	-	discernable
r25	ratio band 2 to band 5	moderate	-	-	discernable
r57	ratio band 5 to band 7	discernable	-	-	-
ds15	diff/sum ratio bands 1, 5	high	discernable	-	discernable
ds25	diff/sum ratio bands 2, 5	moderate	-	-	-
ds35	diff/sum ratio bands 3, 5	discernable	-	-	-
ds57	diff/sum ratio bands 5, 7	discernable	-	-	-
chue123	cylindrical hue bands 1, 2, 3	discernable	-	-	discernable
chue125	cylindrical hue bands 1, 2, 5	moderate	discernable	-	moderate
rhue123	rectangular hue bands 1, 2, 3	discernable	-	-	-
rhue125	rectangular hue bands 1, 2, 5	moderate	-	-	discernable

Many of the variables in Table 5.3 involve a comparison between, on the one hand one or other of the relatively short visible wavelengths (bands 1, 2, 3) and on the other hand the longer infrared wavelengths (bands 4, 5, 7). The strongest discriminators were all ratios involving

band 1 in the numerator and band 5 in the denominator. Water, which reflects energy in the visible wavelengths but not in the infrared (Harrison and Jupp, 1989), has the highest values on these variables. Vegetation and bare soil generally have higher reflectances in the band 5 range than in the band 1 range, and hence have relatively low values on these variables. This is apparent in the pseudocolour image of variable ds15 (Image 5) where water is shown as red (high values) and the rural areas as blue (low values). The commercial and industrial areas of the urban area are also red, indicating a preponderance of constructed surfaces such as bitumen and roofing materials which have relatively strong reflectivity at shorter wavelengths (Forster, 1980; Curran, 1985), whilst in the residential areas intermediate (yellow) values predominate, indicating a mixture at sub-pixel scale of built and natural surfaces. Many of these variables were also observed to exhibit a more mottled texture in residential areas than in other areas, which is consistent with a mixture of built and natural elements at the scale of 30m pixels.

5.4.3 Spatial domain transformations

To explore the aspect of inter-pixel variation more closely, four texture measures (spatial filters) were calculated for each of the 6 TM bands and the 14 derived variables in Table 5.3.

These measures were spatial variance, spatial standard deviation (SD), spatial coefficient of variation (COV), and the pairwise difference texture index (PDTI) defined and discussed in Section 2.5. All were based on a 3×3 pixel neighbourhood. Each texture measure was displayed in a pseudocolour image in which the level of the displayed texture variable represented the amount of local variation in the underlying variable.

The spatial variance and standard deviation are standard and widely used texture measures. The coefficient of variation, by expressing the standard deviation in proportional terms, removes any scale dependence. The pairwise difference measure was used because simulation trials (see Section 2.5) had indicated that it might distinguish residential areas better than the other measures.

Examination of the 80 (4×20) resulting pseudocolour images showed that none of the coefficients of variation were useful as discriminators, but some measure of discrimination was achieved by one or more of the other texture measures in the case of 10 variables: TM bands 1 and 2 and eight of the derived variables (see Table 5.3 and Image 6).

As the variance and standard deviation contain equivalent information and either can be generated from the other, only the standard deviations were stored for future use. Hence, 20 (2×10) spatial transformation variables were added to the 20 variables from Section 5.4.2., resulting in 40 candidate discriminator/predictor variables in all. These were calculated and stored for each pixel in the full image.

5.4.4 Provision for non-linearity and interactions

The 40 variables identified in Section 5.4.3 were augmented in two ways. Firstly, because the residential class was represented by intermediate values on most of the derived variables, it was decided to incorporate square terms for each of the 6 basic bands and the 14 spectral transformations at the regression stage.

Secondly, because residential areas were characterised by a combination of both the level and the spatial variation or texture of some of the variables, it was decided to allow for the possibility of interactive effects. Within the framework of linear modelling, the standard procedure is to incorporate cross product terms (see Section 2.10), and so 20 such terms were calculated, with each of the 10 variables for which texture measures had been calculated being multiplied by its two texture measures.

This completed the suite of 80 variables, of which 40 were calculated for the whole image, a further 20 (the interaction terms) for both discriminant analysis and regression analysis samples, and the final 20 (the square or quadratic terms) only for the regression analysis samples, since linearity of between group differences is not a requirement in discriminant analysis.

5.5 CLASSIFICATION USING TRANSFORMED EXPLANATORY VARIABLES

5.5.1 Selection of classification variables

The final selection of classification variables was made by applying stepwise discriminant analysis (see Section 2.7) to a 1 in 10 sample of pixels (7486 pixels) from the training sets described in Section 5.3. The sample size was designed to ensure that the smallest classes, such as roads, were adequately represented.

Five subsets of the 60 variables described in the previous section were explored in a series of stepwise discriminant analyses. Decisions made in the light of emerging trends resulted in five classification structures being explored. These variable subsets and classification structures are described and classification results summarised in Table 5.4. The variables selected from subset D are listed in Appendix F.

In each case, after stepwise selection of variables and determination of linear discriminant functions, a maximum likelihood classification was made of all pixels in the sample set. Since the actual class of each training set pixel was known, the allocated classes could be compared with the actual classes in a classification or confusion matrix. Table 5.4 shows that the initial analysis based on the 6 TM bands and the twelve classes as defined resulted in 90% of all pixels and 80% of residential pixels being correctly classified.

Table 5.4 Summary of Stepwise Discriminant Analyses

Variable subset		Classification structure											
		1: 12 classes			2: 11 classes			3: 10 classes			4: 10 classes		
		v	C_A%	C_R%	*v*	C_A%	C_R%	*v*	C_A%	C_R%	*v*	C_A%	C_R%
A	*6*	*6*	90	80	*6*	91	85	*6*	92	84	*6*	93	84
B	*20*	*16*	91	76	*20*	92	83	*20*	91	85	*20*	93	85
C	*14*	*10*	80	47	*14*	89	71	*14*	87	67	*14*	88	67
D	*40*	*37*	92	87	*37*	93	92	*37*	92	91	*37*	93	90
E	*60*	*48*	93	86	*50*	93	92	*50*	93	91	*50*	94	91

Key:

Variable subsets

A 6 TM bands
B As for A + 14 spectral transformations
C 14 spectral transformations only
D As for B + 20 spatial transformations
E As for D + 20 interaction cross-product terms

Classification structures

1 12 classes as defined in Table 4.10
2 Public use class omitted
3 Commercial, industrial & public use combined
4 Comm. & ind. combined, public use omitted

v = number of variables selected by the stepwise procedure
C_A% = % of all classes correctly classified (rounded to nearest %)
C_R% = % of residential class correctly classified (rounded to nearest %)

Progressive inclusion of transformed variables resulted in incremental improvements to both classification rates. The lowest rates occurred with model C, the only one not in the hierarchical sequence. This suggests that between them the most discriminating 10 of the 14 spectral transformations do not contain as much information as the 6 basic bands.

It was observed that much of the misclassification occurred between the four main "built environment" classes: commercial, industrial, public use and residential. The public use class in particular, consisting of hospitals, schools, and Sovereign Hill Historical Park, was not well discriminated, being almost entirely misclassified into one of the other three classes. There was also considerable confusion between industrial and commercial. Accordingly three variants of the classification structure were investigated; firstly the public use category was omitted altogether, and the pixels in that class regarded as "unknowns" to be classified; secondly, commercial, industrial and public use were combined into a single category; and thirdly, commercial and industrial were combined and public use omitted. Table 5.4 shows that whilst these variations did not change the overall classification rate very much, the classification of residential pixels was improved in each case by something in the order of 5 percentage points.

However, as in any classification scheme, there is a tradeoff between two types of misclassification. The aim of maximising the correct classification of residential pixels has to be balanced against the aim of minimising the incorrect classification of non-residential pixels as residential. Examination of the confusion matrices (an example is given in Appendix F) showed that most of the increase in the correct residential classifications were pixels which under structure 1 had been classified as road or grass. Conversely, there was an increase in the already substantial proportions of commercial, industrial and public use pixels classified as residential.

In addition, a number of bare ground pixels were classified as residential under structures 2-4, which was not the case under structure 1.

These considerations suggested that the increase in the classification rate for residential pixels under structures 2-4 was probably spurious (the pixels concerned probably were actually road or grass within residential areas), whilst the cost, in terms of misclassification of other pixels in industrial, commercial and rural areas as residential, was real. Accordingly, it was decided to proceed on the basis of the original 12-class structure.

The object of this exercise was to maximise the separation of one class from all of the others. However, the discriminant analysis criterion, like the maximum likelihood classification criterion, is to maximise the separation of all the classes. The best overall separation may not give the best separation of residential vs. the rest. As a long shot, in the light of the discussion of spectral characteristics in Section 5.4.2, a fifth scheme was tried, with just three classes: "Residential", "Water" and "Other". However, as expected, averaging the other 10 classes in multivariate space resulted in an over-arching class which could not be distinguished from the residential class at all, and so the 12-class structure was proceeded with.

There were strong similarities in the lists of variables chosen by each of the stepwise analyses for the hierarchy of models B, D, and E. In particular, the first 4 variables selected in each case were ds25, b5, b7 and b4. This being so, it was decided to proceed to the next stage (classification of the full image) not on the basis of sets of variables arising out of all of the analyses of models A-E, but rather to use a cumulative sequence of subsets of variables from model D (see Appendix F for a listing of the variables). Since the initial set of discriminators was the set of 6 TM bands, it was decided by way of comparison to examine the first 6 variables from model E, then the first 10, 15, 20 and so on. The classification rates for these models are shown in Table 5.5.

It is interesting to note that of the four most consistently selected discriminating variables, the first two (ds25 and b5) were identified in the preliminary visual screening (Section 5.4.2) as having discriminating potential, but the other two (b7 and b4) were not, illustrating the point previously made that univariate screening cannot predict how variables will interact in a multivariate context.

The remaining 21 variables are a jumbled mixture of most types of variable. Again, as has previously been discussed, the exact composition and disposition of variables in the list would probably be quite sample-dependent, but it was hoped that with such a large training sample the discriminating capability would be reasonably robust.

Table 5.5 Summary of the Selected Sequence of Discriminant Analyses

Variables	C_A%	C_R%
6 TM bands	**90.3**	**79.7**
First 6 variables	**88.4**	**73.9**
First 10 variables	90.7	79.4
First 15 variables	**91.4**	**81.5**
First 20 variables	92.1	83.9
First 25 variables	**92.5**	**87.4**
First 30 variables	92.5	86.9
First 37 variables (limit of stepwise selection)	92.6	86.9

C_A% = % of all classes correctly classified
C_R% = % of residential class correctly classified

Another issue in classification is that of the incorporation of prior probabilities (see Section 2.6.1). In the absence of any prior information about the relative abundance of the various classes, the default is to assume that they are equally likely to occur. A common alternative approach is to set the prior probabilities proportional to the sample sizes in the training set. This approach was examined but discarded for three reasons: firstly, in all cases except the residential class, the training samples were visually chosen convenience samples, not random samples nor proportional samples from the classes; secondly, the relative preponderance of the various classes is not the same for different images; and thirdly, the rates and patterns of misclassification were not substantially nor consistently changed. It was decided henceforth to use equal prior probabilities.

5.5.2 Classification of the image

Clearly there was little change in classification rates beyond 25 variables. Consequently, only the first 25 selected variables were calculated and stored for each pixel in the full image.

For each of the subsets of these variables listed in Table 5.5, each pixel in the image was classified into one of the 12 land use categories, using a maximum likelihood classification based on the basis of the selected training sets.

The resulting classifications were displayed using a 12 colour pseudocolour display. It was decided that the only discernible change relative to the classification based on the 6 TM bands occurred when 25 variables were employed, at which point there was a noticeable reduction in the number of pixels in particular rural areas which were incorrectly classified as residential (see Image 4). Nevertheless, it was decided to proceed to the regression stage using the 4 classifications shown in boldface in Table 5.5.

5.6 REGRESSION USING TRANSFORMED EXPLANATORY VARIABLES

5.6.1 Selection of regression models

Each of the 4 pixel classifications was saved as a new data band. In each case a 1 in 50 random sample of pixels classified as residential was selected for regression analysis (sample sizes of 1402, 1499, 1340 and 1364). Each time, a count was made of the number of residential pixels in each CD, and an imputed ground truth population p_g was calculated for each pixel by dividing the estimated CD population P_g by the number of residential pixels. As before, this was based on an assumption of constant population density throughout the residential pixels of each CD.

Stepwise regression analyses of estimated pixel population p_g on various subsets of the 80 candidate predictor variables were carried out. The results are summarised in Table 5.6.

Table 5.6 Summary of Stepwise Regression Analysis on Pixels Classified as Residential

Variable subset		Classification variables							
		6 TM bands		First 6		First 15		First 25	
		v	R^2	*v*	R^2	*v*	R^2	*v*	R^2
A	*6*	***5***	**.444**	*5*	.462	*6*	.400	***6***	**.386**
B	*20*	*10*	.489	*7*	.503	*7*	.440	*5*	.406
C	*40*	*8*	.491	*7*	.502	*8*	.449	*5*	.406
D	*40*	*10*	.532	*14*	.559	*10*	.473	*8*	.452
E	*60*	*17*	.559	*14*	.580	*11*	.485	*11*	.469
F	*80*	***13***	**.569**	*19*	.599	*11*	.492	***13***	**.489**
G	*100*	*14*	.576	*24*	.605	*12*	.489	*13*	.490

Key:
v = number of variables selected by the stepwise procedure
Variable subsets
A 6 TM bands
B As for A + 14 spectral transformations
C As for B + 20 squared terms
D As for B + 20 spatial transformations
E As for D + 20 interaction cross-product terms
F As for E + 20 squared spatial transformations (variances)
G As for F + 20 squared terms from B

The first model in Table 5.6 is the initial classification/regression model of Section 5.3.3. The evidence of Table 5.6 suggests that the investment in spectral and spatial transformations has paid modest dividends. Proceeding down the first column, we see that with a classification based on the simplest set of variables (the 6 TM bands), increasing the range of potential predictors in the regression step leads to a modest increase in R^2. Proceeding across the table, as the number and complexity of variables used in the classification phase is increased, so the R^2 of the subsequent regression phase is decreased. After some preliminary diagnostic examinations of the regression outputs, it was decided to postpone a consideration of issues such as transformation of the dependent variable, and instead proceed to the next step of

applying the regression equations obtained to the full image. To explore the impact of the refinements in either or both of the classification and regression phases it was decided to carry forward the 4 models shown in boldface in Table 5.6. With regard to the first of these, based on the 6 TM bands, it was decided for reasons which are discussed in Section 5.7.1 to retain all 6 variables, including band 4 which did not contribute significantly to the fit. (For details of the models, see Appendix F).

5.6.2 Application to the full image: population density estimates

The 4 regression equations obtained in Section 5.6.1 were used in turn to calculate a population estimate for every pixel classified as residential in the full image. All pixels classified as non-residential were assigned zero population. The resulting data bands were displayed as pseudocolour displays. The resulting images were similar to Image 7 based on the first classification/regression model, but appeared to show some improvement in the form of less spurious population in rural areas.

5.6.3 CD population estimates

Using the CD identification band, the pixel population estimates for each of the 4 models were aggregated to produce satellite-based population estimates P_S for each CD. These were divided by CD areas to produce CD population density estimates D_S. The two sets of estimates were compared with the corresponding ground truth figures P_G and D_G using the graphical and analytic methods of Section 4.6. Whilst the plots were reasonably linear in each case, the width of the spread and the magnitude of the R^2 values both at the pixel level and the CD density level made it clear that for all the refinement to date, the methodology was inherently limited by the quality of the imputed ground truth pixel populations which formed the base on which everything else was built. It was decided to investigate the iterative refinement procedure discussed in Section 2.9. A discussion of the regression results is postponed until after the next section, where the uniterated and iterated models are compared.

5.7 ITERATIVE REFINEMENT OF THE REGRESSION MODELS

5.7.1 Application of the iterative procedure

Each of the four chosen regression models (referred to in this section as 6/6, 6/13, 25/6 and 25/13 reflecting the number of variables used in the classification and regression phases) was iterated using the procedure described in Section 2.9. A program was written which took, as input from Minitab, the current values of the DV (the current imputed population of each pixel in the tests set) and the residuals of the fitted model, summed the residuals for each CD, and

output new values of the DV (new imputed populations) calculated using the adjustment formula (described in Section 2.9.1 and derived in Section 7.2.2).

With regard to the first of these, based on the 6 TM bands, it was decided to retain all 6 variables including band 4 which did not contribute significantly to the fit, the rationale being that it may become significant when the data was adjusted i.e. there may be an underlying relationship between population and band 4 which was masked by the poor quality of the initial raw data estimates. This was borne out in the implementation; after 2 iterations all 6 bands were significant at the .01 level and remained so from then on.

The values of R^2 and the regression coefficients of the 6/6 model after each iteration are shown in Table 5.7. This illustrates two features that were common to all four models. Firstly, in each case the R^2 value increased sharply after the first iteration, and continued to increase monotonically but at an ever diminishing rate. After about 6 iterations the magnitude of the increase had diminished to something in the order of one-tenth of a percentage point per iteration. Secondly, the magnitude of all of the regression coefficients in this model but one (that of b1) also increased at an ever diminishing rate (this was also true of most but not all the coefficients in the other models). This was not unexpected, considering the geometry of the bivariate case discussed in Section 2.9, and might be expected to lead to estimates which better reflect the range of variation in the population densities of individual pixels. A corollary of this increased range was that the proportion of negative estimates increased with each iteration. Trading this off against the extent of convergence, it was decided at this point to work with six iterations. (This was later reviewed – see Chapter 7.)

Table 5.7 Iterative Refinement of a Regression Model for Pixel Population based on the 6 TM bands

	Iteration										
Coefficient	0	1	2	3	4	5	6	7	8	9	10
Constant	1.033	1.161	1.357	1.556	1.736	1.889	2.019	2.126	2.215	2.289	2.350
b1	0.087	0.119	0.128	0.129	0.127	0.124	0.120	0.117	0.115	0.112	0.110
b2	0.114	0.164	0.187	0.199	0.206	0.211	0.214	0.217	0.219	0.221	0.222
b3	-0.136	-0.192	-0.215	-0.225	-0.229	-0.230	-0.231	-0.231	-0.232	-0.232	-0.232
b4	-0.003	-0.005	-0.008	-0.009	-0.011	-0.013	-0.014	-0.015	-0.016	-0.017	-0.017
b5	-0.039	-0.058	-0.068	-0.074	-0.079	-0.082	-0.084	-0.086	-0.088	-0.089	-0.090
b7	0.064	0.096	0.113	0.124	0.131	0.137	0.142	0.146	0.149	0.152	0.154
R^2	0.444	0.751	0.819	0.838	0.845	0.849	0.851	0.852	0.854	0.854	0.855
% increase		*69.1*	*9.1*	*2.3*	*0.8*	*0.5*	*0.2*	*0.1*	*0.2*	*0.0*	*0.1*

Table 5.8 summarises the fit of the four models without iteration and after 6 iterations. Note that in the case of the two 13-variable models, the iterative procedure was applied to the set of

variables originally selected rather than the selection process being repeated at each iteration. The latter procedure was explored, but resulted in negligible further improvement.

Table 5.8 Coefficients of Determination for Four Regression Models for Pixel Population: With and Without Iterative Refinement

	Model			
	6/6	6/13	25/6	25/13
R^2 without iteration	.444	.569	.386	.492
R^2 after 6 iterations	.851	.877	.834	.846

It can be seen that whilst the iteration procedure substantially improved the fit of all four models, the improvement was greatest in the simpler models which utilised only the basic 6 TM bands in the regression phase. Of course it must be emphasised that the substantiveness of these improvements was purely speculative at this stage, based as they were on optimal manipulations of incompletely determined data values.

5.7.2 Application of iterated models to the full image: population density estimates

As was described in Section 5.6.2 for the raw (uniterated models), the 4 regression equations obtained after iteration were each used to calculate a population estimate for every pixel classified as residential in the full image. All pixels classified as non-residential were assigned zero population. Again, when these were displayed as pseudocolour images (see Image 8) there was some evidence of further improvement in the form of less spurious population in rural areas.

5.7.3 Some technical issues

As with the aggregate models considered in Chapter 4, some problems were encountered at this stage with the distribution and range of estimated pixel populations. The populations assigned to urban pixels under both classification schemes were symmetrically distributed with a range from about 0.2 to 6 persons per pixel. However, for various reasons (the extended nature of rural dwellings and associated outbuildings and the radiometric similarity of country roads and residential streets being obvious ones) under both classifications the number of residential pixels was overestimated in the low population density rural areas, resulting in assigned populations in the range 0.1-0.4 persons per pixel. Because these areas were so extensive they contributed a substantial proportion of pixels to the sample, and the resulting mixture distribution was very positively skewed. Logarithm and square root transformations of the dependent variable were considered but because of the difficulty of implementing the iterative

process (based on additive constraints) in the context of nonlinear transformations to the data, it was decided to postpone closer examination of this strategy until the iterative procedure had been evaluated.

Another related issue was that, even with the raw linear models, the fitted population estimates for some pixels were negative. As was discussed above, the iterative process increases the spread of the assigned pixel populations, which further exacerbated this rather counter-intuitive result. These negative populations were essentially caused by the over-estimation in the number of residential pixels in rural areas, and it was felt that in statistical terms they might constitute a self-compensating correction. However, it was decided to examine the effect of setting a zero threshold i.e. setting all negative estimates to zero.

5.8 CD POPULATION AND POPULATION DENSITY ESTIMATES: COMPARISON OF THE MODELS

Table 5.9 and the accompanying Figure 5.1 summarise the performance of the four linear regression models developed, each in four variant forms: raw; iterated; thresholded; iterated and thresholded, in terms of their capacity to accurately estimate the population densities of the 138 CDs in the primary study area. Table 5.10 and Figure 5.2 present a parallel summary based on total population rather than population density.

The 16 models were compared with regard to accuracy of estimates of CD density and CD population. The criteria included:

- extent of bias indicated by intercepts and slope coefficients;
- strength of relationship, indicated by coefficient of variation R^2;
- accuracy of estimation for individual CDs, indicated by error standard deviation *s*, and mean and median relative errors;
- accuracy of estimation of total population for the study region and the urban area;
- overall visual assessment of plots.

Four models stood out from the rest as performing consistently well on all criteria. These were the iterated and iterated + thresholded variants of the 6/6 and 25/6 models. In summary, the best predictive performance was achieved by a classification phase based on either the 6 untransformed TM bands or an extended set of 25 variables, followed by an iterated regression based on the 6 untransformed TM bands only, and optionally followed by an adjustment of negative pixel population estimates to zero.

It was concluded that whilst it is possible to model more complex relationships between population and reflectance characteristics of the individual residential pixels within a particular sample by the use of a range of spectral and spatial transformations, the refinements identified

and implemented in this analysis were not robust and did not translate into improved estimates of CD or regional aggregates.

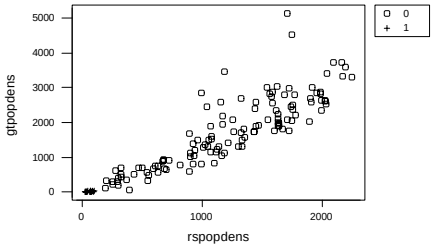
In particular, the spectacularly worst results were obtained for the 6/13 model, which was extremely volatile, producing large negative population estimates for a number of CDs. On investigation, these were due in each case to a very small number of pixels with pathologically negative fitted values (the worst case being –192!). Further examination showed that the most heavily weighted terms in this model in both positive and negative directions were some of the spatial variation terms, particularly those involving the pairwise difference measure. These terms contributed much to the fit of the model to the data from the training set, but were also its downfall when distributional tails not represented in the sample data were revealed in the full image. The pixels most effected were on shorelines. They were classified as residential because of their spectral makeup without reference to spatial characteristics, but were then rather too emphatically assigned low population on the basis of the strong linear spatial feature. The fact that this problem did not occur with the 25/13 model is probably due to the inclusion of spatial characteristics at the initial classification stage.

Table 5.9 Summary of Selected Models for Estimating Census Collection District Population Densities: Based on a Two-phase Pixel Classification and Regression Procedure

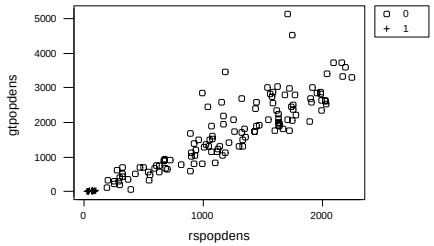
Model type	Number of classifiers and predictors	Variant of model	D vs. $\hat{D}$ Regression coeffts.* (unforced & forced)	R^2	s
2	6/6	Raw	-138 + 1.54; 1.45	.80	493
		Thresholded	-138 + 1.55; 1.45	.80	493
		Iterated	**-41 + 1.14; 1.12**	**.82**	**476**
		Iterated & thresholded	**-52 + 1.14; 1.12**	**.82**	**461**
2	6/13	Raw			
		Thresholded	Not calculated		
		Iterated	See text		
		Iterated & thresholded			
3	25/6	Raw	-150 + 1.51; 1.41	.79	504
		Thresholded	-150 + 1.51; 1.41	.79	504
		Iterated	**-51 + 1.14; 1.11**	**.81**	**490**
		Iterated & thresholded	**-65 + 1.15; 1.11**	**.80**	**491**
4	25/13	Raw	268 + 1.47; 1.67	.73	577
		Thresholded	0 + 1.50; 1.50	.79	504
		Iterated	634 + 1.01; 1.37	.67	642
		Iterated & thresholded	132 + 1.10; 1.17	.79	506

* Intercept + slope; slope when forced through origin.

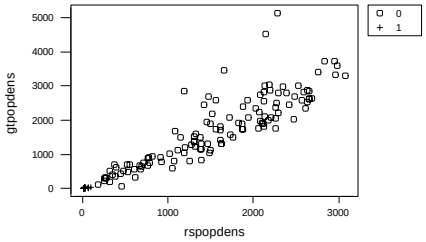
Figure 5.1 Population Density Estimates for 138 Census Collection Districts in the Primary Study Area: Ground Truth vs. Remote Sensing Estimates from Four Variants of Four Models



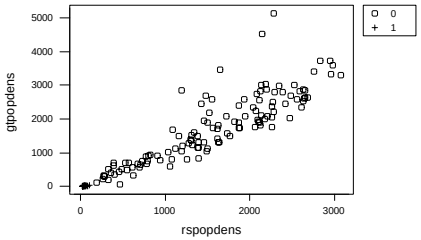
6/6 raw



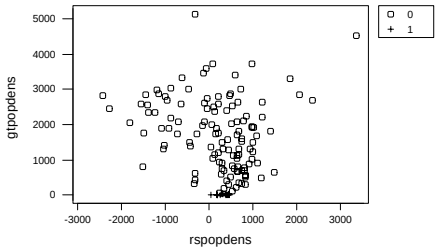
6/6 thresholded at zero



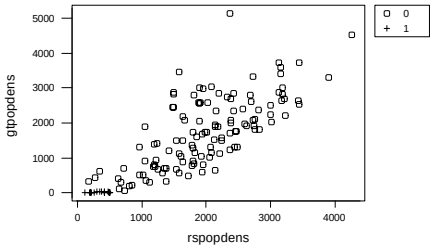
6/6 iterated



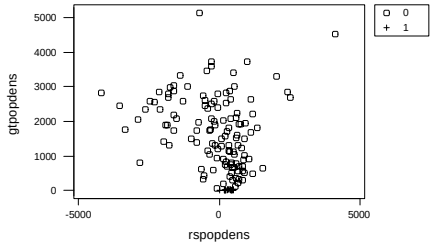
6/6 iterated & thresholded at zero



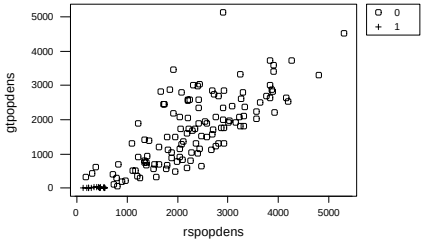
6/13 raw



6/13 thresholded at zero

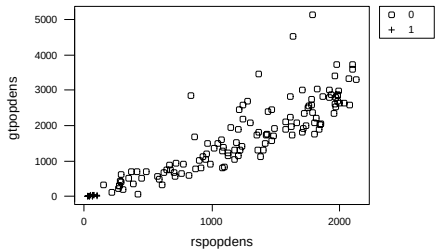


6/13 iterated

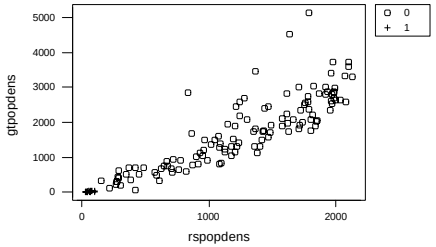


6/13 iterated & thresholded at zero

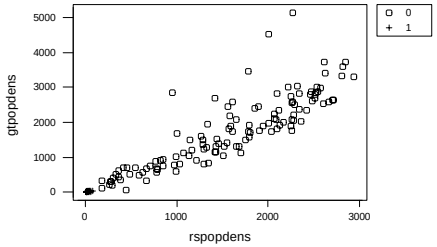
Figure 5.1 Population Density Estimates for 138 Census Collection Districts in the Primary Study Area: Ground Truth vs. Remote Sensing Estimates from Four Variants of Four Models (continued)



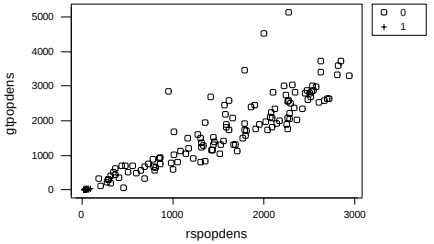
25/6 raw



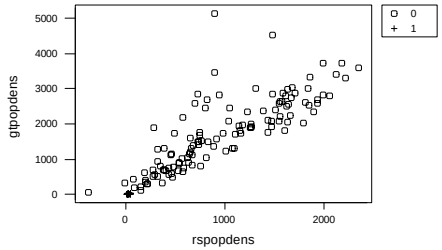
25/6 thresholded at zero



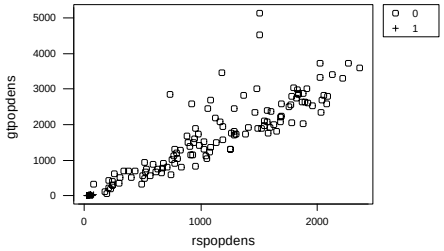
25/6 iterated



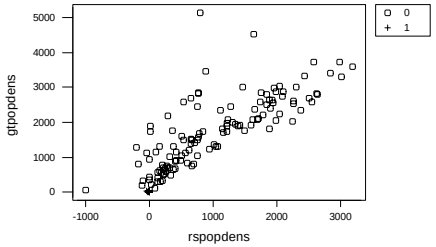
25/6 iterated & thresholded at zero



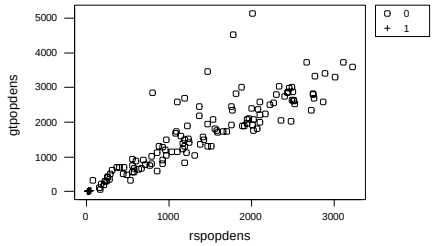
25/13 raw



25/13 thresholded at zero



25/13 iterated



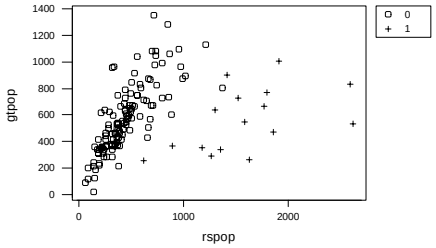
25/13 iterated & thresholded at zero

Table 5.10 Summary of Selected Models for Estimating Census Collection District Populations Based on a Two-phase Pixel Classification and Regression Procedure

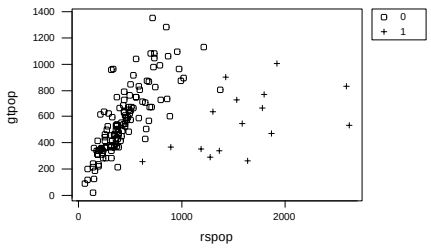
Model type	Number of classifiers and predictors	Variant of model	Ballarat Statistical District (n=138)							Ballarat urban area (n=122)						
			D vs. $\hat{D}$ Regression coeffts.[1] (unforced & forced)	R^2	s	Mean % error	Median % error	Est. total pop.	Total % error[2]	D vs. $\hat{D}$ Regression coeffts.[1] (unforced & forced)	R^2	s	Mean % error	Median % error	Est. total pop.	Total % error[2]
1	6/6	Raw	443 + .225; .693	.16	237	53.7	29.2	80173	+1	196 + .845; 1.18	.60	166	32.7	26.0	54848	-22
		Thresholded	444 + .224; .691	.16	237	53.8	29.2	80245	+1	196 + .845; 1.18	.60	166	32.7	26.0	54850	-22
		Iterated	**203 + .646; .932**	**.50**	**184**	**29.5**	**14.9**	**79160**	**-0**	**134 + .806; 1.01**	**.65**	**155**	**24.7**	**14.0**	**66824**	**-5**
		Iterated & thresholded	**281 + .473; .825**	**.37**	**205**	**36.6**	**17.0**	**85370**	**+8**	**140 + .790; 1.00**	**.65**	**156**	**24.9**	**14.3**	**67315**	**-4**
2	6/13	Raw														
		Thresholded	Not calculated													
		Iterated	See text													
		Iterated & thresholded														
3	25/6	Raw	442 + .224; .686	.17	237	51.4	28.3	81147	+2	214 + .780; 1.14	.55	176	32.2	25.0	56549	-19
		Thresholded	443 + .223; .684	.17	237	51.6	28.3	81239	+3	214 + .780; 1.14	.55	176	32.2	25.0	56551	-19
		Iterated	**235 + .577; .897**	**.44**	**194**	**31.3**	**17.3**	**80888**	**+2**	**158 + .757; .99**	**.59**	**167**	26.1	**15.3**	**67262**	**-4**
		Iterated & thresholded	**325 + .394; .780**	**.31**	**215**	**38.8**	**17.8**	**87062**	**+10**	**165 + .738; .98**	**.59**	**168**	26.3	**16.2**	**67869**	**-3**
4	25/13	Raw	460 + .412; 1.07	.22	229	52.7	41.4	56174	-29	204 + 1.10; 1.57	.59	168	48.3	41.2	41134	-41
		Thresholded	407 + .333; .875	.21	230	43.9	32.5	69037	-13	170 + .985; 1.31	.62	162	32.4	28.9	50271	-28
		Iterated	505 + .300; 1.18	.11	245	73.6	51.4	31869	-60	378 + .663; 1.52	.27	224	63.1	42.3	36340	-48
		Iterated & thresholded	214 + .701; 1.04	.46	191	26.0	16.2	70745	-11	122 + .934; 1.14	.64	158	21.7	15.4	59254	-16

1 Intercept + slope; slope when forced through origin 2 Ground truth populations are: BSD 79179; Urban 70222

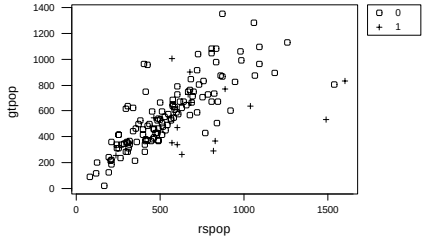
Figure 5.2 Population Estimates for 138 Census Collection Districts in the Primary Study Area: Ground Truth vs. Remote Sensing Estimates from Four Variants of Four Models



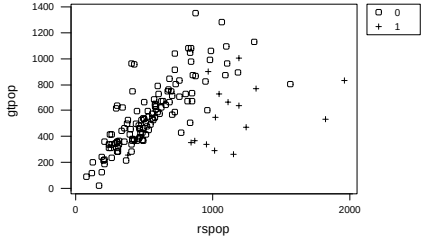
6/6 raw



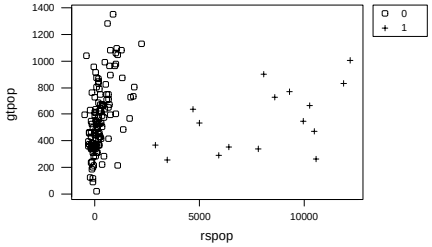
6/6 thresholded at zero



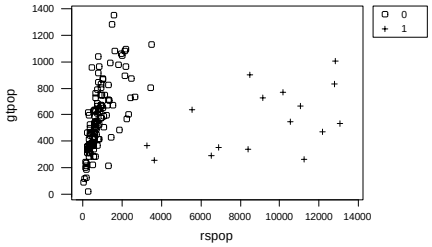
6/6 iterated



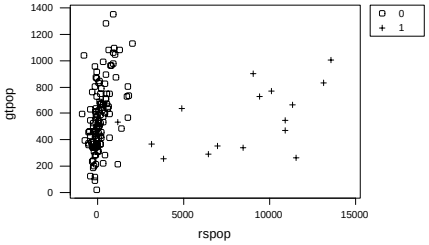
6/6 iterated & thresholded at zero



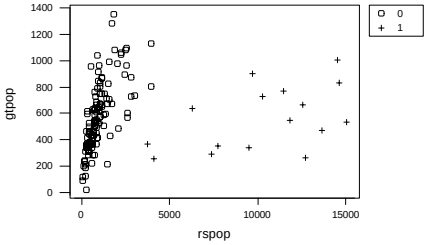
6/13 raw



6/13 thresholded at zero

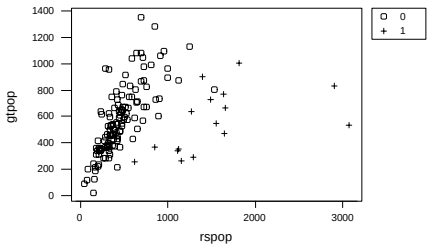


6/13 iterated

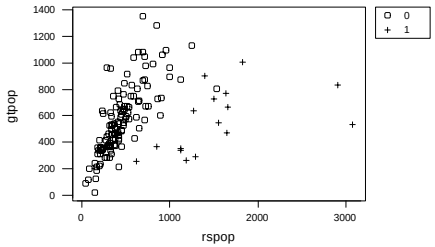


6/13 iterated & thresholded at zero

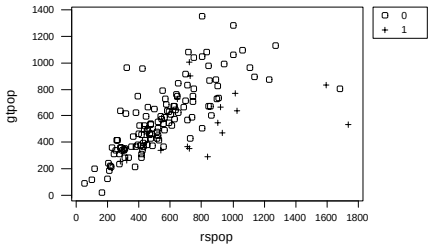
Figure 5.2 Population Estimates for 138 Census Collection Districts in the Primary Study Area: Ground Truth vs. Remote Sensing Estimates from Four Variants of Four Models (continued)



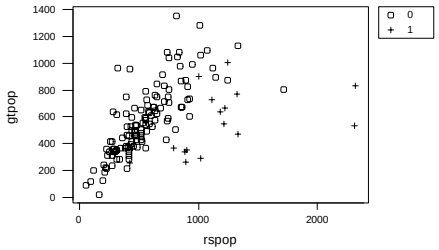
25/6 raw



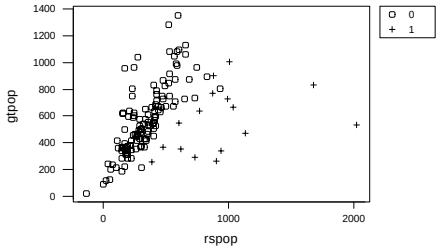
25/6 thresholded at zero



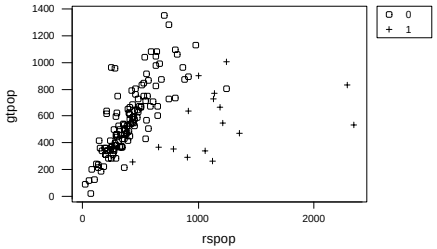
25/6 iterated



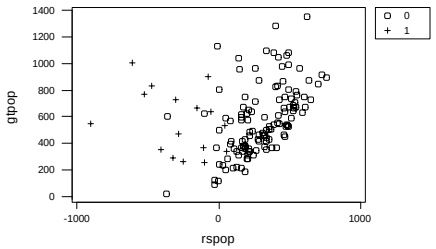
25/6 iterated & thresholded at zero



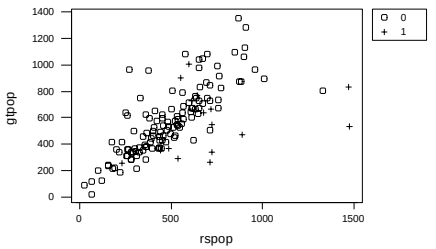
25/13 raw



25/13 thresholded at zero



25/13 iterated



25/13 iterated & thresholded at zero

With regard to the classification phase, the use of a range of spectral and spatial transformations did improve the rate of correct classifications of residential pixels in the training set, and a visual assessment (and the point in the previous paragraph) suggested that this did translate into an incrementally better classification for the whole image. Ultimately however this too failed to produce improved estimates of population or population density.

On closer inspection of the four models, it was clear that zero thresholding had the clear effect of inflating non-urban population estimates.

Considering all of this, together with the principle of parsimony and in particular the notion that simplicity is more likely than complexity to lead to robustness, it was decided to proceed with the simplest 6/6 procedure i.e. to use only the 6 untransformed TM bands for both classification and regression.

The regression equation is summarised in Table 5.11A. This equation, in conjunction with a maximum likelihood classification based on the 6 TM bands, produced estimated population densities of individual CDs with R^2=.82, and estimated CD populations with R^2=.50 overall and R^2=.65 in the urban areas. The mean and median relative errors in CD estimates were 29.5% and 14.9% respectively for the region overall, and 22.7% and 14.0% for the urban area. The total population of the study area was estimated to within 1% accuracy, and that of the urban area was underestimated by 5%.

5.9 TRANSFORMATIONS OF THE DEPENDENT VARIABLE

When the linear model was examined more closely (see Chapter 7) it became clear that its performance was best in the middle range of population densities and worst at the extremes, where it tended to underestimate high densities and overestimate low densities. In the light of these shortcomings, and considering the positively skewed distribution of the imputed pixel populations, it was decided to investigate the use of square root and logarithmic transformations of the pixels.

The iterative refinement procedure was appropriately modified to accommodate the transformations. Linear models were estimated for the square root of the pixel population, and for the natural logarithm of the pixel population. Both initial models had R^2=.47 and R^2_b=.41. Each model was refined iteratively. Convergence to within .002 in R^2_b was achieved in 5 iterations for the square root model (R^2_b=.79) and 3 iterations for the logarithmic model (R^2_b=.77). The resulting regression equations are summarised in Table 5.11.

Table 5.11 Alternative Regression Models for Estimating the Population Associated with a Pixel Classified as Residential

A. Linear model

$$\hat{p}_{pixel} = c_0 + \sum_{i=1}^{nbands} c_i b_i$$

Predictor (TM band)	Coefficient	Standardised coefficient	Standard deviation	*t*	*p*
Constant	2.0187		0.1361	14.83	0.000
b1	0.12024	0.647	0.00521	23.10	0.000
b2	0.21398	0.647	0.01407	15.21	0.000
b3	-0.23112	-0.928	0.00816	-28.31	0.000
b4	-0.01381	-0.069	0.00265	-5.20	0.000
b5	-0.08401	-0.809	0.00241	-34.85	0.000
b7	0.14158	0.732	0.00471	30.05	0.000

R^2=.82

B. Linear model (based on zero threshold)

$$\hat{p}_{pixel} = c_0 + \sum_{i=1}^{nbands} c_i b_i$$

Predictor (TM band)	Coefficient	Standardised coefficient	Standard deviation	*t*	*p*
Constant	1.5767		0.1580	9.98	0.000
b1	0.11343	0.683	0.00604	18.77	0.000
b2	0.17801	0.603	0.01633	10.90	0.000
b3	-0.18380	-0.826	0.00947	-19.40	0.000
b4	-0.01582	-0.088	0.00308	-5.13	0.000
b5	-0.05599	-0.603	0.00280	-20.01	0.000
b7	0.08772	-0.508	0.00547	16.04	0.000

R^2=.75

C. Square root model

$$\hat{p}_{pixel} = (c_0 + \sum_{i=1}^{nbands} c_i b_i)^2$$

Predictor (TM band)	Coefficient	Standardised coefficient	Standard deviation	*t*	*p*
Constant	0.70326		0.07347	9.57	0.000
b1	0.06782	0.754	0.00281	24.13	0.000
b2	0.10343	0.646	0.00760	13.62	0.000
b3	-0.11272	-0.935	0.00441	-25.58	0.000
b4	-0.00307	-0.032	0.00143	-2.15	0.032
b5	-0.03164	-0.629	0.00130	-24.32	0.000
b7	0.05175	0.553	0.00254	20.35	0.000

R^2_b=.79

D. Logarithmic model

$$\hat{p}_{pixel} = \exp(c_0 + \sum_{i=1}^{nbands} c_i b_i)$$

Predictor (TM band)	Coefficient	Standardised coefficient	Standard deviation	*t*	*p*
Constant	-2.1388		0.2633	-8.12	0.000
b1	0.16268	0.771	0.01003	16.21	0.000
b2	0.21593	0.578	0.02737	7.89	0.000
b3	-0.25917	-0.920	0.01606	-16.13	0.000
b4	0.00510	0.023	0.00511	1.00	0.318
b5	-0.05622	-0.482	0.00472	-11.91	0.000
b7	0.09912	0.456	0.00923	10.74	0.000

R^2_b=.77

As a corollary of these analyses, because of the counter-intuitive nature of the occurrence of negative population estimates, the modified zero-threshold iterative refinement procedure was also applied in the linear case. The resulting modified linear equation is also summarised in Table 5.11.

The similar form of the central linear function in each of these four models is consistent with the spectral relationships discussed in Section 5.4.2. The standardised regression coefficients take into account the range of variation of each predictor and hence give a better indication of the relative contribution of each variable. Each formula can be thought of as beginning with a benchmark or datum (around 2 persons per pixel in the case of the linear equation), and adjusting up or down according to a fairly evenly weighted comparison between on the one hand bands 1, 2 and 7, and on the other hand bands 3 and 5, with band 4 playing a more minor role. Within pixels classified as residential, higher population is associated with higher levels of bands 1, 2 and 7, and with lower levels of bands 3 and 5.

As for the earlier models examined, the regression equations were each used to calculate a population estimate for every pixel classified as residential in the full image. All pixels classified as non-residential were assigned zero population. When these were displayed as pseudocolour images there was some evidence of further improvement in the models with a transformed population variable, in the form of less spurious population in rural areas.

The population densities and total populations of the 138 CDs in the primary study area were also estimated. The results are summarised in Table 5.12 and Figure 5.3.

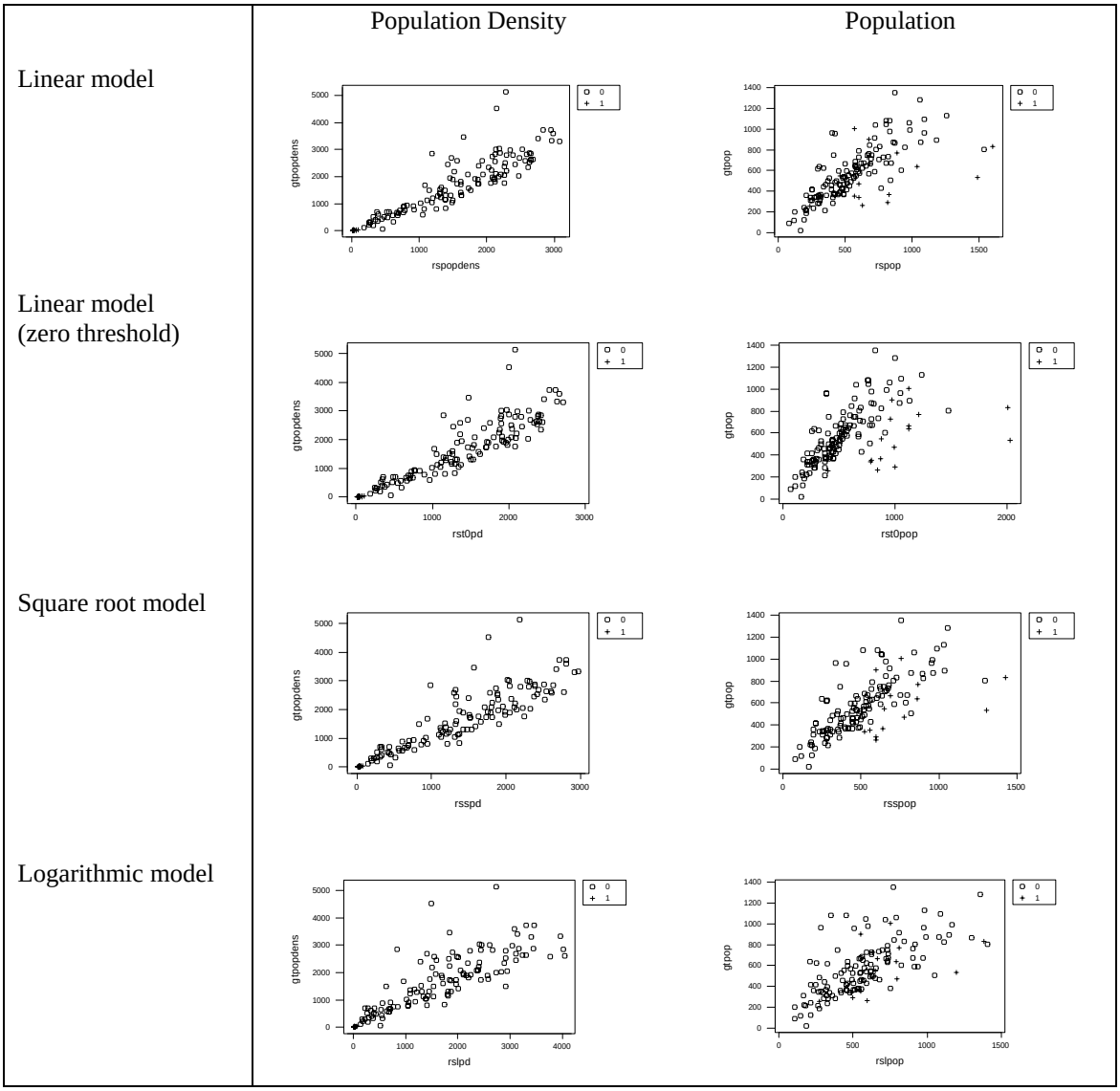
The original linear model performed better than the modified linear model and the other two models on such criteria as higher R^2 values, slope coefficients near unity, and accurate estimates of total population. However, when this model was applied to the secondary study area some questions of robustness arose, and in particular some issues relating to the negative population estimates it produced in areas of low population density. The modified linear model, though it was developed with zero thresholding, could still produce negative estimates when applied to a full image, and so it held no obvious advantage over the original linear model and was used no further at this stage (it was revisited later in the context of a broader theoretical examination – see Chapter 7). The two models involving transformations were investigated further by applying them to the secondary study area. A more detailed evaluation and comparison is carried out in that context in Section 6.3.

Table 5.12 Comparison of Estimated Population Densities and Populations for Census Collection Districts[1]: Based on a Two-phase Procedure of Classification with Various Regression Models

Model	Statistical District								Urban areas							
	D vs. $\hat{D}$ Regression coeffts.[1] (unforced & forced)		R^2	*s*	Mean % error	Median % error	Est. total pop.	Total % error[2]	*D* vs. $\hat{D}$ Regression coeffts.[1] (unforced & forced)		R^2	*s*	Mean % error	Median % error	Est. total pop.	Total % error[2]
Population density																
Linear	-41 + 1.14	1.12	.82	476					-59 + 1.15	1.12	.75	507				
Linear (zero threshold)	-78 + 1.27	1.22	.81	482					-112 + 1.29	1.22	.75	512				
Square root	39.6 + 1.15	1.18	.79	506					72 + 1.14	1.18	.72	538				
Logarithmic	232 + .856	.956	.71	595					355 + .803	.956	.63	624				
Population																
Linear	203 + .646	.930	.50	184	29.5	14.9	79160	-0	134 + .806	1.01	.65	155	24.7	14.0	66824	-5
Linear (zero threshold)	302 + .471	.867	.36	207	34.3	22.5	79620	+1	155 + .820	1.06	.64	158	32.8	20.6	62498	-11
Square root	192 + .725	1.02	.49	185	29.0	17.5	72644	-8	134 + .886	1.11	.61	163	25.9	16.2	60854	-13
Logarithmic	219 + .608	.914	.43	196	36.0	18.0	80595	+2	207 + .648	.943	.47	192	26.3	15.9	69406	-1

1 *n*: Ballarat Statistical District 138; Ballarat urban 122
2 Intercept + slope; slope when forced through origin
3 Ground truth populations are: Ballarat Statistical District 79179; Ballarat urban 70222

Figure 5.3 Population Density and Population Estimates for Census Collection Districts[1]: Ground Truth vs. Remote Sensing Estimates



1 Ballarat Statistical District: n=138

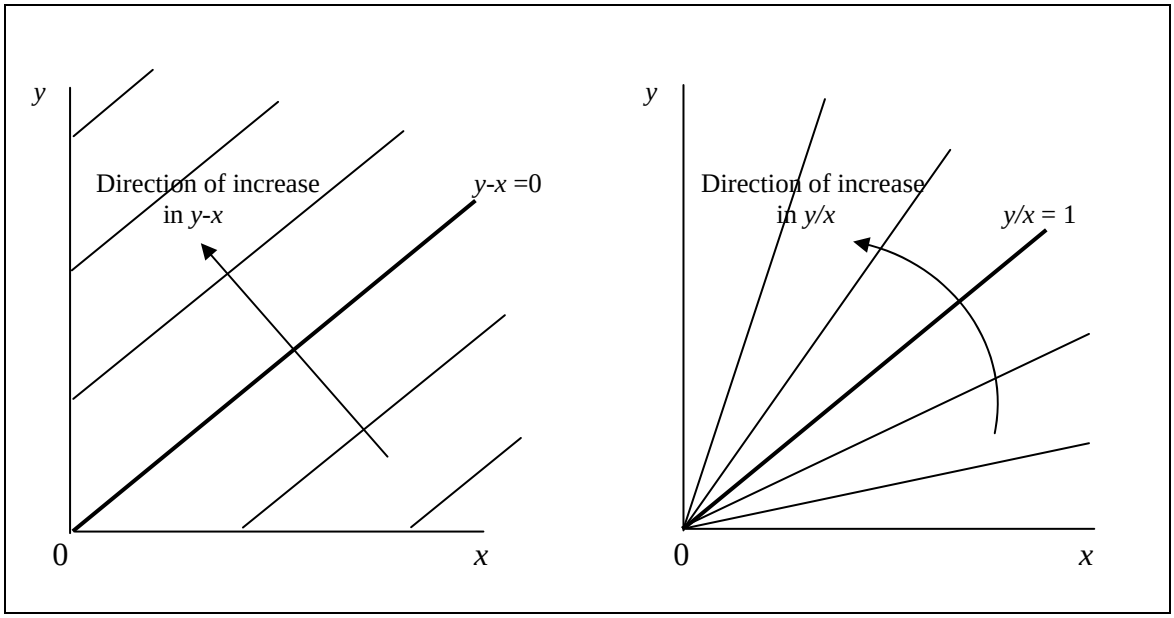
5.10 THE FORM OF THE MODEL

It is also worth considering why, after examination of so many spectral and spatial transformations, a simple linear form of model was ultimately selected.

With regard to the spectral dimensions, there is a conceptual and computational relationship between band differences embodied in the central linear function and the band ratios which were seen in Section 5.4.2 to correlate to some extent with human habitation. Figure 5.4 portrays the relationship between the level curves (contours) of a ratio and a difference of two positive variables. The difference (y-x) ranges over ($-\infty,\infty$) with a central value of zero. The ratio ranges over ($0,\infty$), with a "central" value of 1. The level curves of the difference are

parallel lines, those of the ratio are radial lines. In a restricted region of the positive *x*-*y* quadrant, the two measures provide similar but subtly different alternative frameworks for quantifying differences. Of course, in the logarithmic model, subtraction corresponds to division of the untransformed data, and so differences in the linear function represent ratios of powers of the band values.

Figure 5.4 Contours of Differences and Ratios



The multivariate linear functional form can provide a first order approximation to almost any spectral transformation over a limited range. It is conjectured that the relationship between population and spectral characteristics may be too approximate and noisy for any higher order refinements to be robust.

As to spatial transformations, none of the statistical measures employed emerged either as strong discriminators of residential pixels or as strong surrogates of population among residential pixels. This was an unexpected result considering the visual perception of a characteristic mottled texture in residential areas, the promising simulation results (Section 2.5), and the successful use of texture measures by Webster (1996). However, Webster's measures were not localised texture measures, but rather were measures of homogeneity and pattern regularity (relating to urban street grids) calculated over larger areas (as in Chapter 4 of this report).

With regard to the failure of the simulation results to be reflected at the image classification stage, the problem would seem to have two aspects. Firstly, in a real image, linear features such as shorelines and rural roads are often extremely noisy and ill-defined at the local pixel level, and only become manifest perceptually because of their extension in one direction. Hence even

with high levels of noise, the simulation conditions may have been a very simplistic approximation to the reality. Secondly, even if a particular measure is able to discriminate these particular features well, it may be that this very specific focus is lost in the broad brush statistical approach of discriminant analysis and maximum likelihood classification. The problem is to detect a particular class of pixels which is relatively small in number overall but which potentially has a disproportionate misclassification impact on the residential class, and hence on population estimates.

In the present study, a methodology was developed for effectively reclassifying pixels at a later stage of the analysis using contextual information about average population density (see Section 6.3).

5.11 SUMMARY

In this chapter a two-phase procedure of classification followed by regression modelling has been investigated. Starting from the baseline of a single regression equation for all pixels, and with ground truth population estimates based on an assumption of uniform population density within each CD, two crucial steps for improving the estimation model were developed. The first was an initial classification of pixels into residential and non-residential, and the second was the iterative refinement of the initial estimates of pixel population during the regression modelling stage.

After extensive examination, it was found that the use of spectral and spatial transformations of the six TM bands at both classification and regression stages led to improved model fits within the training set data. However these improvements were not robust, in that when they were applied to the whole image, they did not result in any improvement in the estimated population densities or populations of individual CDs, or of urban or regional totals.

Three candidate models were selected for further investigation, all involving at their core a simple linear function of the six TM bands, but in two cases utilising square root and logarithmic transformations of the dependent variable. The efficacy of these dependent variable transformations is considered further in Section 6.3.

Chapter 6

Application of Estimation Algorithms to a Second Geographical Area

6.1 INTRODUCTION

The validity of the procedures developed in Chapters 4 and 5 was investigated by applying them to a second image. The CD aggregate method of Chapter 4 is discussed (and dispensed with) in Section 6.2. The individual pixel method of Chapter 5 is considered in Section 6.3. In response to problems which were identified, a number of refinements were made, which are reported in Section 6.4.

The secondary study area was thought to be geographically, culturally and temporally similar to that of the primary Ballarat study area, being an urban/rural area centred on the neighbouring provincial city of Geelong (see Section 3.2), taken from the same Landsat TM scene of February 14, 1988. Consequently, it was expected to provide a reasonable test of validity but only the most moderate test of robustness (see Section 2.11.1).

Ground truth population and dwelling data values were calculated for this second area using the methodology of Section 3.6 (see Appendix D). The image was radiometrically corrected for haze and geometrically corrected for earth rotation skew, as outlined in Chapter 3. It was then co-registered to the CD boundaries using a cubic warp based on 27 ground control points (see Figure 3.2) with nearest neighbour resampling.

As for the primary image, the CD boundaries were overlaid in vector form on the TM image for purposes of display. For purposes of analysis, they were also used to define 225 regions on the image, and hence, using the ER Mapper IF INREGION( ) function, a data band containing the CD identification of each pixel was also defined. This band formed the essential link between the remote sensing data based on pixels and the ground truth data based on CDs.

6.2 ESTIMATION BASED ON CENSUS COLLECTION DISTRICT AGGREGATES

Twelve representative models which used CD aggregates for estimating population (six models) and dwelling densities (six models) were chosen from those considered in Chapter 4. The results of applying these models to the 225 CDs of the secondary image are summarised in Tables 6.1 to 6.4. Table 6.5 shows some key indicators of comparative performance of these models on the primary and secondary image. In Figure 6.1 plots of the first and the sixth model for each of population density and dwelling density are compared for the primary and secondary study areas.

For the six population models, the R^2 values for Ballarat increased monotonically from .537 to .843; for Geelong the range was .453 to .741, with the value for model 6 being lower than that for model 5. Correspondingly, the residual standard deviation *s* was in each case larger for Geelong than for Ballarat. These figures indicate a general degradation of performance when applying any of the estimating equations to the secondary image. A similar pattern of reduced correlation and accuracy was observed for the dwelling density models.

As well as reduced correlation, there was also evidence of bias, with consistent underestimation of both population and dwelling densities in the secondary study area. Eleven of the twelve slope coefficients were considerably higher than unity, indicating that the ground truth values tended to be larger than the remote sensing estimates.

Nevertheless, the plots for model 6 (Figure 6.1) confirm that the relationship between D and $\hat{D}$ remained linear, indicating that the underlying form of the link between population density and the particular linear combination of remote sensing characteristics chosen in Ballarat remained valid for the Geelong data. The reduction in the level of correlation and the associated broader spread of points on the Geelong plots is to be expected when validating any procedure (see discussion in Chapter 2). However the slope changes suggested some more systematic calibration problem requiring investigation.

As to the relative robustness of the 6 model types, the first two models, which were only included as a starting point, and which were quite inadequate even on the primary data, were also the least robust with regard to the slope coefficient. At the other extreme, the slope coefficient of .885 for the dwelling density model 6 stands out as an obvious anomaly. On closer inspection, models 6 for both population and dwelling densities are anomalous in other ways also. The population model 6 had a lower R^2 and larger *s* than model 5 (Table 6.1), in contrast to the result for Ballarat (Table 4.10), where model 6 was clearly better in both respects. Whilst the dwelling model 6 did have the largest R^2 and smallest *s* (Table 6.2), the margins were much less than for Ballarat (Table 4.11).

