Masters Program in Geospatial Technologies

Investigating the use of dasymetric techniques for assessing employment containment in Melbourne, Australia

Christabel McCarthy

Dissertation submitted in partial fulfilment of the requirements for the Degree of *Master of Science in Geospatial Technologies*

UNIVERSIDADE NOVA ISEGI Instituto de Estatística e Gestão de Informação
UNIVERSITAT JAUME·I
WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

Investigating the use of dasymetric techniques for assessing employment containment in Melbourne, Australia

Dissertation supervised by

Edzer Pebesma, PhD

Institute for Geoinformatics

Westfälische Wilhelms-Universität, Münster, Germany

Co-supervised by

Jorge Mateu, PhD

Dept. Mathematics

Universitat Jaume I, Castellón, Spain

Ana Cristina Costa, PhD

Instituto Superior de Estatística e Gestão da Informação

Universidade Nova de Lisboa, Lisbon, Portugal

February, 2012

ACKNOWLEDGEMENTS

I would like to thank my supervisors Prof. Edzer Pebesma, Dr. Jorge Mateu Mahiques and Dr. Ana Cristina Costa, and others within the Institute for Geoinformatics and the entire Erasmus Mundus Geospatial Technologies consortium for their advice and guidance in the development of this project.

My colleagues in the Sustainability Analysis group at the Department of Planning and Community Development, Victoria, particularly Christine Kilmartin, provided the theoretical impetus for this work and, more importantly, kindly provided the data that made this work possible.

My fellow Geospatial Technologies students have provided invaluable moral support throughout our Masters studies and particularly during this thesis writing time. Thank you to my family and friends at home who have encouraged me from afar. And last but very far from least, thank you to Karl for your constant support and encouragement.

ABSTRACT

This project studies employment containment in Melbourne, Australia. Employment containment is a measure of the proportion of people that work in a location close to their home. Recent urban planning policies in Melbourne have aimed to improve employment containment in the city's suburbs. While there has been analysis of the rates at which people both live and work within broadly defined 'local areas', little work has been done to investigate employment containment using smaller and more uniform catchment areas as the unit of analysis. This research attempts such a finer scale analysis using dasymetric downscaling techniques. A regression modelling approach supported by land use data, alongside a binary dasymetric method, is used to develop fine scale estimates of employment distribution, while binary and population-density weighted methods are used to develop a fine scale estimate of working population distribution. For the employment distribution estimate, the Poisson model that distributed employment to employment-related land use classes produced the smallest error. However, the error produced by this model is still high. For the working population distribution estimate, the population-density weighted estimate is the more accurate of the approaches, and overall produced low error. For the employment containment analysis, a number of employment centres were randomly selected and an employment containment catchment has been derived from a 5 km^2 commuting distance catchment. Commuting flows from an origin-destination matrix were area-weighted to estimate flows into the employment centre from the 5 km^2 catchment. The method is found to be potentially useful; however inspecting the results of this employment containment calculation highlighted flaws in the current estimates that should be addressed before the measures can be used to further analyse employment containment in Melbourne. Improvements to this method would support urban strategic and transport planning analyses at a metropolitan-wide scale.

AUTHOUR'S DECLARATION

I declare that the work in this dissertation was carried out in accordance with the Regulations of Westfälische Wilhelms-Universität, Münster. The work is original except where indicated by special reference in the text and no part of the dissertation has been submitted for any other degree. Any views expressed in the dissertation are those of the author and in no way represent those of the Westfälische Wilhelms-Universität, Münster. The dissertation has not been presented to any other University for examination either in Germany or overseas.

SIGNED: .....

DATE:.....

TERMS AND ACRONYMS

ABS	Australian Bureau of Statistics
CBD	Central Business District
COSP	Change of Support Problem
GWR	Geographically Weighted Regression
LGA	Local Government Area
MAUP	Modifiable Areal Unit Problem
OLS	Ordinary Least Squares
RMSE	Root Mean Square Error
SLA	Statistical Local Area

CONTENTS

ACKNOWLEDGEMENTS II

ABSTRACT III

AUTHOUR'S DECLARATION IV

TERMS AND ACRONYMS V

CONTENTS VI

LIST OF FIGURES VIII

LIST OF TABLES IX

1. INTRODUCTION 1
- 1.1 Rationale 1
- 1.2 Related Work 4
  - 1.2.1 Journey to Work and Employment Containment 4
  - 1.2.2 Areal interpolation 6
  - 1.2.3 Kriging and Geostatistics 7
  - 1.2.4 Dasymetric methods 8
  - 1.2.5 Areal interpolation and journey to work 12
- 1.3 Formulation of this study 13