Table 6.1 Application of Population Density Models based on Primary Study Area CD Aggregates to Secondary Study Area

Model type	Class of predictors	Number of predictors	Regression equation	D vs. $\hat{D}$ Regression coeffts.* (unforced & forced)	R^2_b	s
1	Band mean	4	72.3 – 135.6 b5 + 332.0 b7 -151.0 b3 + 61.6 b4	-329+1.59 ; 1.40 (Fig 7.1B1)	.453	878
2	Band mean	4	$(-1.4 – 2.33\ b5 + 5.37\ b7 + 1.22\ b4 – 2.04\ b3)^2$	147+1.43 ; 1.52	.448	882
3	Mean, $(mean)^2$	6	$(-171.34 – 0.140\ s1 + 9.220\ b1 + 0.0344\ s7 + 4.874\ b4 – 1.952\ b5 – 0.0359\ s4)^2$	-148+1.27 ; 1.20	.599	752
4	Ratios & difference to sum ratios	4	$(345.35 – 68.41\ r57 – 275.56\ r14 + 226.89\ d14 + 120.30\ r15)^2$	40+1.17 ; 1.19	.601	750
5	Mean, std dev, variance, coefft of variation	9	$(75.20 – 2.19\ b5 – 245.20\ b7c – 70.36\ b4c + 0.171\ b7v + 0.851\ b4 + 2.88\ b7 – 0.124\ b1v + 69.57\ b1c + 70.37\ b5c)^2$	345+1.04 ; 1.19	.741	605
6	Mean & std dev of spectral transformations at pixel level	6	$(530.10 + 0.278\ rh123 – 92.34\ r14 – 60.81\ r57 +165.91\ ds35 – 1.308\ rh125 – 0.370\ rh125s)^2$	64.3+1.09 ; 1.11 (Fig 7.1B2)	.718	630

* Intercept + slope; slope when forced through origin

Table 6.2 Application of Dwelling Density Models based on Primary Study Area CD Aggregates to Secondary Study Area

Model type	Class of predictors	Number of predictors	Regression equation	D vs. $\hat{D}$ Regression coeffts.* (unforced & forced)	R^2_b	s
1	Band mean	4	89.7 - 49.77 b5 + 110.49 b7 + 23.81 b4 - 44.18 b3	-228+1.69 ; 1.34 (Fig 7.1B3)	.564	286
2	Band mean	4	$(-0.879 - 1.451\ b5 + 3.119\ b7 + 0.784\ b4 - 1.051\ b3)^2$	-5.9+1.54 ; 1.53	.556	289
3	Mean, $(mean)^2$	7	$(-176.63 + 0.0101\ s5 + 12.147\ b2 - 0.266\ s2 + 0.019\ s7 + 6.540\ b4 - 0.050\ s4 - 3.035\ b5)^2$	17.3+1.11 ; 1.13	.717	231
4	Ratios & difference to sum ratios	5	$(160.0 - 27.9\ r57 - 153.6\ r14 + 58.8\ d17 - 33.3\ r47 + 146.6\ r15)^2$	10.9+1.15 ; 1.16	.730	225
5	Mean, std dev, variance, coefft of variation	9	$(59.21 - 1.29\ b5 - 169.32\ b7c + 0.130\ b7v - 34.37\ b4c + 0.508\ b4 + 35.46\ b5c + 1.42\ b7 - 0.0715\ b1v + 37.68\ b1c)^2$	125+.985 ; 1.13	.796	197
6	Mean & std dev of spectral transformations at pixel level	8	$(312.05 - 72.85\ r14 + 0.260\ b3 - 0.520\ b2s + 244.92\ ds15 - 30.45\ r57 - 165.42\ ds25 - 0.716\ rh125 - 0.158\ rh125s)^2$	-33.7+.916 ; .885 (Fig 7.1B4)	.815	187

* Intercept + slope; slope when forced through origin

Table 6.3 Summary of Estimated Census Collection District Populations
Based on Application of Population Density Models based on Primary Study Area CD Aggregates to Secondary Study Area

Model type	Class of predictors	Geelong Statistical District (*n*=225)						Geelong urban area (*n*=214)					
		Slope (forced)	R^2	*s*	Mean % error	Median % error	Est. tot. (% error*)	Slope (forced)	R^2	*s*	Mean % error	Median % error	Est. tot. (% error*)
1	Band mean	.06	.00	267	202	37.7	221132 (+50%)	.67	.05	254	59.6	36.5	134990 (-5%)
2	Band mean	.14	.00	268	141	40.7	179991 (+22%)	.86	.04	255	55.3	39.5	114035 (-20%)
3	Mean, $(mean)^2$	.12	.00	267	151	26.7	222618 (+51%)	.60	.04	256	57.2	25.7	147294 (+4%)
4	Ratios & difference to sum ratios	.17	.00	267	142	25.8	208987 (+41%)	.65	.03	256	54.2	24.4	142035 (-0.2%)
5	Mean, std dev, variance, coefft of variation	.10	.00	267	188	23.0	179135 (+21%)	.78	.06	252	41.9	22.1	119238 (-16%)
6	Mean & std dev of spectral transformations at pixel level	.08	.00	267	126	18.4	217264 (+47%)	.89	.21	231	39.9	17.1	138354 (-3%)

* Ground truth populations are: GSD 147910; Urban 142250

Table 6.4 Summary of Estimated Census Collection District Dwelling Counts Based on Application of Dwelling Density Models based on Primary Study Area CD Aggregates to Secondary Study Area

Model type	Class of predictors	Geelong Statistical District (*n*=225)			Geelong urban area (*n*=214)		
		Mean % error	Median % error	Est. tot. (% error*)	Mean % error	Median % error	Est. tot. (% error*)
1	Band mean	223.5	36.8	75249 (+47%)	70.8	35.3	50916 (+3%)
2	Band mean	137.4	42.4	56914 (+11%)	58.9	41.0	40154 (-19%)
3	Mean, $(mean)^2$	150.2	21.5	71017 (+39%)	57	20.9	51345 (+4%)
4	Ratios & difference to sum ratios	121.2	24.1	63983 (+25%)	55.7	22.7	49917 (+1%)
5	Mean, std dev, variance, coefft of variation	176.6	20.8	60062 (+18%)	39.15	19.8	43532 (-12%)
6	Mean & std dev of spectral transformations at pixel level	164.5	21.7	96547 (+89%)	53.1	20.7	62537 (+27%)

* Ground truth dwelling numbers are: GSD 51078; Urban 49411

Table 6.5 Comparison of Twelve Population and Dwelling Density Estimation Models for Primary and Secondary Study Areas

Model type	Class of predictors	Primary study area: Ballarat							Secondary study area: Geelong						
		D vs $\hat{D}$ Coefft (forced)	R^2_b	s	Region Median % error	Region Total % error	Urban Median % error	Urban Total % error	D vs $\hat{D}$ Coefft (forced)	R^2_b	s	Region Median % error	Region Total % error	Urban Median % error	Urban Total % error
	Population density estimates														
1	band mean	1.00	.537	739	31.0	-	28.5	+26	1.40	.453	878	37.7	+50	36.5	-5
2	band mean	1.05	.557	739	32.9	+91	30.5	+7	1.52	.448	882	40.7	+22	39.5	-20
3	mean, $(\text{mean})^2$	1.04	.755	550	24.1	+13	20.0	+7	1.20	.599	752	26.7	+51	25.7	+4
4	ratios & difference to sum ratios	1.03	.762	541	20.9	+16	19.4	+2	1.19	.601	750	25.8	+41	24.4	-0
5	mean, std dev, variance, coefft of variation	1.01	.780	521	23.0	+41	21.0	-2	1.19	.741	605	23.0	+21	22.1	-16
6	mean & std dev of spectral transformations at pixel level	1.02	.843	441	17.4	+14	13.6	+1	1.11	.718	630	18.4	+47	17.1	-3
	Dwelling density estimates														
1	band mean	1.00	.560	265	32.7	-	28.0	+30	1.34	.564	286	36.8	+47	35.3	+3
2	band mean	1.06	.584	258	37.5	+49	29.5	+9	1.53	.556	289	42.4	+11	41.0	-19
3	mean, $(\text{mean})^2$	1.02	.817	171	22.9	+22	19.6	+4	1.13	.717	231	21.5	+39	20.9	+4
4	ratios & difference to sum ratios	1.02	.832	164	18.7	+17	16.0	+3	1.16	.730	225	24.1	+25	22.7	+1
5	mean, std dev, variance, coefft of variation	1.00	.878	139	21.6	+37	18.5	-0	1.13	.796	197	20.8	+18	19.8	+12
6	mean & std dev of spectral transformations at pixel level	1.01	.924	111	16.0	+9	15.4	+1	.885	.815	187	21.7	+89	20.7	+27

This most complex of all the models, based on a hybrid methodology of spectral transformation at pixel level followed by aggregation and averaging at CD level, did not live up to its considerable promise in the primary analysis phase. It would appear that this procedure might be over-fitted, over-engineered and over-tuned, attaining high performance on the primary image at the cost of a lack of robustness. The intermediate model types 3, 4 and 5 were consistent with regard to slope bias, but model 5 had much higher R^2 values (the highest in the case of population density).

Turning to the final measures of performance, the estimated populations of each CD and the population and dwelling totals for the entire study area and for the urban section, it is apparent from Table 6.4 that the estimates for individual CDs were even worse than in the case of Ballarat. The median percentage errors were higher in Geelong, both for the overall region and for the urban area, in all but two cases, both involving model 5. As in the case of Ballarat the Geelong region totals were extravagantly overestimated in all cases. As has been discussed in Chapter 4, in Ballarat this was primarily due to the overestimation of the low densities in the large rural CDs. In the case of Geelong, this effect is further exacerbated by the presence of some large industrial sites in medium sized non-urban CDs (an oil refinery, an aluminium smelter, a car assembly plant, a cement works and a salt works), all of which were assigned large spurious populations by the remote sensing algorithms. Both Iisaka and Hegedus (1982) and Lo (1995) reported similar problems with a few non-residential or otherwise anomalous study units.

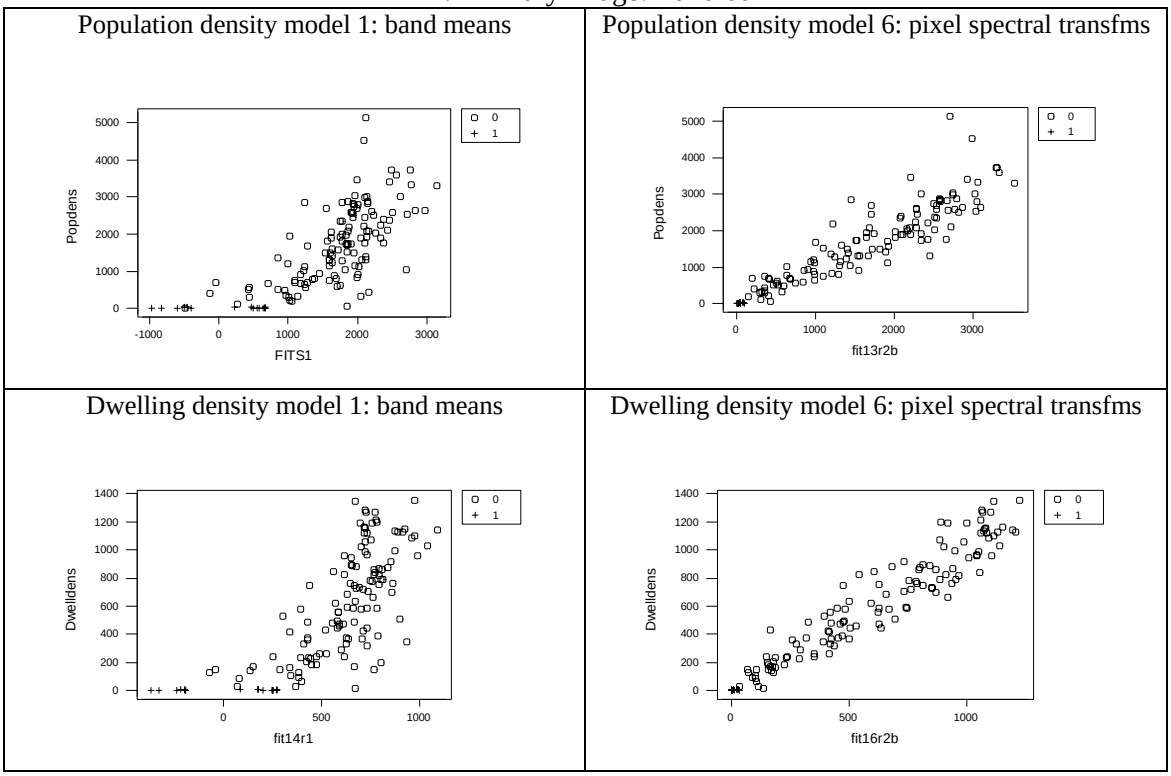
The combined effects of scale and nonlinearity lead to some non-intuitive results. For example, the (forced) slope coefficients for density are greater than 1.0 for all models but one, whilst the slope coefficients for urban population counts were considerably less than 1.0. However, in spite of all this, both population and dwelling totals for the urban area were quite accurately estimated by some of the intermediate models, particularly model 4.

These results prompted a closer examination of the demographic characteristics of the two study areas, which are summarised in Table 6.6. Some unexpected differences emerged.

The first four rows of Table 6.6 show the basic measures and counts. Then follow four characteristics: population density, dwelling density, persons per dwelling and percentage of dwellings which are non-separate houses, each calculated in two ways: as the ratio of the relevant regional totals, and as the mean of the corresponding ratio for each CD. In the case of the two densities, the two methods lead to very different results because of the great variations among CDs both in size and density. For persons per dwelling and percentage of non-separate dwellings there is no such variation, and the two results are almost identical. The two study areas were compared by forming ratios of each of the indices described.

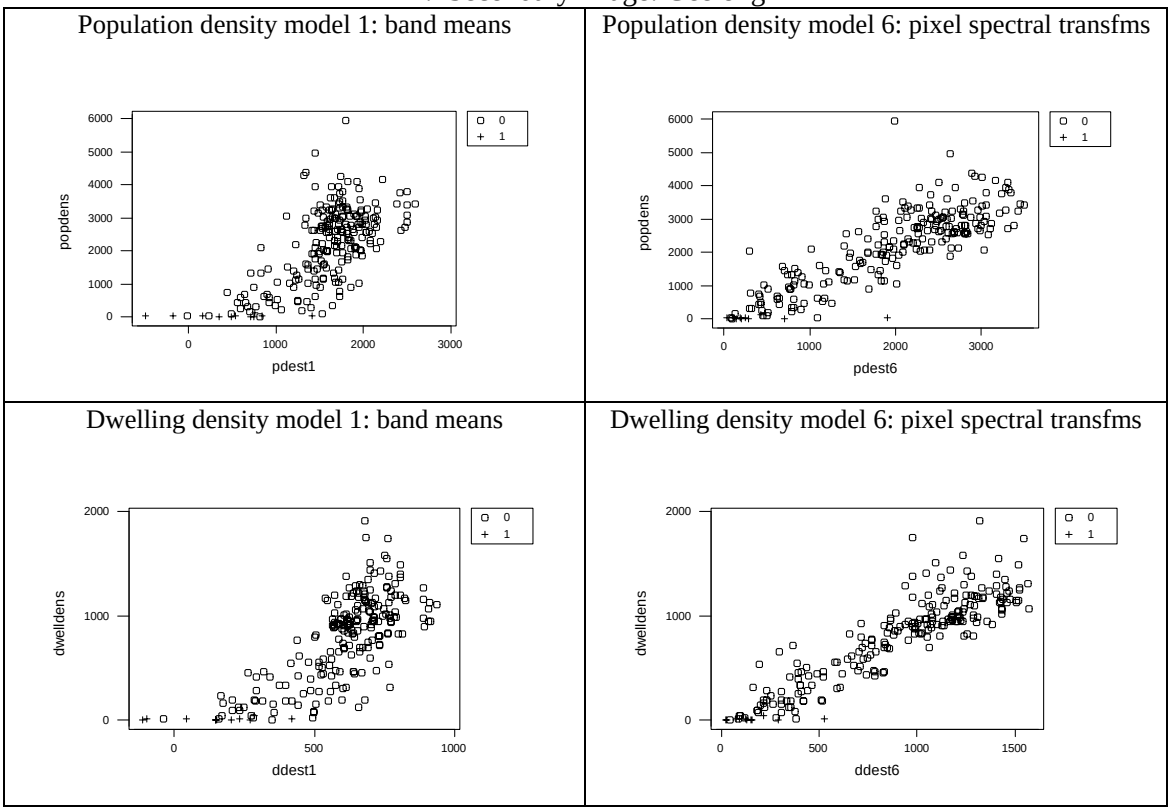
Figure 6.1 Population and Dwelling Density Estimates for Census Collection Districts: Ground Truth vs. Remote Sensing Estimates from Base and Enhanced Models

A. Primary image: Ballarat



0 Urban + Rural

B. Secondary image: Geelong



0 Urban + Rural

Table 6.6 Demographic Characteristics of Primary and Secondary Study Areas

Demographic characteristic	Primary study area: Ballarat		Secondary study area: Geelong		Ratio Geelong: Ballarat	
	BSD*	Urban	GSD**	Urban	Region	Urban
Area (sq.km.)	613.89	74.85	351.9	107.15	0.57	1.43
Population	79179	70222	147910	142250	1.87	2.03
Number of dwellings	26971	24368	51078	49411	1.89	2.03
Number of non-separate houses	5154	4721	9989	9586	1.94	2.03
Population density (persons/sq.km.)	129.0	938.2	420.3	1327.6	3.26	1.42
Mean CD population density	1556.9	1758.3	2133.3	2241.3	1.37	1.27
Dwelling density (dwellings/sq.km.)	43.93	325.56	145.15	461.14	3.30	1.42
Mean CD dwelling density	562.6	635.6	798.2	807.2	1.42	1.27
Persons per dwelling ratio	2.94	2.88	2.90	2.88	0.99	1.00
Mean CD Persons/Dwelling ratio	2.96	2.90	2.93	2.91	0.99	1.00
% Non-separate houses	19.11	19.37	19.56	19.40	1.02	1.00
Mean CD % non-separate houses	18.94	20.01	19.76	19.71	1.04	0.99

* Ballarat Statistical District
** Geelong Statistical District

The closest points of similarity were the ratio of persons to dwellings (occupancy ratio), which was almost identical at around 2.9 persons per dwelling for both areas overall and for both urban areas; and the percentage of non-separate dwellings, which again was almost constant at 19-20%.

There were however a number of differences. GSD had little more than half the area of BSD, but almost double the population. A much larger proportion of GSD than of BSD was urban, but even within the urban areas, the population density was 42% higher overall and the mean CD population density was 27% higher in Geelong than Ballarat. Since the occupancy ratios were the same for the two areas, the comparative figures for dwelling density were the same.

The similarity of the proportions of non-separate dwellings suggests that the higher densities in Geelong are not associated with more multi-dwelling structures, but rather with houses which are either smaller or closer together, or some combination of both, perhaps in different areas.

Figure 6.2 shows plots of the discrepancies $(\hat{D} - D)$[1] against *D* for models 4 and 6. In the case of population for both models and in the case of dwellings for model 6, there is a strong tendency in the case of Geelong for higher densities to be underestimated and lower densities overestimated. The same effect can be observed in a less pronounced form in the Ballarat data, mainly in association with a few high density outliers. The same problem was noted by Langford et al. (1991) and it is also apparent in the reported results of Iisaka and Hegedus (1982) and Webster (1996) though it was not discussed by them.

There is an attenuation or lack of sensitivity whereby the extremes of variation in density are not reflected in the remote sensing estimates. This may be an inherent consequence of the use of CD aggregates, related to the sensitivity issue discussed in the context of iterative refinement of the regression models in Section 5.6.1. Be that as it may, the effect has been exaggerated in the case of Geelong by the overall higher density. The nascent tendency in the Ballarat plots has become more pronounced in the Geelong plots because of the large number of CDs either with characteristics similar to the outliers in Ballarat (mainly associated with institutions or public housing), or with densities over 3000 persons/sq.km. As always, extrapolation beyond the range of the training data, in this case to higher range of population densities, is fraught with risk.

For no obvious reason, the trends described are not in evidence in model 6 for dwelling density, where the pattern of overestimation and underestimation is less dependent on density in the case of Ballarat, and where in the case of Geelong the trend is reversed by a dense cluster of CDs with densities around 1000 for which the density is overestimated.

In the light of these patterns of over- and under-estimation, it is possible to explain the apparent contradiction between the fact that on average individual CD densities are underestimated by all but one of these models, and yet in a number of cases reasonably accurate estimates are produced of total population and dwelling numbers for the whole Geelong urban area.

The contradiction is an artifact of the use of CDs as the unit of aggregation. CDs are designed to have approximately equal populations (within fairly broad tolerances), and so CD area is inversely related to population and dwelling density. As a result, when population totals are calculated by multiplying the density estimate for each CD by its area and summing, there is a tendency for a counterbalance between slight overestimation of density over relatively large areas and more substantial underestimation of density over relatively small areas.

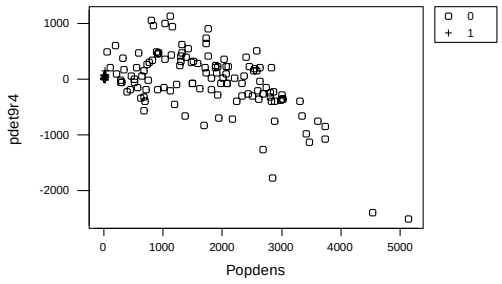
---

[1] The usual convention of defining a residual as $(D - \hat{D})$ has been reversed because the emphasis is on assessing the remote sensing estimate with reference to the benchmark of the ground truth figure. Technically, these discrepancies are not in fact residuals (see Section 2.12.2) Note also that the scales have been chosen to maximise the spread of points on each plot, to facilitate visual comparison of the shapes of the plots.

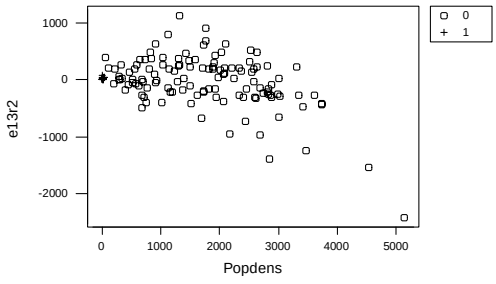
Figure 6.2 Estimation Error vs. Population or Dwelling Density

A. Primary image: Ballarat

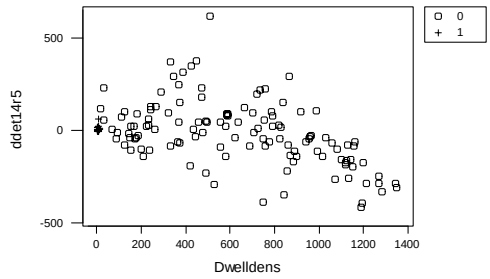
Population density model 4: ratios of band means



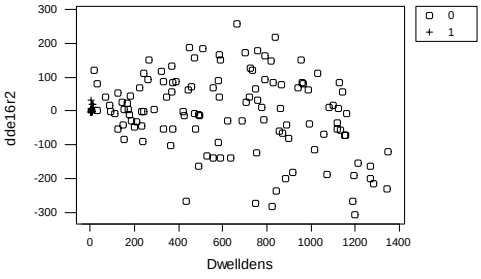
Population density model 6: pixel spectral transfms



Dwelling density model 4: ratios of band means



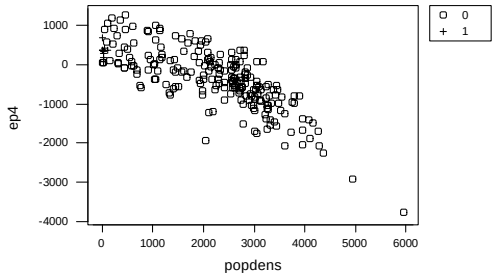
Dwelling density model 6: pixel spectral transfms



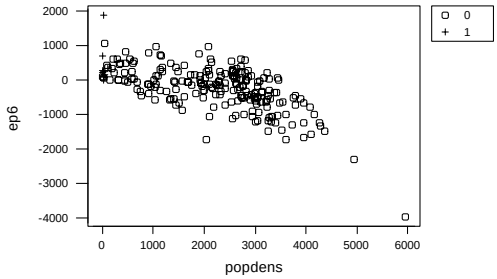
0 Urban + Rural

B. Secondary image: Geelong

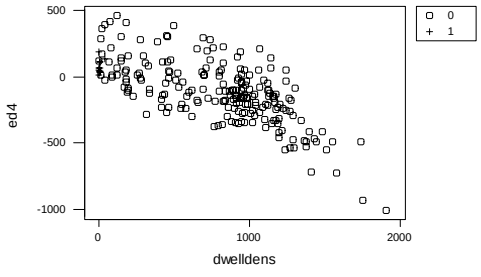
Population density model 4: ratios of band means



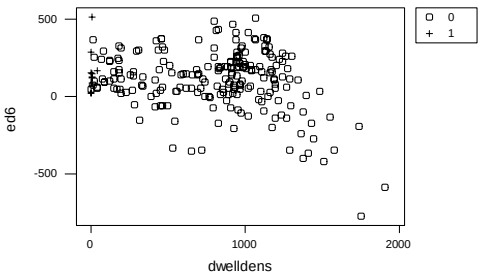
Population density model 6: pixel spectral transfms



Dwelling density model 4: ratios of band means



Dwelling density model 6: pixel spectral transfms



0 Urban + Rural

In summary, it seems that a methodology based on CD means and ratios of CD means can be tuned to model small area population and dwelling densities quite well in a single study area. Performance of models on the initial study area increases with complexity, but there would also appear to be some tradeoff between complexity and robustness. In general, the models are not robust to large variations in population density within the study area, or to moderate demographic differences between different study areas. In particular, there seems to be an attenuation effect, whereby the remote sensing estimates of density vary over a smaller range than the actual densities. This may be an artifact of the CD size distribution. A weighted least squares approach might bring about some improvement. A wider range of training data might also result in greater robustness.

Notwithstanding these limitations, considering the accuracy of the estimates of urban totals produced by the "intermediate" models, it is tempting (though risky from a sample size of 1!) to speculate that these results are not fortuitous. The linearity of the relationships exhibited in the secondary study area suggest there might be a kernel which is relatively robust above a threshold density, but which is then degraded by mechanisms associated with the multilevel analysis. It is conceivable that the use of CDs as the unit of aggregation both causes the problem of bias, but also provides the obverse mechanism for removing the bias in estimates of larger aggregations. It may be that the procedure is inherently more robust with respect to larger aggregations than it is with respect to the unit of spatial analysis, in this case CDs.

If this were the case, and if it remained the case at different scales, then analogously the methods of Chapter 5 based on individual pixels might be expected to be more robust at the level of CDs.

It was concluded that modelling at the level of CD aggregates had not produced a methodology which was either very accurate or very robust for estimating population at the level of CDs. Nor was there any obvious avenue for dealing with the problems identified, which were inherently at sub-aggregate level, whilst working with aggregates. Consequently, the remainder of the study was focussed on modelling at the level of individual pixels.

6.3 ESTIMATION BASED ON INDIVIDUAL PIXELS

A two-stage procedure for estimating population on the basis of the characteristics of individual pixels was developed in Chapter 5. This procedure consists of a classification step followed by an estimation step for pixels classed as residential. Three different estimation formulae (which estimated population, square root of population and logarithm of population respectively) were derived by regression modelling. The procedure was first applied to the Geelong image as a "black box" with no further training or refinement.

A maximum likelihood classification was carried out, using the covariance structure of the 6 TM bands from the Ballarat training sets to define the classes. The land cover classes in the Geelong image were not the same as Ballarat, perhaps the most extreme difference being the presence of the sea waters of Corio Bay, which were largely classified as industrial (which in a sense they probably are!). However, the residential class appeared to have been well discriminated.

The three different population estimation formulae were then applied. In each case, all non-residential pixels were set to zero population, and the final regression equation from Chapter 5 was applied to the values of the 6 TM bands for all residential pixels.

As before, aggregate figures for each CD were derived from these images, and compared with the ground truth populations and population densities. The results are compared with the corresponding results from the primary Ballarat study area in Tables 6.7 and 6.8, and Figure 6.3.

With regard to the linear model for population density, Figure 6.3 shows moderate to strong positive correlations in the Geelong figures, with rather more spread than for Ballarat at the high density end, and with rather more outliers for which the population density was markedly underestimated. Also as with Ballarat, most of the CDs for which the population density was overestimated have very low densities, and so do not stand out visually, although there are one or two moderately overestimated points at higher densities. On detailed examination, much of the underestimation was associated with the presence of large institutions, as was the case with the Ballarat data, but in addition in Geelong there was considerable underestimation in a pocket of contiguous inner city CDs containing many small old workers' cottages and terraced houses. This is a neighbourhood which has no parallel in Ballarat. Geelong is a port city with more large heavy industry; Ballarat was founded on gold mining, and is laid out more spaciously. Whilst there are many old miners' cottages, they tend to be on large allotments on the eastern side of the urban area, which is relatively sparsely settled to this day. It was established in Section 6.2 that the population density in the Geelong study area was considerably higher than that of the Ballarat study area, the ratios on two different measures between the two urban areas (which dominate the population density regressions because of the higher densities) being 1.42:1 and 1.27:1 (see Table 6.6).

The greater spread in the Geelong data was borne out by the somewhat lower values of R^2 (.74 overall, .69 for the urban area, and .85 when the 13 most extreme outliers were removed, compared to .82, .75 and .91 respectively for Ballarat). These moderate reductions of 8, 6 and 6 percentage points respectively indicated a substantial degree of robustness in the underlying form of the relationship established from the Ballarat data.

Table 6.7 Comparison of Estimated Population Densities for Census Collection Districts[1] in Primary and Secondary Study Areas: Based on a Two-phase Pixel Classification and Regression Procedure

Model	Statistical District				Urban areas				Statistical District: outliers omitted			
	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)		R^2	*s*	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)		R^2	*s*	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)		R^2	*s*
Linear												
Ballarat	-41 + 1.14	1.12	.82	476	-59 + 1.15	1.12	.75	507	- 52.9 + 1.10	1.07	.91	312
Geelong	168 + 1.42	1.51	.74	603	239 + 1.38	1.51	.69	616	70.3 + 1.42	1.46	.85	433
Geelong (local classification)	- 23.7 + 1.21	1.20	.73	615	0 + 1.20	1.20	.68	630	- 39.0 + 1.18	1.16	.81	485
Square root												
Ballarat	39.6 + 1.15	1.18	.79	506	72 + 1.14	1.18	.72	538				
Geelong	349 + 1.51	1.74	.70	650	46 + 1.45	1.74	.65	660				
Geelong (local classification)	365 + 1.10	1.28	.61	745	491 + 1.04	1.28	.54	754				
Logarithmic												
Ballarat	232 + .856	.956	.71	595	355 + .803	.956	.63	624				
Geelong	827 + 1.12	1.62	.53	815	984 + 1.03	1.62	.47	810				
Geelong (local classification)	1382 + .385	.819	.25	1025	1575 + .326	.819	.20	995				

1 *n*: Ballarat SD 138; Ballarat urban 122; Ballarat SD with outliers omitted 133;
Geelong SD 225; Geelong urban 214; Geelong SD with outliers omitted 212 (216 with local class).

2 Intercept + slope; slope when forced through origin

Table 6.8 Comparison of Estimated Populations for Census Collection Districts[1] in Primary and Secondary Study Areas: Based on a Two-phase Pixel Classification and Regression Procedure

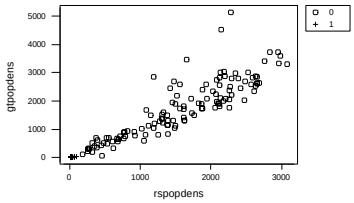
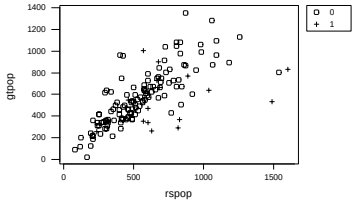
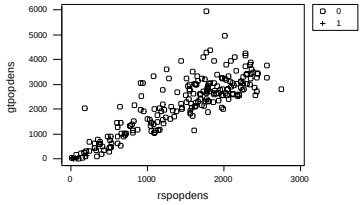
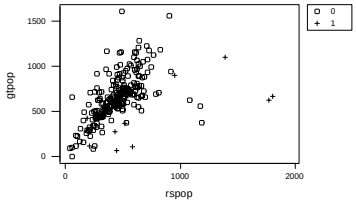
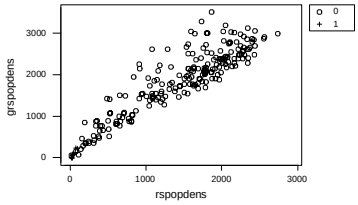
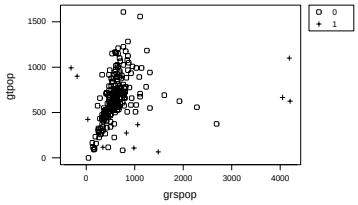
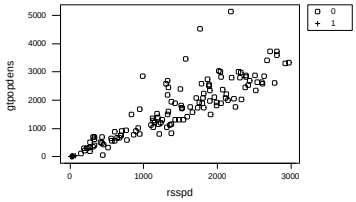
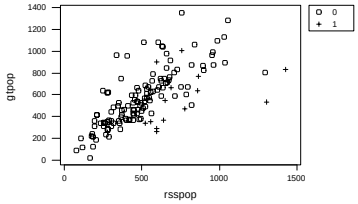
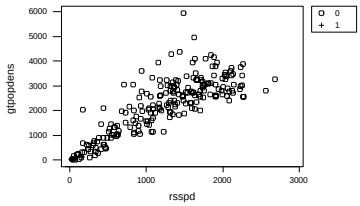
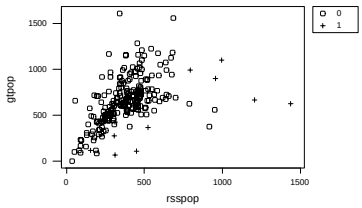
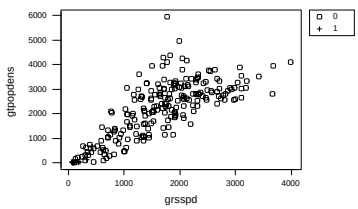
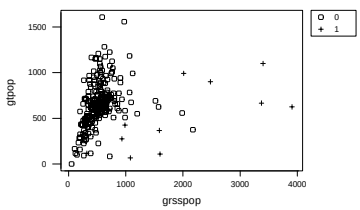
Model	Statistical District							Urban areas						
	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)		R^2	*s*	Mean % error	Median % error	Est. total pop.	% error[3]	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)		R^2	*s*	Mean % error	Median % error	Est. total pop.	% error[3]
Linear																
Ballarat	203 + .646	.93	.50	184	29.5	14.9	79160	-0	134 + .806	1.01	.65	155	24.7	14.0	66824	-5
Geelong	388 + .573	1.24	.25	231	41.2	32.3	105655	-29	281 + .849	1.38	.38	206	36.2	32.2	96729	-32
Geelong (local classification)	583 + .114	.66	.05	261	55.2	22.1	147361	-0	468 + .332	.93	.15	241	34.5	21.3	130754	-8
Square root																
Ballarat	192 + .725	1.02	.49	185	29.0	17.5	72644	-8	134 + .886	1.11	.61	163	25.9	16.2	60854	-13
Geelong	383 + .700	1.50	.23	234	46.1	43.1	87806	-41	305 + .958	1.66	.31	216	43.3	32.6	115723	-19
Geelong (local classification)	590 + .110	.711	.04	263	59.5	30.8	137412	-7	500 + .304	1.04	.10	247	38.9	29.8	137412	-3
Logarithmic																
Ballarat	219 + .608	.914	.43	196	34.3	22.5	80595	+2	207 + .648	.943	.47	192	32.1	20.6	69406	-1
Geelong	461 + .535	1.54	.14	249	46.9	46.6	82678	-44	449 + .608	1.65	.15	241	45.9	46.9	96729	-32
Geelong (local classification)	616 + .059	.593	.02	266	71.4	41.6	156777	+6	615 + .078	.734	.02	258	54.2	40.8	130754	-8

1 *n*: Ballarat SD 138; Ballarat urban 122; Geelong SD 225; Geelong urban 214

2 Intercept + slope; slope when forced through origin

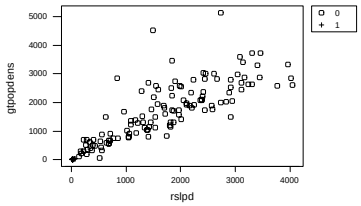
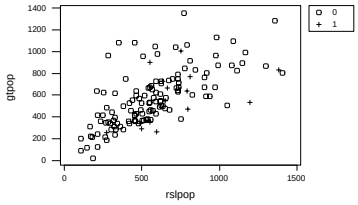
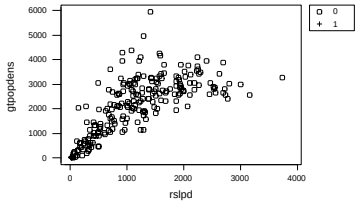
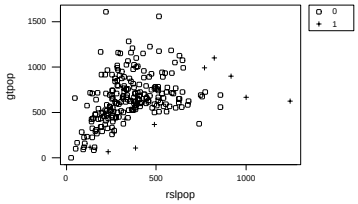
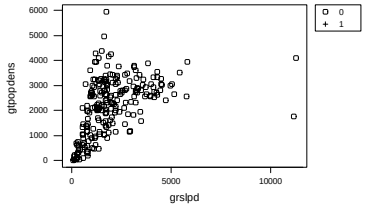
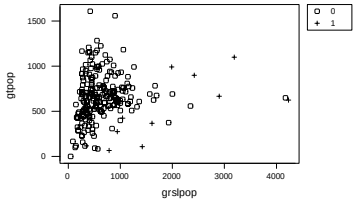
3 Ground truth populations are: BSD 79179; Ballarat urban 70222; GSD 147910; Geelong urban 142250.

Figure 6.3 Population Density and Population Estimates for Census Collection Districts[1]: Ground Truth vs. Remote Sensing Estimates for Primary and Secondary Study Areas

Model	Population Density	Population
Linear		
Ballarat		
Geelong		
Geelong (local classification)		
Square root		
Ballarat		
Geelong		
Geelong (local classification)		

1 *n*: Ballarat SD 138; Geelong SD 225

Figure 6.3 Population Density and Population Estimates for Census Collection Districts[1]: Ground Truth vs. Remote Sensing Estimates for Primary and Secondary Study Areas (continued)

Model	Population Density	Population
Logarithmic		
Ballarat		
Geelong		
Geelong (local classification)		

1 *n*: Ballarat SD 138; Geelong SD 225

However, there was a problem of scale. The slopes of the various lines of best fit range from 1.07 to 1.15 for Ballarat, and from 1.38 to 1.51 for Geelong – an increase of some 30-35%. This seemed to be related to the overall higher population density in Geelong, since the differential fell midway between the two population density ratios quoted in the previous paragraph. It is dealt with in the next section.

With regard to CD populations, the fit was much worse for the Geelong data than for Ballarat, both with regard to R^2 values (.25 overall and .38 for the urban area, compared with .50 and .65 respectively for Ballarat), and regional and urban totals (underestimations of 29% and 32% respectively, compared with underestimation errors of <1% and 5% respectively in Ballarat). The underestimation of the totals was obviously related to the similar underestimation of population densities. The low correlation levels were largely attributable to a relatively small number of quite extreme outliers. As for Ballarat, populations of many of the large rural CDs were greatly overestimated, because of the propensity for overclassification of rural pixels as residential.

As has been discussed in Section 6.1, this was further exacerbated in the case of Geelong by the presence of some large industrial sites in medium sized non-urban CDs. One inner urban largely industrial CD, and two outer suburban CDs also stand out as having had their populations over-estimated. In the case of the two CDs on the urban fringe, it is conjectured that this may have been associated with residential development works.

It was these dual problems of overestimation at low densities and underestimation at high densities which prompted the exploration of curvilinear models. However, the pattern of Geelong results for the square root model, with regard to both population density and population, was marginally worse than that of the linear model, as it had been for the Ballarat data. The logarithmic model was the least robust of all, with greater reduction in R^2 values, and least accurate population estimates. These trends are also apparent from Figure 6.3. Since it now seemed that neither of these models was going to provide a feasible alternative to the linear model, further thought was given to how the linear model might be made more robust.

6.4 MODIFICATIONS TO THE PIXEL-BASED ESTIMATION PROCEDURE

6.4.1 A proposed explanation of scaling error

After some consideration, it was decided that the scale problem with the linear model was occurring at the classification stage rather than the regression stage. To illustrate why this is so, consider a hypothetical pair of CDs of equal area (let us say 1 sq. km. for simplicity and without loss of generality) – one in Ballarat with a population of say 1000 persons and hence a population density of 1000 person/sq. km., and the other in Geelong with a population of 1300 persons and hence a population density of 1300 person/sq. km. Suppose initially that all pixels in each CD are classified as residential. In each case, population associated with each pixel is assigned using the “population formula” - the linear combination of TM bands developed in Chapter 5. Pixels with high values (i.e. those at the “built environment” end of the scale) will be assigned higher populations and pixels at the lower end of the scale (the “natural materials” end) will be assigned lower populations.

Now the greater number of people in the Geelong CD than the Ballarat CD must be accommodated in some combination of 3 ways:

(1) more dwellings of the same average size

(2) the same number of dwellings but of larger average size

(3) the same number of dwelling structures of the same average size.

Cases (1) and (2) imply similar average population densities within dwellings in the two areas – this would seem to be a reasonable expectation in two culturally similar areas in close physical proximity. Case (3) implies more people per structure, which could come about in four ways:

(3a) more crowded private dwellings

(3b) a higher density type of accommodation – more multi-dwelling structures (flats, townhouse etc)

(3c) in particular, multi-storey multi-dwelling structures

(3d) multi-storey single dwellings.

Working backwards from the end of the list, (3d) is regarded as the least likely to lead to anomalies in remote sensing estimates, since residents in multi-storey houses are likely to be reasonably affluent, and the land area saved by "going up" is likely to be utilised for larger carports and other outbuildings which will have the same spectral signature as a larger house.

Case (3b) and more especially (3c) do provide a mechanism for housing extra population which is "hidden" from the satellite sensor. However, except in the case of institutions, this was expected to make a relatively minor contribution in the present instance. It is considered further in Section 6.3.4.

Case (3a) provides a mechanism for anomalies within both study areas, notably the underestimation of population density in areas of public housing, but there is no reason to expect a differential effect between Ballarat and Geelong.

All of which leaves a mix of cases (1) and (2) as the most likely mechanism. Each of these should lead in the Geelong scenario to more pixels with spectral characteristics at the high population end of the scale and fewer at the low population end. To the extent that the populations of individual pixels remain within the range where the regression equation can be validly applied, a proportionately higher population estimate for the CD should result.

And yet that appeared not to be the case.

The above argument was conditional on two propositions:

(1) that all pixels were classified as residential, and

(2) that the populations of individual pixels remained within the range where the regression equation could be validly applied.

The methodology of Section 6.3 used both a classification scheme and a regression equation developed on the Ballarat training sets. The higher levels of population density encountered in Geelong would have been under-represented if represented at all in Ballarat, which is relevant to both of these propositions, but especially the first of them.

In practice, even in a "purely" residential CD, not all pixels were classified as residential. It was conjectured that in Geelong CDs with high population densities, many pixels might be assessed by the Ballarat-trained classification algorithm as belonging to another class (commercial and industrial would seem to be the most likely classes to be thus confounded). In this way, some of the pixels in fact associated with high contributions to the CD population would be assessed as having zero population.

The issue of the validity of the regression equation seemed both harder to address and, fortuitously, less serious. It was judged to be less serious because any reduction in validity would be a gradual process at each end of the scale, unlike the misclassification problem, which is sudden and profound in its effect. It was harder to address because it was hoped to establish a reasonably robust relationship which did not require re-estimation, with the need for ground truth population data, at each implementation.

Classification, on the other hand, requires only qualitative ground truthing in the form of training areas representative of the residential class and an appropriate set of other classes (they need not be the same classes for different areas). This is a routine image processing activity in the remote sensing context, and requiring it to be done afresh for each image does not limit or invalidate the general approach.

Accordingly, it was decided to perform a second classification of the Geelong image, based on local training sets.

6.4.2 Local classification of the secondary study area

Twelve broad categories listed in Table 6.9 were defined. Nine corresponded to a greater or lesser degree with classes used in the primary study area. The other three - sea water, salt works and quarry – related to prominent features which had no parallel in Ballarat. It was decided not to include the remaining three Ballarat classes, one of which (pine plantation) was not present in the Geelong image and the other two of which had been relatively small in extent and not well discriminated in the Ballarat image.

For all categories, training sets were selected visually using local knowledge and a quasi-natural colour RGB image of the study area. A maximum likelihood classification based on the 6 TM bands was carried out and the results displayed in a pseudocolour image, which was then visually compared with the corresponding image based on the Ballarat training sets. As expected, a greater number of pixels, particularly in inner urban areas, appeared to have been classified as residential. Unfortunately, the same was true of many rural areas, which was to exacerbate the problem of overestimation in these areas.

Table 6.9 Categories of Land Use and Land Cover: Comparison of Classes Used in the Two Study Areas

Ballarat	**Geelong**
Residential	Residential
Industrial	Industrial
Commercial	Commercial
Bare ground: dark coloured soils	Dark soils
Bare ground: light coloured soils	
Dry grass, pasture, crops	Light soils & dry grass, pasture & crops
Green grass, pasture, crops	Light green vegetation: grass, pasture, crops
Native eucalypt forest and scrub	Dark green vegetation: forest and scrub
Pine plantation	
Fresh water	Fresh water
Public use	
Road	Road
	Sea water
	Salt works
	Quarry

The population algorithm was then applied to this classification, and the results again displayed in a pseudocolour image, which was then visually compared with the corresponding image based on the Ballarat training sets. As expected, population values seemed higher overall, and the peaks seemed more intense.

As before, aggregate figures for each CD were derived from this image, and compared with the ground truth populations and population densities. The results are compared with the results from the primary Ballarat image and the earlier Geelong image in Tables 6.7 and 6.8 and Figure 6.3.

With regard to population density, the improvement over the Ballarat-trained model was as anticipated. The slope coefficients ranged from 1.16 to 1.21 (down from 1.38 to 1.51), leaving a margin of inconsistency between Geelong and Ballarat figures (1.07 to 1.15) of only 5-10%. This was attained at the cost of marginal reductions in the R^2 values. The residual discrepancy may be due to bias in the regression formula developed from the lower density Ballarat training set, but it may also relate to cases (3b) and (3c) of Section 6.4.1 – the effect of multi-dwelling or multi-storey structures. This is considered in the next section.

With regard to CD populations, the effect of the locally trained classification was more mixed. Population estimates for some of the low density outlier CDs exhibited great volatility. Some swung in the direction of overestimation and some swung in the opposite direction – the estimated populations of two CDs were actually negative. This problem is due to a combination of the tendency towards overclassification of rural pixels as residential, combined with the fact that the linear population estimation formula can produce negative estimates at the low density

end of the scale. As a result of these few quite extreme outliers, the R^2 values for this model plummeted to negligible levels.

It seems that the accuracy of estimates in low density areas is quite sensitive to the range of variation in the residential class. The overclassification of rural pixels as residential was perhaps exacerbated by the fact that the selection of residential training sets in the secondary study area was less precise, not being informed by either the detailed local knowledge nor the supporting objective information than had been the case in the primary study area.

Notwithstanding this, the estimates of the total populations for both region and urban areas were much improved (underestimation by <1% and 8% respectively), and were comparable in accuracy with those obtained for the primary study area.

Whilst the square root and logarithmic models also produced reasonably accurate estimates of the overall regional and urban populations, the accuracy of estimates for individual CDs was substantially degraded, as is apparent from the appearance of the plots as well as the reduced R^2 values. As a result, it was decided at this stage to continue with the linear model only.

6.4.3 Adjustments for under-estimation and over-estimation

In the results from the linear model for both study areas there remained, at all but the lowest densities, a bias towards under-estimation. This was manifested both in the estimated population densities of individual CDs, and also in the estimates of the total populations of the regions and of their urban areas. In terms of the regional totals, this tendency was to some degree counterbalanced by a tendency toward overestimation in the (generally non-urban) areas with the lowest densities, though this was coupled with an instability which led in some cases to quite substantial negative estimates. It had been hoped that logarithmic and square root transformations of population might overcome these problems, but this did not come to fruition because of the overall volatility and lack of robustness in the resulting models. It was decided instead to explore ways to apply corrections to the linear model at the extremes of density.

6.4.4 CD-based adjustment for multiple dwelling structures

As discussed in Section 6.4.1, it was considered that the overestimation might be due to the presence of a “hidden” component of population associated with multiple dwelling structures. Whilst some such structures were present in the suburban CDs constituting the Ballarat training set, it was conjectured that the population formula might not be robust to higher densities. Since 1986 census figures were available for the proportion of dwellings in each CD which were of other types than separate houses, it was decided to investigate the effect of upwardly adjusting the estimated population of each CD by this figure or a proportion of it. Considering

that such structures were present in the training set it was conjectured that adjusting by the full amount would over-compensate and hence lead to overestimates. It was decided to first try this on the Ballarat data, and if it did lead to overestimation, then adjust by some lower proportion until the estimation bias was removed.

The results are shown in Tables 6.10 and 6.11 and Figure 6.4. The initial adjustment was to multiply the remote sensing population density of each CD in the Ballarat study area by the factor $(1+ p_{nsh})$, where p_{nsh} = proportion of non-separate houses. This is referred to in Table 6.5 as model 1. As anticipated, this resulted in the population density being overestimated, with a ground truth vs. remote sensing regression coefficient of 0.93.

Since the target slope coefficient of 1.0 was about midway between this value and the unadjusted slope value, the second adjustment factor tested was $(1+ 0.5p_{nsh})$. The results (model 2 in Table 6.10) show regression coefficients close to 1.0. No separate calculations were done for the urban area in Table 6.10. This is because the urban area dominates density calculations and so the results are similar to those for the whole region, with a somewhat lower R^2 value as a result of reducing the range of the data by omitting the low density rural CDs (see Table 6.7).

The same adjustment was applied to the Geelong data, resulting in the figures shown in Table 6.10 (Geelong model 2).

The residual discrepancies which remain between the slope coefficients for Geelong and Ballarat (3-4% or 6-7% forced through origin) may be due to bias in the regression formula due to its being derived from the lower density Ballarat training set.

Table 6.10 Comparison of Estimated Population Densities for Census Collection Districts[1] in Primary and Secondary Study Areas: Based on a Two-phase Pixel Classification and Regression Procedure with Adjustments for Extreme Densities

Model	Statistical District			Statistical District: outliers omitted		
	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)	R^2	*s*	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)	R^2	*s*
1 Ballarat (CD adj 1)[4]	-18.4 + .94 .93	.80	499			
2 Ballarat (CD adj 2)[4]	-36.4 + 1.03 1.02	.81	484	-56.9 + 1.00 .98	.91	308
3 Ballarat (CD adj 2+3)[4]	-29.2 + 1.03 1.02	.81	484	-49.9 + 1.00 .98	.91	308
4 Ballarat (Pixel adj)[5]	-14.3 + 1.05 1.04	.82	460	-28.9 + 1.01 1.00	.91	309
2 Geelong (CD adj 2)[4]	47.4 + 1.07 1.09	.69	665	13.5 + 1.04 1.04	.78	513
3 Geelong (CD adj 2+3)[4]	62.0 + 1.06 1.09	.69	664	28.3 + 1.03 1.05	.78	512
4 Geelong (Pixel adj 4)[5]	-21.0 + 1.10 1.09	.73	618	-53.4 + 1.07 1.05	.82	467

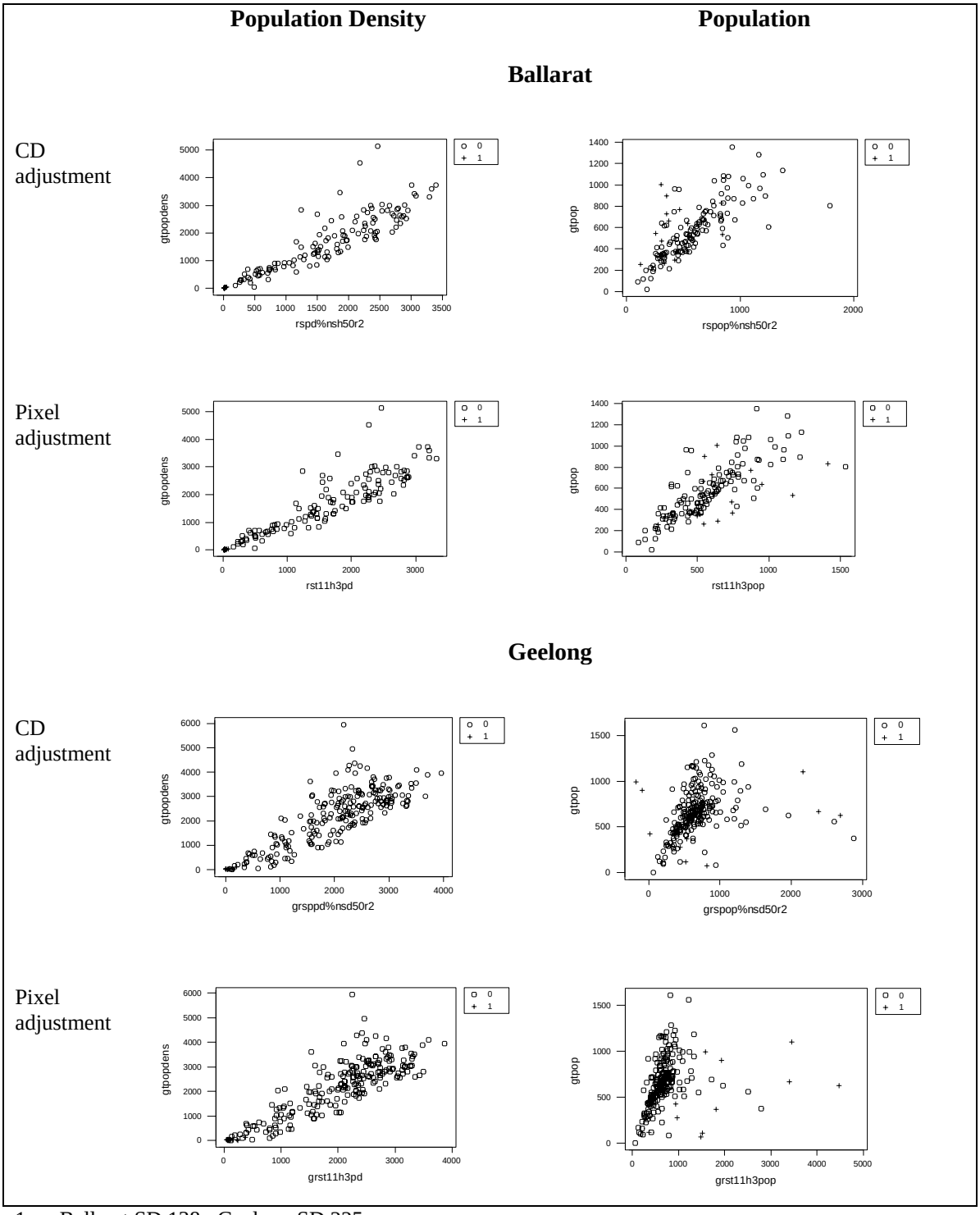
1 *n*: Ballarat SD 138; Ballarat SD with outliers omitted 135; Geelong SD 225; Geelong SD with outliers omitted 214, 215.
2 Intercept + slope; slope when forced through origin
3 Ground truth populations are: BSD 79179; Ballarat urban 70222; GSD 147910; Geelong urban 142250.
4 CD population density adjustments
Adjustment 1: CD population densities multiplied by (1 + proportion of non-separate houses)
Adjustment 2: CD population densities multiplied by (1 + 0.5 × proportion of non-separate houses)
Adjustment 3: Population densities of rural CDs halved
5 Pixel population adjustments: zero threshold; low density threshold coefficients 1.0, 1.0; power coefficient 1.07

Table 6.11 Comparison of Estimated Populations for Census Collection Districts[1] in Primary and Secondary Study Areas: Based on a Two-phase Pixel Classification and Regression Procedure with Adjustments for Extreme Densities

Model	Statistical District								Urban areas							
	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)		R^2	*s*	Mean % error	Median % error	Est. total pop.	% error[3]	G. truth v Rem. sens. Regression coeffts.[2] (unforced & forced)		R^2	*s*	Mean % error	Median % error	Est. total pop.	% error[3]
3 Ballarat (CD adj 2+3)[4]	192 + .66	.93	.53	177	28.5	16.3	79745	+1	151 + .71	.91	.61	164	27.4	15.1	73212	+4
4 Ballarat (Pixel adj)[5]	149 + .75	.93	.57	170	28.7	16.1	81437	+3	117 + .80	.97	.66	153	25.4	14.2	70187	-0
3 Geelong (CD adj 2+3)[4]	523 + .20	.77	.09	256	47.2	21.6	152344	+3	489 + .26	.85	.12	245	57.8	19.9	142866	+0
4 Geelong (Pixel adj)[5]	561 + .13	.65	.06	259	37.6	20.8	164642	+11	458 + .31	.87	.15	240	34.4	17.2	142717	+0

1 *n*: Ballarat SD 138; Ballarat urban 122; Geelong SD 225; Geelong urban 214
2 Intercept + slope; slope when forced through origin
3 Ground truth populations are: BSD 79179; Ballarat urban 70222; GSD 147910; Geelong urban 142250.
4 CD population density adjustments
Adjustment 2: CD population densities multiplied by (1 + 0.5 × proportion of non-separate houses)
Adjustment 3: Population densities of rural CDs halved
5 Pixel population adjustments: zero threshold; low density threshold coefficients 1.0, 1.0; power coefficient 1.07

Figure 6.4 Population Density and Population Estimates for Census Collection Districts[1]: Ground Truth vs. Adjusted Remote Sensing Estimates for Primary and Secondary Study Areas



1 *n*: Ballarat SD 138; Geelong SD 225

6.4.5 CD-based adjustment for inflated counts of residential pixels in rural areas

In Section 6.4.2 accurate population estimates were obtained for the whole of each study area (see Table 6.8). However, the urban area populations were somewhat underestimated (by 5% in

Ballarat and 8% in Geelong), and by implication the non-urban area populations were correspondingly overestimated, for reasons which have been discussed. In fact, since the rural sections contributed only a small proportion to the population of the total region (11% in Ballarat and 4% in Geelong) the over-estimation of the small rural populations was very great in proportional terms (around 50% in Ballarat and 200% in Geelong), though not so in absolute terms.

A mechanism for downward adjustment of the population estimates for the non-urban areas was now considered.

When a regression analysis of ground truth vs. remote sensing population density was carried out on the rural CDs of the primary Ballarat study area, the regression coefficient was close to 0.5. Considering this and the over-estimation figures above, the adjustment of halving the estimated population densities in rural areas was explored.

As anticipated, this made little difference to the results for population density (model 3 in Table 6.10). With regard to the regression results for population (model 3 in Table 6.11), whilst the volatility of the outliers (both positive and negative) was somewhat damped by this adjustment, there was still enough volatility to render the standard statistics meaningless. However, the final resulting estimates of total population for both regions and both urban areas were all accurate to within 4%.

6.4.6 Pixel-based high and low density adjustments

The two adjustments described in the previous sections will not suffice if the goal is to have a population estimation algorithm which is at least in principle able to be applied in a reasonably automated fashion to a TM image, without the need for ancillary structural information such as CD boundaries and without the need for human judgements about urban/non-urban delineations. The first adjustment required ancillary data; the second required human intervention and was both ad hoc and post hoc, being informed by known ground truth values. Furthermore, if we specify that estimates for any area, however small, must be feasible, then negative pixel estimates are not acceptable.

The final modifications investigated at this stage were designed to automate both the high and low density corrections, at least to the extent of requiring only an initial calibration step, and at the same time to overcome the problem of negative estimates.

An immediate and pragmatic solution to the problem of negative estimates is very simple – reassign negative values to zero i.e. apply zero thresholding. This was considered in Chapter 5 but rejected on the grounds that it would further inflate the already high estimates of population in low density areas. The problem is that both the negative estimates and the overall

overestimation have the same root cause - the overclassification of pixels as residential in low density areas. Attempting to fix one aspect of the problem makes the other worse. The alternative is to go to the root cause and essentially redress the overclassification by carrying out what is termed a contextual re-classification (Barnsley and Barr, 1996) i.e. a reclassification based on information about not only the particular pixels, but neighbouring pixels as well. Contextual information in the form of statistical texture measures were tried without success at the initial classification stage (see Section 5.3). However, having derived initial population estimates, there is now more contextual information available of a different sort, making it possible to apply a rule-based approach to re-classification (see Section 2.10.2).

The method is essentially to re-classify some of the residential pixels in low density areas as non-residential, by reassigning them zero population. Prima facie, the most appropriate pixels to reassign to zero are those which have been assigned the lowest (positive) values. Thus the problem becomes one of identifying pixels in low density areas with low positive estimates.

To do this, areas of low average population density (typically the non-urban areas) had to be identified, preferably in a way that did not require direct human intervention. This was done by calculating average population density using a mean-based averaging filter. A 7×7 pixel window was chosen, which represents an 210m×210m square with an area of 4.4 ha or around 10 acres, which is around the scale of inner urban CDs and towards the lower end of rural-residential block size.

It was observed that, whilst there is a continuum of densities at the high end of the scale, the boundary between urban and non-urban is characterised by a sharp discontinuity in average density. Because of this, the average density band was essentially equivalent to an urban/non-urban dichotomy overlaid on the residential/non-residential classification which had already been made.

The procedure adopted was as follows. Firstly, all negative pixel estimates were reset to zero. Secondly, the average population density was calculated and saved as a new band (see Image 9). The third step was to identify pixels in low density areas which also had low individual values. This was done by setting thresholds on both the average population density band and the (individual pixel) population band. All pixels which fell below both thresholds had their population reset to zero.

These corrections were applied to the Ballarat image. After some experimentation values of 1.0 were chosen for both low density thresholds. The thresholds can be interpreted as follows: in areas where the average population density over an area of 210m×210m is less than 1 person per 30m×30m (i.e. 1111 persons/sq.km.), any pixel with an estimated population of less than 1 person is regarded as having been misclassified and has its population reset to zero. Whilst this

threshold is much higher than the ABS criterion of 200 persons/sq.km. used to classify CDs as urban, many urban CDs have non-residential sections which lower the overall population density. In purely residential suburban CDs population densities are generally above 1000 persons/sq.km.. In the primary study area, this was the case in 88 out of the 122 CDs classified as urban.

The final correction, targeted mainly at the high density end of the scale, was very straightforward by comparison. Since population estimates ranged between 0 and around 5, it was reasoned that raising these estimates to a power fractionally larger than unity would slightly but progressively increase estimates above 1, and at the same time slightly reducing estimates below 1, which would have the desired effect of increasing population estimates in areas of high density, and provide an additional (though minor) corrective effect at the low density end of the scale.

The power coefficient was selected to "tune" the model to produce an accurate total for the urban area of the primary image. After some experimentation the value of 1.07 was selected.

When the algorithm with these chosen settings was applied to the full test image in each area, the peppering of small non-zero populations in rural areas was, as expected, noticeably reduced (see Image 10). Population estimates were derived for the CDs as previously described, and compared with ground truth values. The results are shown as model 4 in Tables 6.10 and 6.11 and in Figure 6.4.

In the case of the primary Ballarat study area, the results for population density were closely comparable with those of model 3, being slightly better on some criteria and not quite as good on others. The results for population compared to model 3 were: better correlations; higher estimate for regional population; and a much more accurate estimate for urban population (as it was calibrated to produce). Enlarged views of the urban area are shown in Images 11 and 12.

In the case of the secondary Geelong study area, the results for population density were again mixed, with higher correlations but greater bias in slope. The urban population total was also very accurately estimated, but that was also the case for model 3.

Overall regional population was the only area in which model 4 performed noticeably worse than model 3, in particular by overestimating the regional population by 11%, or 8% more than model 3. This was not a surprising result, considering that much of the overestimation in non-urban Geelong was associated with industrial rather than rural areas. The confounding of industrial with residential areas is more likely to take the form of relatively high (spurious) population densities over relatively small areas compared to the rural pattern of slightly inflated low densities over larger areas. A low density filter will not help in detecting these misclassifications.

That being the case, in the light of the comparative performance of models 3 and 4, and considering the methodological advantages associated with model 4, the final assessments of this phase of the analysis are now made with reference to model 4, which in summary involved:

- classification of each pixel in the image into residential and non-residential, using maximum likelihood classification based on the 6 untransformed TM bands and local training sets selected from the image;
- a linear regression equation for residential pixel population, trained on a sample of residential pixels in the primary study area, based on the 6 untransformed TM bands and incorporating iterative re-estimation;
- contextual reclassification as non-residential (zero population) of low population pixels (< 1 person per pixel) in areas of low average population density (< 1 person/pixel over a 7×7 pixel area);
- high density adjustment via a power coefficient of 1.07

6.4.7 Examination of remaining discrepancies

The right half of Table 6.10 (with outliers omitted) and Figure 6.4 show that in the primary study area, in all but a few CDs there was a strong concordance between ground truth CD population density and the estimates derived from TM data using this model. The same can be said of the secondary area, although the relationship is not quite so strong. With regard to actual CD population, the relationship is weaker again, although again most of the substantial disagreement occurs in relatively few CDs (less than 10% of the total in each case). However with regard to total urban population, the results in both primary and secondary study areas were extremely accurate.

The errors from model 4 (discrepancies between the estimates from model 4 and the ground truth values) for both population density and population of each CD are plotted in Figure 6.5 against both population density and population. These plots confirm that in all respects, the gross discrepancies are associated with a small number of CDs.

The first pair of plots (error in population density vs. population density) show that in both study areas, in spite of the adjustment for high densities incorporated in model 4, there remained a tendency towards underestimation at higher densities. In each case, this was most pronounced in the anomalous CDs which were identified previously, and which are summarised in Table 6.12. The second pair of plots (error in population density vs. population) show that in general, there was not such a strong relationship between error in estimated CD population density and

CD population. In particular, the outlier CDs with underestimated high densities did not necessarily have particularly high populations.

With regard to error in population, in the first pair of plots (error in population vs. population density) many of the same high density outliers stand out, not surprisingly, with underestimated populations. However, equally large discrepancies (larger in the case of Geelong) occurred amongst CDs with low density. Here, there was a difference in the pattern observed in the two study areas.

In the case of Ballarat, whilst there was a tendency for low density populations to be overestimated, there were also a few low density CDs whose population was considerably underestimated. In particular, the populations of 5 of the 16 non-urban CDs were underestimated, 2 of them considerably so. This suggests that the remaining problem of overestimation in low density areas will not be uniformly improved by raising the low density threshold adjustments.

In the case of Geelong, there remained very large overestimation discrepancies in a number of low density CDs, both urban and non-urban. As has already been discussed, the larger industrial base of Geelong is a major contributing factor to the worst of these (see Table 6.12). Conversely, just as in the case of Ballarat, underestimation was generally associated either with institutional anomalies, with public housing estates or with more recent outer suburban residential development which has occurred in the absence of and probably in lieu of large scale public housing development. Housing estates of this type, which are much more extensive in Geelong than Ballarat, have been periodically developed in the post World War 2 period essentially to house industrial workers. Hence it would appear that the presence of large-scale industry is the root cause of much of both the overestimation (in industrial areas) and the underestimation (in associated residential areas).[2]

---

2 The fact that the two effects tend to cancel one another out over the whole region recalls the words of the old Tennessee Ernie Ford song:

You load 16 tons and what do you get
Another day older and deeper in debt
Saint Peter don't you call me 'cause I can't go
I owe my soul to the company store.

There is a certain grim irony in the fact that from the heavenly perspective of the orbiting satellite, many workers are in effect assigned to the "company store" rather than to their homes!

Table 6.12 Census Collection Districts with Discrepant Estimates of Population Density and/or Population

A. Primary study area: Ballarat

Underestimation

CD number	Population density discrepancy[1,2]	Population discrepancy[2]	Characteristics
68	-2678		Includes geriatric hospital
91	-2266		Public housing
8	-1676		Includes general hospital
88	1601	-543	Public housing
89	-1136		Public housing

Overestimation

CD number	Population density discrepancy[1,2]	Population discrepancy[2]	Characteristics
73		+582	Rural, includes racecourse, horse breeding & training establishments
74		+630	Extensive rural, includes motels, some industrial, and airport incorporating former defence forces camp facilities
63		+728	Mixture of residential, industrial/commercial, municipal saleyards, extensive park with sports facilities

1 Persons/sq.km.

2 CDs are listed because of extreme discrepancies on one or either criterion (or both criteria). Only the extreme values are listed. Cutoff points for inclusion in table are: population density ±1000; population ±500

Table 6.12. Census Collection Districts with Discrepant Estimates of Population Density and/or Population (continued)

B. Secondary study area: Geelong

Underestimation

CD number	Population density discrepancy[1,2]	Population discrepancy[2]	Characteristics
200	-3702	-568	Includes geriatric institution
16	-2485	-578	Public housing
28	-2082		Public housing
42	-1949	-517	Public housing
50	-1948		Adjacent to public housing
52	-1848		Public housing
219	-1590		Recently developed outer suburban residential
218	-1590		Recently developed outer suburban residential
68	-1486		Public housing
48	-1450	-556	Public housing
81	-1320		No obvious cause
221	-1255		Recently developed outer suburban residential
67	-1196		Adjacent to public housing
31	-1185		Adjacent to public housing
25	-1110		Public housing
109	-1085		Includes general hospital
45	-1081		Public housing
36	-1030	-790	Public housing
8	-1030		Close to public housing

1 Persons/sq.km.

2 CDs are listed because of extreme discrepancies on one or either criterion (or both criteria). Only the extreme values are listed. Cutoff points for inclusion in table are: population density ±1000; population ±500

Table 6.12. Census Collection Districts with Discrepant Estimates of Population Density and/or Population (continued)

B. Secondary study area: Geelong

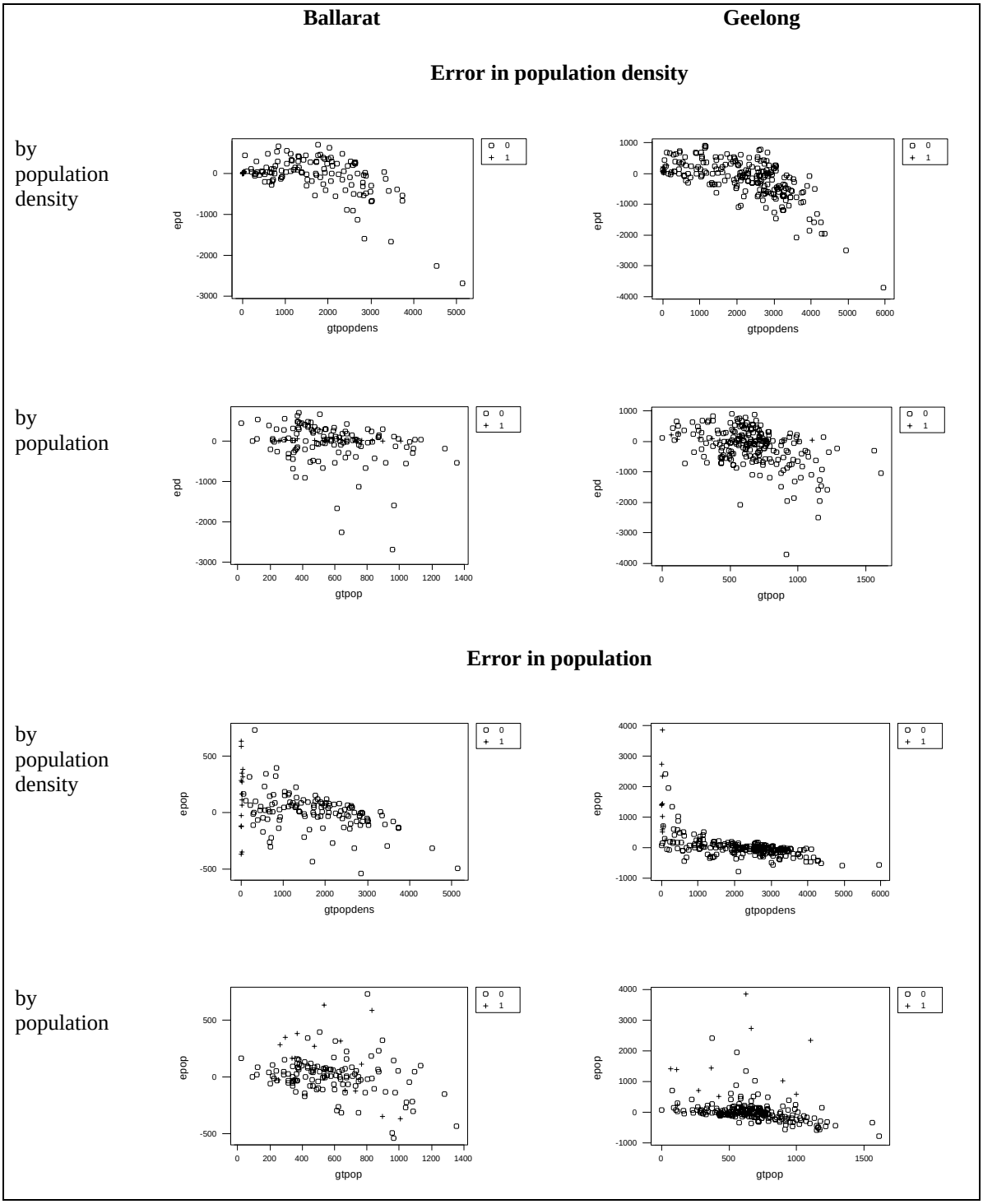
Overestimation

CD number	Population density discrepancy[1,2]	Population discrepancy[2]	Characteristics
22		507	Rural residential
61		509	Mixed industrial, residential, schools & churches
106		524	Concentration of hotels, motels & sports facilities
95		580	Cement works, market gardens
3		593	Rural
122		608	Large park, residential
4		698	Rural residential
57		701	Mixed industrial, motels, highway, residential
89		873	Industrial + railway yards
79		1017	Industrial
222		1027	Rural: marsh
18		1330	Mixed industrial, residential, sports facilities
1		1402	Rural: river flats
32		1415	Industrial: salt evaporating pans
21		1443	Rural: marsh
121		1941	Industrial + racecourse
34		2338	Rural residential
59		2410	Industrial: vehicle assembly plant
43		2726	Rural: river flats
20		3848	Rural + former salt works and aluminium smelter

1 Persons/sq.km.

2 CDs are listed because of extreme discrepancies on one or either criterion (or both criteria). Only the extreme values are listed. Cutoff points for inclusion in table are: population density ±1000; population ±500

Figure 6.5 Estimation error in Census Collection District population density and population for each study area: by population density and by population



1 *n*: Ballarat SD 138; Geelong SD 225

The nature of these remaining discrepancies suggests two things. Firstly, population density is consistently underestimated in areas of concentrated public and similar housing. However, these are not necessarily the areas with the highest population densities; other CDs with similar or higher density are more accurately estimated. This suggested that public housing areas were

not in spectral terms "out of range" of the regression model, but rather that the relationship between the spectral characteristics and population was not correctly calibrated for these areas.

It is postulated that this is because of relative overcrowding in these areas – small rooms in small houses on small blocks. The same spectral mix of built and natural elements which would elsewhere be associated with moderate population density, is in these areas associated with somewhat higher population density. Because of the relative scarcity of such areas in Ballarat, they were underweighted relative to Geelong in the regression training set. If the postulated explanation is true, then giving such areas higher weighting in the training set would presumably improve the fit of estimates in such areas. However this may well occur at the expense of accuracy in other areas i.e. the same total uncertainty may just be more evenly distributed across all residential areas. An alternative strategy which should improve the accuracy of fit overall, would be to classify residential areas into two strata rather than just one, and establish separate regression relationships for each. However, this would involve an additional cost in procedural complexity, and would require extra "on the ground" knowledge.

Secondly, with regard to the other sources of discrepancy, essentially the problem is one of classification. Other approaches to classification, such as those outlined in Section 1.3.2, may produce some improvement, but there is likely to be a limit beyond which further gains are difficult to achieve within a parametric modelling framework. With regard to the residual variation of estimates for rural areas, considering the sparse habitation of Australian rural areas in comparative global terms, it is arguable that the signal-to-noise ratio may be too low to enable reliable discrimination between the spectral signature of human residency and that of other human artifacts such as roads and sheds and the patchwork of agricultural and pastoral activities. With regard to the problem of overestimation in industrial areas and other anomalous non-rural areas, and in the light of earlier discussions about underestimation of population concentrations associated with institutional accommodation and the like (including metropolitan high rise accommodation), there must come a point where further refinement of parametric procedures leads to such diminishing returns that a final recourse to the incorporation of ancillary information about known anomalies becomes a more effective way to proceed.

6.5 SUMMARY

The CD aggregate models developed in Chapter 4 and the pixel-based models developed in Chapter 5 were tested on the secondary image of Geelong.

It was concluded that some of the improvement which had been achieved in the CD aggregate models in the case of the primary image through increased complexity, were not robust to the transition to the secondary image. Overall, the model which performed best on the urban areas

of both images was the “middle ranking” model based on ratios of CD band means, which produced very accurate estimates of the total urban population in both cases. Nevertheless, at the level of individual CDs, this model, like the rest, tended to underestimate the higher densities and overestimate the lower densities, and this was particularly the case in the secondary study area with its higher average density. Like the other models, it grossly overestimated the regional totals for both primary and secondary study areas. It was concluded that procedures based on CD aggregates were not robust to variation in density either within or between study areas, and that it would be difficult to address this shortcoming given the aggregated nature of the data.

When the pixel-based models were applied to the secondary image, it was found that the logarithmic and square root models were not at all robust, in that very variable results were produced. The performance of the linear model was much more consistent, but the estimates produced were badly biased. The problem of bias was largely overcome by retraining the initial pixel classification on the secondary image, which is a routine and straightforward task in remote sensing analysis. As with the CD aggregate methods, there remained a residual tendency to underestimate population in high density areas and to overestimate it in low density areas, but because of the disaggregated basis of the analysis, it was possible to devise methods to overcome these problems to some degree.

When the population density and population estimates produced by this model for individual CDs were compared with ground truth data, there was a strong underlying concordance in the urban sections of both study areas, overlaid by a scattering of problematic cases the nature of which have been identified and explained, and for which remedies have been proposed.

Notwithstanding the limitations of the model with respect to individual CDs, it seemed quite robust with respect to large urban aggregates. Having been tuned to produce an extremely accurate estimate of total urban population in the primary study area, it produced an almost equally accurate estimate of total urban population in the secondary study area.

With this basic framework established, it was decided to undertake a further exploration of the properties of the iterative re-estimation procedure with a view to optimising its performance before proceeding to submit the methodology to a wider range of validation testing.

Chapter 7

The Iterative Re-Estimation Algorithm

7.1 INTRODUCTION

An iterative algorithm for refining regression estimates from incompletely determined data was introduced in Section 2.9.

The algorithm was applied to training data from the primary image (Section 5.7) where it produced the best results of the many models tested. Consequently it was adopted as a key step in the methodology developed and evaluated throughout the remainder of Chapters 5 and 6.

In applying this algorithm, ad hoc decisions were made about sampling strategies and sample sizes, and about the number of iterations of the algorithm to be used. It was decided, before proceeding to a wider range of validation testing, to explore the sampling variation and convergence properties of the algorithm with a view to devising a more considered strategy for its implementation. These investigations are reported in Sections 7.3 and 7.4.

We first show in Section 7.2 that the iterative re-estimation algorithm is a normal-based OLS approximation to an EM (expectation-maximisation) algorithm for maximum likelihood estimation.

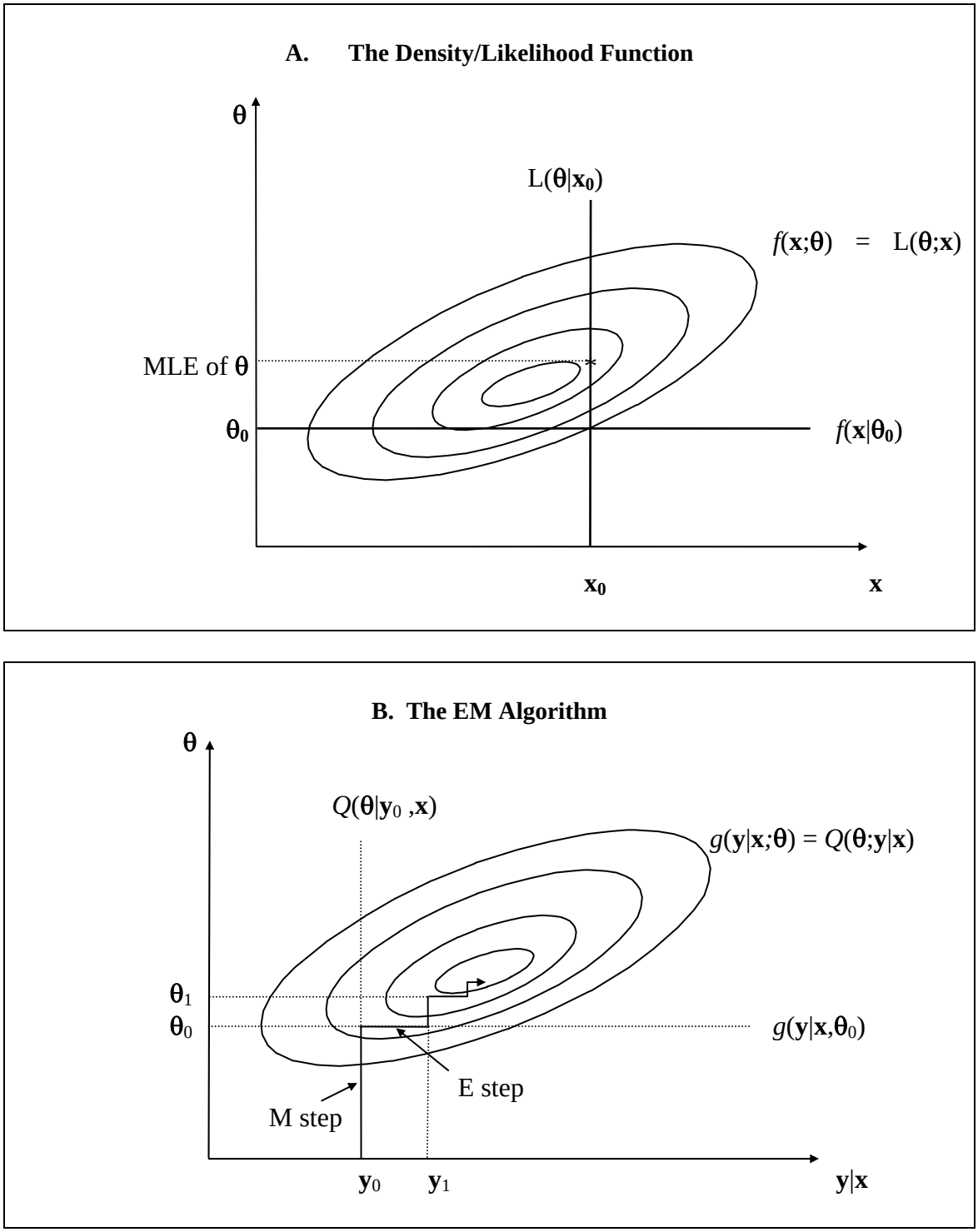
7.2 THE RELATIONSHIP OF THE ITERATIVE RE-ESTIMATION ALGORITHM TO THE EM ALGORITHM

7.2.1 The EM algorithm

The general EM algorithm can be represented schematically as in Figure 7.1. The contours (which in practice would be far more complex than those illustrated) represent a function of two sets of variables, each set represented schematically by a single dimension – a vector of parameters θ, represented by the vertical dimension, and a vector of data values $\mathbf{x}$ represented by the horizontal dimension. The function of $\mathbf{x}$ defined by conditioning on a particular value of θ, say θ_0, is the joint probability (density) function $f(\mathbf{x}|\theta_0)$. The function of θ defined by conditioning on a particular value of $\mathbf{x}$, $\mathbf{x}_0$, is the likelihood function $L(\theta|\mathbf{x}_0)$. The value of θ at

which this function is maximised is the maximum likelihood estimate of θ given the particular sample data $\mathbf{x_0}$.

Figure 7.1 Schematic Representation of the EM Algorithm



The EM algorithm is used in situations where $f(\mathbf{x};\theta) = L(\theta;\mathbf{x})$ is difficult to define or analyse, but where some extra information (usually some extra detail about the data) would simplify the

formulation and analysis. We represent the embellished or augmented data by **y** and the density/likelihood function by $g(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = Q(\boldsymbol{\theta};\mathbf{y}|\mathbf{x})$.

The conditional notation reflects the fact that the (wholly or partially) unobserved **y** values are constrained throughout to be consistent with the observed **x**. The algorithm proceeds by alternately re-estimating the conditional expectation of **y** (the E step), and maximising the likelihood conditional on the current value of **y** (the M step).

7.2.2 The iterative re-estimation algorithm

In the case under consideration, the elements of the **y** vector are the unknown notional populations of each pixel classified as residential, and **x** is the vector of known totals of these for each CD.

We postulate a linear regression model with independently and identically distributed N(0, σ^2) random errors:

$\mathbf{y} \sim N(\mathbf{s}\boldsymbol{\beta}, \boldsymbol{\Sigma})$

where **s** is a vector of explanatory remote sensing variables (augmented by a constant term)

$\boldsymbol{\beta}$ is a vector of regression parameters to be estimated

$\boldsymbol{\Sigma}=\sigma^2\mathbf{I}$ is a diagonal error covariance matrix.

In this case

$$L(\beta : \mathbf{y}) = const - \ln|\Sigma| - (\mathbf{y} - \mathbf{s}\beta)'\Sigma^{-1}(\mathbf{y} - \mathbf{s}\beta)$$
$$= const - \ln|\Sigma| - \Sigma^{-1}(\mathbf{y} - \mathbf{s}\beta)'(\mathbf{y} - \mathbf{s}\beta) \quad \text{since } \Sigma = \sigma^2\mathbf{I}$$

For a given value of **y**, the likelihood is maximised when the sum of squared residuals

$$(\mathbf{y} - \mathbf{s}\beta)'(\mathbf{y} - \mathbf{s}\beta) = \sum (y_i - \mathbf{s}_i\beta)^2$$

is a minimum – hence the least squares estimate of $\boldsymbol{\beta}$ is also the maximum likelihood estimate.

Each iteration of an EM algorithm involves:

The E step: Calculate the expected value of the likelihood as a function of $\boldsymbol{\beta}$, conditional on the current estimate $\boldsymbol{\beta}^j$ and on the known values of **x**,

$$E[L(\beta : \mathbf{y}) \mid \beta^j, \mathbf{x}] = E[\{const - \ln|\Sigma| - (\mathbf{y} - \mathbf{s}\beta)'\Sigma^{-1}(\mathbf{y} - \mathbf{s}\beta)\} \mid \beta^j, \mathbf{x}]$$
$$= const - \ln|\Sigma| - \Sigma^{-1}(\mathbf{y} - \mathbf{s}\beta)'(\mathbf{y} - \mathbf{s}\beta) \quad \text{since } \Sigma = \sigma^2\mathbf{I}$$

For distributions of the exponential family, the E step essentially involves replacing **y** in the expression for $L(\boldsymbol{\beta}:\mathbf{y})$ with its expected value, conditional on the current value of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^j$ and on **x** (Navidi, 1997).

i.e. we require $E(\mathbf{y}|\mathbf{x})$, given that $\mathbf{y} \sim N(\mathbf{r}\boldsymbol{\beta}^j, \boldsymbol{\Sigma})$

The M step: Maximise the expected value $E[L(\boldsymbol{\beta}:\mathbf{y})]$ with respect to $\boldsymbol{\beta}$, and thereby obtain an updated estimate $\boldsymbol{\beta}^{j+1}$.

The procedure

In the iterative re-estimation algorithm (Section 2.9), the M step consists of re-estimating the least squares regression line. Subject to the assumptions of normality, constant variance and independence, and the equivalence of least squares and maximum likelihood, this is equivalent to a standard M step.

However, the E step was originally conceived heuristically in terms of perturbing the **y** values to reduce the residuals from the currently fitted line. In effect, rather than $E(\mathbf{y}|\mathbf{x})$, we find the conditional mode of **y**, i.e. the value of **y** which maximises the likelihood $L(\boldsymbol{\beta}:\mathbf{y})$ conditional on $\boldsymbol{\beta}^j$ and on **x.** This is done by minimising the sum of squared residuals

$$\sum (y_i - \mathbf{s}_i \beta)^2$$

subject to the CD totals **x.**

From this perspective, the problem can be characterised as follows: find a set of adjusted *y* values within each CD subset which minimises the sum of squares of the residuals subject to the sum of the residuals remaining constant. This can be equivalently formulated in terms of residuals thus: given the set of residuals r_i, find a set of adjusted residuals a_i such that

$$\sum_{i=1}^{n} a_i^2 \text{ is a minimum, subject to the constraint } \qquad \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} r_i = K$$

Using the method of Lagrange multipliers, we seek to minimise

$$f(a_1, a_2, \ldots\ldots a_n, \lambda) = \sum_{i=1}^{n} a_i^2 - \lambda \left(\sum_{i=1}^{n} a_i - K \right)$$

$$\frac{\partial f}{\partial \lambda} = 0 \Rightarrow \qquad \sum_{i=1}^{n} a_i = K$$

$$\frac{\partial f}{\partial a_i} = 0 \Rightarrow \qquad 2a_i - \lambda = 0$$

$$\Rightarrow \quad a_i = \frac{\lambda}{2}, \ \textit{all } i$$

$$\Rightarrow \quad \sum_{i=1}^{n} a_i = \frac{n\lambda}{2}$$

$$\Rightarrow \quad \frac{n\lambda}{2} = K$$

$$\Rightarrow \quad \lambda = \frac{2K}{n}$$

$$\Rightarrow \quad a_i = \frac{K}{n} = \bar{r}$$

i.e. the solution is to make all the residuals equal by redistributing the total of the original residuals (*K*) equally. This can be achieved by adjusting the values of the dependent variable as follows:

$$y_{i(adj)} = \hat{y}_i + \bar{r} = y_i - r_i + \bar{r}$$

7.2.3 Relationship to the EM algorithm

For the foregoing procedure to be equivalent to an exact EM algorithm, two conditions must be met:

The least squares estimate must be equivalent to a maximum likelihood estimate (in both E and M steps);

The maximum likelihood estimate of **y**|**x** (i.e. the mode) must be equivalent to the expected value (in the E step).

The first of these requires that the covariance matrix of the *conditional* distribution of **y**|**x** must be diagonal. The second requires that the mode and the mean of the distribution of **y**|**x** be equal.

It will be shown (Proposition 1) that under the original assumptions, the distribution of each element of **y**|**x** is normal, with expected value equal to the modal value calculated by least squares, but (Proposition 2) that the covariance matrix of the conditional distribution of **y**|**x** is not diagonal.

Proposition 1:

Under the original assumptions, the conditional distribution of each element of **y**|**x** is normal, with

$$E(Y_{ip} \mid C_i) = E(Y_{ip}) + \sum_{j=1}^{n_i} \left\{ Y_{ij} - E(Y_{ij}) \right\} / n_i = E(Y_{ip}) + \left\{ C_i - \sum_{j=1}^{n_i} E(Y_{ij}) \right\} / n_i$$

where Y_{ij} is the population of pixel *j* in CD *i*

$C_i = \sum_{j=1}^{n_i} Y_{ij}$ is the known population of CD i (or a known proportion of the CD population)

n_i is the number of pixels contributing to C_i

This is equal to the value calculated by least squares.

Proof

For simplicity and without loss of generality, we consider a particular CD, and omit the i suffix.

Let Y_j be the notional population of pixel j.

$Y_j \sim N(\mu_j, \sigma^2)$ where at the mth iteration, we postulate

$$\mu_j = \beta_0^m + \sum_{k=1}^{\text{nvars}} \beta_k^m s_k$$

Let X be the total population of the CD (or a known proportion of it)

$$X = \sum_{j=1}^{n} Y_j \quad \text{and} \quad X \sim N(\sum_{j=1}^{n} \mu_j, n\sigma^2)$$

Consider the pth pixel, with notional population Y_p

Let Z be the total population of all the pixels except the pth

$$Z = \sum_{j \neq p} Y_j \quad \text{and} \quad Z \sim N(\sum_{j \neq p} \mu_j, (n-1)\sigma^2) \text{ and independent of } Y_p$$

The conditional probability density of Y_p given that $X = c$ is

$$f(y_p \mid X = c) = \frac{f(y_p, X = c)}{\int_{-\infty} f(y_p, X = c) dy_p} \propto f(y_p, X = c) \qquad \text{(Kendall et al., 1987)}$$

Now

$$f(y_p, X = c) = f(y_p, Z = c - y_p)$$
$$= f_y(y_p) f_z(c - y_p) \quad \text{since } Z \text{ and } Y_p \text{ are independent.}$$

Hence

$$f(y_p, X = c) \propto \exp \frac{(y_p - \mu_p)^2}{2\sigma^2} \exp \frac{(c - y_p - \sum_{j \neq p} \mu_j)^2}{2(n-1)\sigma^2}$$

$$= \exp \frac{1}{2(n-1)\sigma^2} \left[(n-1)(y_p - \mu_p)^2 + (y_p - (c - \sum_{j \neq p} \mu_j))^2 \right]$$

$$= \exp \frac{A}{2(n-1)\sigma^2}$$

where

$$A = (n-1)(y_p - \mu_p)^2 + (y_p - (c - \sum_{j \neq p} \mu_j))^2$$

$$= (n-1){y_p}^2 - 2(n-1)y_p\mu_p + {\mu_p}^2 + {y_p}^2 - 2(c - \sum_{j \neq p} \mu_j)y_p + (c - \sum_{j \neq p} \mu_j)^2$$

$$= n{y_p}^2 - 2y_p[(n-1)\mu_p + c - \sum_{j \neq p} \mu_j] + const$$

$$= n\left[y_p - \frac{(n-1)\mu_p + c - \sum_{j \neq p} \mu_j}{n} \right]^2 + const$$

Hence

$$f(y_p, X = c) \propto \exp \frac{n}{2(n-1)\sigma^2} \left[y_p - \frac{(n-1)\mu_p + c - \sum_{j \neq p} \mu_j}{n} \right]^2$$

Thus the conditional distribution of $(Y_p \mid X = c)$ is normal, and

$$\text{var}(Y_p \mid X = c) = \frac{(n-1)\sigma^2}{n}$$

$$E(Y_p \mid X = c) = \frac{(n-1)\mu_p + c - \sum_{j \neq p} \mu_j}{n}$$

$$= \frac{n\mu_p + c - \sum_{j=1}^{n} \mu_j}{n}$$

$$= \mu_p + \frac{c - \sum_{j=1}^{n} \mu_j}{n}$$

$$= \mu_p + \frac{\sum_{j=1}^{n} Y_j - \sum_{j=1}^{n} \mu_j}{n}$$

$$= \mu_p + \frac{\sum_{j=1}^{n} (Y_j - \mu_j)}{n}$$

Thus the conditional expected value is found by incrementing the current expected value (the fitted value from the previous regression step) by a constant equal to the mean of the residual values for all pixels in the CD. This is the same result as that obtained by least squares above.

Proposition 2:

The covariance matrix $\boldsymbol{\Sigma}$ is not diagonal under the conditional constraint of fixed CD totals.

Proof

This is most obvious in the case of n=2 where

$$Y_1 + Y_2 = c \text{ where } c \text{ is constant}$$

We have

$$\text{cov}(Y_1, Y_2) = \text{cov}(Y_1, c - Y_1) = \text{cov}(Y_1, c) - \text{cov}(Y_1, Y_1) = 0 - \text{var}(Y_1)$$

In general, with

$$\sum Y_i = c$$

we have

$$\text{cov}(Y_i, Y_j) = \text{cov}(Y_i, c - \sum_{k \neq j} Y_k)$$
$$= \text{cov}(Y_i, c) - \sum_{k \neq j} \text{cov}(Y_i, Y_k) = 0 - \text{var}(Y_i) - \sum_{k \neq i,j} \text{cov}(Y_i, Y_k)$$

It follows that either all of the covariances in the summation are zero, in which case $cov(Y_i,Y_j) = -var(Y_i)$ is not zero, or at least one of the other covariances is not zero.

Hence, in general the populations of pixels within a particular CD are mutually correlated, given that we know the CD totals. If we assume that the populations of pixels from different CDs are uncorrelated, then the covariance matrix of the conditional distribution will exhibit blocks of local correlation down the diagonal within each CD group of pixels. The extent of departure from diagonality will diminish in relative terms with increasing sample size, both within each CD, and overall[1] .

As a quite separate issue, if there is spatial correlation between the populations of neighbouring pixels, then the original assumption of independently distributed random errors will not be met, and there will be departures from diagonality in the unconditional covariance matrix, which will further contribute to lack of diagonality in the conditional covariance matrix.

7.2.4 Conclusion

This algorithm, which might be characterised as an approximate EM or perhaps MM (Maximisation Maximisation) algorithm, is closely related to the ECM (Expectation Conditional Maximisation) algorithm (Meng and Rubin, 1993), which also has multiple conditional maximisation steps, and in spirit to the gradient algorithms of Titterington et al. (1985) and

[1] This is reminiscent of the concept of m-aymptotics and n-asymptotics in the analysis of categorical data (Hosmer and Lemeshow, 1989).

Lange (1995). It its most general form, ECM has both multiple M steps and an E step, whereas the present algorithm can be characterised as having either, but not both. In the context of the normal distribution, such distinctions are not so marked as in the more general case.

It is conjectured that the iterative re-estimation algorithm is asymptotically equivalent to the EM algorithm. As such, for large sample sizes, the sequence of parameter estimates obtained might be expected to converge to the maximum likelihood estimates based on the true (unknown) *y* values. However, Titterington et al. (1985) cite a number of examples of slow convergence and multiple maxima, and Lange (1995) also alludes to "multiple modes of the likelihood surface". In the presence of multicollinearity, sensitivity of individual parameter estimates to sampling variations in the data corresponds to a complex likelihood surface with flat topped ridges and ill-defined maxima. This would be expected to lead to just such convergence problems. In the present study this has been demonstrated to be the case both by Monte Carlo simulation and empirically (see following sections and Chapter 8). As will be demonstrated, and has been discussed in Section 2.11.5 on multicollinearity, this does not necessarily invalidate the use of the procedure for producing improved population estimates.

7.3 SAMPLING VARIATION

The initial regression analysis of data from the primary image (Section 5.2) was carried out on a random sample of 1402 (2%) pixels classified as residential. It was decided on the basis of the rate of convergence of R^2 values to iterate 6 times.

It was now decided to take replicate samples to examine the extent of sampling variation in both the initial regression results and in the iterated results. This was extended to include a systematic examination of the effects of varying three factors:

- data source (residential class or residential classification training set)
- sample size
- number of iterations

The data source previously used had been the full residential class, because it was more broadly representative in scope than the residential training set, which was sampled from relatively homogeneous suburban residential areas. However the residential class was subsequently found to include, at the low population density end, many pixels misclassified as residential, the inclusion of which might be expected to bias the regression. It was thus worth examining the tradeoff between the effects of under-representation in the more restricted residential classification training set data and spurious representation in the full residential class.

Initially, a sample of 14270 (20%) pixels classified as residential was randomly subdivided into 10 disjoint subsamples each of 1427 pixels, and the complete set of 6345 pixels in the residential training set was subdivided into 5 disjoint subsamples each of 1269 pixels.

Linear regression models using the 6 TM bands as explanatory variables were fitted to the two larger samples and to the 15 subsamples. In each case the procedure was iterated 30 times. Appendix G shows a sample of the regression coefficients and R^2 values after 0, 1, 6, 29 and 30 iterations.

A number of points were apparent from these results. Firstly, there were both similarities and differences between the results for the two data sets. In each case the values of R^2 increased monotonically for all subsamples, sharply after the first iteration, and thereafter at a diminishing rate. In each case, the values of R^2 were reasonably consistent across subsamples.

As to differences, the initial fit was not so good in the training set data, and the initial regression coefficients (apart from the constant) tended to be considerably smaller in magnitude. The rate of convergence of R^2 values was also slower, as is evidenced by the magnitude of the difference between the 29^{th} and 30^{th} iterations. These differences are consistent with a data source which is more homogeneous with respect to the dependent variable. Further examination confirmed that whilst the residential training set encompassed the full range of spectral responses, the initial imputed populations fell into a narrower band than was the case for the full residential class.

With regard to the effect of the re-estimation procedure, the levels of R^2 reached by the 30^{th} iteration were somewhat higher in the case of the training set data. This is again consistent with a more homogeneous set of CDs, where the potential for linearising by re-allocation notional population between pixels might be expected to be greater. Of course it does not follow that such a model necessarily has greater predictive power – it may just indicate a greater potential for capitalising on chance in the more homogeneous data set.

Whilst the initial regression coefficients (apart from the constant) tended to be considerably smaller in magnitude in the training set data, these differences had greatly diminished by the 30^{th} iteration, the increases in the magnitudes of the coefficients being much more marked in the training set data than in the residential class data.

It was also the case that whilst the range of initially imputed populations was much less for the training set data than for the residential class data, this difference too had diminished by the 30^{th} iteration, with both sets of estimates having a similar range, and in both cases including negative values.

Finally, it was noticeable that whilst the initial coefficients were reasonably consistent between subsamples within each data set, the iterative procedure produced much greater divergence between subsamples. This was not unexpected in the presence of multicollinearity (see Section

7.2.3), but again (as discussed in Section 2.11.5) it need not necessarily detract from predictive performance.

In summary, it seemed that using the residential training set rather than the residential class as the source of training data might ultimately lead to similar prediction equations, although convergence would be rather slower. Alternatively, it may be the case that both sets of data were being over-linearised, and that the earlier results from the training set data, with lower magnitude coefficients suggesting less sensitivity of population to differences in spectral reflectance, might ultimately be more appropriate.

To explore this issue further it was decided to undertake a simulation study.

7.4 SIMULATION

The simulation study included a systematic examination of the effects of varying six factors:

- underlying relationship between population and spectral reflectances
- level of random error in the relationship
- data source for estimating the relationship
- sample size
- number of iterations
- suppression of negative estimates at each iteration

7.4.1 Simulated populations

The empirical spectral data from the two sets of pixels described in the previous section were used as the basis of several simulated statistical populations of pixel populations. (The two terminologies come into unavoidable conflict here!)

Three underlying relationships between pixel population and spectral reflectances were simulated, based on the regression equations obtained from the full "class" data set (n=14270) and "training set" data set (n=6345) after 10 iterations.

Because the "class" data had led to negative estimates from the very first iteration, two versions of the population relationship based on this data were used: one obtained when negative estimates were not readjusted to zero at each iteration, and the other obtained when negative estimates were readjusted to zero (see Section 2.9.2). This distinction was not drawn for the relationship derived from the training set data because few negative values were produced by the tenth iteration in that case.

The three underlying population relationships are summarised in Table 7.1.

Table 7.1 Underlying Pixel Population Models Used for Simulation

Model	Source	Band Coefficients						
		Const	b1	b2	b3	b4	b5	b7
1A	Res. class	2.13808	0.13243	0.17399	-0.17622	-0.03143	-0.05826	0.08553
1B	Res class with –ve adjustment	2.98236	0.13959	0.18791	-0.20670	-0.02571	-0.09284	0.13577
2	Res. training set	3.49571	0.00294	0.11197	-0.12913	-0.01534	-0.07410	0.16400

To each of these relationships was added three levels of normally distributed random error: none, moderate (σ=0.5) and high (σ=1.0). The standard deviations in the context of estimated pixel populations ranging from zero to maxima in the range 5 to 9.

Thus there were 9 simulated population relationships in all: 3 basic equations each with 3 levels of error superimposed.

Finally, all negative populations thus generated were reset to zero. Following the addition of random error and the adjustment of the negative populations, the underlying relationships in each of the 9 scenarios were re-estimated by OLS for later use (see Section 3.4.3).

7.4.2 Monte Carlo sampling from the simulated populations

It was considered that a sample size of at least 1000 was necessary to ensure a reasonable sample size within each of the 138 CDs represented in the residential class and the 25 CDs represented in the training set. The re-estimation procedure in its most exact from is predicated on all the residential pixels from each CD being used. The bias and/or loss of precision introduced by using only a sample of pixels from each CD will be greater for small samples, which are more likely to be unrepresentative of the CD as a whole. Small samples also limit the extent to which population can be redistributed between pixels. It was decided to compare sample sizes of 1000 and 5000 in the training set population and 1000, 5000 and 10000 in the class population. Because the largest sample size in each case was approaching the size of the population, measures based on samples of this size would give an indication of the upper limit to the capacity of this procedure to recover the true pixel populations.

The iterative re-estimation algorithm was applied, both with and without the readjustment of negative estimates at each iteration described in the previous section.

A subset of the complete factorial combination of factors, considered sufficient to illuminate the various issues, was selected and implemented. For each selected combination, 10 simulation

runs of 10 iterations were implemented. Regression coefficients and R^2 values were recorded after each iteration. The procedure was implemented in a Minitab macro.

7.4.3 Assessment criteria

The results of the simulations were compared on the basis of four criteria, each calculated after iterations 0, 1, 5 and 10. Three of these were averaged over the ten replications and the fourth was a measure of sampling variation between replications.

Regression coefficients

In theory, an EM algorithm should converge to the maximum likelihood estimate of the relevant parameters, in this case the regression coefficients. However, as discussed in Section 7.2.4, there were reasons for not expecting that to be the case. The preliminary results of Section 7.3 made it clear that there would be substantial sampling variation in the regression coefficients. Nevertheless it was of interest to investigate whether there was any underlying relationship between the "average" or expected rate and extent of convergence of the algorithm and the different settings of the various factors.

In the context of normal errors the maximum likelihood estimates in the present context are the OLS estimates found by regression analysis on the true populations of the pixels in the sample. The true population parameters could be regarded as the coefficients of the equations listed in Table 7.1, but because of the readjustments to the negative values it was decided to replace these parameters by the coefficients of a regression equation fitted to the final population values assigned to all pixels in the population.

Two criteria were chosen to address the issue of how accurately the estimates of the regression coefficients generated by the iterative refinement algorithm converged to (i) the OLS estimates and (ii) the true population parameters.

It was decided to exclude the constant term from consideration on the ground that since the origin of the explanatory variables was distant from the range of the observed spectral data in both data sets, in the presence of multicollinearity the constant term would be expected to be quite volatile.

Accordingly, the first measure calculated was the root mean square average of the discrepancies between the estimated value of each regression coefficient and its "target value" (OLS estimate or population parameter), averaged across the 6 TM bands (but excluding the constant) and the 10 simulations.

The second measure was the same measure calculated for only TM bands 4, 5 and 7, these being least affected by multicollinearity and hence being more likely to converge in a well-behaved fashion to the target values.

These are indicators of average or expected performance, not measures of sampling variation. Inspection of the results of individual simulations confirmed that as expected, and as reported in Section 7.3 for the empirical data, variation between the 10 replications under each condition increased in each case as the number of iterations increased. Sampling variation decreased as sample size increased, as would be expected, particularly since the largest sample size in each case was close to the finite population size.

Values of these two measures are tabulated in Table 7.2. Table 7.2 is in four parts. Part A and B show the first measure, based on the RMSE across the 6 band coefficients. Part A shows the results for models 1A and 1B, derived from the residential class with and without adjustments for negative values at each iteration. Part B shows the corresponding results for model 2, derived from the residential training set data. Parts C and D show corresponding results for the second RMSE measure, based on the coefficients for bands 4, 5 and 7.

Population Estimates

The other two criteria were designed to bypass the effects of multicollinearity and directly address the question of how well the true populations for the sample of pixels were recovered by applying the estimated regressions.

Under each of the combinations of factors examined, the estimated population of each pixel in the sample was compared with the true population of the pixel, and a root mean square average value calculated for the sample. This was done after iteration 0, 1, 5 and 10 iterations. The RMS deviation from the mean population of all pixels in the sample (i.e. the standard deviation of the populations, without the correction for degrees of freedom) was also incorporated for a baseline comparison. The RMS averages of these 5 sets of deviations, and the standard deviations of these RMS errors, calculated across the 10 replications, are tabulated in Table 7.3. Mean values for the 10 replications were also calculated, and in every case were almost identical to the RMS averages.

In this case the standard deviations are of interest because it was conjectured that notwithstanding the sampling variation in the regression coefficients in the presence of multicollinearity, the accuracy of population estimation should be consistent from sample to sample. Of course these results are indicative only, since they are based only on the sample data, with no external validation.

7.4.4 Interpretation and conclusions

The results of the simulations are now assessed with respect to the 6 factors listed at the beginning of Section 7.4, on the basis of the 4 criteria defined in Section 7.4.3, which for convenience in this section are referred to as follows:

6-coefficient:	RMSE based on estimates of 6 TM band regression coefficients
3-coefficient:	RMSE based on estimates of 3 TM band regression coefficients
population accuracy:	RMSE based on population estimates for individual pixels
sampling variation:	sample to sample variability in the population accuracy

Whilst R^2 values are not reported in Table 7.2, similar patterns were observed as in Section 7.3. Initial R^2 values were in the range .45-.55 for samples from the residential class and .25-.35 for samples from the more homogeneous residential training set. In all cases R^2 rose monotonically but at a diminishing rate. By the tenth iteration, values ranged from around .9 to .99. In each case, the level reached can be related to the combination of form of model, source of sample and sample size. However, the R^2 values after iteration have no intrinsic diagnostic use, since they reflect the effect of capitalisation on chance within the samples. Indeed in many cases, the levels reached after 10 iterations were much higher than the value in the simulated population, and hence were quite spurious and misleading.

Turning to the more informative measures presented in Tables 7.2 and 7.3, first a disclaimer! Viewed as a designed experiment this study involved 6 factors each at 2 or 3 levels, with 4 dependent variables. Considering the processing required for each observation (simulation run), it was not feasible to implement a complete design, or to undertake a formal and rigorous inferential analysis of the main effects and the many interactions. Rather, the investigation was strategically targeted, incremental and exploratory, with a view to obtaining some guidance as to an appropriate combination of settings to apply when implementing the algorithm.

As expected, rapid convergence was not observed with respect to the two coefficient criteria. As expected, the values of the 3-coefficient criterion were generally lower than those of the 6-coefficient criterion, but the rates of convergence seemed to be similar. Convergence was generally monotonic up to 5 iterations, but in many cases there was little further improvement or even divergence between 5 and 10 iterations. Convergence was generally marginally better with respect to the true population parameters than the OLS sample estimates. In general improved convergence was observed with larger samples. Interestingly, the level of error in the population data did not have a discernible effect under any conditions.

Table 7.2 Summary of Simulation Results: Estimates of Regression Coefficients[1]

A. Models derived from residential class data; RMS discrepancies in 6 regression coefficients

Simulated population relationship based on	Neg adjust	Est. from Neg adjust		Residential class												Residential training set							
				No						Yes						No				Yes			
	Neg	Sample size		1000		5000		10000		1000		5000		10000		1000		5000		1000		5000	
	adjust	σ	Iter	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop
Residential class	No	0	0	0.0509	0.0513	0.0495	0.0494	0.0499	0.0497	0.0509	0.0513	0.0495	0.0494	0.0499	0.0497	0.1152	0.1102			0.1152	0.1102		
			1	0.0265	0.0265	0.0216	0.0216	0.0206	0.0206	0.0268	0.0268	0.0219	0.0219	0.0209	0.0209	0.1059	0.1005			0.1059	0.1005		
			5	0.0267	0.0263	0.0295	0.0297	0.0265	0.0267	0.0223	0.0222	0.0186	0.0188	0.0158	0.0160	0.0947	0.0887			0.0953	0.0894		
			10	0.0316	0.0313	0.0320	0.0322	0.0258	0.0261	0.0234	0.0234	0.0179	0.0180	0.0141	0.0143	0.0927	0.0868			0.0941	0.0882		
		0.5	0	0.0492	0.0507	0.0492	0.0487	0.0496	0.0491	0.0492	0.0507	0.0492	0.0487	0.0496	0.0491	0.1152	0.1083			0.1152	0.1083		
			1	0.0253	0.0260	0.0214	0.0210	0.0206	0.0201	0.0255	0.0262	0.0216	0.0212	0.0208	0.0203	0.1061	0.0987			0.1061	0.0987		
			5	0.0254	0.0247	0.0275	0.0277	0.0249	0.0251	0.0219	0.0214	0.0188	0.0187	0.0161	0.0160	0.0954	0.0875			0.0960	0.0881		
			10	0.0298	0.0292	0.0288	0.0291	0.0236	0.0238	0.0233	0.0228	0.0169	0.0167	0.0131	0.0128	0.0938	0.0860			0.0951	0.0871		
		1.0	0	0.0486	0.0502	0.0477	0.0483	0.0481	0.0487	0.0486	0.0502	0.0477	0.0483	0.0481	0.0487	0.1138	0.1045			0.1138	0.1045		
			1	0.0273	0.0262	0.0213	0.0208	0.0207	0.0202	0.0273	0.0262	0.0213	0.0208	0.0207	0.0202	0.1050	0.0952			0.1050	0.0952		
			5	0.0269	0.0226	0.0248	0.0233	0.0227	0.0208	0.0256	0.0212	0.0204	0.0189	0.0181	0.0162	0.0950	0.0846			0.0954	0.0850		
			10	0.0307	0.0268	0.0247	0.0233	0.0205	0.0185	0.0278	0.0236	0.0176	0.0161	0.0135	0.0115	0.0937	0.0835			0.0948	0.0843		
	Yes	0	0	0.0449	0.0453					0.0449	0.0453					0.1057	0.1020			0.1057	0.1020		
			1	0.0223	0.0224					0.0224	0.0225					0.0994	0.0956			0.0994	0.0956		
			5	0.0231	0.0228					0.0199	0.0198					0.0932	0.0892			0.0934	0.0894		
			10	0.0284	0.0282					0.0222	0.0222					0.0930	0.0891			0.0936	0.0897		
		0.5	0	0.0426	0.0473					0.0426	0.0473					0.1057	0.1022			0.1057	0.1022		
			1	0.0207	0.0242					0.0208	0.0243					0.0997	0.0959			0.0997	0.0959		
			5	0.0217	0.0210					0.0198	0.0198					0.0941	0.0898			0.0942	0.0900		
			10	0.0266	0.0259					0.0228	0.0226					0.0942	0.0899			0.0947	0.0904		
		1.0	0	0.0418	0.0418					0.0418	0.0418					0.1045	0.0928			0.1045	0.0928		
			1	0.0232	0.0209					0.0232	0.0209					0.0989	0.0869			0.0989	0.0869		
			5	0.0238	0.0200					0.0236	0.0195					0.0939	0.0817			0.0940	0.0818		
			10	0.0280	0.0247					0.0272	0.0235					0.0943	0.0824			0.0947	0.0826		

1 Ten simulations were carried out for each indicated combination of function/population sampled from/sample size/negative adjustment. Each simulation was iterated 10 times. Table entries are the RMS discrepancy between the regression coefficients for the 6 TM bands (excluding the constant) obtained after iterations 0, 1, 5 and 10 and the (1) the coefficients obtained by OLS applied to the true pixel populations in the sample and (2) the population parameters.

Table 7.2 Summary of Simulation Results: Estimates of Regression Coefficients[1] (continued)

B. Model derived from residential training set data; RMS discrepancies in 6 regression coefficients

Simulated population relationship based on	Neg adjust	Est. from Neg adjust Sample size		Residential class												Residential training set							
				No						Yes						No				Yes			
				1000		5000		10000		1000		5000		10000		1000		5000		1000		5000	
		σ	Iter	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop
Residential training set	No	0	0	0.0731	0.0765					0.0731	0.0765					0.0850	0.0984	0.0840	0.1100	0.0850	0.0984	0.0840	0.1100
			1	0.0615	0.0628					0.0615	0.0628					0.0738	0.0910	0.0719	0.1002	0.0738	0.0910	0.0719	0.1002
			5	0.0484	0.0535					0.0484	0.0535					0.0524	0.0808	0.0496	0.0849	0.0524	0.0808	0.0496	0.0849
			10	0.0445	0.0538					0.0445	0.0538					0.0437	0.0792	0.0421	0.0821	0.0439	0.0792	0.0422	0.0821
		0.5	0	0.0714	0.0715					0.0714	0.0715					0.0855	0.0841	0.0843	0.1059	0.0855	0.0841	0.0843	0.1059
			1	0.0603	0.0601					0.0603	0.0601					0.0744	0.0730	0.0723	0.0961	0.0744	0.0730	0.0723	0.0961
			5	0.0480	0.0474					0.0480	0.0474					0.0536	0.0521	0.0501	0.0813	0.0536	0.0521	0.0501	0.0813
			10	0.0444	0.0435					0.0444	0.0435					0.0451	0.0439	0.0428	0.0791	0.0453	0.0440	0.0429	0.0791
		1.0	0	0.0682	0.0696					0.0682	0.0696					0.0845	0.0815	0.0832	0.1124	0.0845	0.0815	0.0832	0.1062
			1	0.0582	0.0588					0.0582	0.0588					0.0738	0.0707	0.0715	0.1029	0.0738	0.0707	0.0715	0.0967
			5	0.0477	0.0468					0.0477	0.0468					0.0536	0.0505	0.0498	0.0876	0.0536	0.0505	0.0498	0.0820
			10	0.0451	0.0434					0.0451	0.0434					0.0453	0.0424	0.0426	0.0845	0.0454	0.0425	0.0426	0.0796

1 Ten simulations were carried out for each indicated combination of function/population sampled from/sample size/negative adjustment. Each simulation was iterated 10 times. Table entries are the RMS discrepancy between the regression coefficients for the 6 TM bands (excluding the constant) obtained after iterations 0, 1, 5 and 10 and the (1) the coefficients obtained by OLS applied to the true pixel populations in the sample and (2) the population parameters.

Table 7.2 Summary of Simulation Results: Estimates of Regression Coefficients[1] (continued)

C. Models derived from residential class data; RMS discrepancies in 3 regression coefficients

Simulated population relationship based on	Neg adjust	Est. from Neg adjust Sample size		Residential class												Residential training set							
				No						Yes						No				Yes			
				1000		5000		10000		1000		5000		10000		1000		5000		1000		5000	
	Neg adjust	σ	Iter	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop
Residential class	No	0	0	0.0445	0.0444	0.0442	0.0442	0.0441	0.0441	0.0445	0.0444	0.0442	0.0442	0.0441	0.0441	0.0528	0.0561			0.0528	0.0561		
			1	0.0282	0.0281	0.0264	0.0264	0.0258	0.0258	0.0286	0.0284	0.0267	0.0267	0.0262	0.0262	0.0365	0.0390			0.0365	0.0390		
			5	0.0153	0.0154	0.0090	0.0090	0.0075	0.0075	0.0182	0.0181	0.0154	0.0154	0.0145	0.0145	0.0239	0.0217			0.0230	0.0211		
			10	0.0163	0.0165	0.0063	0.0064	0.0041	0.0042	0.0175	0.0173	0.0154	0.0154	0.0149	0.0149	0.0286	0.0252			0.0230	0.0202		
		0.5	0	0.0435	0.0434	0.0435	0.0430	0.0434	0.0430	0.0435	0.0434	0.0435	0.0430	0.0434	0.0430	0.0529	0.0542			0.0529	0.0542		
			1	0.0277	0.0274	0.0260	0.0255	0.0255	0.0250	0.0278	0.0275	0.0261	0.0256	0.0256	0.0252	0.0366	0.0373			0.0366	0.0373		
			5	0.0153	0.0150	0.0089	0.0085	0.0075	0.0072	0.0170	0.0165	0.0134	0.0129	0.0125	0.0120	0.0239	0.0219			0.0232	0.0213		
			10	0.0161	0.0160	0.0063	0.0062	0.0043	0.0042	0.0162	0.0157	0.0125	0.0120	0.0120	0.0115	0.0286	0.0265			0.0233	0.0212		
		1.0	0	0.0419	0.0414	0.0420	0.0410	0.0419	0.0409	0.0419	0.0414	0.0420	0.0410	0.0419	0.0409	0.0518	0.0508			0.0518	0.0508		
			1	0.0272	0.0262	0.0252	0.0242	0.0248	0.0239	0.0272	0.0263	0.0252	0.0242	0.0248	0.0239	0.0358	0.0343			0.0358	0.0343		
			5	0.0158	0.0144	0.0088	0.0079	0.0076	0.0067	0.0162	0.0146	0.0104	0.0093	0.0094	0.0083	0.0238	0.0219			0.0232	0.0213		
			10	0.0165	0.0154	0.0063	0.0059	0.0044	0.0040	0.0155	0.0140	0.0078	0.0067	0.0068	0.0055	0.0285	0.0277			0.0242	0.0230		
	Yes	0	0	0.0267	0.0266					0.0267	0.0266					0.0365	0.0380			0.0365	0.0380		
			1	0.0158	0.0158					0.0158	0.0158					0.0262	0.0268			0.0262	0.0268		
			5	0.0157	0.0158					0.0118	0.0119					0.0275	0.0254			0.0273	0.0252		
			10	0.0189	0.0190					0.0118	0.0118					0.0324	0.0300			0.0302	0.0278		
		0.5	0	0.0256	0.0275					0.0256	0.0275					0.0360	0.0382			0.0360	0.0382		
			1	0.0153	0.0167					0.0153	0.0167					0.0259	0.0270			0.0259	0.0270		
			5	0.0153	0.0152					0.0128	0.0128					0.0277	0.0256			0.0275	0.0254		
			10	0.0184	0.0181					0.0134	0.0132					0.0327	0.0303			0.0307	0.0283		
		1.0	0	0.0245	0.0236					0.0245	0.0236					0.0345	0.0330			0.0345	0.0330		
			1	0.0152	0.0138					0.0152	0.0138					0.0250	0.0229			0.0250	0.0229		
			5	0.0154	0.0142					0.0147	0.0133					0.0274	0.0259			0.0273	0.0258		
			10	0.0182	0.0173					0.0165	0.0154					0.0323	0.0316			0.0312	0.0304		

1 Ten simulations were carried out for each indicated combination of function/population sampled from/sample size/negative adjustment. Each simulation was iterated 10 times. Table entries are the RMS discrepancy between the regression coefficients for the 6 TM bands (excluding the constant) obtained after iterations 0, 1, 5 and 10 and the (1) the coefficients obtained by OLS applied to the true pixel populations in the sample and (2) the population parameters.

Table 7.2 Summary of Simulation Results: Estimates of Regression Coefficients[1] (continued)

D. Model derived from residential training set data; RMS discrepancies in 3 regression coefficients

Simulated population relationship based on	Neg adjust	Est. from Neg adjust		Residential class												Residential training set							
				No						Yes						No				Yes			
	Neg	Sample size		1000		5000		10000		1000		5000		10000		1000		5000		1000		5000	
	adjust	σ	Iter	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop	OLS	Pop
Residential training set	No	0	0	0.0731	0.0765					0.0731	0.0765					0.0850	0.0984	0.0840	0.1100	0.0850	0.0984	0.0840	0.1100
			1	0.0615	0.0628					0.0615	0.0628					0.0738	0.0910	0.0719	0.1002	0.0738	0.0910	0.0719	0.1002
			5	0.0484	0.0535					0.0484	0.0535					0.0524	0.0808	0.0496	0.0849	0.0524	0.0808	0.0496	0.0849
			10	0.0445	0.0538					0.0445	0.0538					0.0437	0.0792	0.0421	0.0821	0.0439	0.0792	0.0422	0.0821
		0.5	0	0.0714	0.0715					0.0714	0.0715					0.0855	0.0841	0.0843	0.1059	0.0855	0.0841	0.0843	0.1059
			1	0.0603	0.0601					0.0603	0.0601					0.0744	0.0730	0.0723	0.0961	0.0744	0.0730	0.0723	0.0961
			5	0.0480	0.0474					0.0480	0.0474					0.0536	0.0521	0.0501	0.0813	0.0536	0.0521	0.0501	0.0813
			10	0.0444	0.0435					0.0444	0.0435					0.0451	0.0439	0.0428	0.0791	0.0453	0.0440	0.0429	0.0791
		1.0	0	0.0682	0.0696					0.0682	0.0696					0.0845	0.0815	0.0832	0.1124	0.0845	0.0815	0.0832	0.1062
			1	0.0582	0.0588					0.0582	0.0588					0.0738	0.0707	0.0715	0.1029	0.0738	0.0707	0.0715	0.0967
			5	0.0477	0.0468					0.0477	0.0468					0.0536	0.0505	0.0498	0.0876	0.0536	0.0505	0.0498	0.0820
			10	0.0451	0.0434					0.0451	0.0434					0.0453	0.0424	0.0426	0.0845	0.0454	0.0425	0.0426	0.0796

1 Ten simulations were carried out for each indicated combination of function/population sampled from/sample size/negative adjustment. Each simulation was iterated 10 times. Table entries are the RMS discrepancy between the regression coefficients for the 6 TM bands (excluding the constant) obtained after iterations 0, 1, 5 and 10 and the (1) the coefficients obtained by OLS applied to the true pixel populations in the sample and (2) the population parameters.

Table 7.3 Summary of Simulation Results: Estimates of Population of Individual Pixels

A. Models derived from residential class data; RMS errors in pixel populations averaged over samples

Simulated population relationship based on	Neg adjust	Error level σ	Est. from	Residential class						Residential training set			
			Neg adjust	No			Yes			No		Yes	
			Sample size	1000	5000	10000	1000	5000	10000	1000	5000	1000	5000
			Basis of est.										
Residential class	No	0	Mean	1.3227	1.3204	1.3203	1.3227	1.3204	1.3203	1.3319		1.3319	
			Iteration 0	0.6518	0.6536	0.6562	0.6518	0.6536	0.6562	1.1087		1.1087	
			Iteration 1	0.4130	0.3828	0.3800	0.4166	0.3864	0.3836	0.9702		0.9702	
			Iteration 5	0.3013	0.2453	0.2383	0.3106	0.2550	0.2472	0.7518		0.7551	
			Iteration 10	0.3134	0.2400	0.2294	0.3071	0.2477	0.2399	0.6576		0.6689	
		0.5	Mean	1.3596	1.3559		1.3596	1.3559		1.4143		1.4143	
			Iteration 0	0.7581	0.7564		0.7581	0.7564		1.2075		1.2075	
			Iteration 1	0.5746	0.5500		0.5759	0.5513		1.0824		1.0824	
			Iteration 5	0.5035	0.4691		0.5064	0.4715		0.8928		0.8953	
			Iteration 10	0.5106	0.4668		0.5047	0.4668		0.8153		0.8233	
		1.0	Mean	1.4715	1.4649		1.4715	1.4649		1.6222	1.6056	1.6222	
			Iteration 0	1.0061	0.9983		1.0061	0.9983		1.4525	1.4347	1.4525	
			Iteration 1	0.8890	0.8668		0.8891	0.8669		1.3546	1.3350	1.3546	
			Iteration 5	0.8489	0.8220		0.8491	0.8219		1.2144	1.1869	1.2155	
			Iteration 10	0.8527	0.8208		0.8493	0.8189		1.1605	1.1214	1.1640	
	Yes	0	Mean	1.1962			1.1962			1.2032		1.2032	
			Iteration 0	0.5795			0.5795			1.0295		1.0295	
			Iteration 1	0.3576			0.3592			0.9216		0.9216	
			Iteration 5	0.2638			0.2584			0.7407		0.7412	
			Iteration 10	0.2799			0.2567			0.6486		0.6506	
		0.5	Mean	1.2364			1.2364			1.2920		1.2920	
			Iteration 0	0.7059			0.7059			1.1350		1.1350	
			Iteration 1	0.5501			0.5505			1.0400		1.0400	
			Iteration 5	0.4978			0.4955			0.8865		0.8868	
			Iteration 10	0.5059			0.4960			0.8118		0.8129	
		1.0	Mean	1.3657			1.3657			1.5159		1.5159	
			Iteration 0	0.9776			0.9776			1.3934		1.3934	
			Iteration 1	0.8846			0.8846			1.3222		1.3222	
			Iteration 5	0.8565			0.8561			1.2119		1.2120	
			Iteration 10	0.8606			0.8583			1.1608		1.1611	

Table 7.3 Summary of Simulation Results: Estimates of Population of Individual Pixels (continued)

B. Model derived from residential training set data; RMS errors in pixel populations averaged over samples[1]

Simulated population relationship based on	Neg adjust	Error level σ	Est. from	Residential class						Residential training set			
			Neg adjust	No			Yes			No		Yes	
			Sample size	1000	5000	10000	1000	5000	10000	1000	5000	1000	5000
			Basis of est.										
Residential training set	No	0	Mean	0.7162			0.7162			0.7625	0.7645	0.7625	0.7645
			Iteration 0	0.5018			0.5018			0.6048	0.6026	0.6048	0.6026
			Iteration 1	0.4289			0.4289			0.5000	0.4939	0.5000	0.4939
			Iteration 5	0.3488			0.3488			0.3194	0.3029	0.3194	0.3029
			Iteration 10	0.3227			0.3227			0.2517	0.2257	0.2517	0.2261
		0.5	Mean	0.8700			0.8700			0.9114	0.9060	0.9114	0.9060
			Iteration 0	0.7051			0.7051			0.7860	0.7771	0.7860	0.7771
			Iteration 1	0.6561			0.6561			0.7102	0.6985	0.7102	0.6985
			Iteration 5	0.6084			0.6084			0.6016	0.5856	0.6016	0.5856
			Iteration 10	0.5944			0.5944			0.5711	0.5529	0.5712	0.5529
		1.0	Mean	1.1621			1.1621			1.2357	1.2193	1.2357	1.2193
			Iteration 0	1.0555			1.0555			1.1529	1.1343	1.1529	1.1343
			Iteration 1	1.0269			1.0269			1.1063	1.0863	1.1063	1.0863
			Iteration 5	1.0006			1.0006			1.0452	1.0237	1.0452	1.0237
			Iteration 10	0.9934			0.9934			1.0298	1.0076	1.0299	1.0076

1 The figure tabulated is the RMS average value, calculated over 10 samples, of the RMS error in individual pixel population estimates.

Table 7.3 Summary of Simulation Results: Estimates of Population of Individual Pixels (continued)

C. Models derived from residential class data; RMS errors in pixel populations: variation between samples

Simulated population relationship based on	Neg adjust	Error level σ	Est. from	Residential class						Residential training set			
			Neg adjust	No			Yes			No		Yes	
			Sample size	1000	5000	10000	1000	5000	10000	1000	5000	1000	5000
			Basis of est.										
Residential class	No	0	Mean	0.0271	0.0088	0.0052	0.0271	0.0088	0.0052	0.0237		0.0237	
			Iteration 0	0.0295	0.0081	0.0034	0.0295	0.0081	0.0034	0.0304		0.0304	
			Iteration 1	0.0255	0.0082	0.0039	0.0252	0.0081	0.0038	0.0420		0.0420	
			Iteration 5	0.0238	0.0090	0.0030	0.0207	0.0084	0.0033	0.0615		0.0620	
			Iteration 10	0.0257	0.0137	0.0036	0.0197	0.0077	0.0030	0.0580		0.0647	
		0.5	Mean	0.0284	0.0089		0.0284	0.0089		0.0177		0.0177	
			Iteration 0	0.0296	0.0086		0.0296	0.0086		0.0227		0.0227	
			Iteration 1	0.0232	0.0078		0.0232	0.0078		0.0327		0.0327	
			Iteration 5	0.0207	0.0055		0.0197	0.0058		0.0479		0.0483	
			Iteration 10	0.0219	0.0074		0.0194	0.0057		0.0435		0.0487	
		1.0	Mean	0.0392	0.0113		0.0392	0.0113		0.0171	0.0073	0.0171	
			Iteration 0	0.0371	0.0102		0.0371	0.0102		0.0203	0.0075	0.0203	
			Iteration 1	0.0327	0.0091		0.0327	0.0091		0.0268	0.0081	0.0268	
			Iteration 5	0.0304	0.0075		0.0304	0.0077		0.0360	0.0097	0.0362	
			Iteration 10	0.0303	0.0079		0.0299	0.0081		0.0333	0.0103	0.0358	
	Yes	0	Mean	0.0222			0.0222			0.0248		0.0248	
			Iteration 0	0.0241			0.0241			0.0336		0.0336	
			Iteration 1	0.0196			0.0196			0.0445		0.0445	
			Iteration 5	0.0209			0.0174			0.0619		0.0621	
			Iteration 10	0.0271			0.0188			0.0589		0.0622	
		0.5	Mean	0.0233			0.0233			0.0157		0.0157	
			Iteration 0	0.0239			0.0239			0.0246		0.0246	
			Iteration 1	0.0175			0.0175			0.0344		0.0344	
			Iteration 5	0.0151			0.0142			0.0479		0.0480	
			Iteration 10	0.0174			0.0142			0.0438		0.0459	
		1.0	Mean	0.0364			0.0364			0.0149		0.0149	
			Iteration 0	0.0331			0.0331			0.0207		0.0207	
			Iteration 1	0.0290			0.0291			0.0268		0.0268	
			Iteration 5	0.0264			0.0264			0.0345		0.0346	
			Iteration 10	0.0262			0.0258			0.0319		0.0323	

Table 7.3 Summary of Simulation Results: Estimates of Population of Individual Pixels (continued)

D. Model derived from residential training set data; RMS errors in pixel populations: variation between samples[1]

Simulated population relationship based on	Neg adjust	Error level σ	Est. from	Residential class						Residential training set			
			Neg adjust	No			Yes			No		Yes	
			Sample size	1000	5000	10000	1000	5000	10000	1000	5000	1000	5000
			Basis of est.										
Residential training set	No	0	Mean	0.0169			0.0169			0.0185	0.0042	0.0185	0.0042
			Iteration 0	0.0169			0.0169			0.0137	0.0038	0.0137	0.0038
			Iteration 1	0.0168			0.0168			0.0126	0.0041	0.0126	0.0041
			Iteration 5	0.0222			0.0222			0.0207	0.0046	0.0207	0.0046
			Iteration 10	0.0268			0.0268			0.0283	0.0057	0.0283	0.0057
		0.5	Mean	0.0235			0.0235			0.0145	0.0026	0.0145	0.0026
			Iteration 0	0.0197			0.0197			0.0130	0.0028	0.0130	0.0028
			Iteration 1	0.0171			0.0171			0.0131	0.0031	0.0131	0.0031
			Iteration 5	0.0196			0.0196			0.0150	0.0030	0.0150	0.0030
			Iteration 10	0.0227			0.0227			0.0166	0.0031	0.0167	0.0031
		1.0	Mean	0.0320			0.0320			0.0219	0.0046	0.0219	0.0046
			Iteration 0	0.0280			0.0280			0.0214	0.0050	0.0214	0.0050
			Iteration 1	0.0264			0.0264			0.0207	0.0052	0.0207	0.0052
			Iteration 5	0.0277			0.0277			0.0195	0.0051	0.0195	0.0051
			Iteration 10	0.0292			0.0292			0.0200	0.0050	0.0201	0.0050

1 The figure tabulated is the standard deviation, calculated over 10 samples, of the RMS error in individual pixel population estimates.

Reassigning negative estimates after each iteration had little effect on the results for the samples from the residential training set, where negative estimates were less likely to occur. In the case of the samples from the residential class, this did improve the performance of the 6-coefficient criterion, but the reverse was evident with the 3-coefficient criterion.

Considering the population accuracy criterion, the basic functionality of the algorithm was clearly evidenced, in that the accuracy of population estimates after iteration was greater than before. Beginning with the most bland estimate of population, the mean for all pixels (which is functionally equivalent to counting residential pixels, as in some of the models of Langford et al., 1991 and Lo, 1995), an immediate gain was made by the initial regression, which was further enhanced by the iterative algorithm.

The best results were achieved for the combination of a model based on the residential class and samples drawn from the residential class, where the proportional reduction in estimation error by the fifth iteration was almost 80% in the “no error” models, and around 40% in the “high error” models. There was no further improvement, indeed a slight degradation of performance, by the tenth iteration in these cases.

The other three combinations of model source/sample source exhibited less marked, but still substantial, effects. In samples from the residential training set, the results were still improving after the tenth iteration, suggesting that more iterations might be of benefit for estimating the population of more typical suburban pixels, but that this might be achieved at the expense of loss of accuracy with respect to the more atypical pixels found in the broader distribution of the residential class.

Reassigning negative estimates after each iteration had no effect on the accuracy of the population estimates under any circumstances.

The values of the final criterion, which indicates the sample to sample variability in the population accuracy measure, were consistently smaller than those of the population accuracy measure itself, by a factor of at least 10. Although this measure is calculated only with respect to the sample data in each case, it indicates that the accuracy of population estimates produced by these procedures is likely to be quite consistent, in relative terms, from sample to sample. Of course, as the sample size is increased, this measure decreases, the decrease in this case being amplified by the finite size of the population.

In framing this analysis, the central question at issue was whether to train the regression on data from the residential classification training set or from the full residential class.

The subjective and intuitive mental processes which contributed to that decision can be crudely approximated by the following schema.

We argue as follows. Empirical data from the two sources have led to two different models relating population to spectral signature (leaving aside the two variants of the model based on the residential class). We do not know which of the two models better represents the reality of the relationship, but we have now examined the performance of procedures based on the two sampling strategies when applied to both versions of "reality".

We now construct a 2×2 tableau of "sampling strategy" vs. "actuality" (not unrelated to the familiar statistical illustration of Type I & Type II errors). For each of the four criteria we have examined, we rank the four quadrants from 4 to 1, 4 representing the best performance and 1 representing the worst (with ties averaged in the usual manner). The rankings assigned are shown in Table 7.4.

Table 7.4 can be interpreted as follows. If the actual model is as suggested by the residential class data, then sampling from the residential class will lead to the best performance of all four scenarios. A similar level of performance will be attained if the actual model is more like the one suggested by the residential training set data and sampling is from the training set. The main difference emerges when the actuality and the sampling strategy are mismatched. On the evidence of the simulation study its seems that sampling from the broad residential class is likely to be marginally more robust than sampling from the more narrowly defined residential class training set. The margin is not great. However, the former strategy also has the advantage that one can always reduce the scope of a data set by deleting outliers, but one cannot enhance the scope with data that one has not collected.

Because the rankings were fairly consistent on the 4 criteria, the specifics of Table 7.4 could be changed, for example by giving a higher weighting to the 3rd criterion, accuracy of population estimates, without substantially altering the conclusion.

Table 7.4 Assessment of Sampling Strategies

		Sampling strategy	
		Residential class	Residential classification training set
"Actuality": basis of model	Residential class	4+4+3+2.5=13.5	1+2.5+1+1=5.5
	Residential classification training set	2.5+1+3+2.5=9	2.5+2.5+3+4=12
	Total score	22.5	17.5

It was concluded that, on balance, the original intuitive decision to sample from the residential class was justified.

As to the other aspects, the results were clear cut with respect to the efficacy of increased sample size. On the issue of how many iterations to use, the evidence was more mixed, but considering that under the preferred sampling scheme the best results had been obtained with five iterations, it was decided to retain the fixed number of iterations (six) which had been used in the earlier work. This decision was subsequently reviewed (see Section 9.2.2). It was also concluded that on balance, the procedure of reassigning negative estimates at each iteration was not justified.

7.5 SUMMARY

In this chapter, the iterative re-estimation regression algorithm has been placed in the broader theoretical context of the EM and related algorithms.

The properties of the algorithm were examined, first through repeated sampling then by simulation. As a result, it was decided:

- to continue to apply the algorithm to data from the full residential class, rather than the residential classification training set;
- to continue to use six iterations;
- to increase the sample size;
- not to reassign negative estimates at each iteration.

With the tool of the iterative re-estimation regression algorithm better understood, the next step was to broaden the scope of validation to encompass a range of Australian images.

Chapter 8

Normalised Population Estimation Models

8.1 INTRODUCTION

In Chapter 7, the first steps were taken towards exploring the robustness of the population estimation methodologies developed in Chapter 4 and Chapter 5. The CD aggregate approach of Chapter 4 was found to be inadequate. More success was had with the two phase pixel-based classification and regression methodology of Chapter 5, but only when the supervised classification phase was trained locally on the secondary image. This was hardly surprising, for whilst the residential areas of the two images were reasonably similar in character, the range of other landcover and land use classes was rather different. Nevertheless population estimates, of comparable quality to those obtained for the primary image, were obtained for the secondary image, using the same regression equation for estimating pixel populations which had been trained on the primary image.

It had been hoped in the earlier phase of exploring different spectral and spatial data transformations that some relatively invariant surrogate for population might emerge, but that had not been the case. In testing a range of possibilities, nothing more robust was found than a simple linear combination of the 6 TM bands. The fact that the secondary image was closely related, both spatially and temporally, to the primary image was obviously crucial in this successful demonstration of limited robustness of the estimation equation. But clearly, in its raw form such a function could not conceivably be invariant even under moderate seasonal differences in general level and angle of solar illumination, much less under seasonal or climatic differences in vegetation, geographic differences in soils and other ground cover, and cultural differences in intensity and pattern of residential and other land use.

For the procedure to generalise further, some form of image-specific re-calibration or normalisation would be necessary. Chapter 8 is an account of the testing of various normalising procedures on a number of further test images.

Section 8.2 outlines the demographic characteristics of the supplementary test areas, the spectral characteristics of the images, and the preliminary preparatory steps undertaken. In Section 8.3, a number of minor adjustments to the estimation equation are reported, and three potential normalising transformations are outlined and compared. Section 8.4 reports on their application.

8.2 THE SUPPLEMENTARY TEST AREAS

8.2.1 Characteristics of the study areas and images

The primary and secondary images were supplemented by 5 more images, one showing part of the primary Ballarat test area on another occasion, three showing the areas surrounding and including the major urban centres of Sydney, Brisbane and Adelaide and the other being centred on the remote mining town of Kalgoorlie. Images 15, 17, 19, 21, and 23 are RGB images of each area; census collection district boundaries for each area are shown in Figure 8.1. Some technical details of each image together with some demographic information about each area, are given in Table 8.1. Further information about the population distributions is in Appendix H.

In this Chapter, urban areas are defined in terms of population density of Census Collection Districts. Two cutoffs were used: the basic ASGC criterion of 200 persons/sq.km. (ABS, 1998); and a higher value of 500 persons/sq.km, which was intended to exclude some anomalies such as partly developed CDs within urban areas and around the urban fringe.

The issue of time differences between the TM images and the census data has been discussed in Section 3.6. The population change in the cases of Adelaide, Sydney and the second Ballarat image was assessed as being less than 1%, which is comparable to the margin of error for the census process, and so no adjustments were made in these cases. Of course, specific CDs where development had occurred in the intervening time would be more substantially affected, and would be expected to appear as outliers. The issue of temporal mismatch was more problematical with Brisbane and Kalgoorlie (see Section 8.4.2).

8.2.2 Classification of the images

Each of the 5 supplementary images in turn was co-registered to the CD boundaries as described in Chapter 3. For each of the images in turn, a set of landcover/landuse classes was identified and training regions selected for each, as described in Section 5.3.1. Several of these classes were common to all the images, whilst others, particularly those to do with vegetation were more image-specific. The number of classes varied from 12 for the Kalgoorlie image to 22 for the Adelaide image. This was by far the largest image in extent, which had been chosen to include a number of small country towns as well as Adelaide itself.

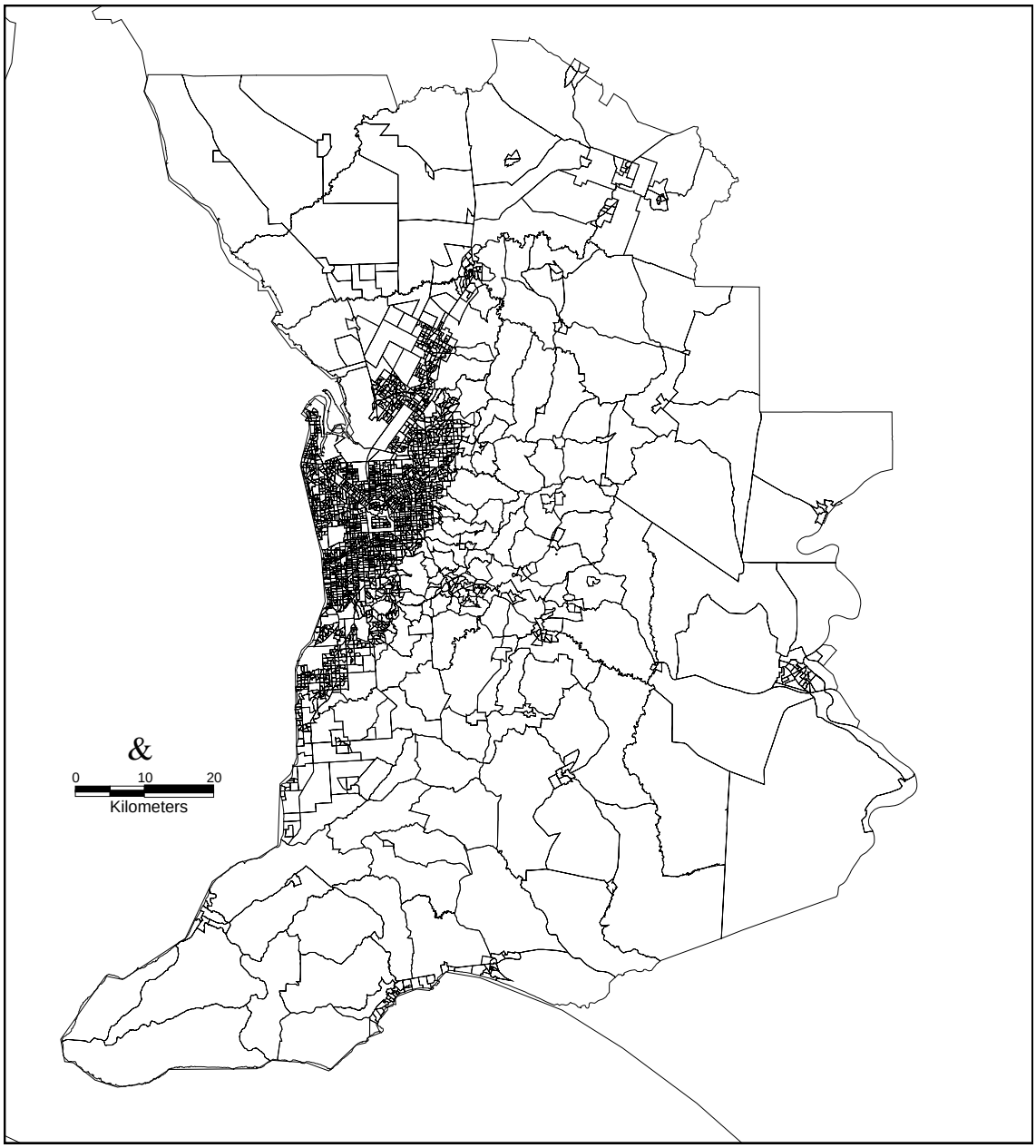
Table 8.1 Characteristics of Study Areas and Images

		Image			Region					Urban section[5]				
Study area	**State**	**Date**	**Size** pixels/line ×lines	**Pixel size** m	**No. of CDs[1]**	**No. of SLAs[2]**	**Pop.[6]**	**Area** sq. km.	**Av. Pop. density** p/sq.km.	**Pop.[6]**	**Area** sq. km	**Av. Pop. density** p/sq.km.	**Pop. %**	**Area %**
Ballarat	Victoria	14/2/88	1350×1008	30	138	6	79,179	634	125	64,564	48	1345	81.5	7.5
Ballarat*	Victoria	15/12/94	616 × 697	30	72	2[3]	35,711	199	179	30,078	27	1114	84.2	13.3
Geelong	Victoria	14/2/88	1119×1174	30	224	8	147,910	352	420	132,366	68	1947	89.5	19.3
Adelaide	South Australia	2/2/97	5010 × 6187	25	2412	47	1,158,625	10735	108	1,001,099	580	1726	86.4	5.4
Sydney	New South Wales	8/12/96	2740 × 3678	25	5628	41	3,283,889	3524	932	3,138,640	1220	2573	95.6	34.6
Brisbane	Queensland	16/9/89	2965 × 3616	30	2605	225[4]	1,488,880	4623	322	1,253,117	770	1627	84.2	16.7
Kalgoorlie	Western Australia	27/9/89	1201 × 1078	30	51	1	30246	62	488	24,686	19	1299	81.6	29.7

1 Census Collection Districts
2 Statistical Local Areas
3 This image included only a section of the primary study area. In addition, the SLA structure in Victoria changed in 1994. Parts of 2 new SLAs corresponded to parts of 5 of the old SLAs in the primary image.
4 SLAs are smaller in the state of Queensland than in the other states.
5 CDs with population density of at least 500 persons/sq.km.
6. For primary Ballarat and Geelong study areas, populations are estimates of residential population as at 14/2/88. For all other areas, populations are 1996 census counts.

Figure 8.1 Census Collection District Boundaries for Supplementary Study Areas

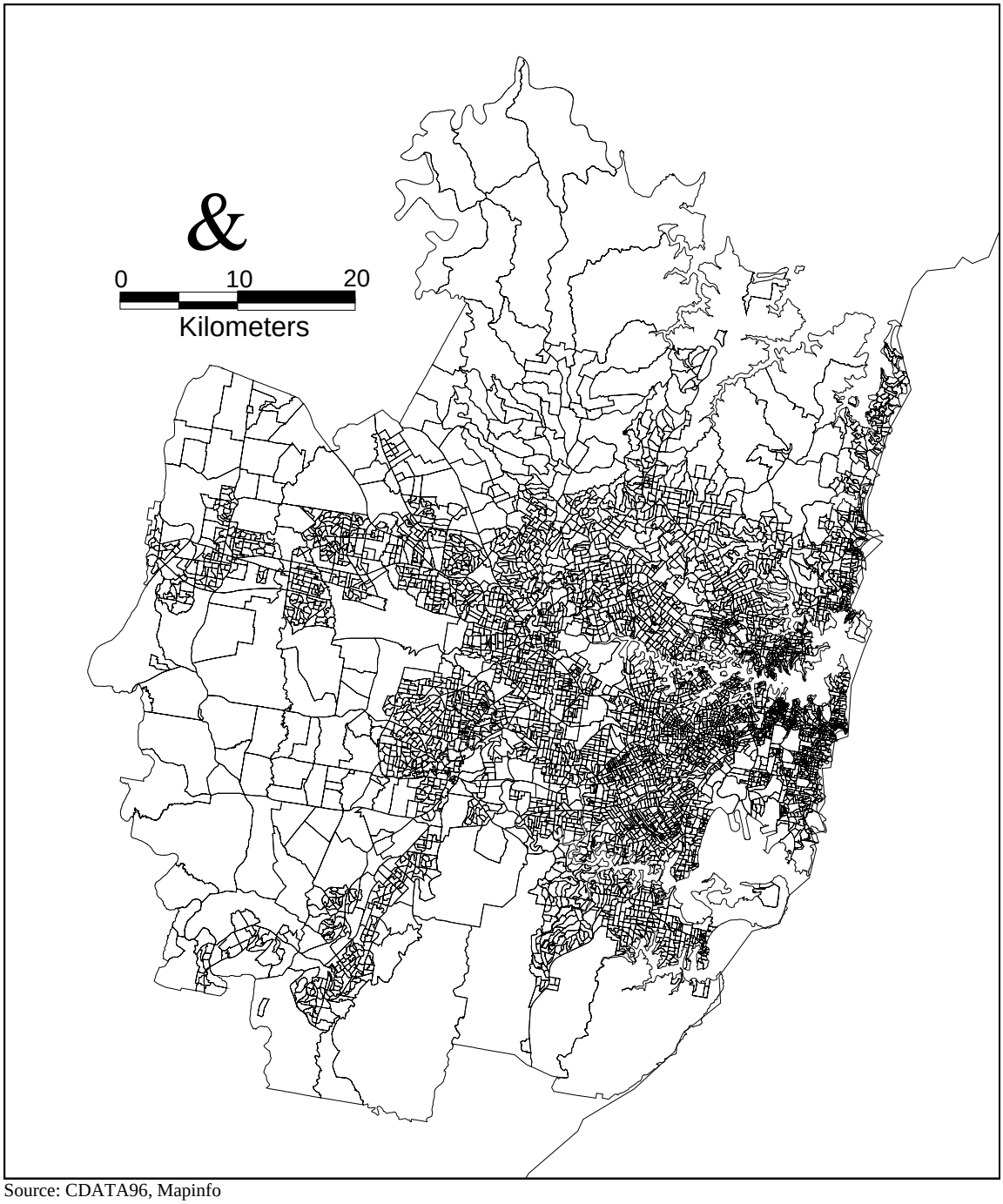
Adelaide study area
CD boundaries and approximate (low resolution) coastline



Source: CDATA96, Mapinfo

Figure 8.1 Census Collection District Boundaries for Supplementary Study Areas
(continued)

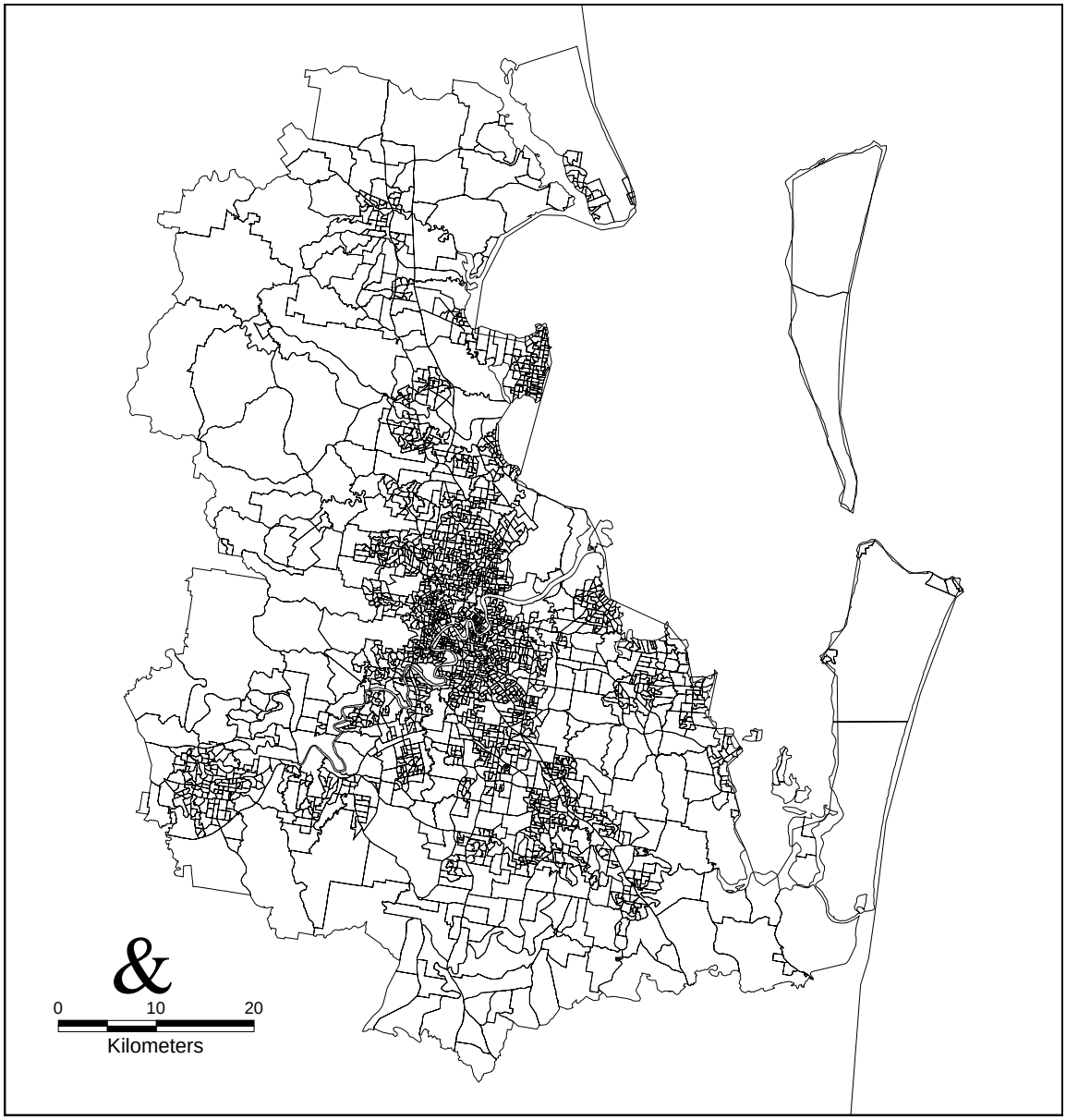
Sydney study area
CD boundaries and approximate (low resolution) coastline



Source: CDATA96, Mapinfo

Figure 8.1 Census Collection District Boundaries for Supplementary Study Areas
(continued)

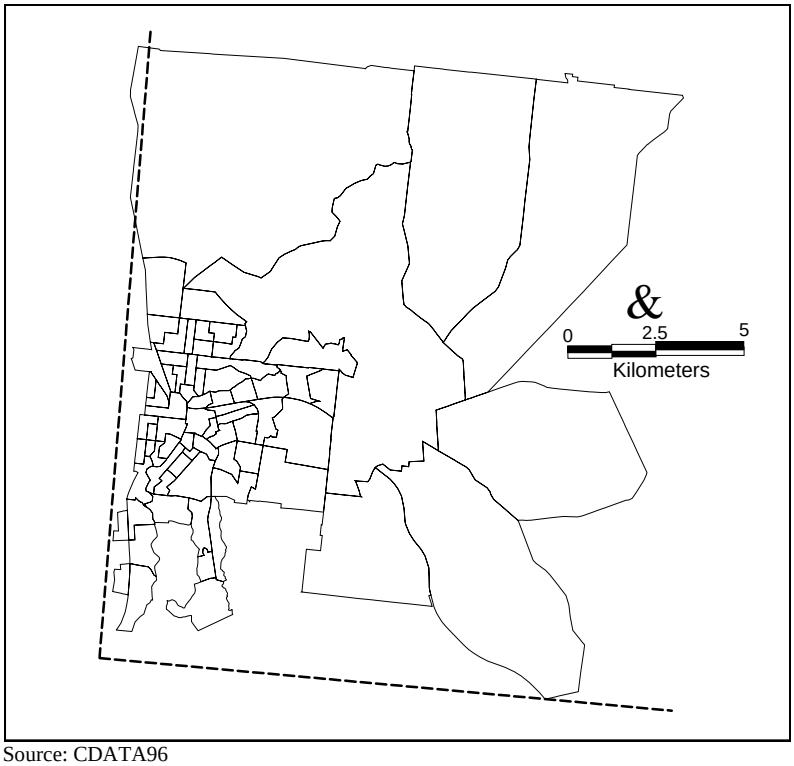
Brisbane study area
CD boundaries and approximate (low resolution) coastline



Source: CDATA96, Mapinfo

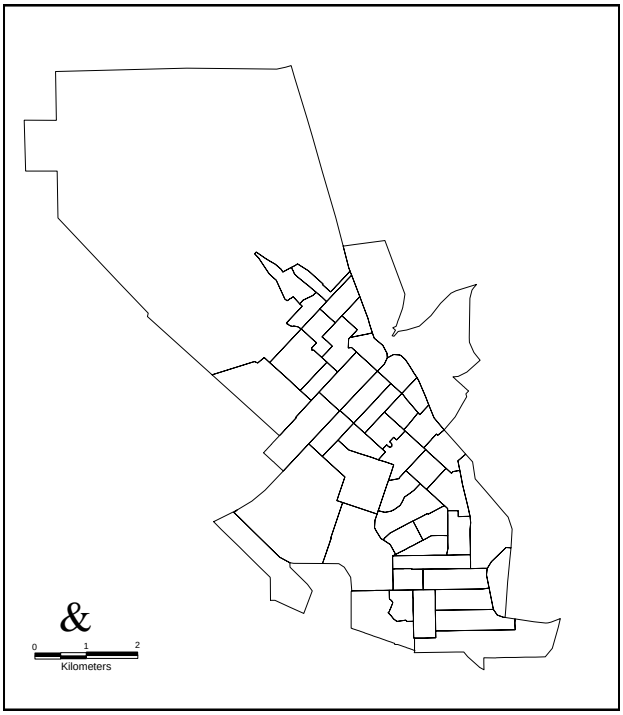
Figure 8.1 Census Collection District Boundaries for Supplementary Study Areas
(continued)

Ballarat: reduced 1994 study area
CD boundaries



Source: CDATA96

Kalgoorlie study area
CD boundaries



Source: CDATA96

As a consequence of its extent, it had by far the largest rural component of any of the images, and included a number of quite different climatic zones and agricultural and pastoral activities. Conversely, the region around Kalgoorlie is so sparsely inhabited that single CDs cover hundreds of square kilometres. In this case the analysis was confined to smaller CDs in the city and its immediate surrounds. Details of the classes are given in Appendix I.

The most crucial class in each image was of course the residential class. A broadly representative set of residential training regions were selected visually. Regions generally consisted of several contiguous blocks of apparently homogeneous residential character, with residential streets included but excluding major arterial roads and visible features such as schools, churches, neighbourhood shopping centres and parks. Checks made using street directories confirmed that judgements made on this visual basis were very accurate.

In each case a maximum likelihood classification of the image was then made using the 6 TM bands. The resulting classifications were displayed as color-coded images and inspected for face validity. As was discussed in Chapter 5, misclassification between non-residential classes was not a matter for concern – the central issue was to discriminate between residential and non-residential pixels. As with the primary and secondary images, some problems of spurious classification of pixels as residential was observed. As then, this problem took two forms: concentrated groups of pixels associated with particular features, natural or constructed; and more extended or diffuse misclassification in rural areas, associated with rural roads and with particular forms of agricultural activity.

Table 8.2 Some Features Misclassified as Residential

Built
Airports (boundaries between runways/aprons and grass)
Livestock markets
Oil refineries
Rural roads
Salt evaporation pans
Sewage treatment works
Agricultural
Market gardens
Orchards
Vineyards
Natural
Forest regenerating after fire
Forest viewed through a diffuse smoke plume
Forested shorelines
Kelp in shallow water
Marshes and mangroves
Partly vegetated sand dunes
Sandy shorelines

Some problem features where many pixels were characteristically misclassified as residential are listed in Table 8.2. What they have in common is a combination at sub-pixel level of two or more of vegetation, pavement, water, and bare ground. Such a mixture is spectrally similar to the mixture of vegetation, pavement, and metal or tile roofing materials which characterises residential pixels. The effect of this misclassification on population estimates is discussed later in this Chapter.

8.2.3 Spectral characteristics of the test images

To compare the spectral characteristics of the seven test images, means, medians and covariance matrices for the 6 TM bands were calculated for: all pixels in the image; pixels in the residential class training sets; pixels in the residential class. The band means are graphed in Figure 8.2 .

Clearly, there were substantial differences between the mean levels of the TM bands, for all three aspects of the images. The differences can be conjecturally related to differences in overall illumination level due to differences in latitude and season, climatic differences in vegetation, cultural differences in roofing materials, extent of paving and so on.

As would be expected with a maximum likelihood classification the mean of the residential class is in each case close to the mean of the training set; hence the similarity in class and training set profiles. There are however differences, and also differences with respect to spread, with the class variances generally being greater than the training set variances, especially in the longer wavelength bands (see Table 8.4). Clearly, there is a need to adapt the population estimation equation trained on Ballarat, if it is to work on this range of images.

8.3 ADJUSTMENT OF THE ESTIMATION EQUATION

8.3.1 Preliminary adjustments

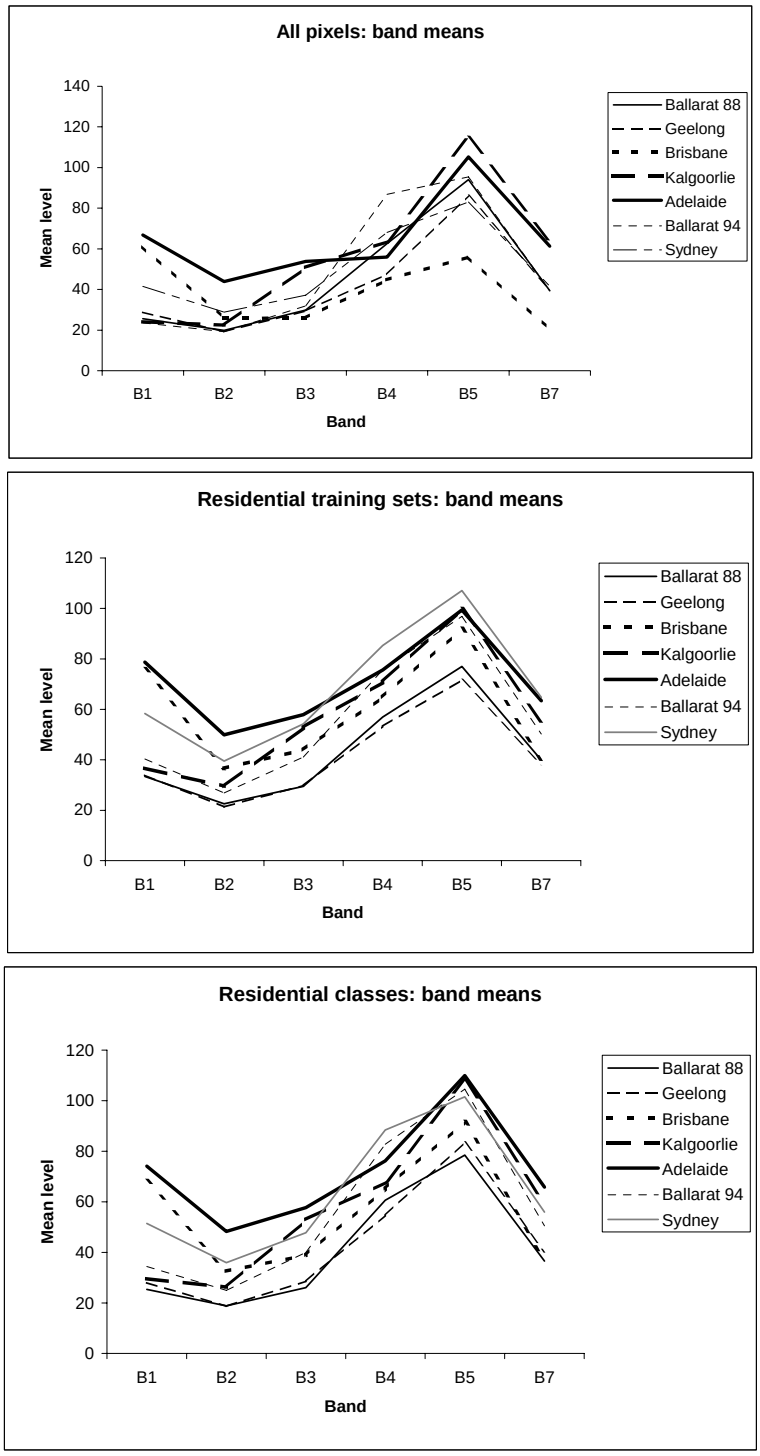
Following the decisions made in the light of the simulation study (see Section 7.5), the regression equation was re-estimated from the larger 20% sample of 14270 pixels from the residential class of the primary Ballarat image. The resulting equation is shown in Table 8.4.

Two other changes were also instituted at this point. Firstly, as was discussed in Chapter 3, the raw Ballarat image was recalibrated against 1996 Census boundaries registered to the AMG grid. It was found that the nominal 30m pixels were in fact slightly larger than 30m in both directions, the estimated error in the area being 2.7%

Since the later images had been resampled to exact 30m pixels, it was decided to rescale the estimation formula to a true 30m pixel size, by proportionally reducing all the regression

coefficients. The rescaled coefficients are also listed in Table 8.4. In the cases of the Adelaide and Sydney images the pixel size was 25m, necessitating a further rescaling of $(25/30)^2 = .6944$.

Figure 8.2 Spectral characteristics of the test images



This is predicated on the reasonable assumption that the same proportional combination of the same materials leading to the same spectral response for the smaller pixel will be associated

with the same population density, or equivalently a proportionally smaller pixel population. This correction factor is not incorporated in the figures in Table 8.4.

Secondly, it was decided not to incorporate the rather ad hoc high density power correction index (see Section 6.4.6) in the subsequent comparative valuations, the reasoning being that whilst this had been a useful gross correction in moving from Ballarat to the slightly higher densities of Geelong, in the context of the much higher densities which would be found in particular sections of the capital cities, a more specifically targeted approach would be more appropriate.

8.3.2 Normalising adjustments

Since it is clear that the linear relationship between pixel population and TM spectral reflectances estimated from the Ballarat image cannot be invariant for other images, we seek to re-express the relationship in some form which is invariant apart from some small number of easily estimated image-dependent parameters.

The basic approach was to recast the linear function of raw TM reflectances as a linear function of reflectances re-expressed in relativity to some reference level for the image. In all three cases, the reference level used was the mean reflectance. Three such normalising or invariance transformations were considered. Each one implied a particular form of invariant relationship: additive or difference-based, multiplicative or ratio-based, and scaled additive or z-score based.

The three models are summarised in Table 8.3.

In the ***untransformed model***, the constant term represents the population that would be assigned to a pixel which had zero reflectance in all bands. Whilst some pixels (such as water) might have such a signature, this is well outside the range of pixels classified as residential. The band coefficients represent the incremental change in population associated with each unit change in a particular band, subject to the other bands not changing.

After ***additive normalisation*** the constant represents the population of a pixel with the mean reflectance level in every band. Again, the band coefficients represent the incremental change in population associated with each unit change in a particular band, subject to the other bands not changing.

The value of this function would remain unchanged for each pixel if the levels of each pixel within a particular band changed by the same amount. Hence, this form of normalisation would be appropriate if the relationship between population and spectral reflectances were invariant apart from a constant translation or shift of level within each band. No obvious mechanism for

such a constant magnitude change from time to time or place to place suggested itself, so this approach was not expected to work very well.

Table 8.3 Normalised Forms of the Population Estimation Model

Transformation	Invariant indicator	Form of equation
Untransformed		$p = c_0 + \sum_{i=1}^{6} c_i b_i$
Additive	Difference from the mean	$p = c_0 + \sum_{i=1}^{6} c_i (b_i - \mu_i)$
Multiplicative	Ratio to the mean	$p = c_0 + \sum_{i=1}^{6} c_i \frac{b_i}{\mu_i}$
z-score (Scaled additive)	z score	$p = c_0 + \sum_{i=1}^{6} c_i \frac{(b_i - \mu_i)}{\sigma_i}$

where
p is the population of a pixel
c_i is the regression coefficient for band i
b_i is the reflectance of the pixel in band i
μ_i is the mean pixel reflectance in band i
σ_i is the standard deviation of the pixel reflectances in band i

With ***multiplicative normalisation***, the constant term again represents the population that would be assigned to a pixel which had zero reflectance in all bands. The band coefficients represent the incremental change in population associated, not with one unit change in a particular band, but with a change equal in magnitude to the mean level.

The value of this function would remain unchanged for each pixel if the levels of each pixel within a particular band changed in the same proportions. This form of normalisation would be appropriate if the relationship between population and spectral reflectances were invariant apart from proportional change within each band. Hence this might be expected to compensate for seasonal and geographic effects relating to overall levels of illumination.

With ***z-score normalisation***, the constant represents the population of a pixel with the mean reflectance level in every band, as for additive normalisation. The band coefficients represent the incremental change in population associated with a 1 standard deviation change in a particular band.

If the standard deviation is constant this is obviously equivalent to additive normalisation. It can also be shown that if the standard deviation is proportional to the mean, it is equivalent to multiplicative normalisation. But it is more generally applicable. So long as the relationship between population and reflectance is consistent relative to position of the pixel in the distribution of reflectances (for example if the greenest pixels always have the lowest

population, regardless of the absolute levels of green involved), this function will be invariant. Prima facie, this seemed to be the most promising transformation, since it is in a sense a generalisation which encompasses the other two methods.

Relationships between the coefficients of the three normalised equations and the raw equation were established (see Appendix J), and these were used firstly to convert the raw Ballarat 1988 equation into the three normalised forms. The inverse transformation was then used to produce raw equations corresponding to each normalisation, for each of the other images.

In calculating relevant means and standard deviations to use for normalisation, the choice had to be made between using the statistics for the residential training sets or the whole residential class for each image. It had previously been decided (see Chapter 7), to sample from the whole residential class to estimate the coefficients of the regression equation, but it did not necessarily follow that the same choice would be correct for this purpose. It was decided to investigate both approaches. The resulting coefficients are listed in Table 8.4.

The methodology described in this section is quite distinct from "band normalisation" (see Sections 2.3.2, 5.4.1). In that case, compensation is made for differences in overall brightness between the individual pixels of an image, by expressing each band level as a proportion of the total of all bands for the pixel. This calculation can be characterised as within pixels and across bands. In the present case, the aim is to compensate for differences between images, by standardising each band value for each pixel by comparison with the values in the same band for other pixels. This calculation can be characterised as across pixels and within bands.

8.4 APPLICATION OF NORMALISED MODELS

8.4.1 Methodology

Using the methodology described in Sections 5.6.2 and 5.6.3[1], the various population estimation equations were applied in turn to the various full images, estimated pixel populations were aggregated to produce CD estimates, and regression analyses were used to compare the remote sensing estimates with ground truth CD populations, with regard to a range of criteria previously described (Section 5.8).

In each case, analyses were performed for the whole region and for the urban areas of the region, defined in terms of the population density of Census Collection Districts. Two cutoffs were used: the basic ASGC criterion of 200 persons/sq.km. (ABS, 1998); and a higher value of

[1] Because of software limitations on vector to raster conversion, the implementation was slightly different to that previously described. Because of the large number of CDs in the capital city images, it was not possible to define a raster CD identification layer for all CDs simultaneously using the ERMapper

500 persons/sq.km, which was intended to exclude some anomalies such as partly developed CDs within urban areas and around the urban fringe.

Two levels of low density thresholding were incorporated: each model was first fitted with no zero thresholding; then all negative pixel populations were reset to zero; then all low pixel populations in areas of low average population density were also reset to zero (see Section 6.4.6). Threshold values of 1.0 persons/pixel had been used previously for both individual pixel population and average pixel population over a 7×7 pixel window, but other settings were also experimented with at this point. Because of the tendency of all models examined to underestimate urban populations, the average population threshold was lowered to a value commensurate with the urban density limit of 200 persons/sq.km. An average population threshold of .27 persons/30 m pixel, which corresponds to 300 persons/sq.km., was chosen. The rationale of the 50% margin was that the very presence of spurious population requiring readjustment would raise the average density above the background level. In contrast, higher individual pixel thresholds of 1.5 and 2 persons per pixel were also tried in combination with the lower average threshold. This was based on the observation that many pixels along rural roads or associated with farm outbuildings were spuriously assigned populations in the range of 1 to 2 persons.

8.4.2 Results

Initially, the re-estimated regression equation (see Section 8.4.1) was applied to the 1988 Ballarat image, without zero thresholding. Two other models were also fitted: one based on the mean of the fitted values from the re-estimated regression model, and the other based on the regression equation trained on the residential classification training set. These were fitted for purposes of elimination – to confirm the conclusion reached in the simulation study (Chapter 7) that they would produce inferior results to the preferred model.

Table 8.5 shows that the alternative models (1 and 2) did as expected produce far worse results than the preferred model on the various criteria. This was also borne out when these models were applied to other images.

The results for the preferred model (3) are comparable (though not identical because of the changes discussed in the previous section) to the results in Tables 7.2 and 7.3 (Section 6.4). As was the case each time it was applied, zero thresholding (model 4) degraded the performance, particularly in the non-urban areas, but this was reversed when low density thresholding was

---

INREGION function. Instead, ERMapper was used to calculate a mean value per pixel for each vector region, which was then multiplied by the CD area to produce an estimate of total CD population.

incorporated. Both approaches to the choice of low density cutoffs, represented in Table 8.5 by models 5 and 6, resulted in similar performance characteristics.

As has been discussed (Chapter 7), quite reasonable results had been obtained for Geelong even without normalisation for the geographically, temporally and culturally similar Geelong image. As a preliminary, the Ballarat 1988 equation was applied to the other images without normalisation. As expected the results were grossly inaccurate with the total populations of the Ballarat 1994 study area being underestimated by 131% (i.e. a negative total population) and the Adelaide study area overestimated by 509% respectively.

The normalisation methods were then explored, initially on the Ballarat 1994, Geelong, Adelaide and Brisbane images.

Whilst none of the six possible combinations - three normalisation methods and two sources of normalisation statistics - performed clearly best on all criteria for all images, it was judged that the best results on most criteria for most images were produced by the z-score normalisation method based on the means and standard deviations of residential training sets. Multiplicative normalisation based on the means of the residential training sets came a close second. The remaining combinations involving additive normalisation and the use of the residential class statistics with any method generally produced lower estimates which were in general less accurate, and frequently produced negative estimates for whole CDs.

A selection of the results obtained after refining the z-transformed models by thresholding are shown in Table 8.5 and Figure 8.3, and in Images 14, 22 and 24. In a number of the test images, normalisation produced estimates which accord to a moderate degree with most criteria.

Considering Ballarat 1994, the regression intercepts are relatively close to zero and the regression coefficients are quite close to unity. The values of R^2 (.89 overall and .82 in urban CDs) are higher, and the median relative errors for individual CDs are lower, than in the primary image. The total urban population is overestimated by 5-6%. The only clear failing is the substantial overestimation of the overall total and an associated rise in the overall mean relative error, indicating a large proportional overestimation of the small population in the few rural CDs. This again raises the recurring theme of the spurious misclassification of rural pixels as residential. In this instance it appears to be related to some extent to seasonal features such as mature potato crops which were not present on the primary image.

This general pattern of performance indications is repeated, with somewhat lower levels of accuracy, in the cases of Geelong and Adelaide. The R^2 levels are somewhat lower, the mean and median relative errors somewhat higher, and the errors in total urban population are around 8%. In the case of the Adelaide image with its especially large rural component (see Table 8.1), the effect of over-estimation in non-urban areas is most marked.

Because the scale of the population plots for Adelaide, Sydney and Brisbane are dominated by a few extreme outliers in each case, these plots have been repeated with the most extreme outliers deleted. In the case of Sydney, the same has been done with the population density plot.

The performance in the case of Brisbane is similar to that of Adelaide with respect to totals, but less accurate with respect to individual CDs, as evidenced by values of R^2 and mean and median relative errors. This can be explained in terms of the 7 year gap between the image (1989) and the population data (1996). The ground truth total populations for the region and the urban area in 1989 were estimated by geometric interpolation from the 1986 and 1991 census totals, and were some 16% lower than in 1996. In Table 8.5, the remote sensing estimates of totals have been compared with these adjusted figures. No such data was available at the level of individual CDs. Many would not have changed in character or population at all. Most of the change would be concentrated on the urban fringe where development had occurred in the intervening period, and perhaps in the inner suburbs also due to urban renewal with higher density developments. Rather than apply a blanket overall correction factor, no adjustment was made to the individual CD populations. As a result, an average underestimation bias of around 16% would be expected to be largely made up of larger errors in a relatively small number of CDs. There is evidence of this on the Brisbane plots, particularly at low estimated densities. This also helps to explain the larger average relative errors in Brisbane than in Adelaide.

In both the Adelaide and Brisbane images, a large proportion of the over-estimation in total population was associated with particular features in a small number of CDs. Appendix K shows the 25 CDs whose populations were over-estimated by the greatest amount by one of the Adelaide models from Chapter 9 (the pattern of results was consistent though the details differed from model to model). This 1% of CDs contributed 6.6% of the estimated population, which was a substantial proportion of the 15% overestimation by that particular model for the whole region. All of the CDs involved are either rural (generally with water or coastline), intensively agricultural, or large scale industrial/commercial. Many of the types of anomaly listed in Table 8.2 and Appendix K could be identified in advance and removed using a binary masking overlay. This strategy is discussed further in Chapter 10.

Appendix K also lists the 25 CDs whose population densities were most under-estimated. These generally occurred in three areas: the central business districts of Adelaide and North Adelaide, where commercial and residential usage is not clearly visibly delineated and where multi-storeyed structures predominate; and the beachside suburb of Glenelg, where both old-established small allotments and substantial modern multi-storeyed developments occur. As was discussed in Section 2.12.1, these CDs are usually small in area and low in total population. Whilst the relative error of under-estimation may be large, the absolute error in population is not. In this case, the combined error in the 25 CDs was .6% of the regional population.

The temporal mismatch in the data was similar in Kalgoorlie to that of Brisbane, exacerbated by the small total population and the small number of CDs. However, there also seems to be an underlying overestimation bias associated with lower average population density (of which see more below).

Table 8.4 Coefficients of the Normalised Models

A. Normalisation by residential class statistics

		Const	B1	B2	B3	B4	B5	B7
Ballarat 88	Coeffts	2.57752	0.14225	0.20537	-0.22033	-0.02290	-0.08650	0.12747
	Rescaled	2.50873	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	M		25.35589	18.81902	26.05240	60.63159	78.51048	36.57259
	SD		7.39498	4.21969	5.77789	6.93751	13.65267	7.45272
Normalised formulae	Ratio	2.50873	3.51061	3.76171	-5.58693	-1.35141	-6.60991	4.53749
	Difference	0.77030	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	0.77030	1.02386	0.84347	-1.23907	-0.15463	-1.14944	0.92464
Geelong	M		27.96414	18.70358	28.60871	54.70186	83.53683	40.27778
	SD		7.24282	4.41075	6.27442	7.42300	18.16413	9.54733
	Ratio	2.50873	0.12554	0.20112	-0.19529	-0.02470	-0.07913	0.11265
	Difference	2.55020	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	1.41509	0.14136	0.19123	-0.19748	-0.02083	-0.06328	0.09685
Brisbane	M		68.35530	32.52362	38.86600	65.23861	91.69204	36.92705
	SD		8.38918	5.12961	7.97117	9.52699	17.22045	9.31466
	Ratio	2.50873	0.05136	0.11566	-0.14375	-0.02071	-0.07209	0.12288
	Difference	-2.26774	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	-3.36508	0.12205	0.16443	-0.15544	-0.01623	-0.06675	0.09927
Kalgoorlie	M		29.62490	26.20624	52.91602	67.69454	108.09001	59.55173
	SD		7.98476	5.82755	10.61907	7.70549	15.98650	10.03744
	Ratio	2.50873	0.11850	0.14354	-0.10558	-0.01996	-0.06115	0.07619
	Difference	5.99873	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	2.99727	0.12823	0.14474	-0.11668	-0.02007	-0.07190	0.09212
Adelaide	M		74.10966	48.31584	57.71413	76.19752	109.92104	65.80383
	SD		11.29171	6.32893	10.08705	10.14706	24.68122	15.04392
	Ratio	2.50873	0.04737	0.07786	-0.09680	-0.01774	-0.06013	0.06895
	Difference	-3.98286	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	-3.06335	0.09067	0.13327	-0.12284	-0.01524	-0.04657	0.06146
Ballarat 94	M		34.43451	24.91465	40.04149	82.78290	104.52559	50.49513
	SD		10.22058	5.80807	9.19648	12.18498	20.18240	11.14098
	Ratio	2.50873	0.10195	0.15098	-0.13953	-0.01632	-0.06324	0.08986
	Difference	3.98991	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	1.91016	0.10018	0.14522	-0.13473	-0.01269	-0.05695	0.08299
Sydney	M		51.41667	35.93813	47.78788	88.34835	101.48724	55.85794
	SD		12.51859	7.32709	12.33321	13.68857	22.13872	17.02770
	Ratio	2.50873	0.06828	0.10467	-0.11691	-0.01530	-0.06513	0.08123
	Difference	0.29929	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	0.46304	0.08179	0.11512	-0.10047	-0.01130	-0.05192	0.05430

Table 8.4 Coefficients of the Normalised Models (continued)

B. Normalisation by residential classification training set statistics

		Const	B1	B2	B3	B4	B5	B7
Ballarat 88	Coeffts	2.57752	0.14225	0.20537	-0.22033	-0.02290	-0.08650	0.12747
	Rescaled	2.50873	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	M		33.43023	22.49828	29.51254	56.94748	76.95767	39.87676
	SD		7.62383	4.25394	5.96737	7.86248	14.89824	7.78165
Normalised formulae	Ratio	2.50873	4.62853	4.49716	-6.32896	-1.26929	-6.47918	4.94743
	Difference	2.50443	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	2.50443	1.05555	0.85032	-1.27970	-0.17525	-1.25430	0.96545
Geelong	M		33.89857	21.26987	29.82603	53.30037	71.68629	38.28784
	SD		8.38123	4.70813	6.30069	8.03055	12.99933	7.51390
	Ratio	2.50873	0.13654	0.21143	-0.21220	-0.02381	-0.09038	0.12922
	Difference	2.42870	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	3.61210	0.12594	0.18061	-0.20310	-0.02182	-0.09649	0.12849
Brisbane	M		76.04295	36.57842	44.01476	64.85020	91.98062	40.23501
	SD		8.35287	4.95409	7.74877	7.55262	14.32829	7.92737
	Ratio	2.50873	0.06087	0.12295	-0.14379	-0.01957	-0.07044	0.12296
	Difference	-1.69912	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	-1.45775	0.12637	0.17164	-0.16515	-0.02320	-0.08754	0.12179
Kalgoorlie	M		36.68581	29.43614	52.91477	70.81511	99.88356	55.41673
	SD		7.26669	4.47676	7.37771	7.13104	11.87222	7.65670
	Ratio	2.50873	0.12617	0.15278	-0.11961	-0.01792	-0.06487	0.08928
	Difference	6.00102	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	6.06810	0.14526	0.18994	-0.17345	-0.02458	-0.10565	0.12609
Adelaide	M		78.77876	49.85804	57.92176	75.56577	99.41751	63.46181
	SD		12.30764	6.77303	10.45570	9.18836	17.94483	12.98379
	Ratio	2.50873	0.05875	0.09020	-0.10927	-0.01680	-0.06517	0.07796
	Difference	-3.76675	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	0.24925	0.08576	0.12554	-0.12239	-0.01907	-0.06990	0.07436
Ballarat 94	M		40.32633	26.85710	41.12811	75.64645	96.85148	50.24763
	SD		10.31695	5.83369	8.94079	9.84239	16.76737	10.37206
	Ratio	2.50873	0.11478	0.16745	-0.15388	-0.01678	-0.06690	0.09846
	Difference	3.97859	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	4.26541	0.10231	0.14576	-0.14313	-0.01781	-0.07481	0.09308
Sydney	M		58.33136	39.46428	54.15352	85.31479	107.03472	64.71005
	SD		10.90405	6.13890	10.31578	11.11162	18.22029	14.58351
	Ratio	2.50873	0.07935	0.11396	-0.11687	-0.01488	-0.06053	0.07646
	Difference	1.03750	0.13845	0.19989	-0.21445	-0.02229	-0.08419	0.12407
	Z	2.53933	0.09680	0.13851	-0.12405	-0.01577	-0.06884	0.06620

Table 8.5 Summary of Selected Models for Estimating Census Collection District Population Densities Based on Local Classification and Normalised Regression Procedures

Model	Basis/ Type	Thresholds T1	Thresholds T2	Region b0 unforced	Region b1 unforced	Region b1 forced	Region R^2	Region s	Region Mean % error	Region Median % error	Region Total[1] % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	Urban Area b1 unforced	Urban Area b1 forced	Urban Area R^2	Urban Area s	Urban Area Mean % error	Urban Area Median % error	Urban Area Total[1] % error
Ballarat 1988																			
1	RTS			-194.3	1.32	1.21	0.80	475.4	90.3	21.4	50.8	-212.5	1.34	1.21	0.67	527.3	20.5	16.8	-9.3
2	Mean			-348.1	2.91	2.47	0.74	548.8	98.1	56.5	-0.3	-417.8	3.00	2.50	0.56	601.9	53.6	55.1	-52.7
3	RC			-28.0	1.06	1.05	0.82	458.4	29.4	16.7	0.9	-6.3	1.05	1.05	0.69	511.4	18.4	14.7	-2.0
4	RC	0.0		-44.7	1.07	1.04	0.82	459.2	36.3	17.3	11.8	-21.3	1.06	1.05	0.68	512.4	18.5	14.9	-1.3
5	RC	1.0	1.00	-15.4	1.06	1.05	0.82	457.9	27.6	16.7	1.1	15.9	1.04	1.05	0.69	511.0	18.2	14.8	-2.8
6	RC	1.5	0.27	-23.2	1.06	1.05	0.82	459.1	27.6	16.3	-0.4	-2.2	1.05	1.05	0.68	512.4	18.4	14.9	-2.0
Ballarat 1994																			
	RC/	1.0	1.00	30.1	0.94	0.96	0.89	263.5	35.0	13.5	17.4	110.5	0.90	0.96	0.82	279.9	16.9	11.7	4.8
	ZNRTS	1.5	0.27	17.1	0.95	0.96	0.89	263.2	30.8	13.6	16.0	86.5	0.91	0.96	0.82	280.1	17.0	12.6	6.1
Geelong																			
	RC/	1.0	1.00	-40.3	1.10	1.09	0.74	609.5	57.5	18.7	11.1	225.6	1.00	1.09	0.56	634.4	20.6	16.1	-7.9
	ZNRTS	1.5	0.27	-69.7	1.11	1.08	0.77	559.9	60.8	18.4	13.7	188.3	1.01	1.09	0.61	578.8	20.1	16.2	-7.1
Adelaide																			
	RC/	1.0	1.00	90.3	0.90	0.94	0.72	528.6	64.6	19.7	26.8	420.9	0.77	0.94	0.47	554.3	23.5	16.5	7.5
	ZNRTS	1.5	0.27	80.1	0.90	0.94	0.72	529.4	78.9	19.6	28.3	406.3	0.77	0.94	0.47	554.8	23.7	16.3	7.9
Sydney																			
	RC/	1.0	1.00	2568.7	1.03	2.28	0.02	5269.7	141.0	39.5	-26.8	3909.4	0.40	2.28	0.00	5362.2	42.5	38.1	-36.6
	ZNRTS	1.5	0.27	2592.0	1.02	2.27	0.02	5271.9	142.8	39.4	-26.0	3963.1	0.37	2.28	0.00	5363.1	42.4	37.9	-36.2
Brisbane																			
	RC/	1.5	1.00	289.7	0.96	1.09	0.54	891.6	291.5	24.6	27.8	671.1	0.81	1.09	0.36	940.7	28.6	21.1	4.6
	ZNRTS	1.5	0.27	196.6	0.99	1.08	0.54	894.7	345.6	24.7	47.2	591.9	0.83	1.08	0.36	943.5	28.3	20.8	8.7
Kalgoorlie																			
	RC/	1.0	1.00	687.9	0.48	0.85	0.35	563.7	37.1	31.7	11.7	1312.0	0.17	0.86	0.07	482.3	34.9	30.6	12.7
	ZNRTS	1.5	0.27	684.8	0.47	0.83	0.35	565.2	39.8	34.5	15.7	1315.3	0.16	0.83	0.07	483.2	36.4	32.6	16.7

1 The total populations for Brisbane and Kalgoorlie have been compared with the estimated total population at the date of acquisition of the image. For more details see text.

Key: RC Equation trained on a sample of pixels from the residential class of Ballarat 1988
RTS Equation trained on a sample of pixels from the residential class training set of Ballarat 1988
ZNRTS Z-score normalisation using the residential class training set statistics

Figure 8.3 Population Density and Population Estimates for Census Collection Districts Ground Truth vs. Remote Sensing Estimates from Normalised Ballarat 1988 Models

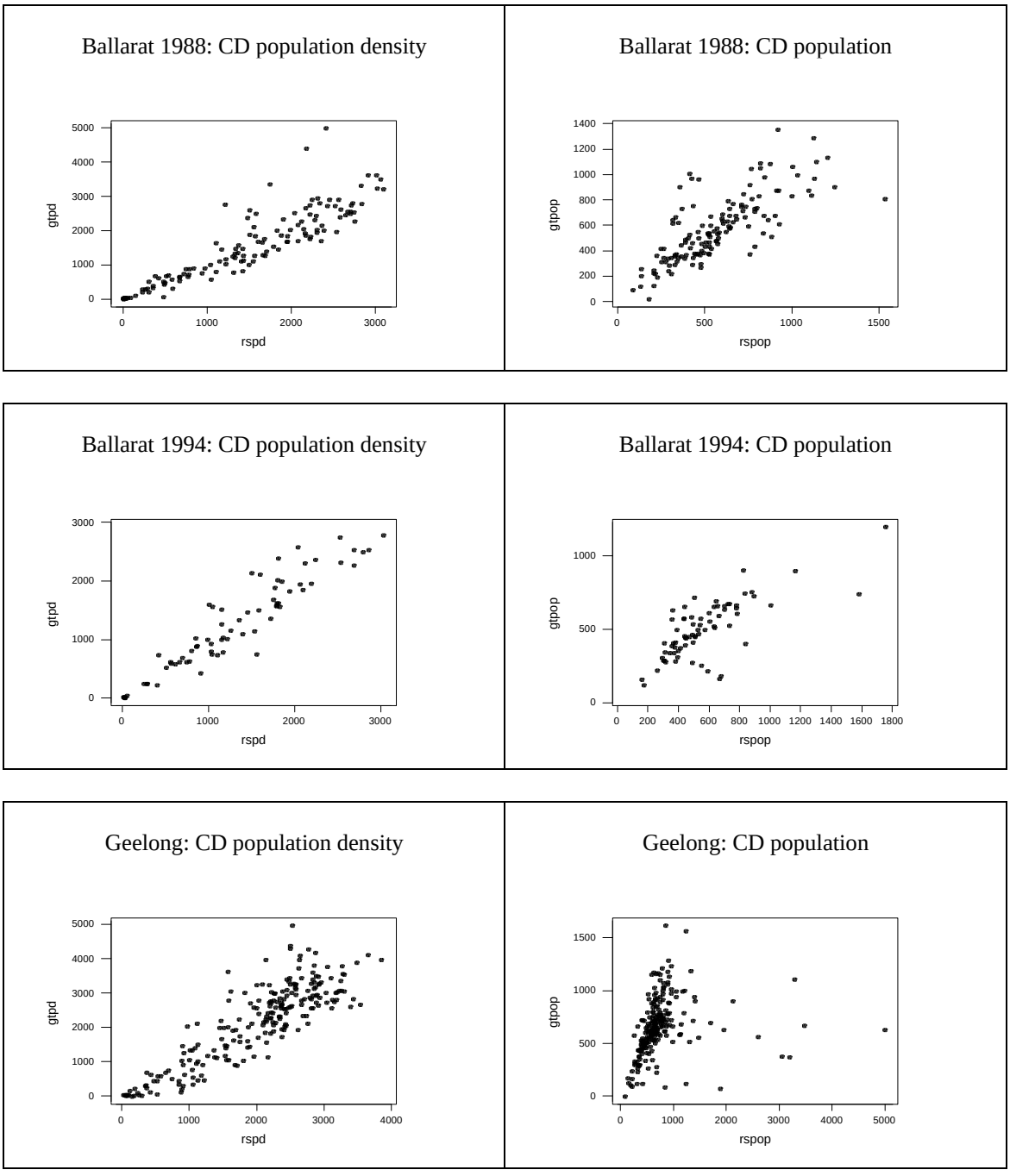
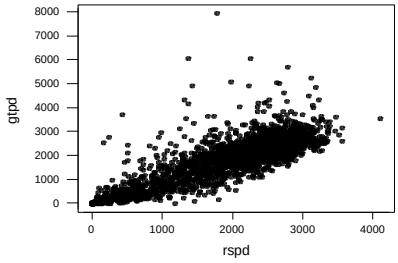
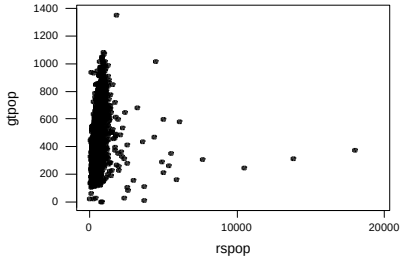


Figure 8.3 Population Density and Population Estimates for Census Collection Districts Ground Truth vs. Remote Sensing Estimates from Normalised Ballarat 1988 Models (continued)

Adelaide: CD population density

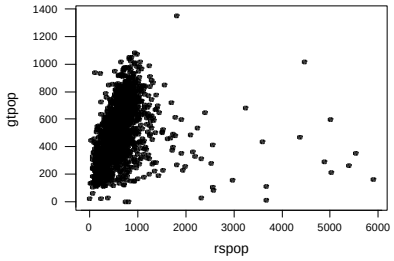


Adelaide: CD population

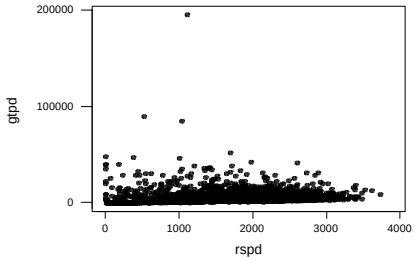


(Intentionally left blank)

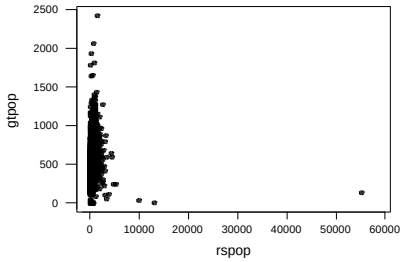
Adelaide: CD population
(5 outliers deleted)



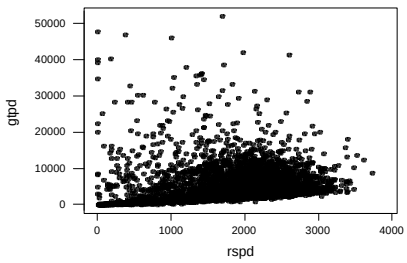
Sydney: CD population density



Sydney: CD population



Sydney: CD population density
(3 outliers deleted)



Sydney: CD population
(3 outliers deleted)

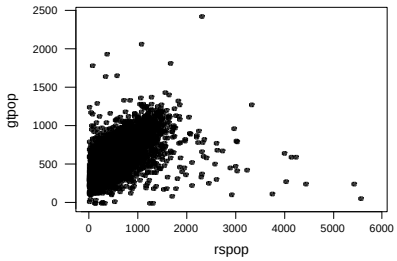
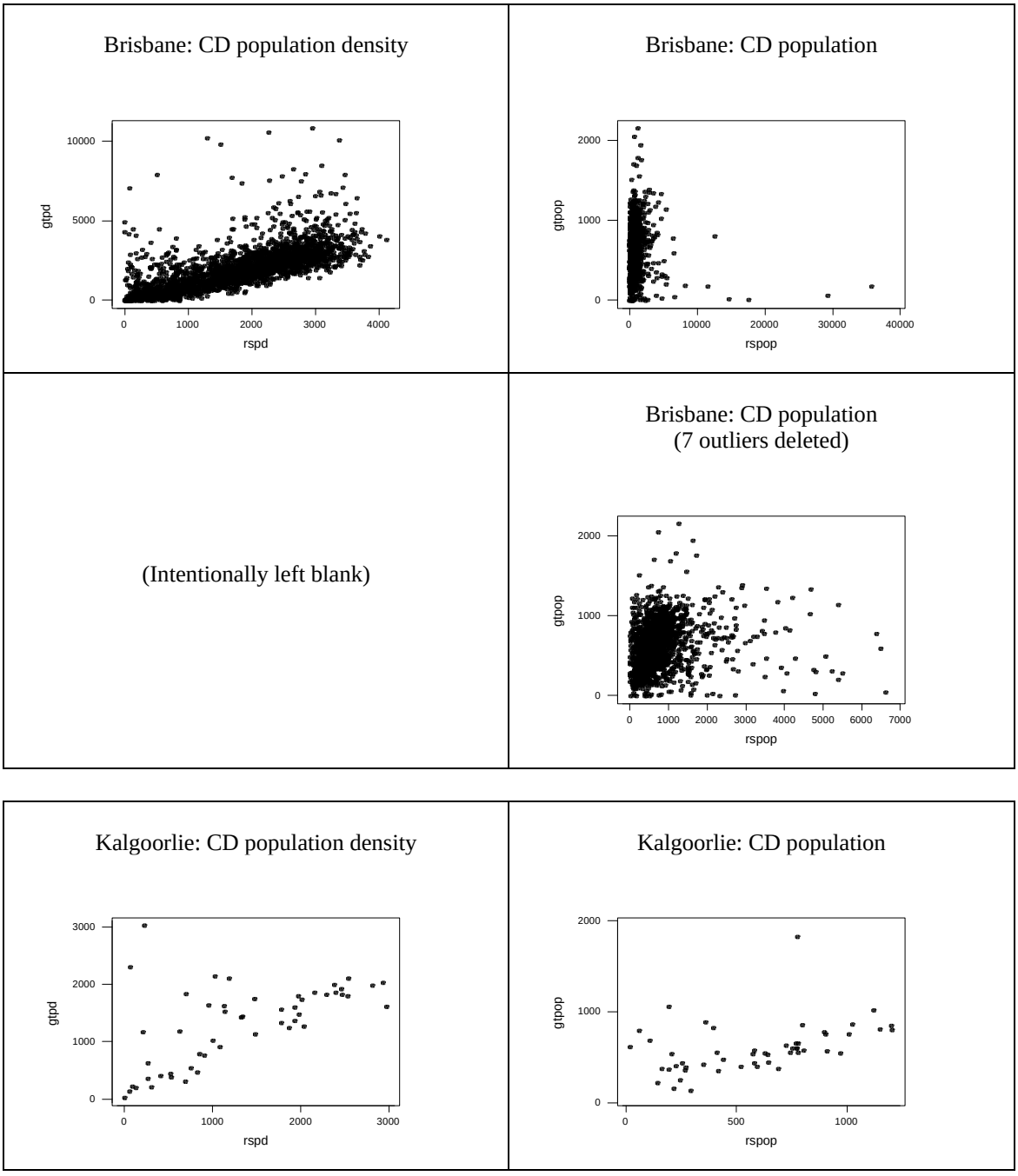


Figure 8.3 Population Density and Population Estimates for Census Collection Districts Ground Truth vs. Remote Sensing Estimates from Normalised Ballarat 1988 Models (continued)



By far the worst results were obtained for Sydney. There was a consistent under-estimation bias which is reflected in most of the indicators. In addition, the values of R^2 were close to zero.

These results can be attributed to a combination of three factors, the third of which is an extreme aspect of the second: average population density; distribution of CD population densities; and the presence of extreme outlying and influential observations.

Changes in population density may or may not be detectable by remote sensing methods, depending on the mechanism of the change. If the population of an area is increased by installing basement flats in existing structures, clearly that change would not be detected. If an increase in population density is brought about by greater crowding in the horizontal plane, then in principle that change might be detectable by remote sensing methods, in terms of higher ratios of constructed surfaces to natural surfaces. Up to a point, multi-level development has elements of both scenarios. The extra population is hidden within the perimeter of the structure, but there are effects on the surroundings such as car parking spaces. But without ancillary physical and/or cultural information, it does not seem possible that the population of residential tower blocks or mixed commercial/residential structures could be accurately estimated by remote sensing alone.

The z-score transformation relates population to z-scores based on the spectral distribution in the training region A, calculates z-scores in the secondary region B, and allocates the appropriate populations. A pixel at the centroid of the spectral distribution in region B will have z-scores of zero, and will be allocated the same population as a pixel at the centroid of the spectral distribution in region A. Similarly, a pixel region B which lies one standard deviation above the centroid in every spectral band will be allocated the same population as a pixel at the same relative position in the spectral distribution in region A. And so on.

This can be brought undone either by a consistent underlying shift in mean population levels (as in the basement case), or by a difference in the shape of the spectral and/or population distributions, so that the population differentials associated with changes in the z-scores are not consistent in the two regions. Both mechanisms would seem to operate in the case of Sydney relative to Ballarat.

Sydney, alone amongst the study areas, has a substantial component of multi-level residential structures. Table 8.1 shows that the average urban population density in Sydney is much higher than in the other areas. But the margin of difference is not alone sufficient to explain the magnitude of the underestimation. Nor, considering the other study areas, is there a consistent relationship between average population density and bias in population estimates.

Secondly, the distributions of all 6 spectral bands are quite symmetrical in the training sets of all 7 test images. But because of geographical concentrations of high density development, the distribution of CD population densities is much more positively skewed in Sydney than in the other areas, suggesting a similar situation at pixel level although these cannot be directly observed. As a result, pixel population estimates allocated on the basis of matching spectral z-scores will tend to underestimate the true populations at both ends of the distribution.

Further to the issue of skew in the Sydney distributions, there are a small number of extreme outliers. The highest CD population density in Sydney is just under 200,000 persons per sq.km., associated with a CD consisting of just one tower block of public housing (not even the surrounding open space is included). Later analysis will show (Chapter 9) that removal of this observation alone results in a substantial increase in R^2.

8.5 SUMMARY

The regression equation for estimating pixel populations derived from the primary Ballarat image was applied to six other test images. Three approaches to normalisation were assessed.

It was concluded that normalisation via z-scores based on the means and standard deviations of the 6 TM bands in the residential class training sets provides a mechanism for training an estimation formula on one image and then applying it to another image. The methodology appears to be moderately robust, particularly so far as urban areas are concerned, to geographical and temporal differences in season and climate, but less robust to differences in average population density or to differences in the shape of the statistical distribution of population densities within an image.

Following the modest degree of success of this enterprise, and considering that it would be difficult to put any error bounds on estimates obtained in this way, it was decided to discontinue the search for a universal “holy grail” of population estimation, and to explore further the issues involved in the less ambitious but more realistic alternative approach of training an estimation equation on a small sample of population data from within an image, and applying the results to the full image.

Chapter 9

Local Training of The Population Estimation Equation

9.1 INTRODUCTION

In this chapter, an approach is described for training an estimation equation on a small sample of population data from within an image, and applying the results to the full image. The aim was to emulate a methodology for estimating a large regional population on the basis of a partial census of relatively small sections of the region, and to evaluate its feasibility.

In Section 9.2 the selection of training samples from four test images is described. In Section 9.3 the estimation methods, the associated refinement of the samples and the results of applying the derived estimation equations to the full image are described and assessed.

In Section 9.4, the many estimation equations derived are examined collectively, and their common characteristics are described and interpreted in terms of the spectral responses of different materials.

9.2 SELECTION OF REGRESSION TRAINING SAMPLES

In this phase of the investigation, it was decided to work with the images of Ballarat, Geelong, Adelaide and Sydney, since each of these had been acquired sufficiently close to the date of the ground truth census data for CD-based training to be feasible.

In each case, samples were taken from pixels classified as residential (see Section 8.2.2). In each case a sample of CDs was selected. In two cases (the original Ballarat 1988 sample and one of the Adelaide samples), a random sample of residential pixels was selected from the selected CDs. In all other cases all residential pixels in the selected CDs were used[1]. For ease

[1] The initial fractional sample had been dictated by software workspace limitations at the time. Full CD samples were considered to be better, both with regard to the theory of the re-estimation algorithm, and

of constructing the rather unwieldy formulae required to sample from the images, the samples of CDs were systematic rather than random. It was reasoned that since CDs are numbered in contiguous blocks, a systematic sample ensured a wide and representative geographic spread, as was subsequently observed to be so. The samples selected are summarised in Table 9.1.

It was considered that for the methodology to be practically useful, the sample should encompass no more than a small fraction, say 10%, of the population to be estimated. It was anticipated that whilst this might produce adequate sample sizes in large cities, it might severely limit the accuracy of estimates obtained for relatively small regional centres.

The first sample from the Ballarat 1988 image was the 20% random sample of all pixels classified as residential in all 138 CDs, which was the basis of the work reported in Chapter 8. Subsequently, two disjoint 10% samples of 13 and 14 CDs were selected. One 20% sample (14 of 72) CDs was selected from the smaller Ballarat 1994 image, and one 5% sample of 11 CDs was selected from the 224 in the Geelong image.

In the case of the more extensive Adelaide image, it was decided to investigate different sized samples. Because sampling from the image was a lengthy and computationally expensive process (2 phases, each involving several hours, regardless of sample size), it was decided to take large samples and then take subsamples from them, rather than selecting separate disjoint smaller samples. A total of 10 samples, three of 5% (around 120 CDs) and seven of 1% (around 25 CDs) were used.

Sample 1 was a 1% sample consisting of a random sample of 20% of residential pixels from 5% of CDs (comparable to sample 1 for Ballarat 1988). Sample 2 (5%) consisted of all the residential pixels from the same 5% of CDs. Samples 3 and 4 were also 5% samples of CDs, disjoint from one another and from samples 1 and 2. From each of samples 2, 3, and 4, two disjoint 1% subsamples (designated 21, 22, 31, 32, 41, 42) were selected.

In the case of Sydney, three samples were used, two of 1% (56 and 57 CDs) and one of 2% (112 CDs).

9.3 APPLICATION OF LOCAL REGRESSION TRAINING

9.3.1 Analysis of the samples

Within each training sample, iterated regression analysis was used to obtain equations for estimating population. After due consideration, 6 iterations had been used in all the empirical work to date based on the Ballarat 1988 image.

---

because this was more likely to correspond to the reality of an operational procedure – having counted the population in particular area, one would want to use all of the information obtained.

Table 9.1 Regression Training Samples

Image Sample	nSLAs	n CDs	n CDs U200	n CDs U500	% of CDs	% of each CD	n pixels	Deletions
BALLARAT 1988	6	138	120	110				
1		138	120	110	100	20	14270	
2		13	12	12	10	100	4312	
3		14	12	12	10	100	7656	
BALLARAT1994	2 (part)	72	65	60				
1		14			20	100	5700	
1r		11					5149	3 CDs with large +ve residuals
GEELONG 1988	8	225	207	194				
1		11	10	10	5	100	9605	
1r1		10					9447	1 CD with large +ve residuals
1r2		9					3300	1 CD with large +ve residuals plus 1 large rural CD
ADELAIDE 1997	47	2412	2151	2001				
1		121	109	97	5	20	13095	
1r		117					12984	4 CDs with large +ve residuals
2		121	109	97	5	100	64797	
2r1		118					64577	3 CDs with large +ve residuals
2r2		117					48647	4 large rural CDs
2r3		114					48427	All of the above
21		25	19	17	1	100	22586	
21r1		24					22471	1 CD with large +ve residuals
21r2		24					14280	1 large rural CD
21r3		23					14165	All of the above
22		24	21	19	1	100	8960	
22r		23					8926	1 CD with large +ve residuals
3		121	108	100	5	100	105715	
3r1		120					73547	1 large rural CD
3r2		118					50059	3 large rural CDs
3r3		115					47283	6 large rural CDs
31		25	23	21	1	100	17848	
31r		24					10490	1 large rural CD
32		24	21	20	1	100	13330	
32r1		22					13033	2 CDs with large +ve residuals
32r2		23					7898	1 large rural CD
32r3		21					7601	All of the above
4		120	110	106	5	100	62823	
4r1		114					62173	5 CDs with large +ve residuals
4r2		118					56255	2 large rural CDs
4r3		112					55605	All of the above

Table 9.1 Regression Training Samples
(continued)

Image Sample	nSLAs	n CDs	n CDs U200	n CDs U500	% of CDs	% of each CD	n pixels	Deletions
ADELAIDE 1997	**(cont.)**							
41		24	21	21	1	100	13074	
41r		23					13019	1 CD with large +ve residuals
42		24	21	21	1	100	10305	
42r		19					9377	3/2 CDs with large +ve/-ve residuals
SYDNEY 1996	41	5628	5451	5323				
1		57	53	50	1	100	24116	
1r		51					23904	6 CDs with large +ve residuals
2		112	107	105	2	100	36277	
2r1		106					36185	6 CDs with large +ve residuals
2r2		98					35894	14 CDs with large +ve residuals
3		56	55	51	1	100	17995	
3r1		42					17080	14 CDs with large +ve residuals
3r2		49					17664	7 CDs with large +ve residuals

However, because the results of simulations had been somewhat ambivalent on this point (Section 7.4.4), and because it was thought that rates of convergence might differ for samples from different images, it was decided at this point to derive two equations in each case, one based on 6 iterations and the other on convergence of R^2 to within 0.001. The number of iterations in the latter case ranged from 5 to 35.

Most of the chosen samples of CDs included examples of the two types of extreme discussed in Section 2.12.1 – viz. small urban CDs with very high population densities and very large rural CDs with low population densities. It was considered that in small samples there was a risk that either of these types of CD might be over-represented which could bias the estimation equation to a significant degree. Accordingly, samples in which such cases occurred were re-analysed after the deletion of each type of extreme case in turn. High density outliers were detected by examining residual plots of the iterated regressions (although these were based on pixels, the parallel striations associated with outlying CDs were clearly discernable). Large low density rural CDs were identified by mapping the sample of CDs. In each case, the average population density of deleted CDs was less than 10 persons/sq.km.

In effect, this selective deletion strategy leads to a middle ground between sampling from the residential training set, which might be too narrow in definition, and the full residential class, which might be too broadly defined, particularly in the rural areas.

Using the methodology previously described (Sections 5.6.2, 5.6.3 and 8.4.1), the various population estimation equations were applied in turn to the various full images, estimated pixel populations were aggregated to produce CD estimates, and regression analyses were used to compare the remote sensing estimates with ground truth CD populations.

In each case, a corresponding regression analysis was also carried out using only the CDs in the training sample. These analyses provided internal validation measures against which the external validation measures, based on all CDs in the image, could be compared.

Both sets of results are listed in Table 9.2. A representative selection of plots, one set for each of the five images, is shown in Figure 9.1. Images 16, 18 and 20 are also representative of this methodology.

9.3.2 Results

It was generally (though not invariably) found that equations based on the convergence criterion produced better results for the urban areas with regard to all the criteria previously described (Section 5.8) than those based on 6 iterations. This was achieved at the cost of an increased level of overestimation in the low density rural areas.

This effect had been anticipated in observations made previously regarding some of the simulation results (Section 7.4.4): “In samples from the residential training set, the results were still improving after the tenth iteration, suggesting that more iterations might be of benefit for estimating the population of more typical suburban pixels, but that this might be achieved at the expense of loss of accuracy with respect to the more atypical pixels found in the broader distribution of the residential class”.

Table 9.2 shows only the results of the equations based on the convergence criterion.

In Chapter 8, the effects of using various different combinations of low density cutoff settings was reported. At this stage, it was decided to standardise for purposes of comparison on settings of 0.27 persons/pixel for smoothed pixel population and 1.5 persons/pixel for individual pixel population respectively (although for some models the results of an individual pixel setting of 2.0 are also reported).

Considering firstly the “whole image” results, taken in the broad, the results of local regression training shown in Table 9.2 are better on most criteria for most of the five images than the corresponding results for the normalised Ballarat 1988 models shown in Table 8.5 In general R^2 values are as high or higher, most slope coefficients are closer to unity, most mean and median relative errors are lower, and most totals for overall regions and urban areas are more accurate.

Table 9.2 Summary of Selected Models for Estimating Census Collection District Population Densities Based on Local Training of Both Classification and Regression Procedures

A. Results for the whole image

Image	Sample	Sampling fraction % of CDs	Thresholds[1] T1	T2	Region b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error
Ballarat 1988																				
	1	20[2]	1.5	0.27	-23.2	1.06	1.05	0.82	459.1	27.6	16.3	-0.4	-2.2	1.05	1.05	0.68	512.4	18.4	14.9	-2.0
			2.0	0.27	-18.6	1.05	1.05	0.82	459.1	27.1	16.7	-2.7	2.4	1.05	1.05	0.68	512.4	18.5	14.9	-2.2
	2	10	1.5	0.27	-36.2	1.12	1.10	0.82	457.8	33.4	15.3	3.7	-11.4	1.11	1.11	0.69	511.7	17.8	13.3	-7.2
			2.0	0.27	-27.0	1.12	1.11	0.82	457.6	30.3	15.4	-0.9	-2.7	1.11	1.11	0.69	511.6	17.9	13.4	-7.6
	3	10	1.5	0.27	-35.9	1.04	1.02	0.81	463.6	30.1	17.3	4.7	-21.3	1.04	1.03	0.68	517.1	19.6	15.8	1.0
			2.0	0.27	-31.5	1.04	1.02	0.81	463.6	29.3	17.0	2.2	-18.0	1.03	1.03	0.68	517.1	19.6	15.8	0.8
Ballarat 1994																				
	1	20	1.5	0.27	18.3	0.96	0.97	0.88	274.8	22.2	12.2	9.6	88.8	0.93	0.98	0.81	288.6	16.7	9.7	4.4
			2.0	0.27	23.5	0.96	0.98	0.88	274.9	22.1	12.2	7.5	94.1	0.93	0.98	0.81	289.1	16.8	9.7	4.2
	1r		1.5	0.27	-1.1	1.14	1.14	0.86	298.1	19.9	13.2	-5.2	52.3	1.11	1.14	0.77	319.7	18.2	12.7	-8.4
			2.0	0.27	3.0	1.14	1.14	0.86	298.2	20.6	13.2	-6.6	56.3	1.10	1.14	0.77	319.8	18.2	12.7	-8.6
Geelong																				
	1	5	1.5	0.27	-71.6	0.97	0.94	0.74	594.3	79.7	20.8	37.6	253.5	0.86	0.95	0.55	616.6	23.6	17.4	7.1
			2.0	0.27	-62.0	0.97	0.94	0.74	593.7	76.0	20.8	35.4	258.0	0.86	0.95	0.55	616.2	23.5	16.9	6.9
	1r1		1.5	0.27	-122.8	1.07	1.02	0.75	583.3	78.9	18.3	30.8	179.7	0.96	1.03	0.57	606.2	20.6	15.3	-0.7
			2.0	0.27	-112.4	1.07	1.02	0.75	582.5	75.4	18.4	28.6	184.2	0.96	1.03	0.57	605.7	20.6	15.2	-0.8
	1r2		1.5	0.27	-324.0	1.13	1.00	0.78	539.4	96.4	17.7	42.7	-156.0	1.07	1.01	0.63	559.9	21.5	14.3	3.2
			2.0	0.27	-320.9	1.12	1.00	0.78	539.4	95.2	17.7	42.0	-154.7	1.07	1.01	0.63	559.8	21.5	14.3	3.2
Adelaide																				
	1	1[3]	1.5	0.27	149.6	0.94	1.01	0.70	544.4	48.5	18.1	13.8	535.9	0.77	1.01	0.44	569.3	20.7	15.2	-0.4
			2.0	0.27	155.7	0.93	1.01	0.70	544.0	45.4	18.0	11.3	539.8	0.77	1.01	0.44	569.0	20.7	15.3	-0.5
	1r		1.5	0.27	162.8	0.95	1.02	0.70	548.9	47.3	18.3	11.2	562.0	0.77	1.03	0.44	572.7	20.6	15.2	-2.6
			2.0	0.27	169.2	0.95	1.03	0.70	548.5	44.2	18.1	8.6	566.0	0.77	1.03	0.44	572.4	20.6	15.1	-2.7
	2	5	1.5	0.27	149.6	0.94	1.00	0.70	543.3	48.5	18.0	13.6	532.3	0.77	1.01	0.44	568.7	20.7	15.3	-0.2
	2r1		1.5	0.27	147.7	0.95	1.02	0.70	543.6	48.6	17.7	13.0	533.5	0.78	1.02	0.44	568.8	20.2	15.2	-1.9
	2r2		1.5	0.27	113.4	0.95	1.01	0.71	535.5	49.9	17.2	14.9	461.2	0.80	1.01	0.45	564.4	20.1	14.7	0.6
	2r3		1.5	0.27	111.1	0.97	1.03	0.71	536.9	50.5	16.8	14.8	465.9	0.81	1.03	0.45	565.6	19.6	14.1	-1.2

1. T1= individual pixel threshold T2 = Average threshold 2. 20% sample of pixels from all CDs 3. 20% sample of 5% of CDs

Table 9.2 Summary of Selected Models for Estimating Census Collection District Population Densities Based on Local Training of Both Classification and Regression Procedures

A. Results for the whole image (continued)

Image	Sample	Sampling fraction %	Thresholds[1] T1	T2	Region b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error
Adelaide (cont.)																				
	21	1	1.5	0.27	183.2	0.92	1.01	0.69	555.1	50.5	18.8	14.9	605.4	0.74	1.01	0.43	576.5	21.2	16.1	-1.2
	21r1		1.5	0.27	184.4	0.95	1.04	0.69	559.9	50.6	18.3	12.8	617.0	0.76	1.04	0.42	580.9	20.7	15.8	-4.2
	21r2		1.5	0.27	136.0	0.94	1.00	0.70	551.0	57.4	18.6	21.4	533.8	0.76	1.00	0.43	577.6	20.8	15.5	1.4
	21r3		1.5	0.27	138.7	0.96	1.03	0.69	560.8	58.9	18.6	14.8	557.5	0.77	1.03	0.41	586.7	20.4	15.1	-1.4
	22	1	1.5	0.27	177.7	0.84	0.91	0.67	571.5	67.2	23.0	25.5	608.8	0.67	0.92	0.40	592.5	26.6	19.4	9.7
	22r		1.5	0.27	172.7	0.88	0.95	0.67	572.3	63.2	20.4	21.9	611.3	0.70	0.96	0.40	592.5	24.4	17.6	5.3
	3	5	1.5	0.27	95.0	1.03	1.08	0.71	537.4	47.9	17.3	7.2	439.1	0.87	1.08	0.45	565.3	20.0	14.7	-6.1
			2.0	0.27	100.8	1.03	1.08	0.71	537.2	45.3	17.2	4.7	442.7	0.87	1.08	0.45	565.2	20.0	14.7	-6.2
	3r1		1.5	0.27	79.4	1.04	1.08	0.72	529.8	48.4	16.7	7.9	402.4	0.89	1.08	0.46	559.1	19.6	14.3	-6.0
	3r2		1.5	0.27	88.8	1.03	1.07	0.72	532.4	51.1	16.9	10.8	428.9	0.87	1.08	0.46	560.9	19.6	14.3	-5.5
	3r3		1.5	0.27	37.1	1.03	1.05	0.75	502.1	47.7	15.0	13.1	289.4	0.91	1.05	0.51	534.5	18.0	12.6	-2.3
	31	1	1.5	0.27	54.1	1.07	1.10	0.72	530.0	49.0	17.0	6.5	351.0	0.93	1.11	0.46	562.0	19.6	14.3	-7.0
	31r		1.5	0.27	21.3	1.06	1.07	0.73	518.7	54.5	16.1	14.5	280.4	0.94	1.07	0.47	552.9	18.7	13.1	-3.4
	32	1	1.5	0.27	0.3	1.03	1.03	0.75	500.2	59.5	15.7	20.6	235.8	0.93	1.04	0.52	530.1	18.9	12.8	-0.8
			2.0	0.27	7.4	1.03	1.03	0.75	500.2	56.3	15.5	17.7	241.1	0.93	1.04	0.52	530.1	18.9	12.8	-0.9
	32r1		1.5	0.27	22.1	1.05	1.06	0.74	509.5	53.1	16.0	12.8	275.0	0.93	1.06	0.50	539.8	19.1	13.4	-3.3
			2.0	0.27	28.5	1.04	1.06	0.74	509.5	51.4	16.3	10.5	279.8	0.93	1.06	0.50	539.8	19.1	13.4	-3.4
	32r2		1.5	0.27	113.4	0.95	1.01	0.71	535.5	49.9	17.2	14.9	461.2	0.80	1.01	0.45	564.4	20.1	14.7	0.6
	32r3		1.5	0.27	10.5	1.05	1.05	0.74	507.3	61.6	15.8	20.5	270.5	0.93	1.06	0.50	536.9	19.0	13.2	-2.6
	4	5	1.5	0.27	48.9	1.01	1.03	0.74	513.0	49.1	15.6	13.4	318.8	0.89	1.03	0.49	545.3	18.8	13.3	-0.8
	4r1		1.5	0.27	66.0	1.02	1.05	0.73	519.9	46.9	16.1	9.6	357.0	0.89	1.06	0.48	551.3	19.0	13.6	-3.2
	4r2		1.5	0.27	27.6	1.01	1.03	0.74	505.0	51.4	15.5	16.3	276.5	0.90	1.03	0.50	537.8	18.5	12.9	-0.2
	4r3		1.5	0.27	46.7	1.03	1.05	0.74	512.0	47.7	15.6	11.1	315.4	0.91	1.05	0.49	544.2	18.6	13.1	-2.8

1. T1= individual pixel threshold T2 = Average threshold 2. 20% sample of pixels from all CDs 3. 20% sample of 5% of CDs

Table 9.2 Summary of Selected Models for Estimating Census Collection District Population Densities Based on Local Training of Both Classification and Regression Procedures

A. Results for the whole image (continued)

Image	Sample	Sampling fraction %	Thresholds[1] T1	T2	Region b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error
Adelaide (cont.)																				
	41	1	1.5	0.27	48.9	1.01	1.03	0.74	513.0	49.1	15.6	13.4	318.8	0.89	1.03	0.49	545.3	18.8	13.3	-0.8
	41r		1.5	0.27	117.8	0.97	1.02	0.68	566.7	59.8	19.4	18.2	525.0	0.79	1.03	0.39	594.6	21.6	16.2	-0.3
	42	1	1.5	0.27	14.3	0.93	0.94	0.74	507.7	65.7	19.4	31.2	248.7	0.84	0.94	0.50	540.9	22.4	16.3	10.2
	42r		1.5	0.27	55.7	0.97	1.00	0.75	504.2	49.5	16.4	14.8	312.0	0.86	1.00	0.51	535.2	19.7	13.7	2.2
Sydney	1	1	1.5	0.27	2185.6	0.73	1.28	0.05	5182.9	140.0	29.1	6.1	2815.5	0.57	1.29	0.03	5294.5	34.6	27.7	-2.6
			2.0	0.27	2186.0	0.73	1.28	0.05	5182.4	138.3	28.8	5.5	2814.5	0.57	1.29	0.03	5294.2	34.6	27.7	-2.7
	1r		1.5	0.27	2489.5	0.68	1.35	0.04	5214.1	119.4	28.8	-1.9	3175.4	0.50	1.35	0.02	5318.6	34.5	28.1	-9.2
			2.0	0.27	2489.1	0.68	1.35	0.04	5213.7	117.5	28.8	-2.5	3173.8	0.50	1.36	0.02	5318.4	34.5	28.2	-9.3
	2	2	1.5	0.27	1541.7	0.96	1.39	0.06	5150.8	161.0	26.5	7.1	2175.2	0.79	1.39	0.03	5274.1	32.4	25.1	-2.2
	2r1		1.5	0.27	1713.7	0.95	1.44	0.06	5168.5	153.1	25.0	2.7	2392.8	0.76	1.44	0.03	5288.6	31.5	23.9	-6.2
	2r2		1.5	0.27	1836.9	0.95	1.52	0.05	5186.1	144.6	23.8	-2.3	2578.6	0.73	1.52	0.02	5303.6	30.9	22.7	-10.6
	3	1	1.5	0.27	1727.1	0.80	1.20	0.07	5133.9	178.1	33.2	23.0	2279.4	0.67	1.20	0.04	5252.7	37.7	31.2	8.1
			2.0	0.27	1729.5	0.80	1.20	0.07	5133.5	176.8	33.2	22.4	2280.1	0.67	1.20	0.04	5252.4	37.7	31.3	8.0
	3r1		1.5	0.27	2074.2	0.85	1.44	0.05	5178.4	134.6	27.8	-1.7	2711.5	0.67	1.44	0.03	5291.3	33.6	26.6	-10.5
			2.0	0.27	2075.1	0.85	1.44	0.05	5177.9	133.0	27.8	-2.2	2710.4	0.67	1.44	0.03	5290.9	33.6	26.6	-10.7
	3r2		1.5	0.27	1962.2	0.81	1.32	0.06	5161.1	153.3	30.0	10.0	2557.6	0.66	1.32	0.03	5275.7	34.8	28.3	-2.4
			2.0	0.27	1964.0	0.81	1.32	0.06	5160.6	151.9	30.1	9.4	2557.4	0.66	1.32	0.03	5275.3	34.8	28.4	-2.6

1. T1= individual pixel threshold T2 = Average threshold 2. 20% sample of pixels from all CDs 3. 20% sample of 5% of CDs

Table 9.2 Summary of Selected Models for Estimating Census Collection District Population Densities Based on Local Training of Both Classification and Regression Procedures

(continued)

B. Results for the regression training sample

Image	Sample	Sampling fraction %	Thresholds[1] T1	T2	Region b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error
Ballarat 1988																				
	1	20[2]	1.5	0.27	-23.2	1.06	1.05	0.82	459.1	27.6	16.3	-0.4	-2.2	1.05	1.05	0.68	512.4	18.4	14.9	-2.0
			2.0	0.27	-18.6	1.05	1.05	0.82	459.1	27.1	16.7	-2.7	2.4	1.05	1.05	0.68	512.4	18.5	14.9	-2.2
	2	10	1.5	0.27	-56.5	1.04	1.01	0.95	202.0	12.5	11.4	3.4	-66.9	1.04	1.01	0.94	211.7	11.5	11.4	2.8
			2.0	0.27	-45.4	1.03	1.01	0.95	202.4	11.7	11.6	2.5	-55.6	1.04	1.01	0.94	212.0	11.4	11.6	2.1
	3	10	1.5	0.27	-0.2	1.02	1.02	0.92	316.7	15.9	10.5	-5.3	-5.9	1.02	1.02	0.87	346.9	12.5	6.2	-1.1
			2.0	0.27	3.0	1.02	1.02	0.92	316.8	15.5	7.1	-7.8	-2.8	1.02	1.02	0.87	347.0	12.6	6.2	-1.2
Ballarat 1994																				
	1	20	1.5	0.27	-101.7	1.03	0.97	0.86	260.3	29.8	12.8	23.6	-20.9	1.00	0.99	0.77	247.7	11.0	9.5	5.6
			2.0	0.27	-98.6	1.03	0.97	0.86	259.1	26.8	12.7	22.3	-18.4	1.00	0.99	0.77	247.7	11.0	9.5	5.6
	1r		1.5	0.27	-141.1	1.23	1.12	0.84	281.4	17.6	10.1	5.1	-171.8	1.27	1.14	0.68	292.6	12.2	9.5	-6.2
			2.0	0.27	-136.9	1.23	1.12	0.84	281.2	16.6	9.0	4.1	-168.1	1.27	1.14	0.68	291.9	12.2	9.5	-6.3
Geelong																				
	1	5	1.5	0.27	-5.4	0.90	0.89	0.91	382.6	68.8	13.1	63.0	95.4	0.86	0.90	0.87	398.9	14.2	11.7	11.9
			2.0	0.27	1.7	0.89	0.89	0.91	382.3	65.9	13.1	60.1	101.2	0.86	0.90	0.87	398.8	14.0	11.7	11.5
	1r1		1.5	0.27	4.0	0.96	0.96	0.90	392.2	63.2	8.2	53.7	112.6	0.92	0.96	0.86	408.0	11.2	7.4	4.7
			2.0	0.27	11.1	0.96	0.96	0.90	391.8	60.7	8.2	50.8	118.0	0.92	0.96	0.86	407.8	11.4	7.4	4.3
	1r2		1.5	0.27	-235.6	1.11	1.01	0.95	272.8	105.9	9.1	92.1	-142.4	1.07	1.02	0.94	281.7	9.4	8.8	1.5
			2.0	0.27	-230.8	1.10	1.01	0.95	273.5	105.5	9.7	91.4	-136.6	1.07	1.02	0.94	282.3	9.5	9.4	1.3
Adelaide																				
	1	5[2]	1.5	0.27	73.9	0.99	1.03	0.79	461.6	48.7	16.8	15.9	464.5	0.83	1.04	0.56	474.3	18.2	14.4	-3.3
			2.0	0.27	80.5	0.99	1.03	0.79	461.1	45.6	16.1	13.5	468.2	0.82	1.04	0.56	473.8	18.2	14.4	-3.4
	1r		1.5	0.27	81.1	1.01	1.05	0.79	463.0	46.8	16.4	12.7	477.2	0.84	1.06	0.56	475.4	18.6	14.8	-5.4
			2.0	0.27	87.9	1.01	1.05	0.79	462.6	43.8	16.1	10.4	481.3	0.84	1.06	0.56	474.9	18.6	14.8	-5.5
	2	5	1.5	0.27	74.3	0.99	1.03	0.79	458.4	48.7	16.2	15.9	457.8	0.83	1.03	0.57	471.2	18.4	14.3	-3.0
	2r1		1.5	0.27	70.7	1.01	1.04	0.79	459.7	48.4	15.9	14.8	457.2	0.84	1.05	0.56	473.6	18.2	13.8	-4.5
	2r2		1.5	0.27	56.0	1.00	1.02	0.79	455.6	53.0	15.8	18.6	420.1	0.84	1.03	0.57	471.0	17.9	14.9	-2.1
	2r3		1.5	0.27	50.9	1.02	1.04	0.79	458.2	52.8	16.0	17.7	420.3	0.86	1.05	0.56	475.2	17.8	14.3	-3.6

1. T1= individual pixel threshold T2 = Average threshold 2. 20% sample of pixels from all CDs 3. 20% sample of 5% of CDs

Table 9.2 Summary of Selected Models for Estimating Census Collection District Population Densities Based on Local Training of Both Classification and Regression Procedures

B. Results for the regression training sample (continued)

Image	Sample	Sampling fraction %	Thresholds[1] T1	T2	Region b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error
Adelaide (cont.)																				
	21	1	1.5	0.27	-73.5	1.14	1.11	0.79	558.7	115.1	19.8	42.2	322.3	0.98	1.13	0.48	628.1	21.8	15.9	-10.4
	21r1		1.5	0.27	-79.4	1.19	1.15	0.79	567.3	115.0	22.6	39.0	321.3	1.02	1.18	0.46	640.7	22.4	18.2	-13.4
	21r2		1.5	0.27	-105.1	1.15	1.10	0.79	560.5	148.6	23.7	57.5	247.4	1.00	1.12	0.45	645.8	19.8	14.1	-7.7
	21r3		1.5	0.27	-106.1	1.20	1.14	0.78	579.2	153.7	26.5	56.5	285.5	1.02	1.16	0.41	670.6	20.3	15.5	-10.5
	22	1	1.5	0.27	323.0	0.82	0.95	0.80	449.3	20.0	12.4	5.9	1501.6	0.37	0.95	0.30	375.3	16.2	12.3	1.7
	22r		1.5	0.27	304.3	0.84	0.96	0.81	440.6	21.0	15.7	6.6	1445.7	0.40	0.96	0.36	358.8	15.6	14.9	0.1
	3	5	1.5	0.27	91.4	1.03	1.08	0.76	474.0	49.1	17.2	14.3	537.3	0.82	1.08	0.46	491.6	19.9	15.3	-4.7
			2.0	0.27	96.1	1.03	1.08	0.76	473.8	45.4	17.5	11.5	539.8	0.82	1.08	0.46	491.6	19.9	15.3	-4.8
	3r1		1.5	0.27	64.6	1.05	1.08	0.78	456.7	53.1	16.6	17.3	459.7	0.86	1.08	0.49	477.5	19.4	14.9	-4.4
	3r2		1.5	0.27	67.5	1.04	1.07	0.78	452.7	61.5	16.0	22.9	461.0	0.85	1.07	0.50	473.2	19.0	13.5	-3.7
	3r3		1.5	0.27	19.9	1.03	1.04	0.82	404.2	52.4	12.4	22.2	301.6	0.90	1.04	0.59	429.0	17.5	10.3	0.1
	31	1	1.5	0.27	-119.0	1.14	1.08	0.84	374.6	75.2	21.7	21.1	-35.5	1.11	1.09	0.72	398.7	19.5	17.9	-1.6
	31r		1.5	0.27	-162.0	1.14	1.06	0.87	345.3	92.9	17.5	29.6	-170.3	1.15	1.06	0.75	371.9	19.1	16.1	1.8
	32	1	1.5	0.27	51.0	1.01	1.03	0.91	311.1	14.5	8.7	4.1	372.6	0.88	1.03	0.60	336.6	8.7	7.4	-1.0
			2.0	0.27	59.3	1.01	1.03	0.91	310.6	11.0	7.8	-0.1	384.4	0.87	1.03	0.60	335.7	8.7	7.4	-1.1
	32r1		1.5	0.27	127.0	1.01	1.06	0.84	406.2	13.8	8.2	-2.8	812.6	0.71	1.06	0.37	423.9	12.7	7.4	-3.9
			2.0	0.27	132.6	1.00	1.06	0.84	405.5	15.7	8.2	-6.1	811.5	0.71	1.06	0.37	423.5	12.6	7.4	-3.9
	32r2		1.5	0.27	42.7	1.01	1.02	0.92	285.6	23.7	9.1	12.3	318.5	0.89	1.02	0.66	309.4	8.6	8.2	-0.8
	32r3		1.5	0.27	69.6	1.02	1.05	0.90	320.2	17.8	8.3	4.0	455.5	0.86	1.05	0.58	343.1	9.5	5.6	-3.2
	4	5	1.5	0.27	53.8	1.00	1.03	0.82	398.6	29.7	15.9	7.2	174.1	0.95	1.03	0.73	415.3	21.0	14.8	-0.9
	4r1		1.5	0.27	75.7	1.01	1.05	0.81	412.2	28.9	17.1	4.0	211.8	0.95	1.05	0.71	427.7	21.3	15.6	-3.2
	4r2		1.5	0.27	23.8	1.01	1.03	0.83	382.1	31.9	16.2	10.2	125.9	0.97	1.03	0.75	399.5	20.4	14.9	-0.2
	4r3		1.5	0.27	50.7	1.02	1.05	0.82	398.4	29.4	16.5	5.5	170.8	0.97	1.05	0.73	414.9	20.8	15.1	-2.7

1. T1= individual pixel threshold T2 = Average threshold 2. 20% sample of pixels from all CDs 3. 20% sample of 5% of CDs

Table 9.2 Summary of Selected Models for Estimating Census Collection District Population Densities Based on Local Training of Both Classification and Regression Procedures

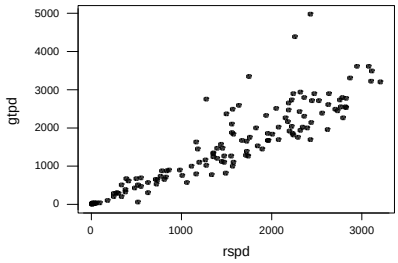
B. Results for the regression training sample (continued)

Image	Sample	Sampling fraction %	Thresholds[1] T1	T2	Region b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error
Adelaide (cont.)																				
	41	1	1.5	0.27	47.6	1.00	1.02	0.84	423.5	30.1	20.7	6.9	143.8	0.96	1.03	0.77	450.4	22.1	17.9	-4.0
	41r		1.5	0.27	11.8	1.02	1.03	0.85	413.5	25.5	18.7	5.5	76.3	1.00	1.03	0.78	442.6	21.2	18.0	-1.8
	42	1	1.5	0.27	51.0	0.97	0.99	0.86	377.0	27.0	12.9	13.8	178.0	0.91	0.99	0.71	402.8	17.5	10.7	1.0
	42r		1.5	0.27	136.7	0.97	1.03	0.80	453.9	23.9	12.3	3.0	400.3	0.86	1.03	0.60	477.0	20.6	11.8	-3.4
1 CD deleted[4]	*31r*		*1.5*	*0.27*	*-198.9*	*1.16*	*1.06*	*0.84*	*351.8*	*21.9*	*16.8*	*6.2*	*-170.3*	*1.15*	*1.06*	*0.75*	*371.9*	*19.1*	*16.1*	*1.8*
Sydney																				
	1	1	1.5	0.27	527.8	1.12	1.25	0.44	2256.1	57.3	31.7	24.3	1204.2	0.97	1.25	0.30	2370.4	32.0	29.1	-6.2
	1r		1.5	0.27	813.0	1.10	1.31	0.39	2366.2	53.0	30.9	14.3	1629.6	0.91	1.32	0.23	2470.6	32.0	28.0	-11.5
	2	2	1.5	0.27	666.4	1.06	1.24	0.15	3639.7	43.7	22.3	12.3	1049.4	0.95	1.24	0.10	3748.7	29.2	21.8	0.7
	2r1		1.5	0.27	740.7	1.07	1.29	0.14	3651.9	41.7	22.4	7.8	1144.9	0.96	1.29	0.09	3760.0	28.3	22.0	-3.3
	2r2		1.5	0.27	814.7	1.10	1.35	0.13	3675.0	39.9	21.5	2.7	1258.7	0.97	1.36	0.08	3782.1	27.5	18.6	-7.8
	3	1	1.5	0.27	604.9	1.18	1.31	0.15	5347.7	53.9	36.8	25.8	1205.4	1.05	1.32	0.10	5591.0	37.0	33.3	8.2
	3r1		1.5	0.27	986.1	1.31	1.58	0.12	5435.4	43.4	26.3	1.8	1723.1	1.13	1.59	0.08	5673.2	32.4	21.6	-11.8
	3r2		1.5	0.27	840.2	1.24	1.45	0.13	5398.2	47.3	31.4	12.7	1517.6	1.08	1.45	0.09	5638.1	33.2	29.4	-3.0
1 CD deleted[4]	*3r2*		*1.5*	*0.27*	*-792.0*	*1.55*	*1.35*	*0.57*	*2306.6*	*46.4*	*31.3*	*12.6*	*-729.7*	*1.54*	*1.36*	*0.51*	*2422.6*	*31.9*	*29.4*	*-3.0*

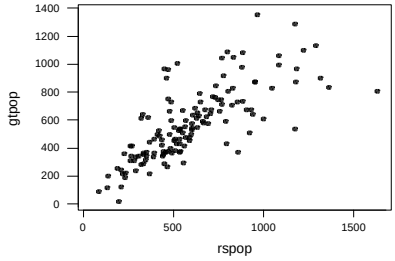
1. T1= individual pixel threshold T2 = Average threshold 2. 20% sample of pixels from all CDs 3. 20% sample of 5% of CDs 4. See text

Figure 9.1 Population Density and Population Estimates for Census Collection Districts Ground Truth vs. Remote Sensing Estimates from Locally Trained Models

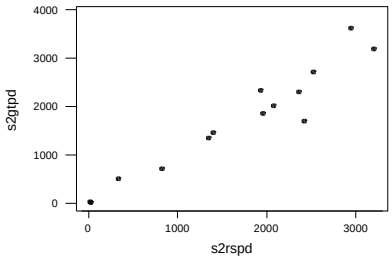
Ballarat 1988: CD population density
Study region



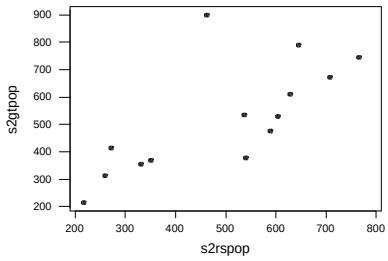
Ballarat 1988: CD population
Study region



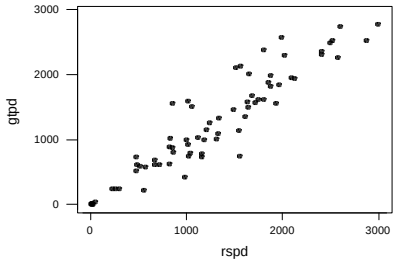
Ballarat 1988: CD population density
Training sample



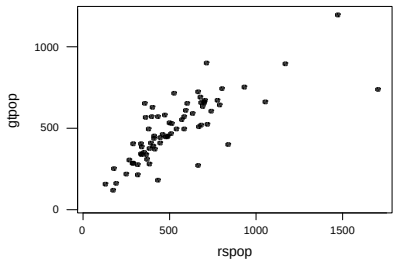
Ballarat 1988: CD population
Training sample



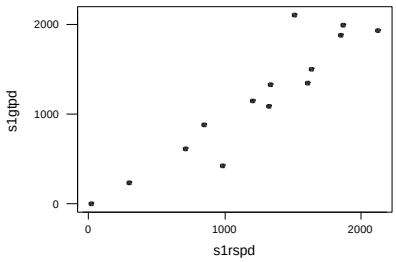
Ballarat 1994: CD population density
Study region



Ballarat 1994: CD population
Study region



Ballarat 1994: CD population density
Training sample



Ballarat 1994: CD population
Training sample

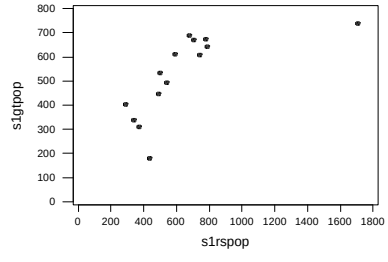
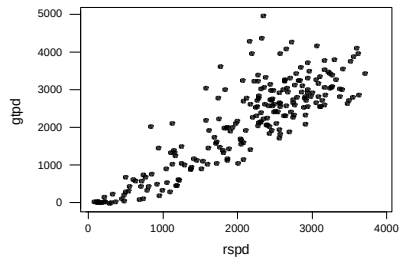
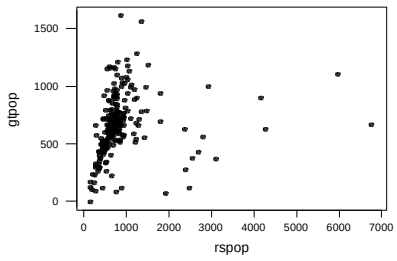


Figure 9.1 Population Density and Population Estimates for Census Collection Districts Ground Truth vs. Remote Sensing Estimates from Locally Trained Models
(continued)

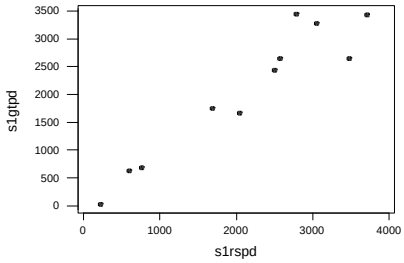
Geelong: CD population density
Study region



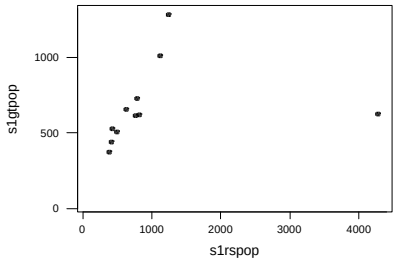
Geelong: CD population
Study region



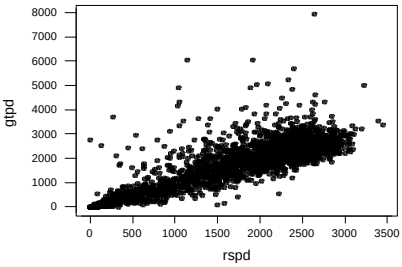
Geelong: CD population density
Training sample



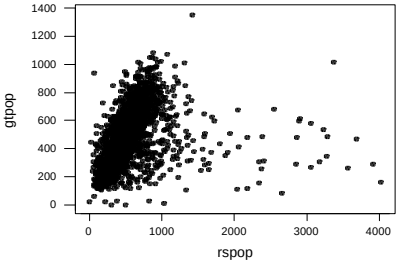
Geelong: CD population
Training sample



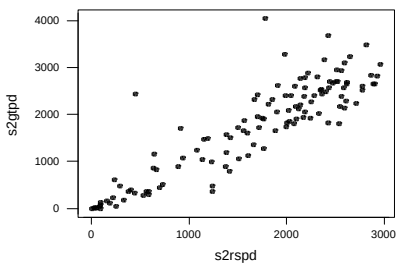
Adelaide: CD population density
Study region



Adelaide: CD population
Study region



Adelaide: CD population density
Training sample



Adelaide: CD population
Training sample

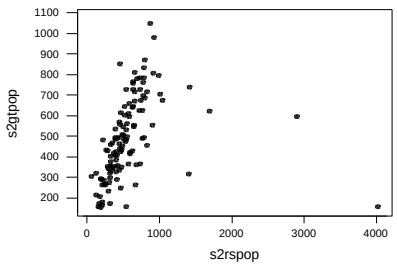
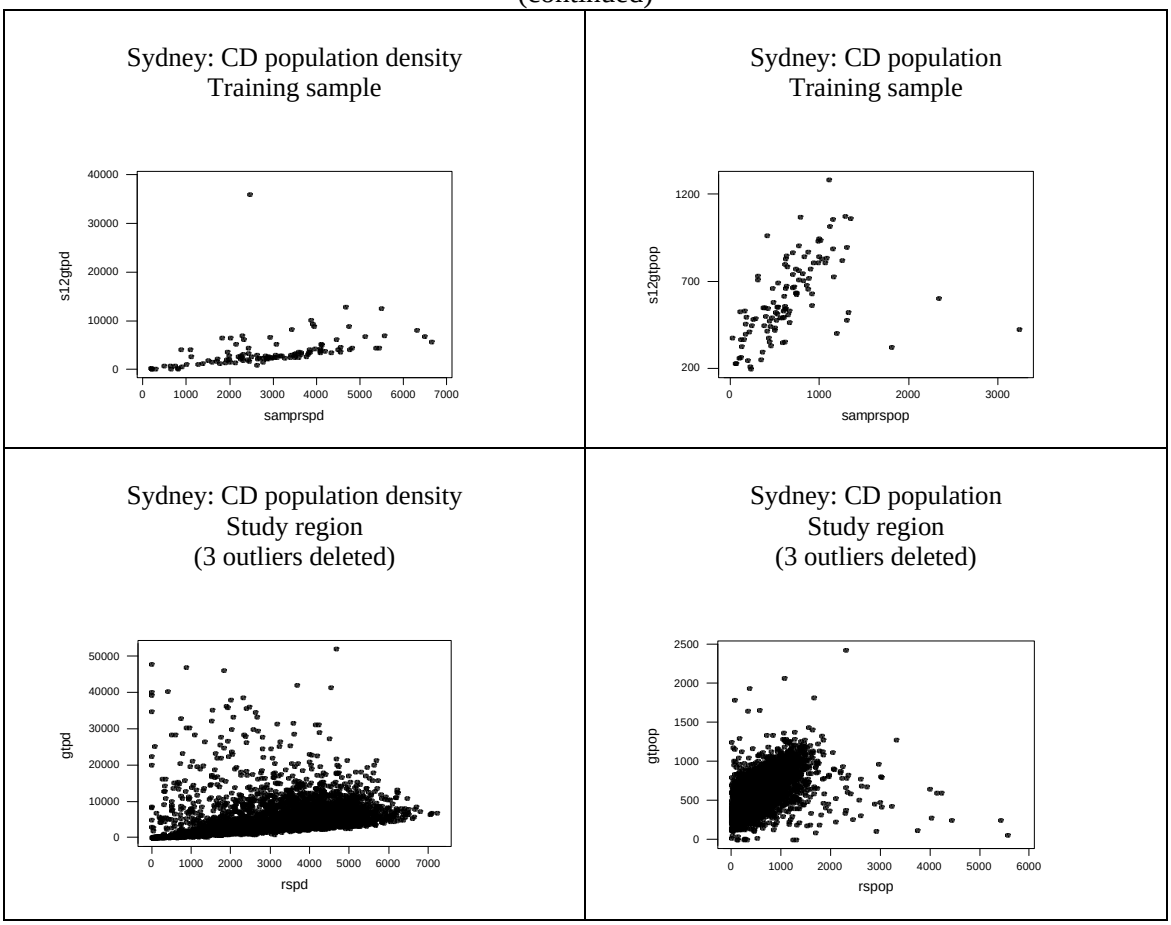


Figure 9.1 Population Density and Population Estimates for Census Collection Districts Ground Truth vs. Remote Sensing Estimates from Locally Trained Models
(continued)



As expected, the results for the smaller images of Ballarat and Geelong were more variable and sensitive to the specifics of the samples than for those of Adelaide. For example, the sample number 2 from Ballarat 1988 included no CDs from the high density tail of the population distribution. As a result, the equation based on this sample produced the smallest mean and median errors for individual CDs, but a consistent underestimation bias was evidenced in the low estimate for the total urban population. In the case of Geelong, the results for the urban area from a training set of only 9 CDs were almost as good as those for Ballarat, but the estimates blew out badly in the problematical industrial and rural CDs which have been discussed at length in Chapter 7.

The results for Sydney, whilst much less biased than the very poor results obtained by normalisation, still exhibited much larger errors at CD level than was the case for Adelaide, and still had almost vanishingly small R^2 values, indicating a continuing problem with multi-level residential structures and extreme population density outliers.

Leaving aside the perennial problem of overestimation in the rural areas, which has been discussed in Section 8.4 and will be again in Chapter 10, with the exception of Sydney and one

of the Adelaide samples, the mean relative errors of estimation for individual CD populations in the urban areas were consistently in the order of 20%, and medians somewhat lower, indicating a positively skewed distribution, with a tail of larger errors in every case.

Again, with almost every sample taken, from all five images including Sydney, it was possible, in many cases after judicious deletion of outliers, to produce an estimate of total urban population that was correct to within a few percent.

9.3.3 Adaptive adjustments to training sets

From an inferential perspective - when attempting to estimate an unknown population on the basis of sample data from a number of small areas - the question arises as to how much can be inferred from a training sample about the accuracy of estimates pertaining to the whole image. Further, since deleting outliers from the training sample can reduce bias, it may also be the case that information from the sample itself may assist in the decisions about which CDs, if any, to delete.

These issues were addressed by comparing three key indicators for sample and image: two indicators of overall accuracy and one indicator of bias. The values of mean relative error, median relative error and relative error in the total population, obtained from all urban CDs in the image (Table 9.2 Part A), were compared with the values obtained for urban CDs in the training sample[1] [2](Table 9.2 Part B). The difference between the two values of the three statistics was calculated in all 45 cases (excluding Ballarat 1988 sample 1, which was drawn from all CDs in the image).

Figure 9.2 shows a number of plots pertaining to the analysis of these indicators and their differences. In the sixth plot of each set, the samples have been grouped into 4 groups of similar sample size – around 10, 25, 50 and 120 CDs respectively.

The pattern of results were similar for both the means and medians of the relative errors (Figures 9.2A and 9.2B). The values for both sample and full image were first plotted against sample size, expressed for clarity of presentation as the square root of the number of CDs in the sample. There were clear differences between images, with the values for Sydney being much larger than for the other images. The values were generally higher for the image than for the corresponding sample, although as one cluster on Figures 9.2A3, 9.2A4, 9.2B3 and 9.2B4 show, this was not always so.

---

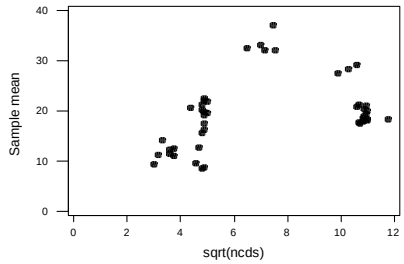
[1] The full set of urban CDs in the training sample was used in each instance – not the reduced set.

[2] Urban criteria were used not only because of the predominantly urban focus of the study, but also because non-urban CDs were small in number in the samples, resulting in greater variation in sample results for non-urban measures.

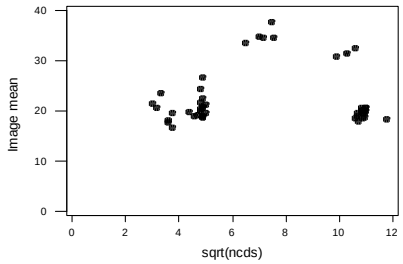
Figure 9.2 Comparison of Three Key Indicators for Training Samples and Whole Images

A. Mean relative error in estimates of urban CD populations

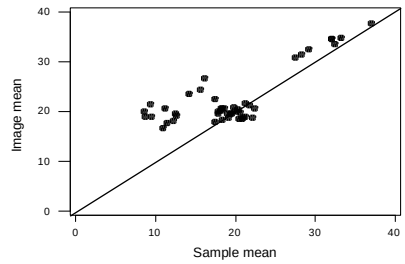
1. Sample mean vs. √(sample size)



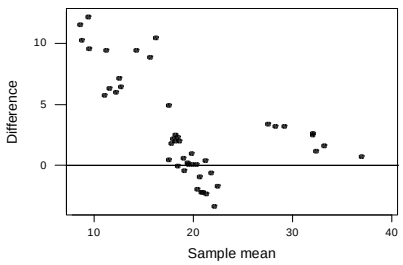
2. Image mean vs. √(sample size)



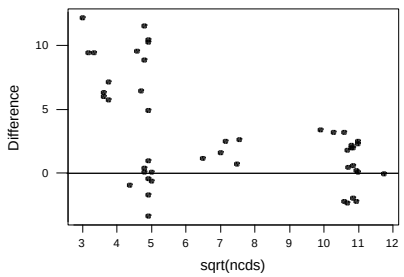
3. Image mean vs. sample mean



4. Difference vs. sample mean



5. Difference vs. √(sample size)



6. Magnitude of difference vs. grouped sample size

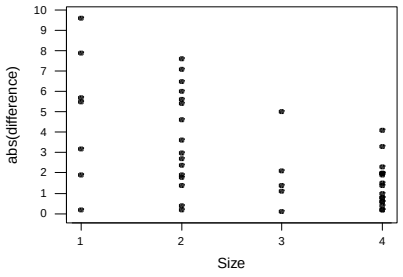
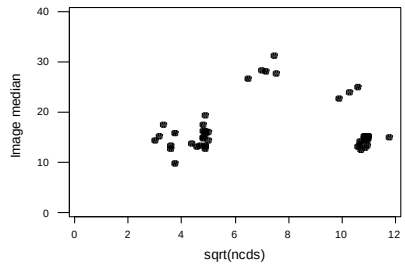


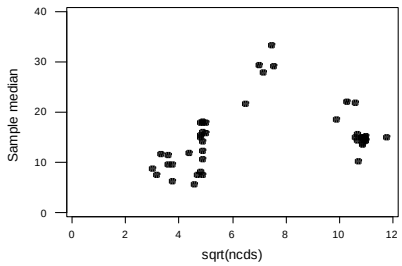
Figure 9.2 Comparison of Three Key Indicators for Training Samples and Whole Images (continued)

B. Median relative error in estimates of urban CD populations

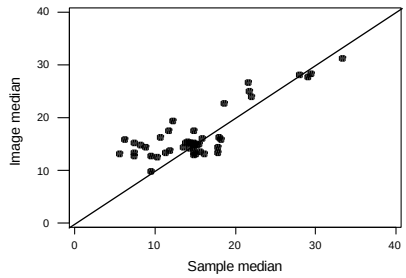
1. Sample median vs. √(sample size)



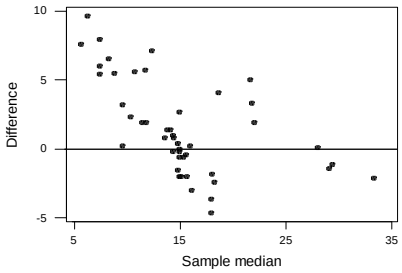
2. Image median vs. √(sample size)



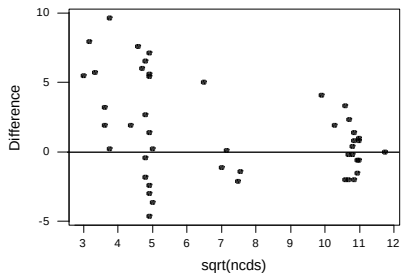
3. Image median vs. sample median



4. Difference vs. sample median



5. Difference vs. √(sample size)



6. Magnitude of difference vs. grouped sample size

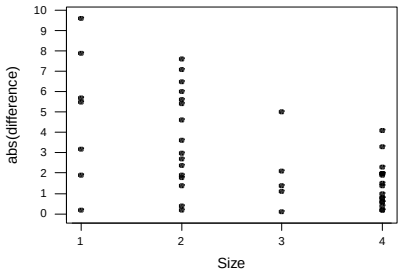
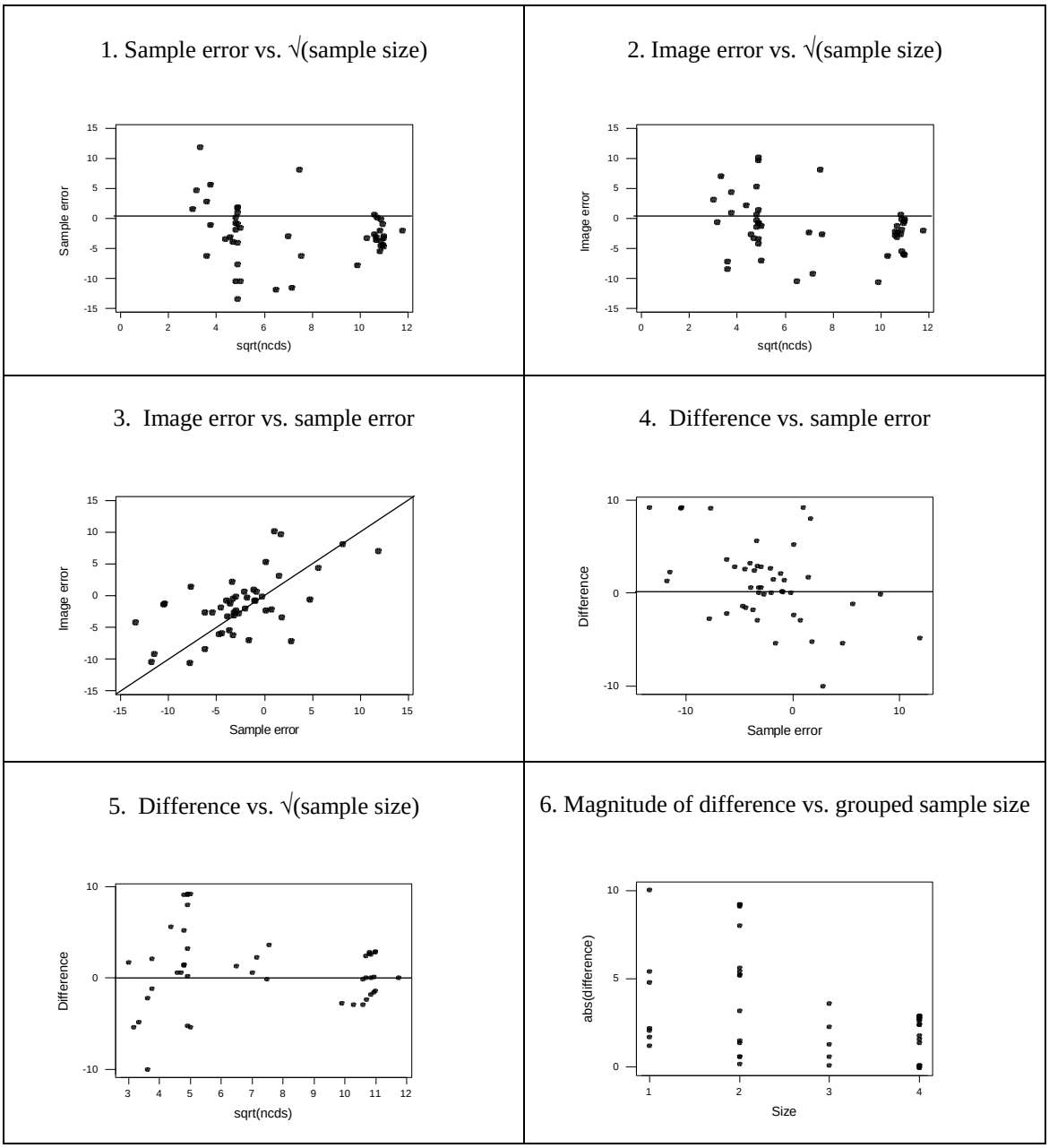


Figure 9.2 Comparison of Three Key Indicators for Training Samples and Whole Images (continued)

C. Relative error in estimate of total urban population



Figures 9.2A3 and 9.2B3 show that for the less accurate Sydney estimates (top right of each plot), the mean and median values for sample and image were similar. For lower values of mean and median, there was a greater differential between the sample values and the values for the whole image, the latter of which almost never fell below about 18% for the mean and 13% for the median. In each case, plots 5 and 6 give some indication that the difference between results for sample and image tend to diminish as sample size increases - a not unexpected result.

Of course these results are not independent of one another, being based on groups of variants of a very small number of independent samples. Nevertheless, there are clear indications that the mean and median of the relative errors in the training sample can provide some guidance as to the mean and median relative errors than can be expected for the overall image. On the basis of Figures 9.2A6 and 9.2B6, one could tentatively conclude that for a training sample made up of around 120 CD-sized areas, the mean and median relative errors for the overall image are unlikely to exceed those observed in the training sample by more than 5 percentage points.

There is some slight evidence in Figures 9.2C1 and 9.2C2 that bias, as indicated by the relative error in the estimate of total population, is reduced as sample size increases. Of more practical interest is the clear evidence in Figure 9.2C3 of a positive correlation ($r=.61$, $p<.0005$) between the relative errors in training sample and image. There is also some evidence of "regression towards the mean", with the relative error tending to be smaller in magnitude for the image than for the sample (Regression equation is: Image error = 0.034 + 0.585 Sample error). Figure 9.2C4 shows no evidence of a relationship between the magnitude of the difference and the size of the relative error ($r=-.14$, $p=.355$).

Figures 9.2C5 and 9.2C6 give some indication, albeit based on limited evidence, that the differential between the relative error in the training sample and the overall image is inversely related to sample size – again a not unexpected result. In all variants of the samples in group 4, with around 120 CDs, the difference between the relative errors for sample and image was less than 3%.

These results suggest a methodology whereby an initial representative training set of CDs is selected and then exploratory deletions are made of two types of CD: those in which imputed pixel populations are substantially underestimated; and those with large areas and very low population densities. The final training set is the one which minimises the relative error in the total for the urban area of the full training set as originally selected. In this way, the training sample can be tuned to minimise bias.

9.3.4 Estimates for Statistical Local Areas

Finally, one representative set of estimates for each of Adelaide and Sydney were further aggregated to the level of Statistical Local Areas (SLAs). The results are shown in Table 9.3 and Figure 9.3.

The Adelaide image encompassed 46 SLAs, with a mean population of 25187 and a mean area of 232.9 km. The Sydney image included 41 SLAs, with a mean population of 80095 and a mean area of 86.0 sq.km.. The Adelaide SLAs are likely to be comparable in size and population with the suburbs of Harare used by Webster (1996). The Leicester census wards

used by Langford et al. (1991) would seem to be intermediate in size between Australian CDs and SLAs, whilst the Tertiary Planning Units of Hong Kong used by Lo (1995) are comparable in area with Australian CDs, although they may well have larger populations.

As expected, on the larger scale of the SLAs, much higher values of R^2 (.97 for Adelaide) and much lower values of mean and median relative errors (9.7% and 5.6% for urban Adelaide) were obtained. However, considering the magnitude of the increase in aggregation level (50-fold in the case of Adelaide, 140-fold in the case of Sydney) the extent of the reduction in error levels might be regarded as modest. This is presumably due to a substantial degree of positive spatial correlation between the errors in neighbouring CDs within an SLA.

On the larger population scale of SLAs, the overestimation of population in particular rural CDs is greatly diminished in impact - perhaps a more realistic perspective. In fact the linear fit was better for population than for population density.

Figure 9.3 Population Density and Population Estimates for Statistical Local Areas Ground Truth vs. Remote Sensing Estimates from Locally Trained Models

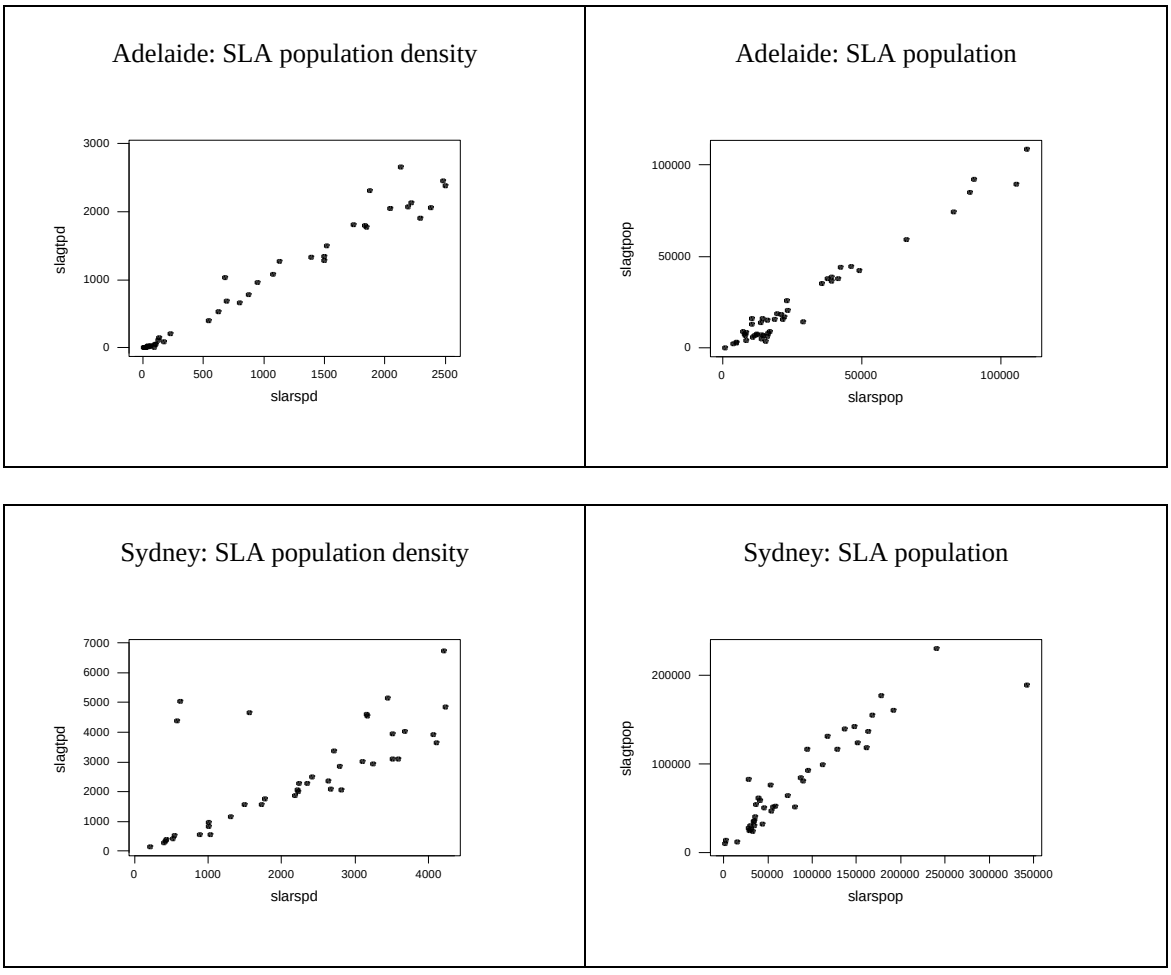


Table 9.3 Summary of Selected Models for Estimating Statistical Local Area Population Densities and Populations Based on Local Training of Both Classification and Regression Procedures

Image	Sample	Sampling fraction %	Thresholds[1] T1	T2	Region b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error	Urban Area (CDs >500 persons/sq.km.) b0 unforced	b1 unforced	b1 forced	R^2	s	Mean % error	Median % error	Total % error
Adelaide																				
SLAs	2r3	5	1.5	0.27	-9.7	0.99	0.98	0.97	152.2	47.5	18.3	14.7	72.1	0.94	0.98	0.89	210.4	9.7	5.6	4.7
SLA population					-2731	0.97	0.92	0.97					1053	0.93	0.94	0.99				
Sydney																				
SLAs	2	2	1.5	0.27	499.6	0.93	1.11	0.50	1184.1	22.5	13.2	7.1	887.6	0.81	1.11	0.35	1240.1	22.3	12.6	4.0
SLA population					19176	0.71	0.80	0.86					19917	0.71	0.85	0.84				
2 CDs omitted[2]					12607	0.81	0.90	0.92					12509	0.83	0.92	0.91				

1. T1= individual pixel threshold T2 = Average threshold
2. Two rural Sydney CDs in which regrowth after forest fire was misclassified as residential contributed very influentially to the regression relationship for population.

The Sydney SLA plot makes very clear the nature of the population estimation problem for that city. Population densities were quite accurately estimated for 34 of the 41 SLAs whose average population densities were below about 4000 persons/sq.km.. Of the eight SLAs with average population densities above that threshold, the errors of underestimation were moderate in one case, substantial in four cases, and gross in three cases. These constituted three inner central SLAs dotted with high-rise residential tower blocks, and a surrounding ring of inner suburban SLAs with a more even spread of lower multi-level residential structures. The distance of the points above the regression line in these cases is a rough indicator of relative building heights in these SLAs.

Notwithstanding these problems, the population contribution of these SLAs is relatively modest, and so in most cases they are not obvious on the population plot, and they do not greatly bias the estimate of total population.

In this connection, it should be noted that the estimates of urban totals in Table 9.3 are higher than for the corresponding models in Table 9.2. This is because the urban cutoff of 500 persons/sq.km. takes in a larger area when applied at SLA level than at CD level.

The extreme outlying point on the right of the Sydney population plot was due almost entirely to two rural CDs within one SLA in which regrowth after forest fire was misclassified as residential. This regression equation was recalculated without the spurious population contributions of these two very influential points (see Table 9.3).

9.4 CHARACTERISTICS OF THE ESTIMATION EQUATIONS

The data dependence of the regression coefficients and the reasons for it have been discussed (Sections 2.11.5, 7.2.4 et seq.). However, the 94 (2×47) sets of coefficients selected from five different images in the preceding sections, were examined for any common features. Because the magnitudes of the coefficients were quite variable, it was decided to examine the signs for consistency.

Table 9.4 Numbers of Positive and Negative Regression Coefficients in 94 Estimation Equations

Band		Basis of equation			
		6 iterations		Convergence criterion	
		+	-	+	-
Constant		37	10	43	4
B1	Blue	45	2	46	1
B2	Green	14	33	10	37
B3	Red	12	35	28	19
B4	Near infrared	38	9	37	10
B5	Mid infrared	1	46	0	47
B7	Mid infrared	46	1	47	0

The results in Table 9.4 show that the regression coefficients have three very consistent features: positive valued coefficients for band 1 and band 7, and negative valued coefficients for band 5. This consistent pattern can be speculatively interpreted in terms of the characteristic TM spectral signatures of different materials.

In Section 5.4.2, it was observed that the strongest discriminators between residential and non-residential areas were ratios involving band 1 in the numerator and band 5 in the denominator. Constructed surfaces such as bitumen and many roofing materials have relatively strong reflectivity at the shorter wavelength of band 1 (Forster, 1980; Curran, 1985), whereas both vegetation and bare ground generally have higher reflectances in the band 5 range than in the band 1 range. At the longer band 7 wavelengths, the reflectance of vegetation drops off somewhat compared to band 5, whilst that of bare ground (and presumably that of clay tiles also) stays fairly constant (Harrison and Jupp, 1989, p6).

Considering these relativities between the responses of the three types of material at the three wavelengths, we can say that in all the regression models fitted to pixels classified as residential, conditional on the other bands remaining constant, increases in population density are consistently associated with higher reflectances in bands 1 and 7 (indicating the predominance of built surfaces), and lower reflectances in band 5 (indicating the predominance of natural surfaces).

Discrimination between spectral responses of these three types of material is much less marked at the intervening wavelengths of TM bands 2-4 (Harrison and Jupp, ibid.). To this fact, together with the substantial level of multicollinearity, can be attributed the less consistent pattern of signs from sample to sample displayed in Table 9.3.

9.5 SUMMARY

In this chapter, an approach has been described for training an estimation equation on a small sample of population data from within an image, and applying the results to the full image.

It is concluded that the results obtained are more accurate and reliable than those obtained by normalisation of an estimation equation trained on another image.

Methods have also been described and demonstrated for tuning the training sample to minimise estimation bias.

Limited evidence has also been presented relating the accuracy of estimates for the full image to comparable measures from the training sample. Potentially, error bounds for the various

estimates for the full image might be estimated in this way, though much more replication would be needed to establish reliable heuristics.

It was also demonstrated that estimates for larger areas had lower relative errors.

Finally, the many estimation equations derived have been examined collectively, and their common characteristics described and interpreted in terms of the spectral responses of different materials.

It is contended that the methodology developed and tested in this chapter could form the basis of a feasible operational procedure for estimating a large regional population on the basis of a partial census of relatively small sections of the region.

The full specification of such a procedure, together with a discussion of remaining problems and limitations, performance in relation to other published results, and directions for further research, are discussed in Chapter 10.

Chapter 10

Conclusions and Recommendations: Towards a Feasible Operational Methodology for Population Estimation from Landsat TM Imagery

In this chapter the study is reviewed and reflected upon. Section 10.1 is a summary of the phases and milestones of the study. Section 10.2 summarises the conclusions in terms of the research questions posed in Section 1.4. In Section 10.3, the advantages of the individual pixel approach in preference to the CD aggregate approach are discussed in more detail. In Section 10.4, the outcomes of the study are assessed in comparison to other related work. Section 10.5 gives a specification of the recommended model and procedure for population estimation. Section 10.6 outlines directions for further research. Finally in Section 10.7 some possible applications of the methodology are suggested.

10.1 SUMMARY OF THE STUDY

Two approaches to population estimation from Landsat TM imagery have been investigated, one based on data aggregated over Census Collection Districts, and the other based on data for individual pixels.

Beginning from the most basic prediction model based on the Collection District means of each TM band, substantial improvements were achieved in the estimation of CD population and dwelling densities on the basis of CD aggregates of more complex remote sensing indicators. From the many models tested, six were chosen for further testing on the secondary image, five utilised the square root of the density, and involved progressively more complex CD aggregate functions of the TM bands: basic means; squares of means, ratios of means, variation measures; and means and variation measures of selected pixel-level spectral transformations. The effective R^2 values of these models ranged from .54 to .84 for population density, and from .56 to .92 for dwelling density.

After testing on the secondary image, it was concluded that some of the improvement which had been achieved in the CD aggregate models in the case of the primary image through increased complexity, was lost in the transfer to the secondary image through lack of robustness. Overall, the model which performed best on the urban areas of both images was a model based on ratios of CD band means, which produced very accurate estimates of the total urban population in both cases. Nevertheless, at the level of individual CDs, this model, like the other models, tended to underestimate the higher densities and overestimate the lower densities, and this was particularly the case in the secondary study area with its higher average density. Like the other models, it grossly overestimated the regional totals for both primary and secondary study areas. It was decided that procedures based on CD aggregates were not robust to variation in density either within or between study areas.

The second approach involved a two phase procedure of classification followed by regression modelling. After extensive examination, no significant benefit was found from the use of spectral and spatial transformations of the six TM bands at either classification or regression stages. Three candidate models were selected for further investigation, all involving at their core a simple linear function of the six TM bands, but in two cases utilising square root and logarithmic transformations of the dependent variable, population.

A crucial step in the estimation of the regression relationship was an algorithm for iteratively re-estimating the imputed ground truth values initially assigned to each pixel. The properties of the algorithm have been examined by repeated sampling and by simulation, and the algorithm has been placed in the broader theoretical context of the EM and related algorithms.

When the pixel-based models were applied to the secondary image, the logarithmic and square root models were found to be not at all robust. The performance of the linear model was more consistent, though there was a problem of bias which was largely overcome by retraining the initial classification phase on the secondary image. As with the CD aggregate methods, there remained a residual tendency to underestimate population in high density areas and to overestimate it in low density areas, but because of the disaggregated basis of the analysis, it was possible to devise methods to overcome these problems to some degree.

Comparing the best performances achieved using the two approaches, and considering both the advantages of simplicity and parsimony at the model development stage, and the advantages of spatial flexibility and GIS-compatibility of the outputs, the pixel-based method was considered to be clearly superior to the CD aggregate method.

The robustness of the model developed and trained on the primary image was further explored by applying it, in various normalised forms, to five more Australian images, included three very extensive images of large cities and their environs.

It was concluded that z-score normalisation provided a methodology which was moderately robust, particularly so far as urban areas were concerned, to geographical and temporal differences in season and climate, but less robust to differences in average population density or to differences in the shape of the statistical distribution of population densities within an image.

Finally, a less ambitious approach was developed and tested for training an estimation equation on a small sample of population data from within an image, and applying the results to the full image. It was concluded that the results obtained are more accurate and reliable than those obtained by normalisation of an estimation equation trained on another image.

Methods have been described and demonstrated for tuning the training sample to minimise estimation bias. Limited evidence has also been presented relating the accuracy of estimates for the full image to comparable measures obtained from the training sample.

10.2 CONCLUSIONS

In terms of the specific hypotheses listed in section 1.4, it was concluded:

- That whilst linear population estimation models based on CD aggregates could be enhanced by the incorporation of spectral and spatial transformations of TM data, and by mathematical transformations of the dependent population variable, this enhancement was to some extent brought about by capitalisation on chance and did not translate into similarly improved performance on a separate validation set.
- That the capability of linear population estimation models was enhanced by modelling the population of individual pixels rather than that of larger spatial aggregates, but only with the incorporation of iterative re-estimation of imputed pixel populations.
- That the capability of linear population estimation models based on individual pixels was enhanced by classification of the pixels into different landcover/landuse classes.
- That discrimination between the residential class and other landcover/landuse classes was not substantially enhanced by the incorporation of spectral or spatial transformations of TM data.
- That classification of pixels in low population density areas was enhanced by the incorporation of a second stage of contextual reclassification.
- That pixel-based models utilising just the 6 TM bands at both classification and regression modelling stages produced aggregate population estimates in the training set which were as accurate as those produced by much more complex aggregate-based models.

Regarding validity and robustness, it was concluded that the pixel-based model was moderately robust to variations in geographical location, land cover, climate, time and season, but much less robust to differences in intensity of human settlement.

The final objective was to specify a feasible operational procedure for estimating population from TM imagery, with respect to the non-remote sensing inputs required, and the nature and extent of human intervention and interpretation required, and the accuracy obtained.

The recommended methodology is fully specified in Section 10.5. Two phases of human intervention and interpretation are required: firstly, a comprehensive suite of landuse/landcover classes must be identified, and representative training sets must be selected for each class; secondly, a representative set of 10-100 small regression training areas (encompassing 1-5% of the population to be estimated) must be selected, and the total population of each area must be obtained.

When this methodology was emulated in the study, the most accurate results were obtained for urban areas of moderate population density. In such areas, population estimates at the macro (major metropolitan centre) level were accurate to within 3%, estimates at the intermediate (Statistical Local Area, provincial city) level had mean errors in the order of 5-10%, estimates at the micro (Census Collection District) level had mean errors in the order of 10-20%. It is conjectured that these levels of accuracy may be close to the limit attainable with this methodology.

Errors were much greater in areas of extremely low or extremely high population density. Approaches to improving the accuracy of estimation in these areas are discussed in Sections 10.5 and 10.6.

10.3 ADVANTAGES OF PIXEL-BASED ESTIMATION

Approaches to population estimation based on the use of remote sensing information aggregated over some extended spatial area suffer from a number of common limitations not shared by pixel-based methods. Furthermore, pixel-based methods offer a number of advantageous features. These, together with the small number of counter aspects, are summarised in Table 10.1.

As Table 10.1 shows, the only clear advantage of the aggregate–based methods is the obvious one – that in developed countries at least, ground truth population data for training is available at this level.

Table 10.1 Comparison of Pixel-based and Aggregate-based Estimation Methods

Feature or aspect	Aggregate-based methods	Pixel-based methods
Model building and training		
• Information about the relationship between population and spectral response	Loss of detailed information about spectral response on individual pixels	Pixel-level population not known, but can be estimated
• Mathematical form of model	Usually complex – problem of capitalisation on chance	Simple and robust
• Sample size and degrees of freedom	Small (relative to pixel-based methods)	Large
• Area (and population) needed for training (converse of sample size)	Large (relative to pixel-based methods)	Small
• Suppression of anomalous spurious population features (masking)	Difficult – areal incompatibility	Routine
• Addition of anomalous concentrations of population	Feasible	Routine
• Classification and stratification	Possible via areal interpolation methods for some forms of model	Routine
• Incorporation of ancillary information e.g. differential weighting by building heights or occupancy ratios	Difficult if available on an incompatible areal basis	Routine
• Statistical texture measures	Larger extent – wider range of measures available e.g. pattern-based	Local neighbourhood measures only
• Morphological approaches to classification	Difficult? Areal incompatibility	Routine via masking layer
Estimation beyond the training set		
• Estimates for incompatible areas defined for the same training region	Difficult – areal interpolation methods required.	Routine
• Estimates for similar areas to training set	Defined comparable areas required	Routine
• Estimates for areas of arbitrary size and shape	Difficult – areal interpolation methods required. Also, models may not be robust to changes in scale.	Routine
• Estimates for other regions	Normalisation generally not feasible	A degree of robustness via normalisation
General		
• Resolution & GIS	Resolution very coarse	Routine - pixel level is finer than administrative units and fine enough for most demographic applications. However, difficult to validate below the scale at which population data is available
• Mapping	Without further analysis, limited to choropleth. Further data and areal interpolation analysis required for dasymetric mapping	As above

However, in this study it has been demonstrated that pixel-based models of simple linear form can be trained on imputed pixel populations, which perform comparably with much more mathematically complex aggregate-based models with respect to training set criteria, and are also more robust. That being the case, the comparison is very one-sided. Because the size of a TM or other remote sensing pixel is below the scale of all administrative units and most demographic applications, the problems of areal interpolation (Goodchild and Lam, 1980; Langford, et al., 1991; Goodchild et al., 1993) do not arise. Looked at another way, this problem is dealt with at the regression modelling stage when the pixel populations are imputed and re-estimated.

In the training phase, the pixel-based approach has advantages with respect to sample size, degrees of freedom and training set flexibility, as well as much greater flexibility to incorporate ancillary information via extra raster or vector layers in the remote sensing image or GIS.

In the validation and implementation phases, the flexibility and the lack of areal interpolation problems is manifest.

It may be for these reasons that much of the work in this area published to date, which has been aggregate-based, has not proceeded beyond the training phase where a relationship is demonstrated at face value, to a more searching validation, implementation and evaluation phase.

10.4 RESULTS, OUTCOMES AND PERFORMANCE

The task of estimating human population from remote sensing imagery differs in three important respects from most other remote sensing applications. Firstly, the phenomenon being investigated (population) is less directly linked to the remote sensing indicators (reflectances of materials) than is the case in most applications in the earth, biological and environmental sciences. Secondly, the aim is to make quantitative estimates across the spatial dimensions of the image, rather than is often the case, qualitative or categorical classifications. Thirdly, many remote sensing analyses are locally focussed and analytically specific, whereas the aim here was to establish a generic framework.

With regard to the first two aspects, there is an expectation that there is an upper limit to the accuracy which is achievable in principle. With regard to the third aspect, it is to be expected that different parts of the estimation process might have different degrees of robustness, depending on what aspect (season, geographical location, culture, etc.) is changed and by how much.

The outcome of this study is a generic methodological framework for population estimation which it is contended could be made to work anywhere, at any time, with the appropriate inputs

of TM imagery, training information about land use, land cover and population, and where available, additional ancillary information. How well it works, how accurately population is estimated, will always depend on the quality of those inputs. Thus inaccuracy will always be partly inherent to the process, and partly amenable to reduction by improved inputs.

In this study, the methodology has been emulated and evaluated, with very little ancillary information, in various contexts. The accuracy of the estimates obtained has varied from image to image, for reasons which have been identified and discussed.

In this section, the results obtained for the areas including and surrounding two of Australia's four largest cities, Sydney and Adelaide, are compared with comparable published results from the last decade or so.

Such comparisons are not extensive or straightforward, since there is a paucity of published work in the area, and reporting has not always been comprehensive. Some of the purported results due to other researchers quoted in this section have been inferred by the author (and in one case calculated by him from published data). These calculations and interpretations have been made in good faith and are believed to be accurate. A comparative summary is set out in Table 10.2.

Table 10.2 includes one indicator of variability or consistency (R^2), one indicator of bias (relative error in the total population) and two indicators of overall accuracy (mean relative error and median relative error), as discussed in Section 2.12.

As the results throughout this study indicate, it is one thing to demonstrate a relationship between some demographic characteristic and a set of remote sensing indicators, possibly quite complicated, in a single data set. It is quite another to find relationships that have genuine predictive capability beyond the data from which they were derived.

The results of Iisaka and Hegedus (1982) and Webster (1996) pertain only to a training set. Langford et al. (1981) undertook a degree of validation by attempting to recover a different set of areal subtotals for the same training area. This is broadly comparable to the use of training set CDs to validate pixel-based models in the present study. Only Lo (1995) trained his models on a subset of data, and applied them to a broader set, which constitutes genuine external validation, as has been undertaken in this study (though it might be said that ideally the validation set would not include the training set). However, Lo's quoted R^2 values referred to the results in the training set.

Bearing in mind this distinction, the final results quoted for Adelaide and Sydney from the present study compare very favourably on almost all points where they can reasonably be compared with the other studies. The exception is the low values of R^2 brought about by the presence of a relatively small number of extreme high population density outliers in the Sydney

data. Some of the quoted training set values of R^2 are quite high, and similarly high values have been obtained for training sets throughout this study, but that does not necessarily translate into accurate predictions, nor even high R^2 values in a validation context.

The relative errors in the urban totals range from –3% to +5%, compared to Lo's which range from –10% to +8%. The mean relative errors for both Adelaide and Sydney CDs are much less than Lo's comparable figures for Hong Kong and Webster's training set dwelling densities (calculated by the author) for Harare, though the Sydney figure exceeds Webster's mean figure for Cardiff training set dwelling densities. The median relative errors for Adelaide are much lower than Webster's training set figures, and those for Sydney are comparable. Lo's research was based on SPOT imagery, which has higher resolution (smaller pixel size) than TM, which may or may not be advantageous (Webster, 1996; Barnsley and Barr, 1996), and also involved a lot of multi-level and multi-purpose structures.

Only Webster's report contained any comparable figures relating to lower density non-urban areas, although Langford et al. discussed the problems of overestimation at low densities. Again, whilst the results for non-urban areas in the present study are not very good, they are more consistent than the only other reported results. There are no available benchmarks with regard to bias in non-urban results, because the only accessible reported total was based on a training set of equal grid squares, in which the total is automatically constrained to equal the correct value by the OLS analysis.

Considering the spread of estimates obtained using many variants of the estimation equation, based on many training sets and applied across seven different test images, it is considered that the median relative errors in the range 10-15% obtained for the urban CDs of the Adelaide, Ballarat and Geelong images, and the median relative error of 6% obtained for the larger Adelaide SLAs, probably come close to the upper limit of accuracy that can be achieved with this methodology.

In each case, the mean value is somewhat higher, as is characteristically the case for the skewed distribution of absolute values of deviations, but in many cases the margin is very large, indicating the influence of a small number of very large discrepancies. These relatively large discrepancies usually fall at the extremes of population density, and should in principle be able to be reduced by incorporating various refinements which will be summarised in Sections 10.5 and 10.6. The methodology as applied in this study is believed to be close to fully functional at moderate levels of population density (200-3500 persons/sq.km.), and straightforwardly amenable to refinement at both high and low density extremes.

Table 10.2 Comparison of Some Results of This Research with Comparable Published Results

Source	Location	Nature of study area	Scale of test unit	Status of test units	Dependent variable	Density or count	Urban area (Moderate to high density[4]) R^2	Urban area Mean % error	Urban area Median % error	Urban area Total % error	Whole region (if different) R^2	Whole region Mean % error	Whole region Median % error	Whole region Total % error
Iisaka & Hegedus (1982)	Tokyo	Urban	Small	Training	Population	Both[5]	.59-.70			0[6]	Not applicable			
Langford et al. (1991)	Leicester	Mixed	Large	Training[2]	Population	Count					.76-.85			
Lo (1995)	Hong Kong	Urban	Small	Validation	Population	Mixture		64-99		-5 .. +8	Not applicable			
	Hong Kong	Urban	Small	Validation	Dwellings	Mixture	.77-.88[3]	63-77		-10 .. +4	Not applicable			
Webster (1996)	Harare	Mixed	Small	Training	Dwellings	Density	.81	57	27			65	30	
	Cardiff	Urban	Small	Training	Dwellings	Both[5]	.86	27	26			410	36	0[6]
Harvey (1999)	Adelaide[6]	Mixed	Small	Validation	Population	Density	.45-.51	18-20	12-14	-2 .. 0	.71-.75	47-51	15-17	+13 .. +16
	Sydney[6]	Mixed	Small	Validation	Population	Density	.03	32-35	25-28	-3 .. -2	.06	140-161	26-30	+7
	Adelaide[1]	Mixed	Large	Validation	Population	Density	.89	10	6	+5	.97	112	18	+15
	Sydney[1]	Mixed	Large	Validation	Population	Density	.35	22	13	+4	.50	23	13	+7

1. SLA estimates based on one representative model
2. Whilst Langford et al. discussed RMS errors in the context of an areal interpolation crossvalidation, the only comparable summary statistics quoted were the R^2 values based on the training data.
3. Lo's reported relative errors were based on a validation set, but the two R^2 values reported were based on the training data.
4. Harvey urban figures are for CDs with population density > 500 persons/sq.km.. Webster figures (calculated by Harvey) are for dwelling densities > 200 dwellings/sq.km. Figures of Iisaka and Hegedus and of Lo are presumed to relate to densities above these thresholds. No figures available for Langford et al.
5. With analyses based on equal grid squares, count and density are equivalent. With least squares analyses on such data, the totals are fixed by the analysis.
6. Values based on whole-image results for models trained on 50 or more CDs, and the variant of each model chosen on the basis of performance on the training sample.

10.5 THE MODEL AND ITS IMPLEMENTATION

10.5.1 The basic model and procedure

The recommended procedure for estimating a genuinely unknown population is now specified. In brief, the procedure involves:

- selection of classification training sets
- performing supervised classification
- selection of regression training areas
- obtaining of ground truth population for the regression training sets
- fitting of the regression model at pixel level
- applying the model to all pixels in the image
- smoothing the population estimates and performing contextual reclassification
- checking for bias by generating aggregated estimates for training areas
- if necessary, reducing training set and refitting the regression model

Step 1

Define a comprehensive set of land use /land cover classes, including a residential class. Select training sets for each class and perform a supervised classification of the 6-band TM image.

Step 2

Assign zero population to all pixels classified as other than residential.

Step 3

Select a training set of around 100 representative small areas, ranging upwards from 20 ha in size, and including a range of residential densities from the highest down to 10 persons/sq.km.

Step 4

Obtain ground truth population figures for this training set.

Step 5

Extract from the image the spectral data for the set of pixels from the population training set classified as residential, and perform an iterated regression analysis on them to obtain an initial population estimation formula

$$L(\mathbf{b}) = c_0 + \sum_{i=1}^{nbands} c_i b_i$$

Step 6

Use this formula to assign population values to all pixels in the image which are classified as residential.

Step 7

Apply a smoothing filter to the resulting population image.

Step 8

Use the smoothed and unsmoothed population bands to make a low density adjustment, by resetting the population of selected pixels to zero, to compensate for over-classification as residential in low density rural areas.

Step 9

Aggregate the pixel estimates for each of the areas that make up population training set, and compare the aggregated estimates with the ground truth values.

Step 10

Make any required deletions from the training set to reduce bias.

Repeat steps 5-10 as required.

Steps 6-8 can be expressed mathematically thus:

$$p_{pixel} = T(L(\mathbf{b}))\left[1-(1-T(S(T(L(\mathbf{b})))-t_A)(1-T(L(\mathbf{b})-t_p)))\right]$$

where

$T(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases}$ is the thresholding function

$S(x) = \frac{\sum_{n\times n} x}{n^2}$ is the smoothing function

t_p is the pixel population reclassification threshold

t_A is the average population density reclassification threshold

The first term in the equation represents the initial thresholding at zero, to remove negative population estimates. The compound term in square brackets represents the resetting to zero of the population of any pixel which falls below both the average population density threshold t_A and the individual pixel threshold t_p. Values must be chosen for these coefficients.

10.5.2 Enhancements: adjustment for anomalies

Two straightforward procedures are outlined here for improving the accuracy of estimates at high and low density extremes by incorporating ancillary information. These are routine in principle, though the second may require some non-standard computational implementation in remote sensing or GIS software. They require a modicum of human intervention and judgement, which is in any case required at the image classification stage. Other potential enhancements requiring further research are considered in the next section.

Suppression of spurious population associated with specific misclassified features

Table 8.2 (Section 8.2.2) shows a number of features – built, agricultural and natural – for which many pixels were wrongly classified as residential and hence assigned spurious population. A number of these features contributed substantially to the over-estimation of population in non-urban CDs. In a practical population estimation exercise, some such features at least should be clearly identifiable. It is a routine matter to construct a binary masking overlay (set to 1 or 0) which when multiplied by the population estimates layer, has the effect of setting population to zero in these areas.

Injection of concentrations of population

This is the converse of the previous step. Again, an overlay can be constructed in which known major anomalous concentrations of population, such as institutions and tower blocks, can be gleaned from ancillary sources and assigned to small regions or single pixels. This layer is then added to the population estimates layer.

With these enhancements, the mathematical specification of the model is as follows:

$$p_{pixel} = T(L(\mathbf{b}))\left[1-(1-T(S(T(L(\mathbf{b})))-t_A)(1-T(L(\mathbf{b})-t_p)))\right]A_m + A_a$$

where

A_m is the multiplicative anomaly adjustment factor, which takes the value 0 to suppress spurious population and 1 elsewhere.

A_a is the additive anomaly adjustment factor, which is used to inject known concentrations of population.

In the present study:

- The coefficients of the central linear equation were estimated by iterated linear regression on a training set of pixels within census collection districts for which populations were known.

- Averaging of population density was performed using a mean filter over a 7×7 pixel neighbourhood.
- Various combinations of low density individual and smoothed thresholds were used, including (1,1), (1.5, 0.27) and (2.0, 0.27).
- Anomaly correction factors were not explicitly included in calculations.

10.6 DIRECTIONS FOR FURTHER RESEARCH

The issues of over-estimation at low density and under-estimation at high density, reported in this and other recent studies, remain problematical. The present methodology can produce reasonably accurate estimates of population within the range of typical Australian suburban densities. But clearly further research is needed if acceptably accurate estimates of scattered rural populations or of inner city populations are to be attained. Directions for further research will be considered under three headings: improving estimation at low population densities; improving estimation at high population densities; and other aspects.

10.6.1 Improving estimation at low population densities

The key here is better classification. A beginning might be to define better targeted training sets for particular confounding features such as country roads and shorelines. An alternative might be to incorporate a second complementary classification stage, using a non-statistical morphological approach for line detection (for example Ton et al., 1989).

Other alternative approaches to classification in general reported during the past ten years include: analysis of fractal dimensions (De Cola, 1989; Lam, 1990); fuzzy set theory (Wang, 1990; Gopal and Woodcock, 1994); mixed pixel or end member analysis (Smith et al., 1990); knowledge-based systems (Wharton, 1987; Moller-Jensen, 1990; Bolstad and Lillesand, 1992); neural networks (Chen et al., 1995; Foody et al., 1995; Foody, 1996); and genetic programming (Riolo and Line, 1995).

The recent literature of contextual reclassification (Treitz et al., 1992; Gong and Howarth, 1992; Van Deusen, 1995; Barnsley and Barr, 1996; Sharma and Sarkar, 1998) may also provide useful insights.

It may be that very specialised and sophisticated methods are needed to distinguish the faint genuine human signal from the imitative noise of a such a sparsely inhabited landscape as an Australian rural area.

10.6.2 Improving estimation at high population densities

At the high density end, the problem is one of hidden population. Landsat TM and similar sensors essentially operate in two dimensions, and to the extent that they can detect population, really do so on a "population per floor level" basis. From this perspective, deleting high density outliers from a training set (Chapter 9) is essentially discarding those pixels for which the value of what might be termed the "floor truth" dependent variable is unknown.

One approach to this problem would be to explicitly incorporate information about numbers of levels, or equivalently building height, into the models. The ground truth pixel populations would be converted to populations per level, which would be used as the dependent variable to train the remote sensing estimation equation. This would then be applied to the image, and the resulting estimates of population per level would be backtransformed using the height information, to produce population estimates. Even very rough suburb-by-suburb estimates of average building height, which would be incorporated as a multiplicative overlay, would be expected to bring about substantial improvement in population estimates such as those of Sydney.

This should not replace the specific pointwise anomaly corrections as described in the previous section, but should complement them. One saving grace of large institutions and tower blocks is their visibility and regularity of shape, which makes such ad hoc adjustments quite feasible. The "average building height" layer would be used to correct for more widely distributed lower profile multi-level structures.

A less precise alternative to modelling population per level would be to employ more than one residential classification, as was done by Langford et al. (1991) and Lo (1995). This could be incorporated into the pixel-based methodology of the present study at the cost of considerable organisational complexity, with separate training sets being selected from within each residential stratum, and separate estimation equations being derived and applied.

To the extent that high density comes about in the vertical dimension, this approach is essentially a discrete approximation to the explicit inclusion of height in the model. With the aggregate-based models used by Langford et al. and Lo, only this less precise discrete approach is feasible. But with pixel-based models there is no such restriction (see Table 10.1). Considering that some judgements about height would in any case probably be used to inform the multi-stratum classification, I am inclined to think that the more direct approach would be just as feasible to apply, and more promising.

A "pie in the sky" footnote is the fact that satellites exist with radar altimetry capability (for example European Space Agency ERS-1 and ERS-2), which in principle could measure the

heights of structures. In practice, these satellites have limited narrow swathe coverage and are designed principally for monitoring the sea surface rather than land.

Potentially more feasible nascent remote sensing methodologies for direct estimation of building heights are the analysis of shadows (Shettigara and Sumerling, 1998) and stereo image matching (Kim and Muller, 1998).

10.6.3 Other aspects

Some other applied lines of enquiry might include:

- exploration of the sensitivity of the results obtained to changes of scale or sensor resolution, by simulation and/or using data from other multispectral sensors, such as MSS, SPOT multispectral and new generation higher resolution sensors (in the manner of Cushnie and Atkinson, 1985; Cushnie, 1986; Ng, 1990, and in the light of the decision frameworks established and applied by Woodcock and Strahler, 1987; Chavez, 1992; Atkinson and Curran 1997);
- combined use of more than one sensor, such as TM plus SPOT panchromatic, or TM plus night-time illumination (Sutton et al., 1997);
- exploration of the sensitivity of the procedure and the trade-offs involved with different classification schemes, different thresholding levels, different smoothing window sizes, etc;
- application of the procedure in other non-Australian cultural settings;
- applicability of the general approach to synthetic aperture radar (SAR) imagery. Some research was carried out in the 1980s into the use of radar imagery for population estimation, using selective imagery generated in Space Shuttle experiments (Harrison and Jupp, 1989; Henderson and Xia, 1997). However this form of imagery has only recently become more generally available from orbital platforms.

Some more theoretical investigations might include:

- modelling with simulated data in order to better understand the relationships between population and the physical properties of different surfaces and materials, which underlie the central linear models identified in this study (Forster, 1980b; Curran, 1986);
- probabilistic assessments of accuracy such as confidence intervals for totals, based on empirical distributions from repeated sampling or resampling methods such as bootstrapping (this would be very computationally intensive, and perhaps very image-dependent);

- the use of generalised linear models, in particular Poisson-based models and spatially correlated error structures.
- in the context of generalised linear modelling, development of an exact EM re-estimation algorithm.

10.7 APPLICATIONS

10.7.1 Direct use of the methodology: estimation of population

In principle, the methodology established in this study, perhaps incorporating some of the embellishments suggested in the previous two sections, could be directly applied for the purposes of estimating population and mapping population distribution in countries where the demographic infrastructure is not well developed (Polle, 1996). It has been demonstrated in this study (if demonstration were needed) that particular regression relationships, even when normalised, do not apply robustly across all geographic locations, seasons or cultural variations, even within a single nation. Such an exercise will always require an initial investment in ground truthing and calibration of the regression component, as well as the usual training aspect of the land cover/land use classification.

10.7.2 Indirect use of the methodology: hybrid methodologies

A less ambitious application than unassisted population estimation might be the use of remote sensing estimates to modify, update or disaggregate other estimates.

For example, in Australia intercensal estimates of resident population (e.r.p.) are currently available for SLAs and larger areas, and are updated annually with a lag of about 1 year. Using image differencing techniques for change detection (Jensen, 1982; Griffiths, 1988; Royer et al., 1988, Martin, 1989; Quarmby and Cushnie, 1989; Martin and Howarth, 1989), the changes in remote sensing population estimates between the last available e.r.p. date (or the last census date) and some later date could be used as an index which could in principle be applied to the last known “true” figure to provide updated estimates which could be more geographically flexible and more timely than those which are currently available.

In the context of GIS and multi-level analysis, CD populations are frequently either assigned to the centroid of each CD or assumed to be uniformly distributed across the CD, for such areal interpolation purposes as assigning population to grid squares, defining catchments or estimating populations exposed to environmental influences.

In the dasymetric approach of Langford et al. (Langford and Unwin, 1994; Fisher and Langford, 1995; Fisher and Langford, 1996) remote sensing imagery was used to classify pixels and hence

to geographically distribute the known census populations within census enumeration districts. A similar approach was used by Lo (1995) in one of his estimation models.

The same approach could be applied using estimates of individual pixel populations as derived in this study. Again, as in the discussion of height data versus multiple residential classifications (Section 10.6.2), from the perspective of pixel populations, the use of a discrete classification is equivalent to assuming a constant (average) pixel population within each class. Hence, whilst the remote sensing population estimates for individual pixels are probably not highly accurate individually, it is conjectured that an allocation of known CD populations to other areas based on these estimates would be an improvement over the assumption of either a single uniform distribution, or a different uniform distribution in each residential class.

To establish the validity of this conjecture, or of any dasymetric allocation, or indeed of any areal interpolation, ideally direct observations should be made at the micro level at which the re-allocations are made, in this case pixels. The indirect cross-validation performed by Langford et al. using grid squares with known population does not establish anything about the accuracy at the micro level. The same can be said about the CD totals used for validation in the present study. Nevertheless, the images of estimated population density exhibit some face validity in that little or no population is assigned to sports grounds, parks, waterways, major roads or commercial areas within residential CDs.

10.7.3 Back to Earth

It is contended that the methodology developed in this study could form the basis for operational procedures for estimating large regional populations using Landsat TM imagery supplemented by population censuses of relatively small areas.

From a different perspective, when combined with data from other sources such as censuses, it provides a more finely honed alternative to existing methods of areal interpolation and dasymetric mapping.

But whilst both types of application are possible in principle, practical questions remain about whether the levels of accuracy obtained are sufficient to be of any real use to anyone; about the extent to which it is possible to improve accuracy, both in principle and in practice; and about whether the procedures would be cost effective. The question of the operationally feasibility and utility of remote sensing methods for population estimation remains open for consideration by geographic and demographic practitioners.

Landsat TM Images

Image 1. Ballarat Study Area
Quasi natural colour RGB - TM bands 3, 2, 1

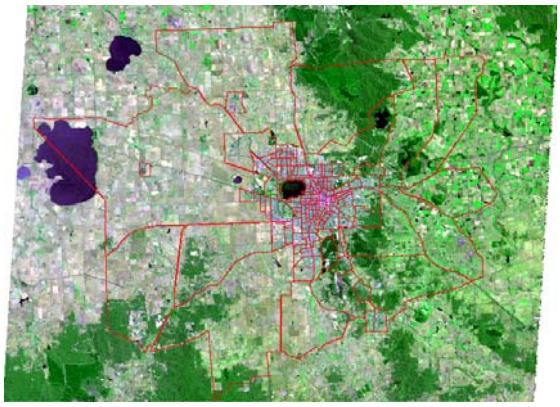


Image 2. Ballarat Study Area
Green-enhanced quasi natural colour RGB
TM bands 3, 2+4, 1
1986 Census Collection District boundaries overlaid

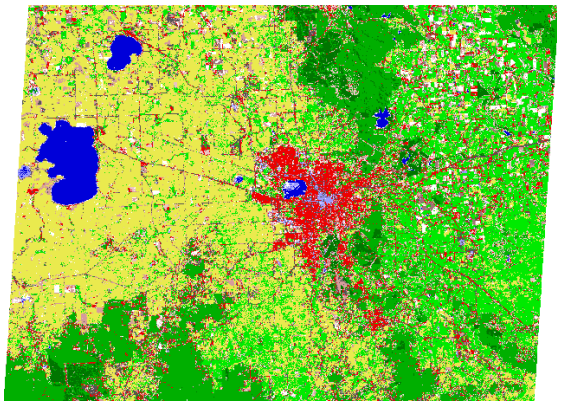


Image 3. Ballarat Study Area
2 class MLC based on 6 TM bands
(Red = residential)

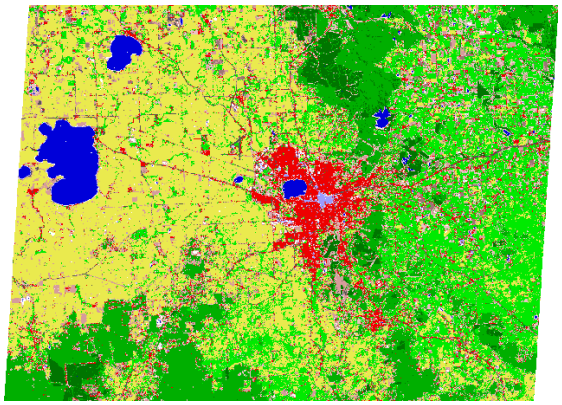


Image 4. Ballarat Study Area
12 class MLC based on 25 spectral and spatial transformations of 6 TM bands
(Red = residential)

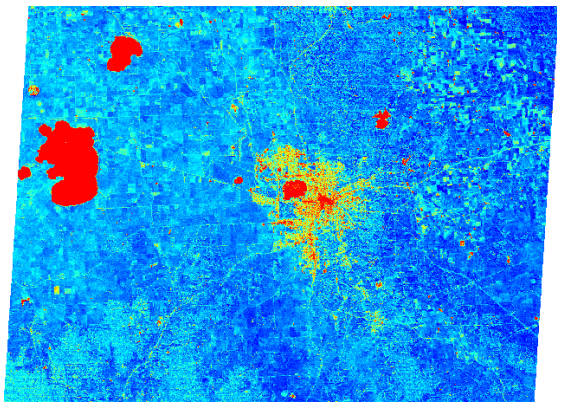


Image 5. Ballarat Study Area
Difference to sum ratio of TM bands 1 and 5
(Pseudocolour: red = high, blue = low)

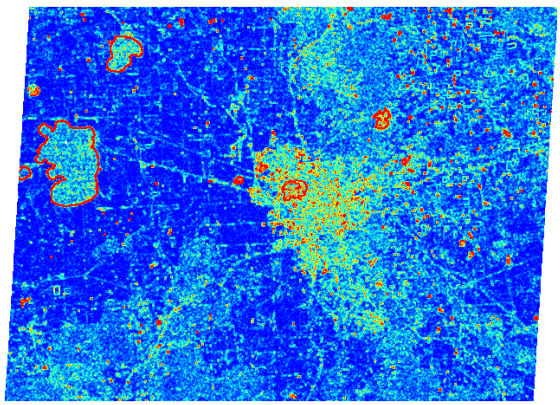


Image 6. Ballarat Study Area
Spatial variability in difference to sum ratio of TM bands 1 and 5
Standard deviation over a 3 pixel × 3 pixel neighbourhood
(Pseudocolour: red = high, blue = low)

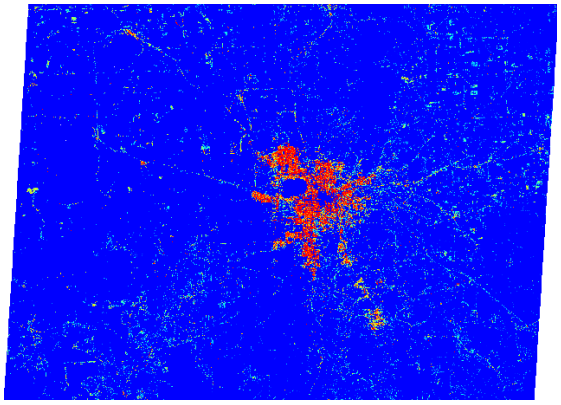


Image 7. Ballarat Study Area
Estimated population density with classification and regression based on 6 TM bands.
(Pseudocolour: red = high, blue = low)

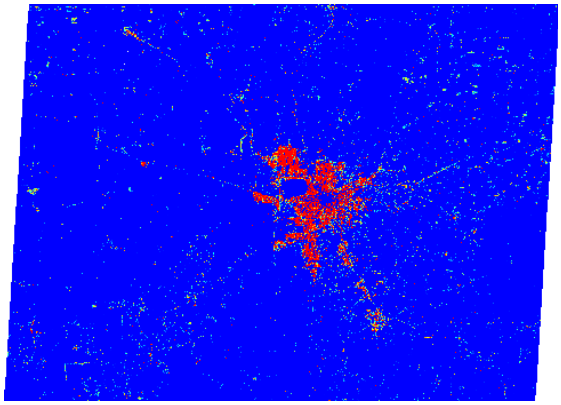


Image 8. Ballarat Study Area
Estimated population density with classification and iterated regression based on 6 TM bands.
(Pseudocolour: red = high, blue = low)

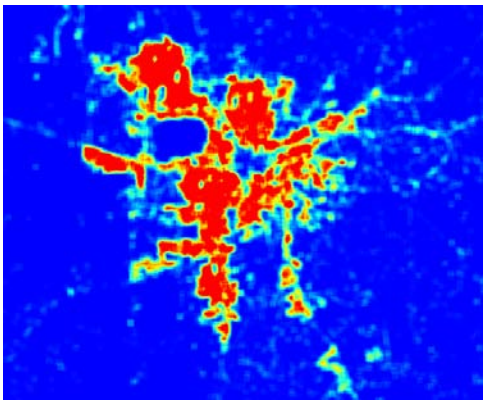


Image 9. Ballarat urban area
Estimated average population density based on 6 TM bands with iterated regression, smoothed with a mean filter over a 7 pixel × 7 pixel neighbourhood.
(Pseudocolour: red = high, blue = low)

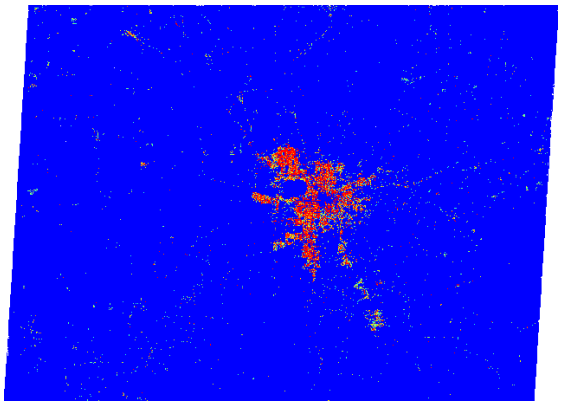


Image 10. Ballarat Study Area
Estimated population density based on 6 TM bands with iterated regression, low density contextual reclassification and high density enhancement.
(Pseudocolour: red = high, blue = low)

Image 11. Ballarat urban area
Quasi natural colour RGB - TM bands 3, 2, 1

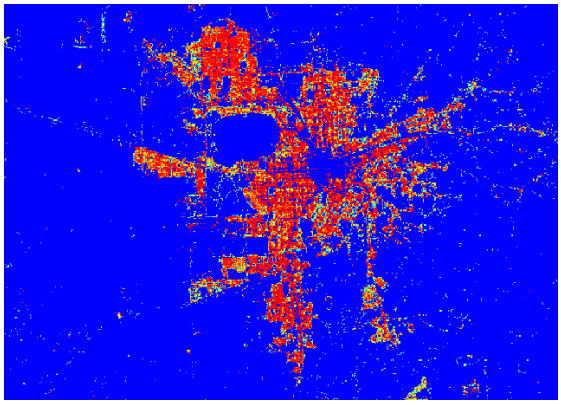


Image 12. Ballarat urban area
Estimated population density based on 6 TM bands with iterated regression, low density contextual reclassification and high density enhancement.
(Pseudocolour: red = high, blue = low)

Image 13. Geelong Study Area
Quasi natural colour RGB - TM bands 3, 2, 1

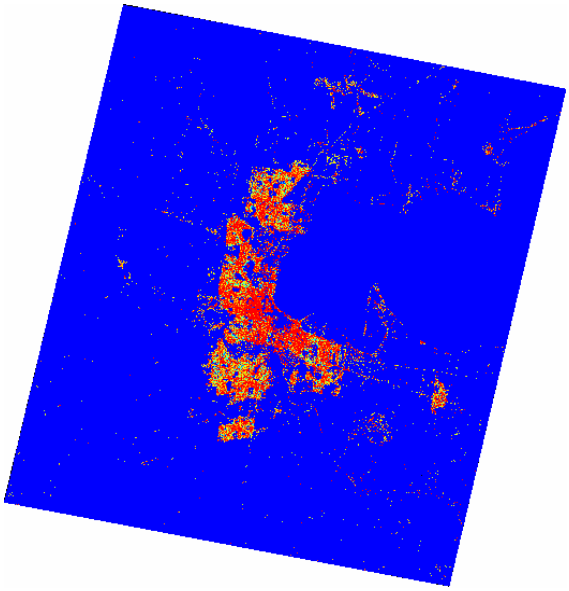


Image 14. Geelong Study Area
Estimated population density based on locally trained classification and normalised Ballarat regression model
(Pseudocolour: red = high, blue = low)

Image 15. Ballarat Study Area (1994)
Quasi natural colour RGB - TM bands 3, 2, 1

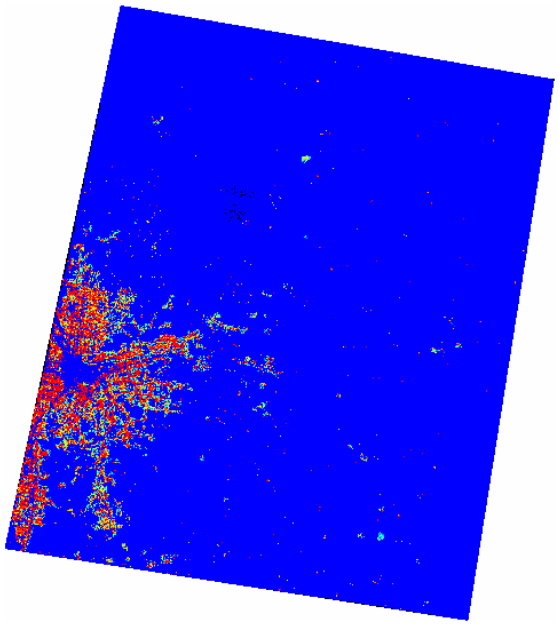


Image 16. Ballarat Study Area (1994)
Estimated population density based on locally trained classification and regression
(Pseudocolour: red = high, blue = low)

Image 17. Adelaide Study Area
Quasi natural colour RGB - TM bands 3, 2, 1

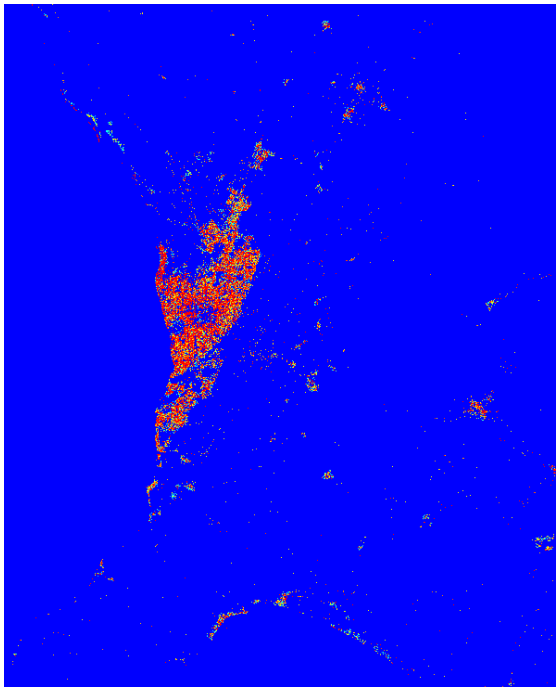


Image 18. Adelaide Study Area
Estimated population density based on locally trained classification and regression
(Pseudocolour: red = high, blue = low)

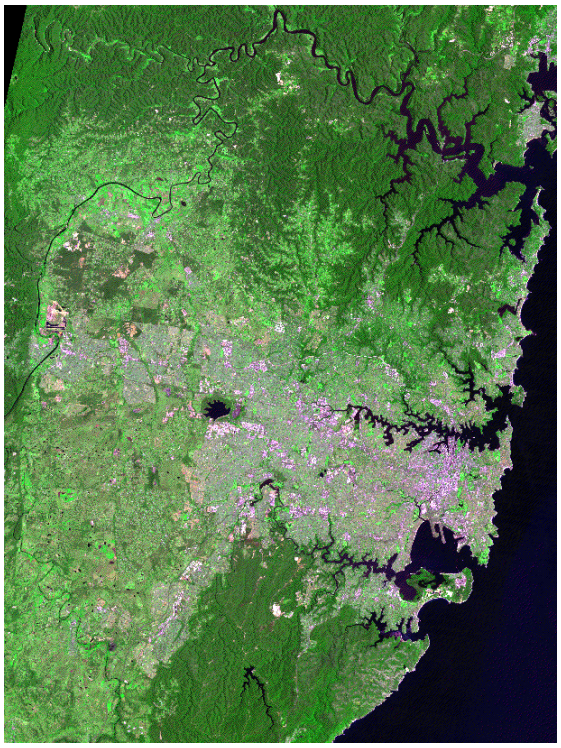


Image 19. Sydney Study Area
Green-enhanced quasi natural colour RGB
TM bands 3, 2+4, 1

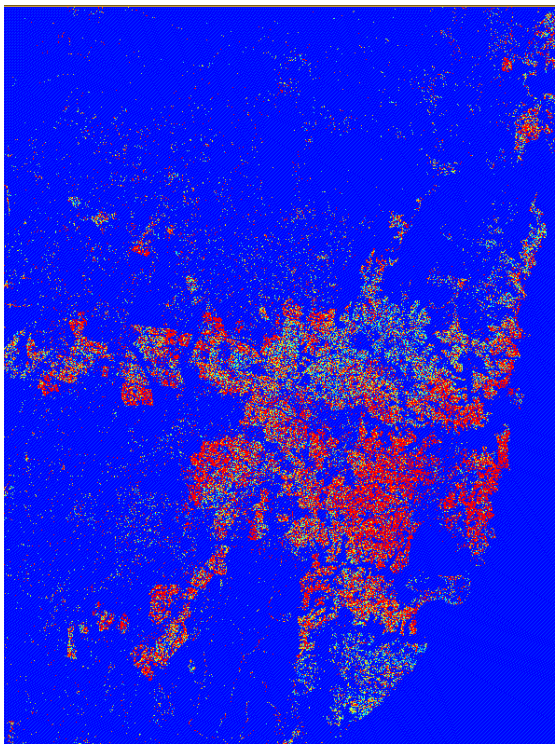


Image 20. Sydney Study Area
Estimated population density based on locally trained classification and regression
(Pseudocolour: red = high, blue = low)

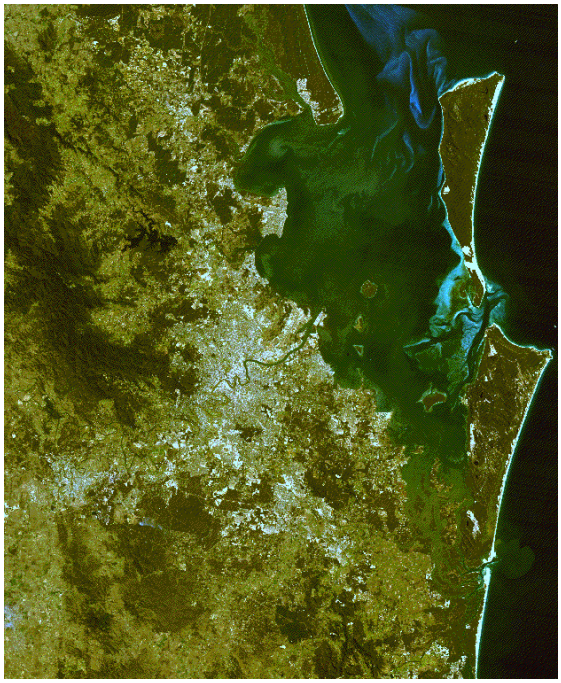


Image 21. Brisbane Study Area
Quasi natural colour RGB - TM bands 3, 2, 1

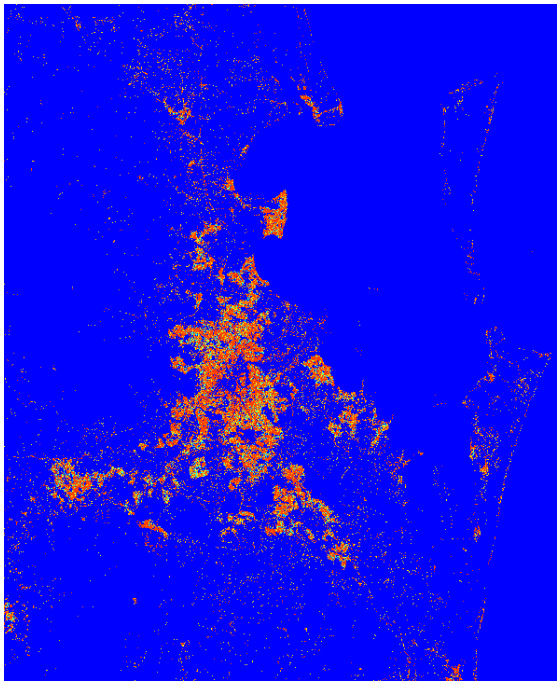


Image 22. Brisbane Study Area
Estimated population density based on locally trained classification and normalised Ballarat regression model
(Pseudocolour: red = high, blue = low)

Image 23. Kalgoorlie Study Area
Quasi natural colour RGB - TM bands 3, 2, 1

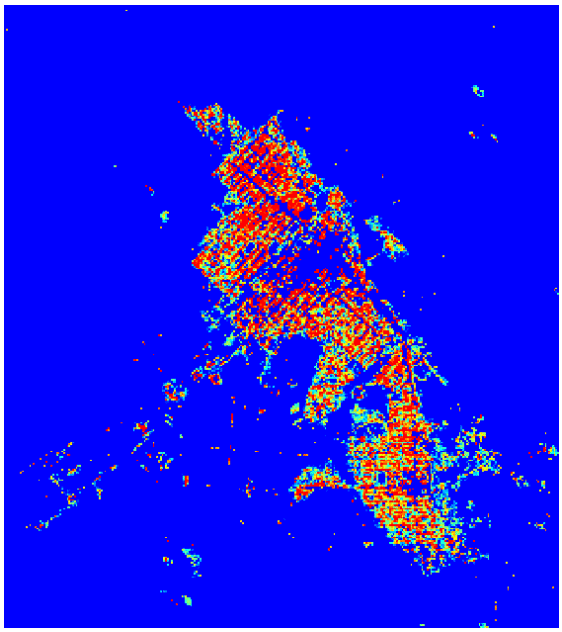


Image 24. Kalgoorlie Study Area
Estimated population density based on locally trained classification and normalised Ballarat regression model
(Pseudocolour: red = high, blue = low)

Appendices

Appendix A

Transformations from RGB to HSI co-ordinates

(Chapters 2, 4 and 5)

RGB co-ordinates	Normalised RGB co-ordinates	HSI co-ordinates
R = red	$r = R/(R+G+B)$	H = hue
G = green	$g = G/(R+G+B)$	S = saturation
B = blue	$b = B/(R+G+B)$	I = intensity

Triangular HSI co-ordinates

$$I = \tfrac{1}{3}(R+G+B)$$
$$S = 1 - \min(r,g,b,)$$
$$\cos H = \frac{(2r-g-b)}{\sqrt{6((r-\frac{1}{3})^2+(g-\frac{1}{3})^2+(b-\frac{1}{3})^2)}}$$

Cylindrical HSI co-ordinates

$$I = \tfrac{1}{3}(R+G+B)$$
$$S = \sqrt{v_1^{\,2}+v_2^{\,2}}$$
$$\tan H = v_1/v_2$$

where

$$v_1 = \tfrac{1}{\sqrt{6}}(R+G-2B)$$
$$v_2 = \tfrac{1}{\sqrt{2}}(R-G)$$

Rectangular HSI co-ordinates

$$I = \max(R,G,B)$$
$$S = 1 - \min(R,G,B)/\max(R,G,B)$$
$$H = \begin{cases} 60(1+(G-B)/R); R = \max(R,G,B) \\ 60(3+(B-R)/G); G = \max(R,G,B) \\ 60(5+(R-G)/B); B = \max(R,G,B) \end{cases}$$

Note:

The H co-ordinate is an angular measure which may range through a full cycle. Both the scale and the zero point are arbitrary. When computing triangular or cylindrical co-ordinates using the inverse cosine and inverse tangent functions, which are not defined over the whole range, appropriate adjustments have to be made quadrant by quadrant.

Appendix B

Development and Comparative Evaluation of Texture Measures by Simulation

Of the four types of texture measures used by Hsu (1978) and discussed in Section 2.5.3, the deviation- and proportion-based measures (types (i) and (iii)) take no account of the spatial pattern within the window. This is the case regardless of the range or degree of quantisation of the data. For this reason, such measures were considered unlikely to be adequate for the task of differentiating between characteristically amorphous residential areas and a number of non-residential features with clear geometric structures.

Considering the pairwise-difference measures and the wave-form measures, the former have the advantage of being simpler and faster to compute. For this reason, it was decided to investigate the performance of several measures of this type by applying them to simulated data, based on a number of geometric patterns characteristically associated with residential and non-residential features, but also degraded by random noise.

The aim was to find one or more pairwise-difference based texture measures with the capacity to differentiate between, on the one hand, the amorphous pattern characteristic of residential land-use, and on the other hand, a number of geometric patterns associated with common non-residential features, which exhibit a similar level of spatial variance. There was an intuitive expectation that such a measure should produce lower scores for the less "busy" geometric patterns than for an amorphous pattern.

Window size

As window size is increased, so too are edge effects, with pixels between two adjacent classes likely to be rejected. Hsu (1978) reported that edge effects are substantial even with a 5×5 window. Also, processing time increases as the square of the window size or faster. Nevertheless, it was decided to model both 3×3 and 5×5 windows, to test whether these disadvantages were offset by substantially improved discrimination using the larger window.

Simulated data

Both 3×3 and 5×5 windows were used. Twelve test patterns were used, eight based on geometric patterns representing non-residential features, and four purely random representing residential land-use. The patterns for a 3×3 window are shown in Figure 3. The ninth pattern in each case formed the constant datum which was the basis of the four random test patterns.

All of the test data simulated was in the range (0,1). Initially, the low and high values were set to 0.25 and 0.75 respectively, except for the last pattern, where the datum was set at 0.5.

On the first eight patterns, there was then superimposed random variation, uniformly distributed with mean 0 and a maximum magnitude r_{max}, which could be varied from 0 to 0.25. Thus the data could be made purely deterministic with values of 0.25 and 0.75, or it could be made to vary randomly over the whole range from 0 to 1, whilst retaining to a controlled degree the underlying geometric pattern.

Figure 2. Test patterns for a 3×3 window

Alt	VE	DE
H L H L H L H L H	L H H L H H L H H	L L H L H H H H H
VL	**DL**	**HVC**
L H L L H L L H L	L L H L H L H L L	L H L H H H L H L
DC	**Dot**	**Con**
H L H L H L H L H	L L L L H L L L L	H H H H H H H H H

Alt Alternating high and low values (chess board)
VE Vertical edge
DE Diagonal edge
VL Vertical line
DL Diagonal line
HVC Horizontal-vertical cross
DC Diagonal cross
Dot High central pixel, surrounded by low pixels
Con Constant - all pixels high

The ninth pattern was used as the basis of four patterns of random variation, each with a mean value of 0.5, but with four different amplitudes.

The four patterns and their amplitudes were:

Pattern	Amplitude
R	r_{max}
R25	0.25
R25+	0.25 + r_{max}
R50	0.50

The first of these represents an underlying homogeneity with a small amount of superimposed random variation. In the second case, the data ranges from 0.25 to 0.75, which is the same as for the geometric patterns with no added randomness. In the third case, the data ranges from 0.25-r_{max} to 0.75+r_{max}, which is the same as for the geometric patterns with added randomness. The fourth case represents the maximum possible random variation, across the whole range (0,1).

In the context of distinguishing random residential patterns of variation from other more geometric features, it was reasonable to assume that the overall range of variation in the residential areas would be as great as the high-low difference in the geometric features. On this basis, the third random pattern, R25+, was regarded as the most appropriate benchmark for comparison.

Texture measures

A total of 33 texture measures were examined, comprising eleven basic patterns of differences, with each being averaged in three ways - mean absolute difference (MA), mean squared difference (MS) and root mean squared difference (RMS). Of the eleven basic patterns, two were deviation-based. These were included for comparison, and to verify that they did fail to discriminate between the two types of pattern.

The eleven basic methods were characterised as follows:

Deviation measures:

MV	Deviation from the mean value of all pixels in the window
CP	Deviation from the central pixel value

Pairwise-difference measures:

AP	All pairs i.e. each pixel is compared with all others
HVnn	Nearest neighbours - horizontal & vertical directions
Dnn	Nearest neighbours - diagonal directions
HVDnn	Nearest neighbours - all directions
HV2nn	Second nearest neighbours - horizontal & vertical
D2nn	Second nearest neighbours - diagonal
HVD2nn	Second nearest neighbours - all directions
Pnnt	Nearest neighbours around the perimeter (trimmed mean)
MinD2nnPnnt	Minimum of D2nn and Pnnt

The first nine measures are self explanatory. The last two are discussed in detail below.

For the deviation measures, the form of the statistics is as follows:

$$MA = \frac{\sum_{i=1}^{N} |x_i - R|}{N\text{-}1}$$

$$MS = \frac{\sum_{i=1}^{N} (x_i - R)^2}{N\text{-}1}$$

$$RMS = \sqrt{MS}$$

where x_i = data value for pixel i

N = number of pixels in window = (window size)^2

R = reference value: either the mean or the central pixel value

For the pairwise-difference measures, the form of the statistics is as follows:

$$MA = \frac{\sum_{i=1}^{N} \sum_{j \in S_i} |x_i - x_j|}{P}$$

$$MS = \frac{\sum_{i=1}^{N} \sum_{j \in S_i} (x_i - x_j)^2}{P}$$

where x_i, x_j = data values for pixels i, j

S_i = set of j values which are paired with a particular i value

N = number of pixels in window = (window size)^2

P = total number of pairs included in the summation

Results

The above specifications were implemented using Turbo Pascal. For each window size, three runs of 100 simulations were undertaken, with r_{max} set to 0.00, 0.10 and 0.25. The criterion was the mean value of the 100 calculated statistics from the simulation run, expressed as a percentage of the "residential benchmark" figure for the R25+ pattern.

With no random noise, as was expected, the deviation based measures MV and CP are higher for all the geometric patterns than for the benchmark R25+ pattern, since the data from the

former are bimodal in the extreme. As an increasing degree of superimposed randomness was introduced this difference is reduced. With r =0.25, there is very little difference between the scores for most patterns, with either MV or CP.

Similar results were obtained for the measures AP, HVnn, HVDnn and HVD2nn. The Dnn measure clearly distinguished the Alt and DC patterns from the R25+ pattern, at all levels of randomness. The HV2nn measure offered further improvement, with clear discrimination of five of the eight geometric patterns. With D2nn, all but two of the geometric patterns were clearly differentiated from the random pattern, the exceptions being VE and DE - vertical and diagonal edges.

Having reached this point, it was decided to try to utilise the D2nn measure in conjunction with some other measure which could distinguish edges. What was required was not an edge detector per se, but rather a complementary measure which would produce a lower score for an edge than for a random pattern - what might be called an "edge pass filter".

An attempt was made to derive a suitable measure by essentially reversing the sign of a standard edge detector - the Sobel gradient operator (Abdou and Pratt, 1979). This approach was rejected for two related reasons:

(i) Because the Sobel operator is sensitive to the orientation of an edge, there is no obvious maximum value, nor is there a clear set of circumstances under which the maximum would occur. Hence the complementary measure has no clear minimum corresponding to the zero of a measure such as D2nn.

(ii) The Sobel operator is based on weighted sums of signed differences, and hence is not directly commensurate with the D2nn measure which is based on unsigned differences.

An alternative measure was formulated which was based on unsigned pairwise differences, which was not sensitive to the direction of an edge, and which had a clearly defined zero. It was however computationally somewhat more expensive than simple edge detectors. This measure is the "perimeter nearest neighbour trimmed average" (Pnnt). Its rationale is as follows. In the case of a pure edge, the differences between nearest neighbours on the perimeter of the window are zero except at the two edge crossings, where the differences are large. In the presence of superimposed random variation, the two edge crossings are still likely to produce larger differences than the other positions. The Pnnt statistic is found by taking the absolute differences between nearest neighbours on the perimeter of the window, discarding the two largest values (trimming), and calculating the average of the remaining values. This measure had the desired property of producing low values in the VE and DE cases.

The final statistic, (MinD2nnPnnt) which achieved the stated objective, is simply the smaller of D2nn and Pnnt. i.e. The value assigned to the central pixel is the lesser of:

- the average of the absolute values of the differences between 2nd nearest neighbours in the diagonal direction (which in the case of a 3×3 window is just the average absolute difference between pixels in diagonally opposite corners), and
- the average of all but the two largest of the absolute values of the differences between nearest neighbours around the perimeter of the window.

Appendix C

Implementation of Procedures for Chapters 3 - 6

(Methods for Chapters 8 & 9 are variants of these)

MAJOR STEP	DETAILS	ERMAPPER (or other) PROCEDURES	SEC.
Preparation: Ch 3			
Set up basic 6-band TM dataset.		IMPORT from tape subsetter via μBRIAN.	3.4
Calculate ground truth CD population and dwelling estimates.	Projections based on census and other available figures.	EXCEL spreadsheet	3.6
Geometric correction of image.	Remove earth rotation skew.	PASCAL	3.7
Co-register CD boundary co-ordinates	Two stages:		
	• Parametric corrections - origin, scale, aspect ratio, alignment.	PASCAL	
	• Polynomial transformation - least squares fit to ground control points.	ERMAPPER - pixel addresses of GCPs. EXCEL & MINITAB - fit polynomial by least squares.	3.7.5
		PASCAL - polynomial transformation.	3.7.6
Set up CD "regions" in 3 forms.	• Vector overlay. • Regions CD1-CD138. • One data band containing CD IDs 1-138.	IMPORT transformed co-ordinates. Edit co-ordinates into .ers header file. FORMULA: IF INREGION ... × 138	3.7.6
CD Regression **- without & with transformations: Ch 5**			
Calculate population and dwelling densities for each CD	Density = CD ground truth figure / CD area (number of pixels × pixel area)	PASCAL	5.1
Calculate CD aggregate measures for spectral variables	• Export dataset including all candidate variables plus the CD ID band. • In 3 stages calculate for each CD: - means of each band - variation measures for each band - means and variation measures for selected band transformations	PASCAL	5.1-5.5
Investigate models	Use stepwise multiple regression analysis to select best regression models for predicting population and dwelling counts. Examine residuals for patterns which might suggest approaches for improvement.	MINITAB	5.1-5.5

Pixel regression **– without transformations: Ch 6.1**			
Select best regression variables by statistical analysis of a sample of pixels.	• Export dataset including all candidate variables plus the band: - CD ID	PASCAL	6.1
	• Sample from all pixels. Assign ground truth population and dwelling estimates by dividing CD estimates by the number of pixels in the CD.	PASCAL	
	• Use stepwise multiple regression analysis to select best regression models for predicting population and dwelling counts. Examine residuals for patterns which might suggest approaches for improvement.	MINITAB	
Assign population and dwelling estimates to each pixel in full image.	New data band:- use chosen regression model	FORMULA	6.1
Test procedure by estimating CD population and dwelling counts.	• Obtain CD pop. and dwelling estimates Export population and dwelling estimates plus the CD ID band. Sum the estimates for each CD.	PASCAL	6.1
	• Compare with ground truth figures for each CD. Calculate correlation, mean squared error, percentage errors.	MINITAB	
Pixel classification and regression **- without transformations: Ch 6.2-**			
Set up supervised classification training sets.	Twelve <u>regions:</u> residential others	Chosen CDs - edit .ers header file . DEFINE REGION command	6.2
Carry out classification of full image.	• Maximum likelihood classification: → 10 regions each represented by a classification overlay.	CLASSIFICATION algorithm	
	• One data band coded with the 10 levels of classification (simpler to incorporate with regression equations into a single population and dwelling density estimation algorithm).		

Select best regression variables by statistical analysis of a sample of residential pixels.	• Export dataset including all candidate variables plus the two bands: - classification - CD ID	PASCAL	
	• Delete non-residential pixels and sample from those remaining. Assign ground truth population and welling estimates by dividing CD estimates by the number of pixels in the CD classified as residential.	PASCAL	
	• Use stepwise multiple regression analysis to select best regression models for predicting population and dwelling counts. Examine residuals for patterns which might suggest approaches for improvement.	MINITAB	
Assign population and dwelling estimates to each pixel in full image.	New data band: residential - use chosen regression model commercial - use a chosen value all else - set to zero.	FORMULA: IF...THEN...ELSE	
Test procedure by estimating CD population and dwelling counts.	• Obtain CD pop. and dwelling estimates Export population and dwelling estimates plus the CD ID band. Sum the estimates for each CD.	PASCAL	
	• Compare with ground truth figures for each CD. Calculate correlation, mean squared error, percentage errors.	MINITAB	
Pixel classification and regression - with transformations: Ch 6.			
Set up supervised classification training sets.	Twelve <u>regions:</u> residential others	Chosen CDs - edit .ers header file . DEFINE REGION command	
Calculate derived variables.	New data bands: band ratios, filters, intensities, texture measures, etc	FORMULA KERNEL OUTPUT TO DATASET.	
Select best classification variables by statistical analysis of a sample of pixels from the training regions.	• New data band coded with 10 training sets, else NULL • Export dataset including this band. Delete NULL pixels and sample from those remaining. • Use stepwise discriminant analysis to select best variables for classification.	FORMULA: IF INREGION...x 10 PASCAL SPSS	

Carry out classification of full image.	• Maximum likelihood classification: → 10 regions each represented by a classification overlay. • One data band coded with the 10 levels of classification (simpler to incorporate with regression equations into a single population and dwelling density estimation algorithm).	CLASSIFICATION algorithm	
Select best regression variables by statistical analysis of a sample of residential pixels.	• Export dataset including all candidate variables plus the two bands: - classification - CD ID	PASCAL	
	• Delete non-residential pixels and sample from those remaining. Assign ground truth population and welling estimates by dividing CD estimates by the number of pixels in the CD classified as residential.	PASCAL	
	• Use stepwise multiple regression analysis to select best regression models for predicting population and dwelling counts. Examine residuals for patterns which might suggest approaches for improvement.	MINITAB	
Assign population and dwelling estimates to each pixel in full image.	New data band: residential - use chosen regression model commercial - use a chosen value all else - set to zero.	FORMULA: IF...THEN...ELSE	
Test procedure by estimating CD population and dwelling counts.	• Obtain CD pop. and dwelling estimates Export population and dwelling estimates plus the CD ID band. Sum the estimates for each CD.	PASCAL	
	• Compare with ground truth figures for each CD. Calculate correlation, mean squared error, percentage errors.	MINITAB	
Application of algorithms to another geographical area: Ch 7			
Prepare secondary test image and ground truth data.	As for Chapter 3.		
CD Aggregate method			7.1
Calculate population and dwelling densities for each CD	Density = CD ground truth figure / CD area (number of pixels × pixel area)	PASCAL	

Calculate CD aggregate measures for spectral variables	• Export dataset including all candidate variables plus the CD ID band. • Calculate for each CD: - means of each band - variation measures for each band - means and variation measures for selected band transformations	PASCAL	
Calculate remote sensed estimates of population and dwelling densities for each CD.	Use chosen regression models	MINITAB	
Evaluate performance of models.	Compare with ground truth figures for each CD. Calculate correlation, mean squared error, percentage errors.	MINITAB	
Pixel-based classification/regression method			7.2
Calculate derived variables.	New data bands: band ratios, filters, intensities, texture measures, etc.	FORMULA KERNEL OUTPUT TO DATASET.	
Carry out classification of full image.	Use the classification structure determined by the analysis of the primary image.	FORMULA	
Assign population and dwelling estimates to each pixel in full image.	New data band: residential - use chosen regression models commercial - use a chosen value all else - set to zero.	FORMULA: IF...THEN...ELSE	
Evaluate procedure by estimating CD population and dwelling counts.	• Obtain CD pop. and dwelling estimates Export population and dwelling estimates plus the CD ID band. Sum the estimates for each CD.	PASCAL	
	• Compare with ground truth figures for each CD. Calculate correlation, mean squared error, percentage errors.	MINITAB	

Appendix D

Collection District resident population and dwelling count estimates as at 14/2/88

The following spreadsheet excerpts show the calculations discussed in Section 3.6, for population and dwellings respectively, for the first 5 of the 138 Census Collection Districts (CDs) in the primary study area (Ballarat Statistical District). Each row corresponds to one CD. Similar calculations were performed for the 225 Census Collection Districts (CDs) in the secondary study area (Geelong Statistical District).

The methodology is as follows.

1. Population

CD population data is only available as raw counts by place of enumeration from 5-yearly censuses - in this case data from 1981 and 1986 was used (the most recently available at the time of calculation in 1991). Statistical Local Area (SLA) data is available from the censuses and from the ABS estimated resident population (e.r.p.) series.

The procedure essentially involves two phases.

Firstly, the 1986 CD e.r.p.s were estimated by comparing the 1986 SLA e.r.p.s with the corresponding 1986 SLA census counts, then applying the resulting differential to the CD counts. (See column 23 of spreadsheet.)

Secondly, the intercensal rate of population change for each CD was compared with that of its SLA. This differential was then used, together with the annual SLA e.r.p. figures, as a basis for extrapolating the e.r.p. of each CD beyond 1986. (See columns 24 and 27 of spreadsheet.)

For convenience of reference, the CDs have been numbered in alphabetical order of SLA, and in ABS field code order within each SLA. These sequence numbers appear in columns 14 and 29.

The columns of the population spreadsheets are as follows:

0 Statistical Local Area (SLA) name
1 SLAC81 SLA census count 30/6/81
2 SLAC86 SLA census count 30/6/86
3 SLAE81 SLA estimated resident population (e.r.p.) 30/6/81
4 SLAE86 SLA e.r.p. 30/6/86
5 SLAE87 SLA e.r.p. 30/6/87
6 SLAE88 SLA e.r.p. 30/6/88
7 Ratio SLAC86/SLAC81 SLA census count multiplier 81→86
8 Ratio SLAE86/SLAE81 SLA e.r.p. multiplier 81→86
9 Ratio SLAE81/SLAC81 SLA ratio of e.r.p. to census count 81

10 Ratio SLAE86/SLAC86 SLA ratio of e.r.p. to census count 86

11 Ratio SLAE87/SLAE86 SLA e.r.p. multiplier 86→87

12 Ratio SLAE88/SLAE87 SLA e.r.p. multiplier 87→88

13 (Ratio 12)$^{0.63}$ SLA e.r.p. multiplier 30/6/87→14/2/88

This period is 0.63 of one year.

14 CD sequence number

15 CD field code 81

16 CDC81 CD census count 30/6/81

CDs with an asterisk (*) in the field code column did not exist in 1981. Whilst CD boundaries are changed as little as possible from census to census, some changes are necessary, the most common being the splitting of a CD into 2 or more CDs where large population increases have occurred. In the study area, 18 new CDs were created for the 1986 census in this way. In these cases, actual 1981 counts for CDs which were subsequently split, were distributed amongst their constituent 1986 CDs in the same proportions as the 1986 counts. In this way imputed 1981 counts were obtained for both the new CDs and the residual reduced CDs. In the absence of actual 1981 data for these areas, the implicit assumption has been made that growth rates were equal in all constituent parts of a split CD.

17 CD field code 86

18 CDC86 CD census count 30/6/86

19 CD area (sq.km.)

20 Ratio c18/c19 Average population density 30/6/86 (persons/sq.km.)

21 Ratio CDC86/CDC81 CD census count multiplier 81→86

22 Ratio c21/c7

$$= \frac{\text{CD census count multiplier } 81 \rightarrow 86}{\text{SLA census count multiplier } 81 \rightarrow 86}$$

This compares the growth rate of each CD during the period 30/6/81 to 30/6/86 with that of the SLA in which it is located.

23 CDE86 CD estimated resident population (e.r.p.) 30/6/86 = CDC86 × c10

Obtained by multiplying the 1986 census count for each CD by the ratio of e.r.p to census count for the SLA in which the CD is located. This assumes that this ratio is constant across all CDs in the SLA.

24 CDE288U Preliminary (unadjusted) CD e.r.p. as at 14/2/88

= CDE86 × ratio11 × ratio13 × ratio22$^{0.326}$

Obtained by applying to the CD e.r.p. as at 30/6/86, the SLA e.r.p. multipliers for 86→87 and 30/6/87→14/2/88, adjusted for each individual CD by the CD/SLA growth rate ratio.

Because this last ratio is based on a five-year period, it is raised to the power 0.326 (=1.63/5) to adjust it to the 1.63 year period under consideration.

This procedure assumes that the relativities amongst CD growth rates observed during the period 81-86, remain unchanged through to 14/2/88.

Regardless of the validity of this assumption, it can be shown that this procedure has a small systematic bias towards overestimating the CD populations. A final damping adjustment was made by rescaling so that sum of the CD estimates for each SLA corresponds to the overall SLA estimate.

25 SLATOT Sum of the preliminary CD estimates for each SLA

26 SLAE288 SLA e.r.p. as at 14/2/88 = SLAE87 × c13

Obtained by multiplying the SLA e.r.p. as at 30/6/87 by the multiplier for the period 30/6/87→14/2/88.

27 CDE288A Final (adjusted) CD e.r.p. as at 14/2/88

= CDE288U × c26/c25

$$= \text{Preliminary CD e.r.p.} \times \frac{\text{E.r.p. for the whole SLA}}{\text{Sum of preliminary CD e.r.p.s for the SLA}}$$

This adjustment ensures that the final estimates sum to the correct SLA figure. The ratio c25/c26 ranges from 1.0006 (Buninyong) to 1.0098 (Sebastopol), indicating that the extent of the bias in the basic estimation procedure is less than 1% in each case.

2. Dwellings

All dwelling calculations were based on CD figures, including population estimates calculated in the population spreadsheet described above.

As defined by ABS, a *household* is a person living alone, or two or more persons who live and eat together, in private residential accommodation, which includes houses, flats townhouses etc. but excludes hotels, motels, boarding houses, hospitals, staff quarters etc. A *private dwelling* is the premises occupied by a household.

The columns of the dwellings spreadsheets are as follows:

0 CD sequence number

1 Occupied private dwellings 30/6/86

2 Caravans etc. in caravan parks 30/6/86

3 Total occupied dwellings 30/6/86

= c1+c2

4 Unoccupied private dwellings 30/6/86

5 Total private dwellings 30/6/86

= c3+c4

6 Separate houses 30/6/86

7 Percentage of non-separate-house structures 30/6/86

$$= \frac{c5 - c6}{c5} \times 100$$

8 Estimated resident population (e.r.p.) 30/6/86

9 Estimated resident population (e.r.p.) 14/2/88

These two figures were obtained from columns 23 and 27 of the population spreadsheet described above.

10 Ratio c9/c8 Population multiplier 30/6/86→14/2/88

11-15 These are the estimates for 14/2/88 corresponding to columns 1-5. They were obtained by multiplying columns 1-5 by ratio 10. This assumes that the occupancy ratio (the ratio of population to number of private dwellings) and the proportions in each category of dwelling remain constant.

16 Percentage of non-separate-house structures 14/2/88

This was set equal to c7, again assuming that the proportion of such dwellings had not changed.

Sample of Ballarat Ground Truth Population Calculations

SLA name	SLA C81	SLA C86	SLA E81	SLA E86	SLA E87		2/1	4/3	3/1	4/2	5/4	6/5	12^0.63
	1	2	3	4	5	6	7	8	9	10	11	12	13
Ballaarat	35681	34806	36700	36790	36860	36830	0.9755	1.0025	1.0286	1.0570	1.0019	0.9992	0.9995
Ballaarat	35681	34806	36700	36790	36860	36830	0.9755	1.0025	1.0286	1.0570	1.0019	0.9992	0.9995
Ballaarat	35681	34806	36700	36790	36860	36830	0.9755	1.0025	1.0286	1.0570	1.0019	0.9992	0.9995
Ballaarat	35681	34806	36700	36790	36860	36830	0.9755	1.0025	1.0286	1.0570	1.0019	0.9992	0.9995
Ballaarat	35681	34806	36700	36790	36860	36830	0.9755	1.0025	1.0286	1.0570	1.0019	0.9992	0.9995

CD No.	CDC81		CDC86				18/16	21/7	CDE86 18*10	CDE288U 23*11*13*22^(0.2* 1.63)	SLATOT =5*13	SLA288 =24*26/25	CDE288A
14	15	16	17	18	19	20	21	22	23	24	25	26	27
1	70901	528	80901	433	3.09	140.1	0.8201	0.8407	458	433	36896	36841	432
2	70902	395	80902	377	0.23	1639.1	0.9544	0.9784	398	396	36896	36841	396
3	70903	432	80903	400	0.31	1290.3	0.9259	0.9492	423	416	36896	36841	416
4	70904	329	80904	303	0.35	865.7	0.9210	0.9441	320	315	36896	36841	314
5	70905	556	80905	561	0.21	2671.4	1.0090	1.0344	593	600	36896	36841	599

Sample of Ballarat Ground Truth Dwelling Number Calculations

CD SEQ	Occ 86	Cara 86	T Occ 86 1+2	Unocc 86	Tot h 86 1+2+4	Sep h 86	% NSH 86 (5-6)/5	CDE86
0	1	2	3	4	5	6	7	8
1	160	0	160	28	188	148	21.28	457.68
2	155	0	155	17	172	138	19.77	398.49
3	165	0	165	21	186	129	30.65	422.80
4	101	0	101	10	111	96	13.51	320.27
5	171	0	171	10	181	160	11.60	592.98

CDE88	9/8	Occ 88 1*10	Cara 88 2*10	T Occ 88 3*10	Unocc 88 4*10	Tot h 88 5*10	% NSH 88 =7
9	10	11	12	13	14	15	16
432.46	0.94	151	0	151	26	178	21.28
395.62	0.99	154	0	154	17	171	19.77
415.63	0.98	162	0	162	21	183	30.65
314.29	0.98	99	0	99	10	109	13.51
599.48	1.01	173	0	173	10	183	11.60

Appendix E

Collection District Aggregate-based Methods: Primary and Secondary Study Area Descriptive Statistics and Model Diagnostics

(Sections 4.3-4.6, 6.2)

Descriptive Statistics

Variable	N	Mean	Median	TrMean	StDev	SE Mean
b1	138	32.744	33.540	32.814	5.714	0.486
b2	138	21.516	21.750	21.592	3.010	0.256
b3	138	29.417	29.215	29.483	4.125	0.351
b4	138	56.491	57.175	56.731	5.729	0.488
b5	138	79.24	76.93	78.39	12.49	1.06
b7	138	39.729	39.705	39.782	4.902	0.417
Pop	138	573.8	538.5	565.6	258.1	22.0
pixcount	138	4943	358	2044	15816	1346
Popdens	138	1556.9	1500.0	1509.1	1105.5	94.1

Variable	Minimum	Maximum	Q1	Q3
b1	17.910	51.590	30.255	35.615
b2	13.290	31.620	20.217	22.965
b3	19.260	41.910	27.225	32.150
b4	37.240	69.810	54.152	60.112
b5	61.42	114.01	69.49	86.54
b7	26.460	52.830	36.508	42.620
Pop	20.0	1353.0	370.0	731.3
pixcount	103	132186	233	1038
Popdens	4.5	5142.2	639.8	2450.2

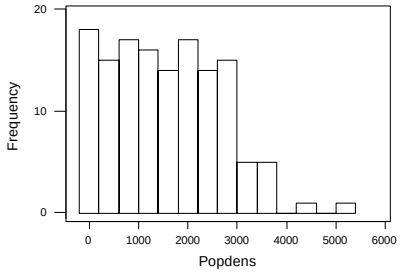
Descriptive Statistics: rural-urban split

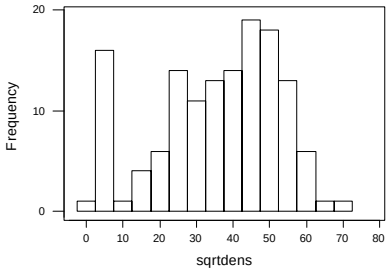
Variable	Rur/urb	N	Mean	Median	TrMean	StDev
b1	0	123	33.744	33.740	33.762	4.954
	1	15	24.54	22.34	24.49	4.98
b2	0	123	21.904	21.880	21.956	2.718
	1	15	18.335	17.440	18.352	3.477
b3	0	123	29.592	29.320	29.635	3.804
	1	15	27.98	26.08	27.96	6.16
b4	0	123	55.774	56.840	56.102	5.439
	1	15	62.37	62.48	62.47	4.67
b5	0	123	77.723	76.050	77.083	10.981
	1	15	91.68	84.37	92.13	17.06
b7	0	123	39.865	39.780	39.852	4.467
	1	15	38.61	35.44	38.77	7.73
Pop	0	123	579.1	539.0	571.7	263.0
	1	15	530.1	534.0	523.1	215.8
pixcount	0	123	1166	324	521	6282
	1	15	35913	27156	30532	30839
Popdens	0	123	1744.1	1733.0	1706.4	1023.4
	1	15	21.76	19.97	21.47	10.88

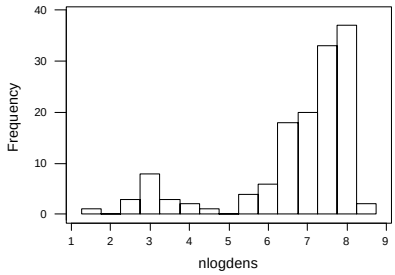
Variable	Rur/urb	SE Mean	Minimum	Maximum	Q1	Q3
b1	0	0.447	19.760	51.590	31.500	35.970
	1	1.28	17.91	31.90	20.31	29.56
b2	0	0.245	13.520	31.620	20.530	23.350
	1	0.898	13.290	23.150	15.490	22.400
b3	0	0.343	19.460	41.910	27.590	32.000
	1	1.59	19.26	36.86	22.80	34.76
b4	0	0.490	37.240	67.080	53.570	59.650
	1	1.21	53.71	69.81	59.48	65.89
b5	0	0.990	61.420	113.650	68.490	85.190
	1	4.40	63.41	114.01	79.91	108.91
b7	0	0.403	27.590	52.830	37.190	42.590
	1	2.00	26.46	48.71	33.00	47.06
Pop	0	23.7	20.0	1353.0	375.0	735.0
	1	55.7	254.0	898.0	337.0	729.0
pixcount	0	566	103	69886	214	804
	1	7963	9589	132186	18669	46568
Popdens	0	92.3	16.0	5142.2	844.2	2555.6
	1	2.81	4.49	42.87	13.12	31.38

Correlations

	Popdens	b1	b2	b3	b4	b5
b1	0.339					
b2	0.179	0.950				
b3	-0.091	0.767	0.919			
b4	-0.395	-0.474	-0.226	0.016		
b5	-0.515	-0.042	0.251	0.577	0.716	
b7	-0.120	0.585	0.783	0.918	0.266	0.757

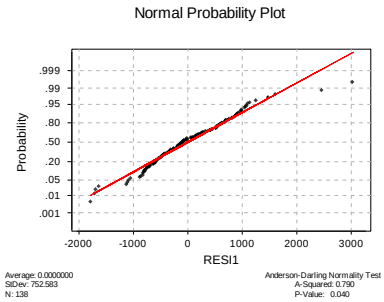


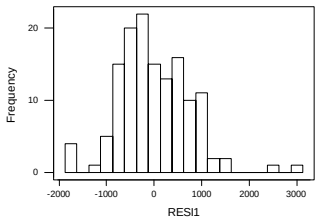


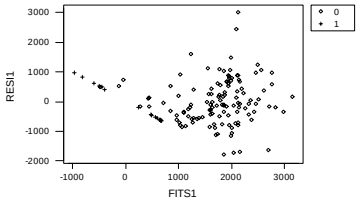


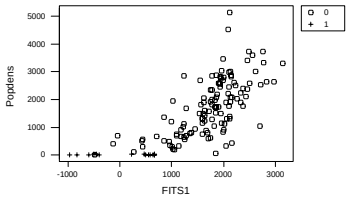
Plots pertaining to Table 4.2 row 1 Ballarat population density

Untransformed

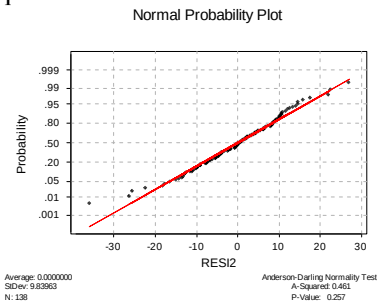


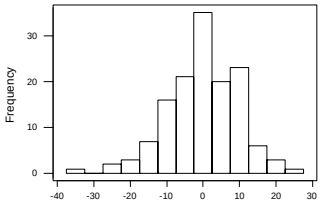


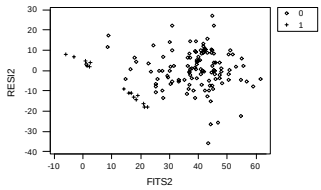


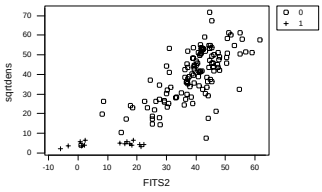


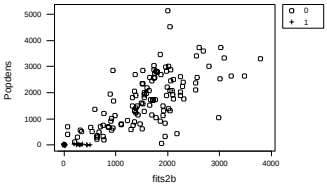
Square root transformation



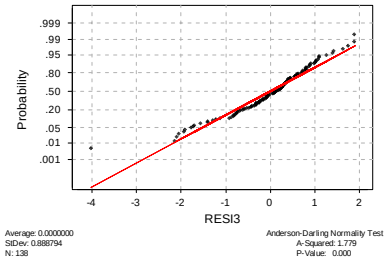


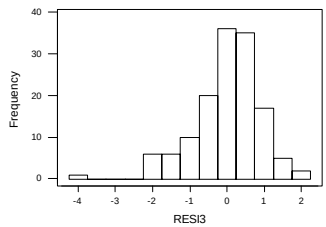


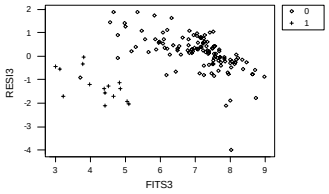


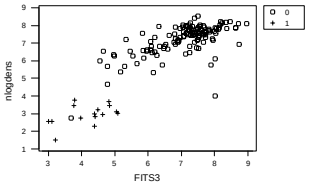


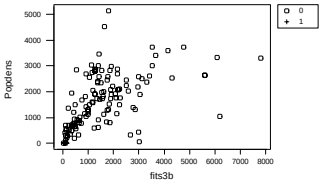
Logarithmic transformation





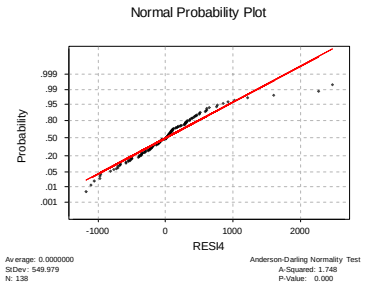


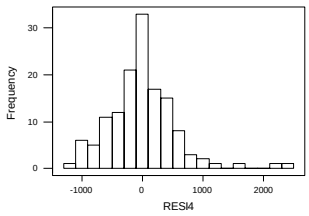


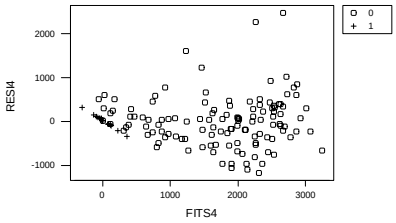


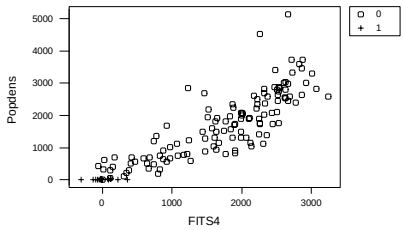
Plots pertaining to Table 4.2 row 5 Ballarat population density

Untransformed

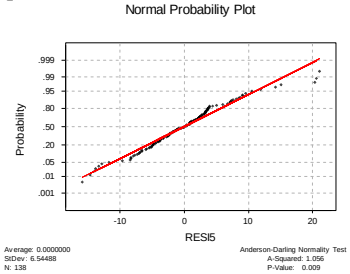


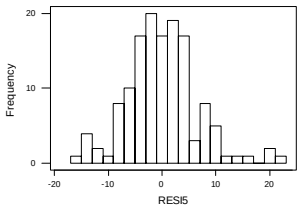


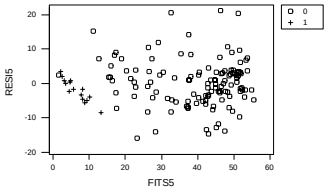


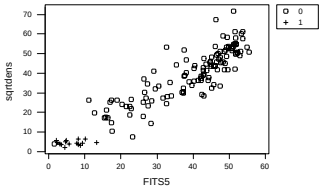


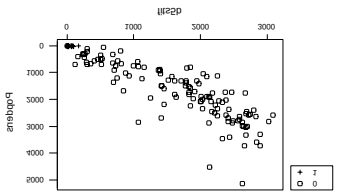
Square root transformation



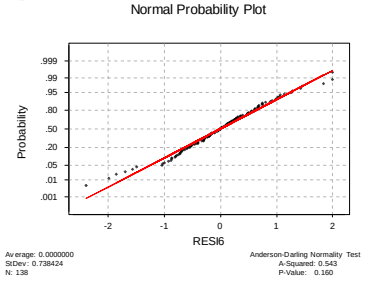


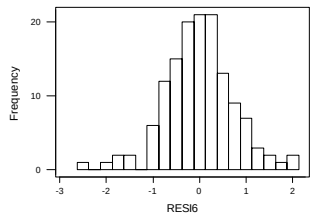


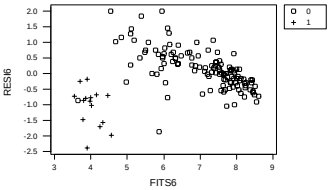


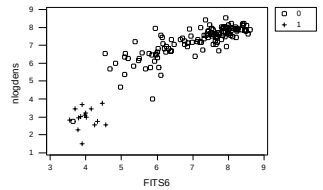


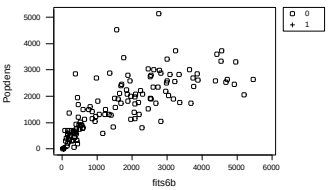
Logarithmic transformation





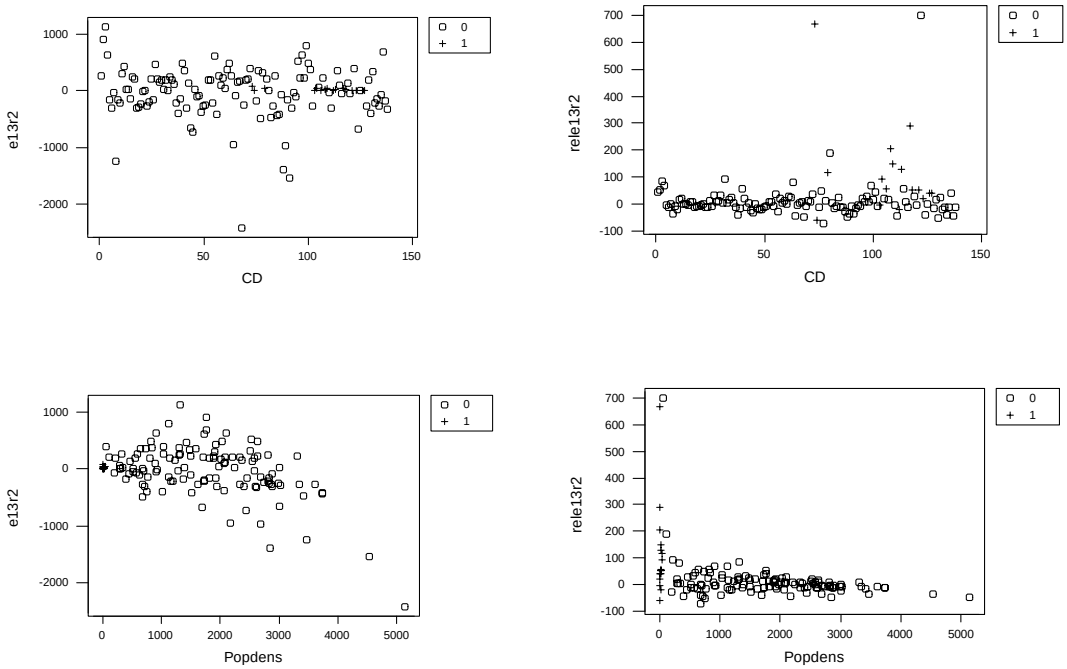






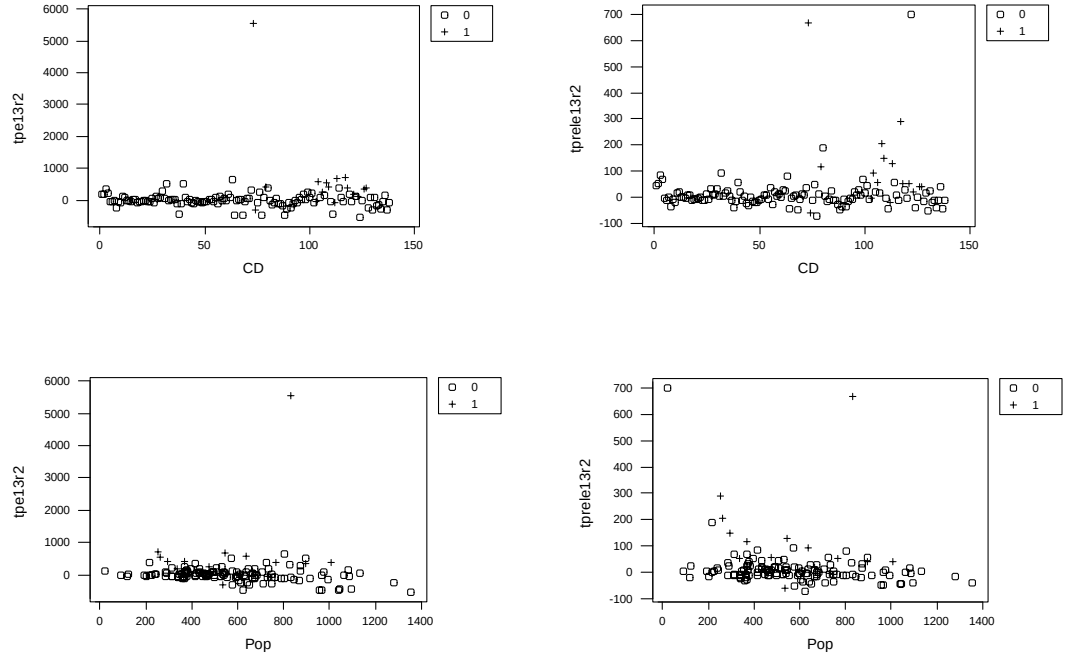
Plots pertaining to Table 4.10 Ballarat population density model 6

Absolute and relative errors in estimated population density by CD number and by ground truth population density.



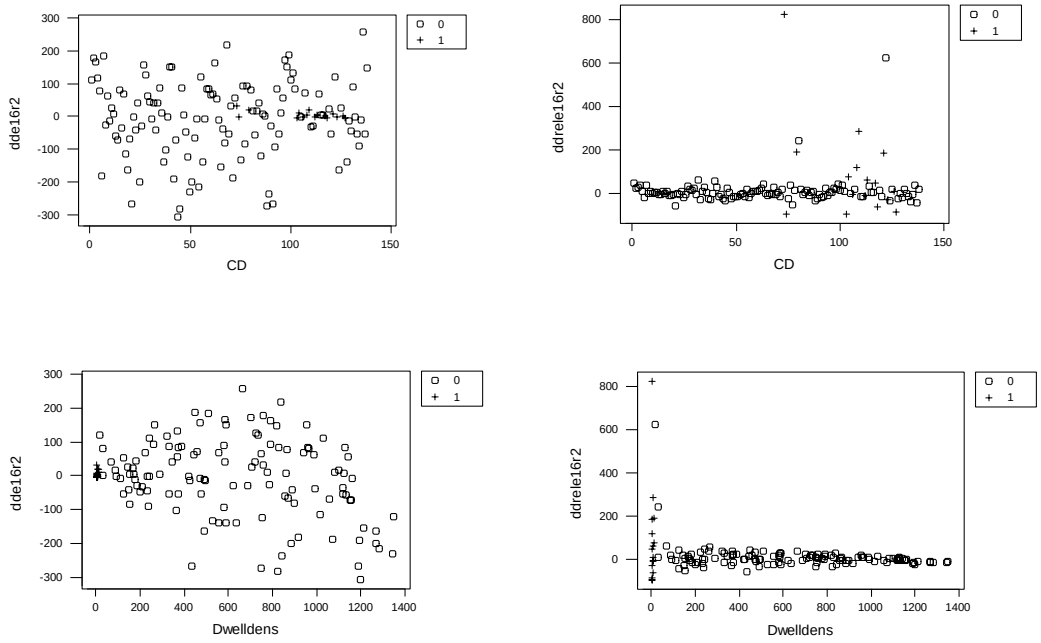
Plots pertaining to Table 4.12 Ballarat population model 6

Absolute and relative errors in estimated population by CD number and by ground truth population.



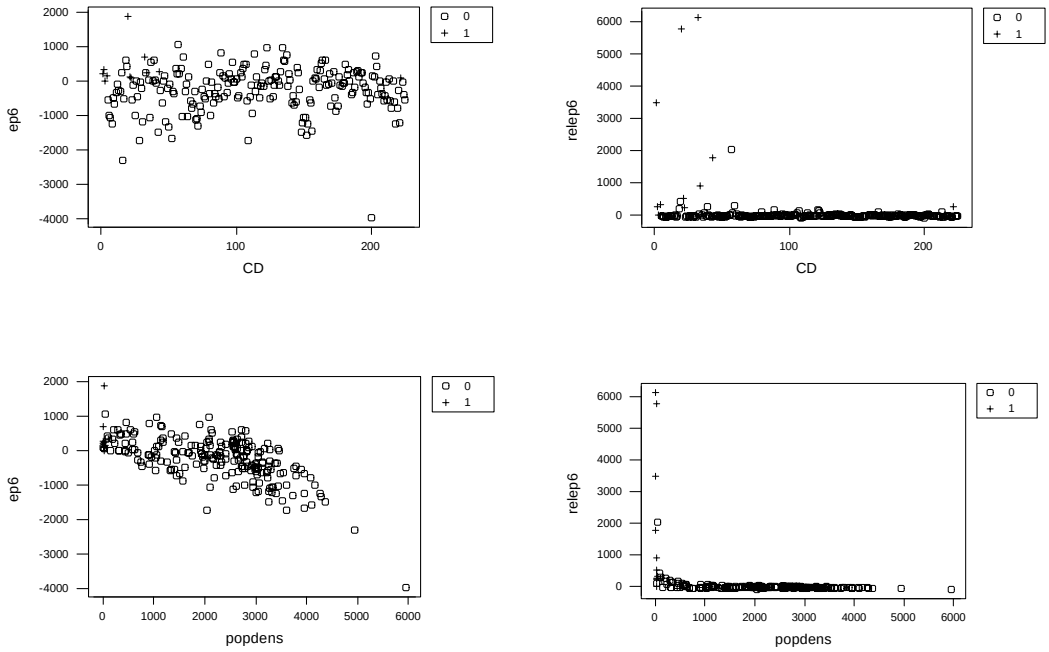
Plots pertaining to Table 4.11 Ballarat dwelling density model 6

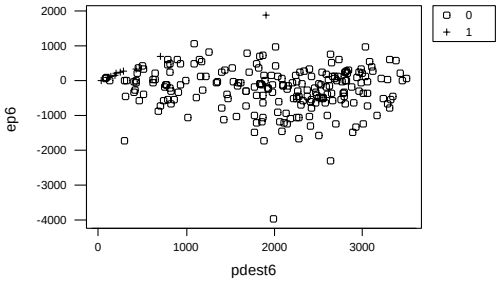
Absolute and relative errors in estimated dwelling density by CD number and by ground truth dwelling density.

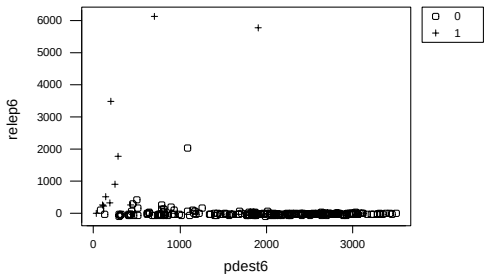


Plots pertaining to Table 6.1 Geelong population density model 6

Absolute and relative errors in estimated population density by CD number, ground truth population density, and estimated population density

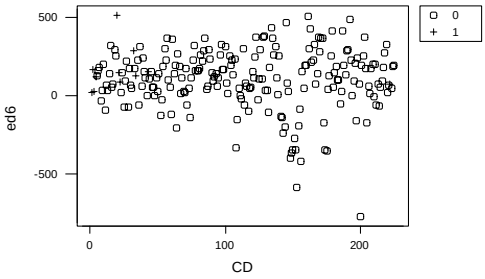


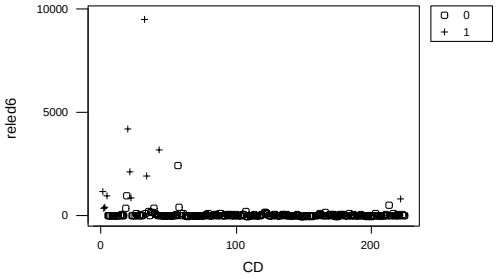


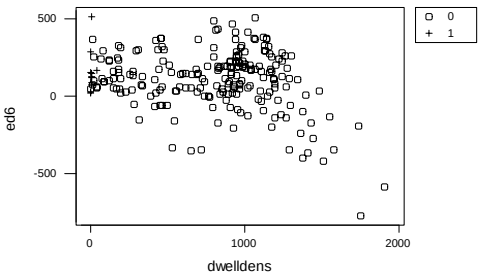


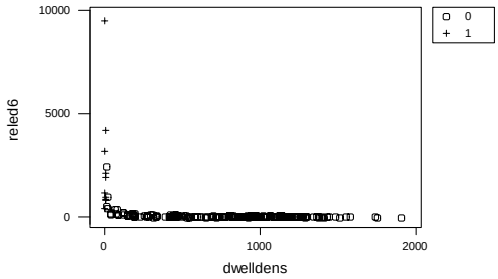
Plots pertaining to Table 6.2 Geelong dwelling density model 6

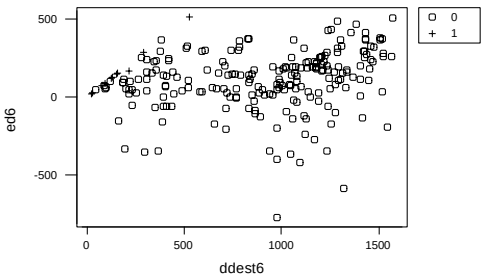
Absolute and relative errors in estimated dwelling density by CD number, ground truth dwelling density, and estimated dwelling density

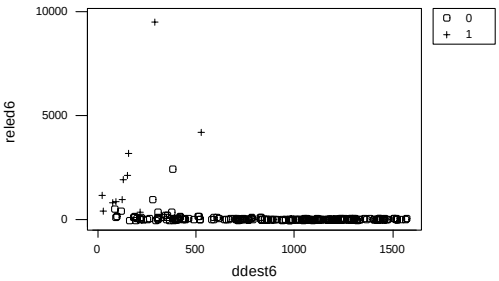






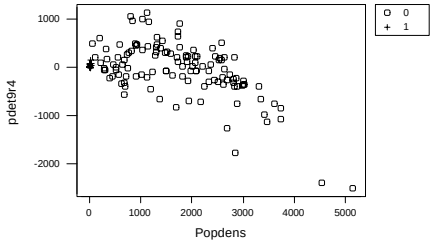






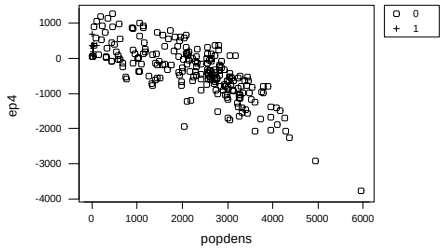
Plots pertaining to Table 4.10 Ballarat population density model 4

Absolute error in estimated population density by ground truth population density.



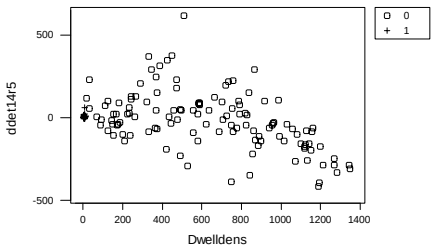
Plots pertaining to Table 6.1 Geelong population density model 4

Absolute error in estimated population density by ground truth population density.



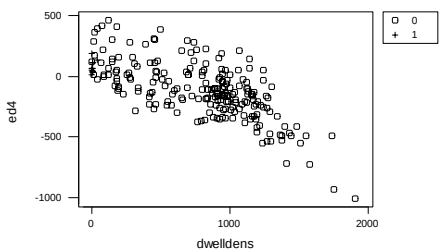
Plots pertaining to Table 4.11 Ballarat dwelling density model 4

Absolute error in estimated dwelling density by ground truth dwelling density.



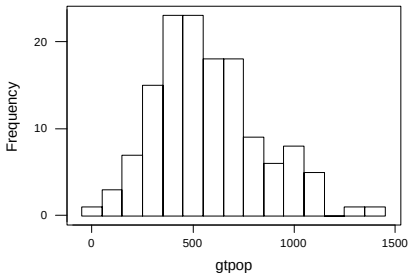
Plots pertaining to Table 6.2 Geelong dwelling density model 4

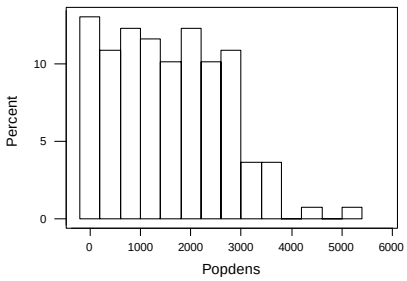
Absolute error in estimated dwelling density by ground truth dwelling density.

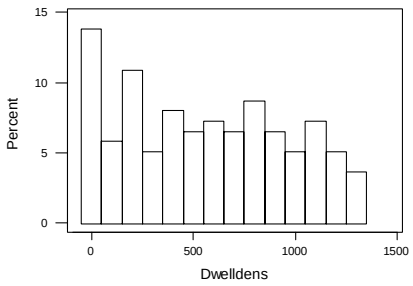


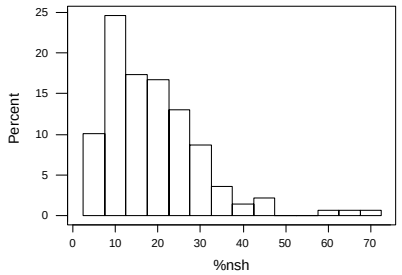
Population and housing indicators

Ballarat

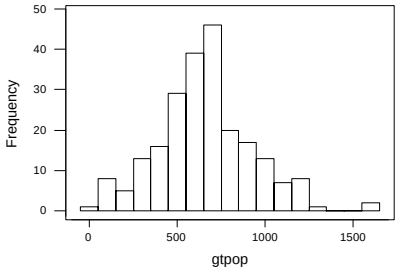


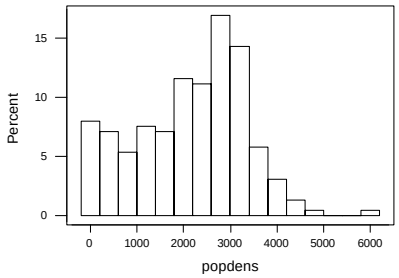


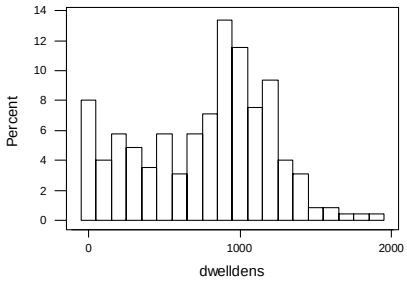


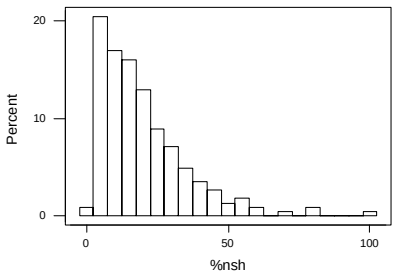


Geelong









Appendix F

Selected Results from Exploratory Discriminant Analysis and Regression Analysis on Samples from the Primary Image

(Sections 5.5 & 5.6)

Final ordered set of variables from which groups variables were selected for use for maximum likelihood classification

Step	Variable	Step	Variable	Step	Variable	Step	Variable	Step	Variable	Step	Variable	Step	Variable	Step	Variable
1	DS25	6	RH125S	11	B3	16	DS57	21	B1	26	B2P	31	B1P	36	NB1S
2	B5	7	B2S	12	NB2	17	NB1	22	DS15S	27	R57	32	R25P	37	CH125S
3	B7	8	R25	13	R15	18	CH123	23	R25S	28	CH125P	33	DS15P		
4	B4	9	B2	14	CH123S	19	DS15	24	R14S	29	CH123P	34	R14P		
5	CH125	10	DS35	15	R14	20	RH123	25	B1S	30	RH125	35	R15S		

Example of a confusion matrix

Classification Results[a]

			Predicted Group Membership											
		CLASS	1 Industrial	2 Commercial	3 Public use	4 Dry grass	5 Green grass	6 Native forest	7 Pine plant	8 Dark soil	9 Light soil	10 Water	11 Road	12 Residential	Total
Original	Count	1 Industrial	2	25	92	3	4	0	0	0	28	0	7	52	213
		2 Commercial	0	60	11	0	0	0	0	0	0	0	0	11	82
		3 Public use	0	4	44	1	9	0	0	0	10	0	0	29	97
		4 Dry grass	0	0	11	1101	23	0	0	0	11	0	0	1	1147
		5 Green grass	0	0	10	42	581	0	0	0	0	0	0	5	638
		6 Native forest	0	0	3	0	4	2182	23	0	17	0	1	5	2235
		7 Pine plant	0	0	1	0	0	1	164	0	0	0	0	0	166
		8 Dark soil	0	0	0	0	0	0	0	25	0	0	0	0	25
		9 Light soil	0	0	1	1	0	0	0	0	53	0	0	0	55
		10 Water	0	25	0	0	0	0	0	0	0	2149	0	0	2174
		11 Road	0	0	0	0	0	0	0	0	0	0	25	0	25
		12 Residential	1	4	33	12	4	0	0	0	8	0	19	539	620
	%	1 Industrial	.9	11.7	43.2	1.4	1.9	.0	.0	.0	13.1	.0	3.3	24.4	100.0
		2 Commercial	.0	73.2	13.4	.0	.0	.0	.0	.0	.0	.0	.0	13.4	100.0
		3 Public use	.0	4.1	45.4	1.0	9.3	.0	.0	.0	10.3	.0	.0	29.9	100.0
		4 Dry grass	.0	.0	1.0	96.0	2.0	.0	.0	.0	1.0	.0	.0	.1	100.0
		5 Green grass	.0	.0	1.6	6.6	91.1	.0	.0	.0	.0	.0	.0	.8	100.0
		6 Native forest	.0	.0	.1	.0	.2	97.6	1.0	.0	.8	.0	.0	.2	100.0
		7 Pine plant	.0	.0	.6	.0	.0	.6	98.8	.0	.0	.0	.0	.0	100.0
		8 Dark soil	.0	.0	.0	.0	.0	.0	.0	100.0	.0	.0	.0	.0	100.0
		9 Light soil	.0	.0	1.8	1.8	.0	.0	.0	.0	96.4	.0	.0	.0	100.0
		10 Water	.0	1.1	.0	.0	.0	.0	.0	.0	.0	98.9	.0	.0	100.0
		11 Road	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0	100.0	.0	100.0
		12 Residential	.2	.6	5.3	1.9	.6	.0	.0	.0	1.3	.0	3.1	86.9	100.0

a. 92.6% of original grouped cases correctly classified.

Details of the four chosen regression models

Model 6/A

Predictor	Coef	StDev	T	P
Constant	1.0333	0.2103	4.91	0.000
b1	0.087223	0.008043	10.84	0.000
b2	0.11376	0.02174	5.23	0.000
b3	-0.13557	0.01261	-10.75	0.000
b4	-0.002823	0.004101	-0.69	0.491
b5	-0.039239	0.003724	-10.54	0.000
b7	0.064379	0.007279	8.84	0.000

S = 0.8623 R-Sq = 44.4% R-Sq(adj) = 44.2%

Model 25/A

Predictor	Coef	StDev	T	P
Constant	1.5870	0.1845	8.60	0.000
b1	0.047860	0.007354	6.51	0.000
b2	0.10985	0.01941	5.66	0.000
b3	-0.119086	0.009827	-12.12	0.000
b4	-0.006780	0.003366	-2.01	0.044
b5	-0.029367	0.002848	-10.31	0.000
b7	0.056081	0.005982	9.37	0.000

S = 0.8017 R-Sq = 38.6% R-Sq(adj) = 38.3%

Model 6/F

Predictor	Coef	StDev	T	P
Constant	-0.9914	0.3698	-2.68	0.007
b3	-0.04956	0.01333	-3.72	0.000
b5	0.023197	0.004671	4.97	0.000
ds57	-3.2204	0.5541	-5.81	0.000
rh123	0.0021593	0.0006978	3.09	0.002
ch123	0.42146	0.05549	7.59	0.000
ch125	-4.0478	0.5159	-7.85	0.000
r15r	72.590	6.555	11.07	0.000
r15ri	-40.514	5.626	-7.20	0.000
r15r2	-81.726	8.349	-9.79	0.000
r14si	-2.0183	0.6609	-3.05	0.002
ds15r	-44.924	4.621	-9.72	0.000
ch123si	-0.21000	0.02485	-8.45	0.000
b2si	-0.0017901	0.0004990	-3.59	0.000

S = 0.7614 R-Sq = 56.9% R-Sq(adj) = 56.5%

Model 25/F

Predictor	Coef	StDev	T	P
Constant	0.9909	0.2474	4.00	0.000
ds57	-3.2853	0.4591	-7.16	0.000
rh123	0.0034466	0.0005456	6.32	0.000
ch123s	0.42689	0.08204	5.20	0.000
r15r	73.021	8.540	8.55	0.000
ds15r	-61.707	9.060	-6.81	0.000
ch123r	-0.3477	0.1416	-2.45	0.014
r14si	-1.6891	0.5022	-3.36	0.001
r25ri	12.980	5.126	2.53	0.011
b2s2	-0.005105	0.001093	-4.67	0.000
ch123s2	-0.07757	0.02877	-2.70	0.007
r15r2	-276.07	42.99	-6.42	0.000
ds15r2	195.98	52.97	3.70	0.000
ch123r2	0.22557	0.07523	3.00	0.003

S = 0.7331 R-Sq = 48.9% R-Sq(adj) = 48.4%

Appendix G

Sampling Variation in Results of Iterative Re-estimation

(Sample of results from Section 7.3)

Sample	n	it0	it1	it6	it29	it30
rtsall	6345	2.60260	2.82460	3.29300	4.89030	4.94940
		0.00660	0.01040	0.01010	-0.02480	-0.02590
		0.01610	0.02990	0.08130	0.24130	0.24760
		-0.02820	-0.04960	-0.10730	-0.21920	-0.22370
		0.00190	0.00240	-0.00540	-0.06080	-0.06260
		-0.02180	-0.03740	-0.06980	-0.06650	-0.06570
		0.03700	0.06560	0.14180	0.19620	0.19660
		15.96070	42.40700	82.13880	92.95730	93.09880
rts1	1269	2.51260	2.68300	3.19110	5.01360	5.07010
		0.00580	0.00910	0.00950	-0.01450	-0.01510
		0.02000	0.03440	0.06590	0.08680	0.08750
		-0.03020	-0.05210	-0.10170	-0.14580	-0.14700
		0.00290	0.00430	0.00020	-0.04000	-0.04110
		-0.02240	-0.03840	-0.07130	-0.07710	-0.07700
		0.03950	0.06940	0.14470	0.20870	0.20980
		15.91700	42.03870	80.07840	90.52680	90.67360
rts2	1269	2.54320	2.68070	2.82850	3.85310	3.88610
		0.01140	0.01730	0.01680	-0.01630	-0.01710
		0.00640	0.01410	0.05430	0.16970	0.17320
		-0.02960	-0.05150	-0.10910	-0.19600	-0.19840
		0.00640	0.01020	0.01150	-0.01790	-0.01880
		-0.02450	-0.04110	-0.07370	-0.07990	-0.07970
		0.04000	0.07020	0.14790	0.20310	0.20350
		18.87300	46.49540	81.27310	89.74130	89.82640
rcall	14270	1.42420	1.68780	2.57750	3.61070	3.62080
		0.07940	0.11580	0.14230	0.13360	0.13360
		0.12330	0.17850	0.20540	0.12000	0.11720
		-0.12730	-0.18530	-0.22030	-0.14790	-0.14550
		-0.00920	-0.01430	-0.02290	-0.02960	-0.02960
		-0.04020	-0.06050	-0.08650	-0.10820	-0.10870
		0.06050	0.09060	0.12750	0.15530	0.15590
		45.18720	80.65890	91.97080	92.59140	92.60360
rc1	1427	1.66870	1.95530	2.39110	2.36650	2.36420
		0.08170	0.11480	0.14110	0.16280	0.16320
		0.12430	0.16570	0.14100	0.02970	0.02810
		-0.13250	-0.18350	-0.20050	-0.15760	-0.15700
		-0.01290	-0.01610	-0.00640	0.01640	0.01660
		-0.04580	-0.06830	-0.10420	-0.13210	-0.13240
		0.07330	0.10910	0.16310	0.19770	0.19790
		46.87010	77.89880	87.14610	88.18450	88.19320
rc2	1427	1.29380	1.40200	1.58310	1.43230	1.42940
		0.07560	0.10910	0.13760	0.14470	0.14480
		0.13310	0.18960	0.22330	0.20440	0.20410
		-0.13680	-0.19650	-0.23960	-0.22960	-0.22940
		-0.00490	-0.00580	0.00110	0.01490	0.01510
		-0.03990	-0.05860	-0.07980	-0.08910	-0.08920
		0.06140	0.08860	0.10980	0.10860	0.10860
		44.41330	76.94710	87.33580	87.76790	87.77060

rts = residential class training set
rc = residential class

Appendix H

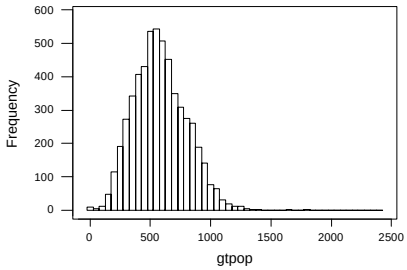
Summary Statistics for the Distributions of CD Population and CD Population Density in the Supplementary Study Areas

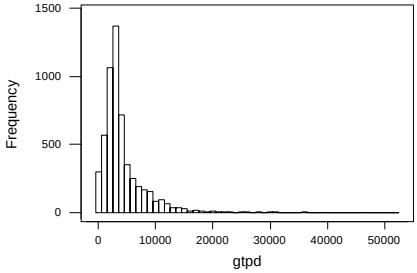
(Chapter 8)

Sydney

Variable	N	Mean	Median	TrMean	StDev	SE Mean
gtpop	5628	583.49	566.00	577.43	221.98	2.96
gtpd	5628	4424	3140	3812	5317	71

Variable	Minimum	Maximum	Q1	Q3
gtpop	0.00	2421.00	426.00	725.00
gtpd	0	194545	2074	5076





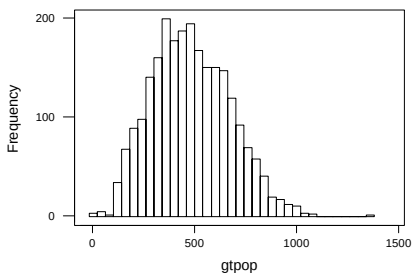
Note: 6 extreme outliers omitted from histograms

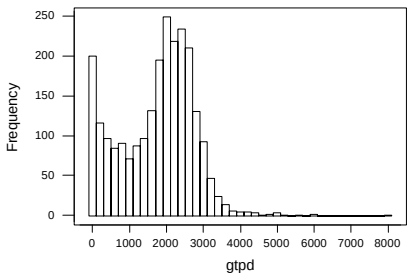
Adelaide

Variable	N	N*	Mean	Median	TrMean	StDev
gtpop	2412	7	480.36	469.50	476.32	190.80
gtpd	2412	7	1742.7	1936.7	1733.3	999.9

Variable	SE Mean	Minimum	Maximum	Q1	Q3
gtpop	3.88	0.00	1349.00	341.00	615.00
gtpd	20.4	0.0	7916.7	952.7	2449.1

MTB >

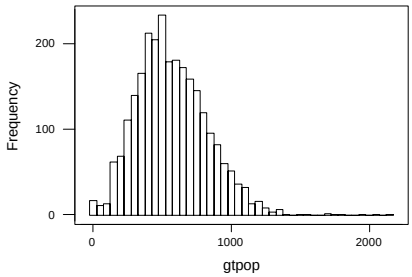


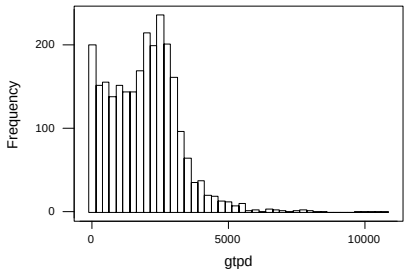


Brisbane

Variable	N	Mean	Median	TrMean	StDev	SE Mean
gtpop	2605	571.55	543.00	562.68	257.89	5.05
gtpd	2605	1914.0	1951.8	1832.2	1320.8	25.9

Variable	Minimum	Maximum	Q1	Q3
gtpop	0.00	2142.00	388.00	731.00
gtpd	0.0	10769.2	888.1	2676.5

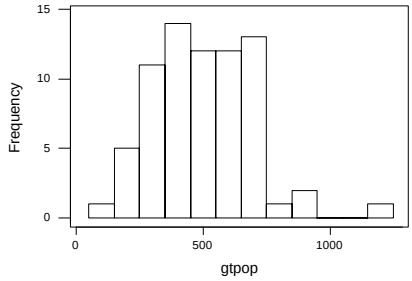


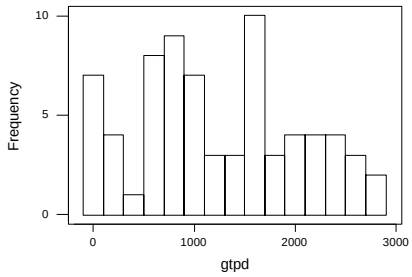


Ballarat 94

Variable	N	Mean	Median	TrMean	StDev	SE Mean
gtpop	72	496.0	493.5	489.9	193.8	22.8
gtpd	72	1234.3	1111.1	1222.2	784.8	92.5

Variable	Minimum	Maximum	Q1	Q3
gtpop	118.0	1196.0	357.2	639.0
gtpd	6.5	2776.4	618.7	1875.2

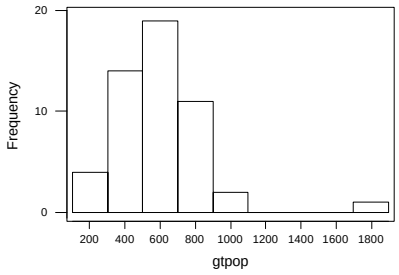


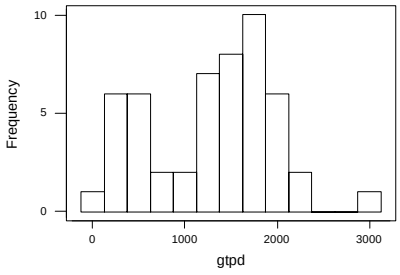


Kalgoorlie

Variable	N	Mean	Median	TrMean	StDev	SE Mean
gtpop	51	593.1	551.0	574.4	271.8	38.1
gtpd	51	1296.2	1439.9	1295.5	693.9	97.2

Variable	Minimum	Maximum	Q1	Q3
gtpop	130.0	1820.0	402.0	751.0
gtpd	26.7	3017.1	619.4	1822.6





Appendix I

Landuse/Landcover Classes in Supplementary Images

(Chapter 8)

Sydney	Brisbane	Adelaide	Kalgoorlie	Ballarat 1994
			Smoke	
Sea water	Sea water	Sea water		
Fresh water	Fresh water	Clear water	Fresh water	Fresh water
		Muddy water (Murray River & lake)		
Sand	Sand	Sand		
Surf & shallows		Surf & shallows		
Mid -green grass/crops	Mid -green grass/crops	Mid -green grass/crops		Mid -green grass/crops
Lush green grass/crops	Lush green grass/crops	Lush green grass/crops		Lush green grass/crops
Dry grass/crops		Dry grass/crops (east)	Dry grass	
		Dry grass/crops (north)		
Crop (low in band 4)				Crop (low in band 4)
Native forest (South)	Native forest	Native forest		Native forest
Native forest (North)				
Native forest (West)				
Conifer Plantation				Conifer Plantation (established)
				Conifer Plantation (young)
	Coastal scrub & mangroves	Coastal scrub & mangroves		
Open scrub		Open scrub - green	Open scrub - light	
		Open scrub -khaki	Scrub - medium	
			Scrub - dark	
Bare ground – development?	Bare ground – light	Bare ground	Bare ground	Bare ground – light (freeway const. & mining)
Bare ground – agricultural	Bare ground – medium			Bare ground - dark
Bare ground – burnt	Bare ground – dark (old coal mine)			
		Quarry		
		Intensive agriculture (vineyards)		Intensive agriculture (potatoes)
Bitumen roads and pavements	Bitumen roads and pavements	Bitumen roads and pavements	Bitumen roads and pavements	Bitumen roads and pavements
Concrete roads and pavements				
Railways				
		Salt pans	Dry lakes / tailings pans	
Commercial	Commercial	Commercial	Commercial	Commercial
Industrial	Industrial	Industrial	Industrial/mining	Industrial
Residential	Residential	Residential	Residential	Residential

Appendix J

Normalisation Formulae

(Chapter 8)

We postulate that there exists an invariant linear relationship between the population p_j of a pixel j and some function of the bands $f(b_{ij})$

$$p_j = c_0 + \sum_{i=1}^{nbands} c_i f(b_{ij}) \text{ for all } j$$

1. *multiplicative normalisation*. Suppose $f(b_{ij}) = \frac{b_{ij}}{\mu_i}$ where μ_i is the band mean for some class of pixels in the image. Then

   $$p_j = c_0 + \sum_{i=1}^{nbands} c_i \frac{b_{ij}}{\mu_i}$$

   Suppose that in a training image we estimate a linear relationship

   $$p_j = t_0 + \sum_{i=1}^{nbands} t_i b_{ij}$$

   It follows that we estimate $c_0 = t_0$ and $c_i = t_i \mu_{iT}$ where the *T* subscript refers to the training image.

   We apply this relationship to a second application image, in the form

   $$p_j = a_0 + \sum_{i=1}^{nbands} a_i b_{ij}$$

   It follows that we estimate $a_0 = c_0 = t_0$ and $a_i = \frac{c_i}{\mu_{iA}} = \frac{t_i \mu_{iT}}{\mu_{iA}}$ where the *A* subscript refers to the application image.

Similar arguments lead to the following relationships in the other two cases.

2. *Additive normalisation* $f(b_{ij}) = b_{ij} - \mu_i$

   $c_i = t_i$ $\qquad$ $c_0 = t_0 + \sum_{i=1}^{nbands} t_i \mu_{iT}$

   $a_i = c_i = t_i$ $\qquad$ $a_0 = c_0 - \sum_{i=1}^{nbands} c_i \mu_{iA} = t_0 + \sum_{i=1}^{nbands} t_i (\mu_{iT} - \mu_{iA})$

3. *Scaled additive normalisation* $f(b_{ij}) = \frac{b_{ij} - \mu_i}{\sigma_i}$

   $c_i = t_i \sigma_{iT}$ $\qquad$ $c_0 = t_0 + \sum_{i=1}^{nbands} t_i \mu_{iT}$

   $a_i = \frac{c_i}{\sigma_{iA}} = \frac{t_i \sigma_{iT}}{\sigma_{iA}}$ $\qquad$ $a_0 = c_0 - \sum_{i=1}^{nbands} \frac{c_i \mu_{iA}}{\sigma_{iA}} = t_0 + \sum_{i=1}^{nbands} t_i (\mu_{iT} - \frac{\sigma_{iT} \mu_{iA}}{\sigma_{iA}})$

Appendix K

Summary of Least Accurately Estimated CDs in Adelaide Image

(Chapter 8)

A. 25 CDs with most overestimated population

CD	GT Pop	GT Pop dens	RS Pop	Pop diff	Area	Description
2210	158	3.38	4024.94	3866.94	Bolivar	Sewage works, mangroves
2103	291	4.63	3921.46	3630.46	Port Gawler	Salt evaporation pans
2264	263	19.6	3565.93	3302.93	Parafield	Industrial, university, airfield
1428	470	105	3688.00	3218.00	Regency Park	Industrial
32	342	3.12	3280.10	2938.10	Williamstown	Reservoir, forest
62	304	2.87	3178.85	2874.85	Port Gawler	Salt evaporation pans
703	482	14.6	3284.85	2802.85	McLaren Vale	Vineyards
257	264	1.35	3060.79	2796.79	Cape Fleurieu	Rural, coast
2102	533	31.6	3231.29	2698.29	Virginia	Vineyards, market gardens
2198	81	6.77	2654.26	2573.26	Elizabeth	Defence research centre
38	288	2.3	2853.39	2565.39	Williamstown	Adjacent to 32
1427	581	46.2	3057.31	2476.31	Port Adelaide	Swamp, rubbish tip
72	480	11	2859.54	2379.54	Kersbrook	Agricultural
1727	1017	156	3376.21	2359.21	City	Parklands, gaol, university, hospital, cemetery
610	597	50.7	2901.55	2304.55	Lonsdale	Oil refinery
21	615	15.3	2909.75	2294.75	Barossa	Vineyards
2412	154	1.82	2343.19	2189.19	Cape Fleurieu	Adjacent to 257
256	256	1.6	2372.40	2116.40	Cape Fleurieu	Adjacent to 257
211	310	3.31	2408.81	2098.81	Goolwa	Wetlands, dunes
202	118	2.33	2178.23	2060.23	Hope Forest	Unknown
701	304	24.4	2339.67	2035.67	McLaren Flat	Vineyards
1173	111	13.4	2044.16	1933.16	Port Adelaide	Industrial, waste land
71	484	8.24	2386.61	1902.61	Kersbrook	Agricultural
1197	682	95.7	2544.53	1862.53	Outer harbour	Industrial, waste ground
11	477	33.9	2184.37	1707.37	Barossa	Vineyards
255	411	5.31	2063.54	1652.54	Cape Fleurieu	Adjacent to 257

B. 25 CDs with most underestimated population density

CD	GT Pop	GT Pop Dens	RS Pop Dens	Pop Dens diff	RS Pop	Pop diff	Area	Description
820	456	7917	2672	5245	152.37	-303.63	Glenelg	Retirement village
791	431	6045	1141	4904	81.60	-349.40	Glenelg	Small allotments
1716	633	6029	1895	4134	201.62	-431.38	Nth Adelaide	Mixed comm/res
802	264	4898	1054	3844	56.28	-207.72	Glenelg	Small allotments
1729	935	3683	255	3427	68.81	-866.19	City	Mixed comm/res
1025	444	5692	2318	3374	187.17	-256.83	Keswick	Small allotments
1734	386	4323	1052	3270	93.91	-292.09	City	Mixed comm/res
1744	725	4148	1056	3092	180.21	-544.79	City	Mixed comm/res
801	280	4912	1860	3052	107.46	-172.54	Glenelg	Small allotments
1743	336	5037	2008	3030	130.79	-205.21	City	Mixed comm/res
793	332	5061	2157	2904	137.40	-194.60	Glenelg	Small allotments
1803	288	5227	2536	2690	128.84	-159.16	Glenside	Leafy suburb
1741	466	4024	1490	2534	173.73	-292.27	City	Mixed comm/res
1746	301	3529	1092	2437	93.53	-207.47	City	Mixed comm/res
1728	447	2535	128	2408	22.85	-424.15	City	Mixed comm/res
1733	240	2948	575	2374	43.60	-196.40	City	Mixed comm/res
2380	739	4178	1833	2345	366.35	-372.65	Gawler East	Small allotments
1721	234	3634	1331	2302	81.90	-152.10	Nth Adelaide	Mixed comm/res
1742	284	3329	1080	2249	89.78	-194.22	City	Mixed comm/res
1723	315	4831	2595	2236	155.05	-159.95	Nth Adelaide	Mixed comm/res
1722	847	3648	1419	2229	328.49	-518.51	Nth Adelaide	Mixed comm/res
1429	507	4085	1879	2207	227.70	-279.30	Regency Park	Small allotments
1710	481	4035	1847	2189	212.71	-268.29	Nth Adelaide	Mixed comm/res
794	403	3112	930	2182	121.37	-281.63	Glenelg	Small allotments
1713	314	4486	2333	2153	157.95	-156.05	Nth Adelaide	Mixed comm/res

References

Abdou, I.E. and Pratt, W.V. (1979) Quantitative design and evaluation of enhancement/thresholding edge detectors, *Proceedings of the IEEE* **67**: 753-763.

Amrhein, C.G. and Flowerdew, R. (1989) The effect of data aggregation on a Poisson model of Canadian migration, in Goodchild, M. and Gopal, S. (eds) *The Accuracy of Spatial Databases*, London: Taylor and Francis.

Anderson, D.E. and Anderson, P.N. (1973) Population estimates by humans and machines, *Photogrammetric Engineering* **39**: 147-54

Atkinson, P.M. and Curran, P.J. (1997) Choosing an appropriate spatial resolution for remote sensing investigations, *Photogrammetric Engineering and Remote Sensing* **63** (12): 1345-1351.

Australian Bureau of Statistics (1987) *Age and Sex of Persons in Statistical Local Areas, Victoria*. Cat. No. 2455.0. Canberra: Australian Bureau of Statistics.

Australian Bureau of Statistics (1989) *Estimated Resident Population in Statistical Local Areas, Victoria, 30 June, 1987 (final) and 1988 (preliminary).* Cat. No. 3203.2. Melbourne: Australian Bureau of Statistics.

Australian Bureau of Statistics (1990) *Estimated Resident Population in Statistical Local Areas, Victoria, 30 June, 1988 (final) and 1989 (preliminary).* Cat. No. 3203.2. Melbourne: Australian Bureau of Statistics.

Australian Bureau of Statistics (1997) *CDATA96*, Australian Bureau of Statistics.

Australian Bureau of Statistics (1998) *Australian Standard Geographical Classification (ASGC).* Cat. No. 1216.0. Canberra: Australian Bureau of Statistics.

Ballarat and Western Victoria Regional Information Bureau (1989) *A Social Atlas of the Central Highlands Region, 1986 Census Edition*. Ballarat: Ballarat and Western Victoria Regional Information Bureau.

Barnsley, M.J. and Barr, S.L (1996) Inferring urban land use from satellite sensor images using kernel-based spatial reclassification, *Photogrammetric Engineering and Remote Sensing* **62** (8): 949-958.

Becker, N. (1995) The EM Algorithm, Modern Methods in Applied Statistics Workshop, Statistical Society of Australia, Melbourne, 9 February, 1995.

Bolstad, P.V. and Lillesand, T.M. (1992) Rule-based classification models: integration of satellite imagery and thematic spatial data, *Photogrammetric Engineering and Remote Sensing* **58**: 965-971.

Brugioni, D.A. (1983) The census: it can be done more accurately with space-age technology, *Photogrammetric Engineering and Remote Sensing* **49**: 1337-39.

Chavez, P.S. Jr. (1992) Comparison of spatial variability in visible and near-infrared spectral images, *Photogrammetric Engineering and Remote Sensing* **58**: 957-964.

Chen, K.S, Tzeng, Y.C., Chen, C.F. and Kao, W.L. (1995) Land-cover classification of multispectral imagery using a dynamic learning neural network, *Photogrammetric Engineering and Remote Sensing* **61** (4): 403-408.

Clayton, C. and Estes, J.E. (1980) Image analysis as a check on census enumeration accuracy, *Photogrammetric Engineering and Remote Sensing* **46**: 757-64.

Cliff, N. (1987) *Analysing Mutivariate Data.* Orlando: Harcourt Brace Jovanovich.

Collins, W.G. and El-Beik, A.H.A. (1971) Population census with the aid of aerial photographs: an experiment in the city of Leeds, *Photogrammetric Record* **7**: 16-26

Congalton, R.G., Oderwald, R.G. and Mead, R.A. (1983) Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques, *Photogrammetric Engineering and Remote Sensing* **49**: 1671-78.

Crockett, R.A. (1990) Deputy Commonwealth Statistician, Victoria. Letter to the author, 25 October 1990.

Curran, P.J. (1985) *Principles of Remote Sensing*, New York: Longman.

Cushnie, J.L. and Atkinson, P. (1985) Effect of spatial filtering on scene noise ansd boundary detail in Thematic Mapper imagery, *Photogrammetric Engineering and Remote Sensing* **51** (9): 1483-1493.

Cushnie, J.L., (1987) The interactive effect of spatial resolution and degree of internal variability within land-cover types on classification accuracies, *Int. J. Remote Sensing*, **8** (1): 15-29.

Dayal, H.H. and Khairzada, B.A.(1976) The first national demographic survey of Afghanistan: the role played by air photos and photo-counting techniques, *The I.T.C. Journal*, No.1: 84-97.

De Cola, L. (1989) Fractal analysis of a classified Landsat scene, *Photogrammetric Engineering and Remote Sensing* **55** (5): 601-610.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. Royal Statistics Society, Series B*, **39**: 1-38.

Duecker, K.J. and Horton, F.E. (1971) Urban change detection systems: status and prospects, *Proceedings of the Seventh International Symposium on Remote Sensing of the Environment*, University of Michigan: Ann Arbor; pp 1523-36.

Durland, R.E. (1975) The remote sensing census projects, *Population Association of America Annual Meeting* (Collected papers), Seattle: Vol 3, p 110-114.

ER Mapper *Release* 2.0 (1991). Perth: Earth Resource Mapping Pty Ltd.

Excel Version 3.0 (1990) Redmond, Washington: Microsoft Corporation.

Eyre, L.A., Adolphus, B. and Amiel, M. (1970) Census analysis and population studies, *Photogrammetric Engineering* **36**: 460-66.

Fisher, P.F. (1989) Knowledge-based approaches to determining and correcting areas of unreliability in geographic databases, in Goodchild, M. and Gopal, S. (eds) *The Accuracy of Spatial Databases*, London: Taylor and Francis.

Fisher, P.F. and Langford, M. (1995) Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation, *Environment and Planning A* **27**: 211-224.

Fisher, P.F. and Langford, M. (1996) Modelling sensitivity to accuracy in classified imagery: a study of areal interpolation by dasymetric mapping, *Professional Geographer* **48** (3): 299-309.

Flowerdew, R. and Green, M. (1989) Statistical methods for inference between incompatible zonal systems, in Goodchild, M. and Gopal, S. (eds) *The Accuracy of Spatial Databases*, London: Taylor and Francis.

Foody, G.M. (1996) Relating the land cover composition of mixed pixels to artificial neural network classification output, *Photogrammetric Engineering and Remote Sensing* **62** (5): 491-499.

Foody, G.M., McCulloch, M.B, and Yates, W.B. (1995) Classification of remotely sensed data by an artificial neural network: issues relating to training data characteristics, *Photogrammetric Engineering and Remote Sensing* **61** (4): 391-401.

Forster, B.C. (1980a) Urban control for Landsat data, *Photogrammetric Engineering and Remote Sensing* **46**: 539-45.

Forster, B.C. (1980b) Urban residential ground cover using Landsat digital data, *Photogrammetric Engineering and Remote Sensing* **46**: 547-58.

Forster, B.C. (1981) *Some Measures of Urban Residential Quality from Landsat Multispectral Data* (Unisurv S18). Sydney: University of New South Wales.

Forster, B.C. (1983) Some urban measurements from Landsat data, *Photogrammetric Engineering and Remote Sensing* **49**: 1693-1707.

Forster, B.C. (1993) The coefficient of variation as a measure of urban spatial attributes, using SPOT HRV and Landsat TM data, *International Journal of Remote Sensing* **14**: 2403-2409.

Forster, B.C. and Jones, C. (1988) Urban density monitoring using high resolution spaceborne systems. *Int. Arch. Photogramm. Rem. Sens*., Kyoto, Vol. 27, Part B9: 189-195.

Forster, B.C. and Xing, C. (1992) Urban spatial attributes from satellite remote sensing using end member analysis and variability measures. *Int. Arch. Photogramm. Rem. Sens.* Vol. 29, Part B7: 930-934.

Fotheringham, A.S. (1989) Scale-independent spatial analysis, in Goodchild, M. and Gopal, S. (eds) *The Accuracy of Spatial Databases*, London: Taylor and Francis.

Gautam, N.C. (1976) Aerial photo-interpretation techniques for classifying urban land use, *Photogrammetric Engineering and Remote Sensing* **42** (6): 815-822.

General Electric Corporation (1977) *Preliminary Design Requirements for Census/Urbanised Area Applications Systems Verification and Transfer*, Final report prepared for NASA, Contract No. NAS5-23412. Goddard Space Flight Center.

Goldstein, H. (1995) *Multilevel Statistical Models* (2nd edition), New York: Wiley.

Gong, P. and Howarth, P.J. (1990a) The use of stuctural information for improving land-cover classification accuracies at the rural-urban fringe, *Photogrammetric Engineering and Remote Sensing* **56**: 67-73.

Gong, P. and Howarth, P.J. (1990b) An assessment of some factors influencing multispectral land-cover classification, *Photogrammetric Engineering and Remote Sensing* **56**: 597-603.

Gong, P. and Howarth, P.J. (1992) Frequency-based contextual classification and gray-level vector reduction for land-use identification, *Photogrammetric Engineering and Remote Sensing* **58**: 423-437.

Goodchild, M.F. and Lam, N.S-N. (1980) Areal interpolation: a variant of the traditional spatial problem, *Geo-processing* **1**: 279-312.

Goodchild, M.F., Anselin, L. and Deichmann, U. (1992) A framework for the areal interpretation of socioeconomic data, *Environment and Planning A* **25**: 383-397.

Gopal, S. and Woodcock, C. (1994) Theory and methods for accuracy assessment of thematic maps using fuzzy sets, *Photogrammetric Engineering and Remote Sensing* **60** (2): 181-188.

Green, N. E. (1957) Aerial photographic interpretation and the social structure of the city, *Photogrammetric Engineering* **23**: 89-96.

Griffiths, G.H. (1988) Monitoring urban change from Landsat TM and SPOT satellite imagery by image differencing, *Proc. IGARSS '88 Symposium, Edinburgh, Scotland, 13-16 September, 1988, Ref ESA SP-284 (IEEE 88CH2497-6)*, ESA Publications Division.

Haack, B.N. (1984) Multisensor data analysis of urban environments, *Photogrammetric Engineering and Remote Sensing* **50**: 1471-1477.

Haralick, R.M. (1979) Statistical and structural approaches to texture, *Proceedings of the IEEE* **67**: 786-804.

Haralick, R.M. (1986) Statistical image texture analysis. In Young, T.Y. and Fu, K.S. (Eds.) *Handbook of Pattern Recognition and Image Processing*. Orlando: Academic Press; pp 247-281.

Harris, P.M. and Ventura, S.J. (1995) The integration of geographic data with remotely sensed imagery to improve classification in an urban area, *Photogrammetric Engineering and Remote Sensing* **61** (8): 993-998.

Harrison, B.A. and Jupp, D.B.L. (1989) *Introduction to Remotely Sensed Data*. Canberra: CSIRO.

Harrison, B.A. and Jupp, D.B.L. (1990) *Introduction to Image Processing.* Canberra: CSIRO.

Harvey, J.T. (1996) The Application of Orbital Remote Sensing to the Estimation of Small-area Population and Dwelling Densities, *Proc. 8th Australasian Remote Sensing Conference,* Canberra, March 1996.

Harvey, J.T. and Taylor, S. (1984) *Central Highlands Land Use Database*. Ballarat: Ballarat and Western Victoria Regional Information Bureau.

Heikkonen, J. and Varfis, A. (1998) Land cover/land use classification of urban areas: a remote sensing approach, *International Journal of Pattern Recognition and Artificial Intelligence* **12** (4): 475-489.

Heikkonen, J., Varfis, A., Kanellopoulos, I., Wilkinson, G., Fullerton, K. and Steel, A. (1997) A method for remote sensing based classification of urban areas, *The 10th Scandanavian Conference on Image Analysis*, Lappeenranta, Finland, June 9-11, 1997.

Henderson, F. M. (1979) Housing and population analyses. In Ford, K. (Ed.) *Remote Sensing for Planners.* New Brunswick: Centre for Urban Policy Research; p 140.

Henderson, F.M. and Xia, Z.G. (1997) SAR applications in human settlement detection, population estimation and urban land use pattern analysis: a status report, *IEEE Transactions on Geoscience and Remote Sensing* **35** (1): 79-85.

Holz, R., Huff, D.L. and Mayfield, R.C. (1973) Urban spatial structure based on remote sensing imagery. In Holz, R.K. (ed). *The Surveillant Science: Remote Sensing of the Environment.* Boston: Houghton Mifflin; pp 375-80.

Hosmer, D.J. and Lemeshow, S. (1989) *Applied Logistic Regression*, New York: Wiley.

Hsu, S. Y. (1978) Texture-tone analysis for automated land-use mapping, *Photogrammetric Engineering and Remote* Sensing **46**: 1051-1058.

Hsu, S.Y. (1971) Population estimation, *Photogrammetric Engineering* **37**: 449-54.

Iisaka, J. and Hegedus, E. (1982) Population estimation from Landsat imagery, *Remote Sensing of the Environment* **12**: 259-72.

Jackson, M.J., Carter, P., Smith, T.F. and Gardner, W.G. (1980), Urban land mapping from remotely sensed data, *Photogrammetric Engineering and Remote* Sensing **46** (8): 1041-1050.

Jensen, J.R., (1982) Detecting residential land-use development at the urban fringe, *Photogrammetric Engineering and Remote* Sensing **48** (4): 629-643.

Johnson, R.A. and Wichern, D.W. (1982) *Applied Multivariate Statistical Analysis*. Englewood Cliffs: Prentice Hall.

Kauth, R.J. and Thomas, G.S. (1976) The Tasseled Cap, a graphic description of agricultural crops as seen by Landsat. *Symposium on Machine Processing of Remotely Sensed Data*. West Lafayette, Ind.: Purdue University.

Kawamura, M., Jayamanna, S. and Tsujiko, Y. (1996) Relation between social and environmental conditions in Colombo, Sri Lanka and the urban index estimated by satellite remote sensing data, *International Archives of Photogrammetry and Remote Sensing* **31** (Part B7): 321-326.

Kendall, M., Stuart, A. and Ord, J.K. (1987) *Kendall's Advanced Theory of Statistics* (5th ed.) Vol. 2. London: Charles Griffin.

Kim, T. and Muller, J-P. (1998) A technique for 3D building reconstruction, *Photogrammetric Engineering and Remote* Sensing **64** (9): 923-930.

Kivell, P.T., Parsons, A.J. and Dawson, B.R.P. (1989) Monitoring derelict urban land: a review of problems and potentials of remote sensing techniques, *Land Degradation and Rehabilitation* **1**: 5-21

Kraus, S.P., Senger, L.W. and Ryerson, J.M. (1974) Estimating population from photographically determined residential land use types, *Remote Sensing of Environment* **3**: 35-42.

Lam, S. (1990) Description and measurement of Landsat TM images using fractals, *Photogrammetric Engineering and Remote Sensing* **56** (2): 187-195.

Lange, K. (1995) A gradient algorithm locally equivalent to the EM algorithm, *J. Royal Statistical Society B* **57** (2): 425-437.

Langford, M. and Unwin, D.J. (1994) Generating and mapping population density surfaces within a geographical information system, *The Cartographic Journal* **31** (1): 21-26.

Langford, M., Maguire, D.J. and Unwin, D.J. (1991) The areal interpolation problem: estimating population using remote sensing within a GIS framework. In Masser, I. And Blakemore, M. (eds.) *Handling Geographical Information: Methodology and Potential Applications*, London: Longman, 55-77.

Lavreau, J. (1991) De-hazing Landsat Thematic Mapper images, *Photogrammetric Engineering and Remote Sensing* **57**: 1297-1302.

Lee, P.M. (1997) *Bayesian Statistics: an Introduction* (2^{nd} ed), London: Arnold.

Lindgren, D.T. (1971) Dwelling unit estimation with color-IR photos, *Photogrammetric Engineering* **37**: 373-8.

Lo, C.P. (1979) Surveys of squatter settlements with sequential aerial photography - a case study in Hong Kong, *Photogrammetria* **35**: 45-63.

Lo, C.P. (1986a) *Applied Remote Sensing*. Harlow: Longman.

Lo, C.P. (1986b) Accuracy of population estimation from medium-scale aerial photography, *Photogrammetric Engineering and Remote Sensing* **52**: 1859-1869.

Lo, C.P. (1989) A raster approach to population estimation using high altitude aerial and space photographs, *Remote Sensing of Environment* **27**: 59-71.

Lo, C.P. (1995) Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach, *Int. J. Remote Sensing* **16** (1): 17-34.

Lo, C.P. and Chan, H.F. (1980) Rural population estimation from aerial photographs, *Photogrammetric Engineering and Remote Sensing* **46**: 337-45.

Lo, C.P. and Welch, R. (1977) Chinese urban population estimates, *Annals of the Association of American Geographers* **67**: 246-53.

Mapinfo Version 4.1 (1996) Mapinfo Corporation.

Martin, R.G. and Howarth, P.J. (1989) Change-detection accuracy assessment using spot multispectral imagery of the rural-urban fringe, *Remote Sensing of Environment,* **30**: 55-66.

Martin, R.G., (1989) Accuracy assessment of Landsat-based visual change detection methods applied to the rural-urban fringe, *Photogrammetric Engineering and Remote Sensing* **55** (2): 209-215.

Martin, R.G., Howarth, P.J. and Holder, G.A. (1988) Multispectral classification of land use at the rural-urban fringe using SPOT data, *Canadian Journal of Remote Sensing* **14**:72-79.

Mendenhall, W. (1979) *Introduction to Probability and Statistics* (5th ed.). North Scituate: Duxbury; p 97.

Meng, X-L. amd Rubin, D.B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework, *Biometrika* **80** (2): 267-278.

Mesev, V. (1988) The use of census data in urban image classification, *Photogrammetric Engineering and Remote Sensing* **64** (5): 431-438.

Mesev, V. (1998) Remote sensing of urban systems: hierarchical integration with GIS, *Computers, Environment and Urban Systems* **21** (3/4): 175-187.

microBRIAN Version 2.2 (1988). Melbourne: CSIRO & MPA Pty Ltd.

Minitab Release 12 (1997). State College, PA: Minitab, Inc.

Moller-Jensen, L. (1990) Knowledge-based classification of an urban area using texture and context information in Landsat-TM imagery, *Photogrammetric Engineering and Remote Sensing* **56**: 899-904.

Morgan, R.W. (1984) Further comments on "The Census" (letter), *Photogrammetric Engineering and Remote Sensing* **50**: 80.

Morrow-Jones, H.A. and Watkins, J.F. (1984) Remote sensing technology and the U.S. census, *Photogrammetric Engineering and Remote Sensing* **50**: 229-232.

Myers, R.H. (1990) *Classical and Modern Regression with Applications*. 2nd ed. Boston: Duxbury.

National Aeronautics and Space Administration (1978) *Application of Satellite Pictures to Census Operations. Bolivian Experience in Census-Taking of Population and Residences*. Translation into English of *Aplicaciones de Las Imagines de Satelite a Operaciones Censales. Experiencia Boliviana en El Censo de Poblacion Y Vivienda*, Rept. Inst. Nacl. De Estadistica, Min. De Planeamiento Y Coord., Rep. of Bolivia, La Paz 1977 p 1-14. Scitran: Santa Barbara. Report No. NASA-TM-75090

Navidi, W. (1997) A graphical illustration of the EM algorithm, *The American Statistician,* **51** (1): 29-31.

Ng, T.K. (1990) *Predicting Residential Housing Density and Housing Size with Satellite Imagery using Simplified Models. Research Project Report, Master of Engineering Science.* Sydney: University of New south Wales.

Niblack, W. (1986) *An Introduction to Digital Image Processing.* Englewood Cliffs: Prentice Hall.

Ogrosky, C.E. (1975) Population estimates from satellite imagery, *Photogrammetric Engineering and Remote Sensing* **41**: 707-12.

Olerunfemi, J.F. (1986) Towards a philosophy of population census in Nigeria: remote sensing inputs, *Remote Sensing Yearbook 1986*: 117-125

Olorunfemi, J.F. (1984) Land use and population: a linking model, *Photogrammetric Engineering and Remote Sensing* **50**: 221-27.

Polle, V.F.L. (1996) Planning urban services in developing countries in developing countries: quantification of community service needs using remote sensing indicators, *ITC Journal* 1996-1: 64-70.

Porter, P.W. (1956) *Population Distribution and Land Use in Liberia*, Ph.D. Dissertation, London School of Economics and Political Science: London.

Quarmby, N.A., and Cushnie, J.L. (1989) Monitoring urban land cover changes at the urban fringe from SPOT HRV imagery in south-east England, *Int. J. Remote Sensing*, **10**: 955-965.

Richards, J.A. (1986) *Remote Sensing Digital Image Analysis*: *an Introduction*. Berlin: Springer-Verlag.

Riolo, R.L and Line, M.P. (1995) Automatic discovery of classification and estimation algorithms for Earth-observation satellite imagery. *Genetic Programming*, Papers from the 1995 AAAI Fall Symposium. (Tech. Report FS-95-01): vii+133, 73-7.

Rosenfeld, A. (1984) Image analysis. In Ekstrom, M.P. *Digital Image Processing Techniques.* Orlando: Academic Press; pp 257-288.

Rosenfeld, A. and Kak, A.C. (1982) *Digital Picture Processing* (2nd ed.). Orlando: Academic Press.

Royer, A., Charbonneau, L., and Bonn, F. (1988) Urbanisation and Landsat albedo change in the Windsor-Quebec corridor since 1972, *Int. J. Remote Sensing*, **9** (3): 555-566.

Ryherd, S. and Woodcock, C. (1996) Combining spectral and texture data in the segmentation of remotely sensed images, *Photogrammetric Engineering and Remote Sensing* **62** (2): 181-194.

Scarpace, F.L. and Quirk, B.K (1980) Land-cover classification using digital processing of aerial imagery, *Photogrammetric Engineering and Remote Sensing* **46** (8): 1059-1065.

Sharma, K.M.S. and Sarkar, A. (1998) A modified contextual classification technique for remote sensing data, *Photogrammetric Engineering and Remote Sensing* **64** (4): 273-280.

Shettigara, V.K, and Sumerling, G.M. (1998) Height determination of extended objects using shadows in SPOT images, *Photogrammetric Engineering and Remote Sensing* **64** (1): 35-44.

Smith, M.O., Adams, J.B. and Gillespie, A.R. (1990) Reference end members for spectral mixing analysis, *Proc Fifth Australian Remote Sensing Conference, Perth, Australia,* Vol. 1: 331-340.

Space-Time Research Pty Ltd (1988) *Supermap Version* 2.0. Melbourne: Space-Time Research Pty Ltd.

SPSS Inc. (1988) *SPSS-X User's Guide* (3rd ed) (1988). Chicago: SPSS Inc.

Stern, M. (1983) Landsat data for population estimates – approaches to inter-censal counts in the rural Sudan, in Carter, W.D. and Engman, E.T. (Eds.) *Remote Sensing from Satellites*: 117-125.

Sutton, P., Roberts, D., Elvidge, C. and Meij, H. (1997) A comparison of nighttime satellite imagery and population density for the continental United States, *Photogrammetric Engineering and Remote Sensing* **63** (11): 1303-1313.

Tabachnick, B.C. and Fidell, L.S. (1996) *Using Multivariate Statistics*, 3rd Edition. New York: Harper Collins.

Takeuchi, S. and Tomita, T. (1988) Evaluation of some spectral and spatial features of satellite images using airborne MSS data for the analysis of urban areas. *Int. Arch. Photogramm. Rem. Sens*., Kyoto, Vol. 27, Part B7: pp 599-606.

Thompson, D. (1975) Small area population estimation using land use data derived from high altitude aircraft photography, *Proceedings of the American Society of Photogrammetry* (Fall Convention), pp. 673-96.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*, Chichester: Wiley.

Titterington, M., (1990) Some problems of inference related to image models, 10th Australian Statistical Conference Special Session on Image Analysis and Processing, Sydney, 6th & 7th July, 1990.

Toll, D.L. (1984) An evaluation of simulated Thematic Mapper data and Landsat MSS data for discriminating suburban and regional land use and land cover, *Photogrammetric Engineering and Remote Sensing* **50**: 1713-1724.

Tom, C.H. and Miller, L.D. (1984) An automated land use mapping comparison of the Bayesian maximum likelihood and linear discriminant analysis algorithms, *Photogrammetric Engineering and Remote Sensing* **50**: 193-207.

Ton, J., Jain, A.K., Enslin, W.R. and Hudson, W.D. (1989) Automatic road identification and labeling in Landsat TM images, *Photogrammetria* **43**: 257-276.

Treitz, P.M, Howarth, P.J. and Gong, P. (1992) Application of satellite and GIS technologies for land-cover and land-use mapping at the rural-rban fringe: a case study, *Photogrammetric Engineering and Remote Sensing* **58** (4): 439-448.

Van Deusen, P.C. (1995) Modified highest confidence first classification, *Photogrammetric Engineering and Remote Sensing* **61** (4): 419-425.

Wang, F. (1990) Improving remote sensing image analysis through fuzzy information representation, *Photogrammetric Engineering and Remote Sensing* **56** (8): 1163-1169.

Wang, L. and He, D.C. (1990) A new statistical approach for texture analysis, *Photogrammetric Engineering and Remote Sensing* **56** (1): 61-66.

Watkins, J.F. (1984) The effect of residential structure variation on dwelling unit enumeration from aerial photographs, *Photogrammetric Engineering and Remote Sensing* **50**: 1599-1607.

Watkins, J.F. and Morrow-Jones, H.A. (1985) Small area population estimates using aerial photography, *Photogrammetric Engineering and Remote Sensin,g* **51**: 1933-35.

Webster, C.J. (1996) Population and dwelling unit estimates from space, *Third World Planning Review* **18** (2): 155-176.

Welch, R. and Zupko, S. (1980) Urbanized area energy utilisation patterns from DMSP data, *Photogrammetric Engineering and Remote Sensing* **46** (2): 201-207.

Wellar, B.S. (1969) The role of space photography in urban and transportation data series, *Proceedings of the Sixth International Symposium on Remote Sensing of the Environment*, Vol II. Ann Arbor: University of Michigan; pp 831-54.

Wharton, S.W. (1987) A spectral knowledge-based approach for urban land cover discrimination, *IEEE Transactions on Geoscience and Remote Sensing*, Vol. GE-25, No. 3: pp 272-282.

Woodcock, C.E. and Strahler, A.H. (1987) The factor of scale in remote sensing. *Remote Sensing of the Environment* **21**: 311-332.

Yuan, Y., Smith, R.M. and Limp, W.F. (1997) Remodeling census population with spatial information from Landsat TM imagery, *Computers, Environment and Urban Systems* **21** (3-4): 245-258.

Zhuang, X., Engel, B.A., Xiong, X. and Johannsen, C.J. (1995) Analysis of Classification results of remote sensed data and evaluation of classification algorithms, *Photogrammetric Engineering and Remote Sensing,* **61** (4): 427-433.