2. DATA AND METHODS 15
- 2.1 Study Area 15
- 2.2 Data 15
  - 2.2.1 Source zones 16
  - 2.2.2 Target zones 17
  - 2.2.3 Ancillary data 17
  - 2.2.4 Validation zones 21
  - 2.2.5 Transport Network 21
  - 2.2.6 Commuting Data 21
- 2.3 Downscaling employment data 22
  - 2.3.1 Regression-based approach to deriving employment densities 22
  - 2.3.2 Generating employment estimates 26
- 2.4 Downscaling working population data 28
- 2.5 Employment containment assessment 29

3. RESULTS 32
- 3.1 Deriving employment density from regressions 32
- 3.2 Producing employment estimates 33
- 3.3 Downscaling residential data 47
- 3.4 Employment Containment 52

4. DISCUSSION 60

5. CONCLUSION 65
- 5.1 Further work 66

6. REFERENCES 67

APPENDIX A: R SCRIPTS FOR REGRESSION MODELLING..... 71

APPENDIX B: ARCPY SCRIPT FOR EMPLOYMENT DISTRIBUTION ESTIMATES FROM REGRESSION COEFFICIENTS 76

APPENDIX C: ARCPY SCRIPT FOR BINARY EMPLOYMENT OR WORKING POPULATION DISTRIBUTION ESTIMATES..... 78

APPENDIX D: ARCPY SCRIPT FOR WORKING POPULATION DISTRIBUTION ESTIMATES FROM TOTAL POPULATION DENSITY 80

LIST OF FIGURES

Figure 1: Map of the study area, showing Source zones (the SLAs) and the ancilliary data (land use classification) used in the study. 16

Figure 2: Schematic diagram of relevant nested Australian Bureau of Statistics data aggregations. 17

Figure 3: Graph showing the area occupied by the different land use classes within the study area, and the count of parcels of each land use type within the study area. 20

Figure 4: Division of study area based on the proportion of Urban or non-Urban (Agricultural and Parkland) covers. 25

Figure 5: Best estimate of employment distribution, based on Poisson model with employment attributed to employment-related land uses. 35

Figure 6: Residual plots comparing the predicted values and residuals from the regression modeling stage to the predicted values and residuals from the employment estimate stage. 38

Figure 7: Residual count in Validation Zones from the employment estimate based on the Poisson employment land uses model. 43

Figure 8: Percentage error in Validation Zones from the employment estimate based on the Poisson employment land uses model. 44

Figure 9: Residual count in Validation Zones from the employment estimate based on the OLS all land uses model that was split by region 45

Figure 10: Percentage error in Validation Zones from the employment estimate based on the OLS all land uses model that was split by region 46

Figure 11: Residual plots comparing the predicted values and residuals from two working population estimates 48

Figure 12: Best estimate of working population distribution, using the density-weighted distribution method. 49

Figure 13: Residual count in Validation Zones from the working population estimate using the density-weighted distribution method. 50

Figure 14: Percentage error in Validation Zones from the working population estimate using the density-weighted distribution method.. 51

Figure 15: The 40 sites randomly selected for the employment containment study. They are overlaid on the best final employment estimate surface (based on the Poisson employment land uses model). 52

Figure 16: Relationship between the estimated number of employees in the selected employment centres, the estimated percentage of employment containment for the centre, and the percentage of workers in the employment centre that live in the catchment. 58

LIST OF TABLES

Table 1: Land use classification derived from Australian Bureau of Statistics Mesh Block data, with notes about features..... 19

Table 2: Results of Ordinary Least Squares regression models to determine relative employment densities of land use classes. 34

Table 3: Results of Poisson regression models to determine relative employment densities of land use classes. 34

Table 4: Comparison of error in the employment estimates 36

Table 5: Comparison of mean square error, etc. for the population estimates produced by different models..... 47

Table 6: Summary of employment and employment containment in selected employment centres..... 55

1. INTRODUCTION

1.1 Rationale

This project studies employment containment in Melbourne, Australia. Employment containment is a component of journey to work analysis, a rich field of study relevant to contemporary transport planning, land use planning and labour market analysis. Employment containment, in particular, is a measure of the proportion of people that work in a location close to their home (studied by the likes of Burke, Li, & Dodson, 2010; Debenham, Stillwell, & Clarke, 2003; Yigitcanlar, Dodson, Gleeson, & Sipe, 2007). This measure is of interest because most urban planning academics and practitioners have long advocated for land use planning strategies that emphasise a mixture of housing and employment land uses. This approach, summarised by the term 'Compact Development' (Frank & Devine 2006) is thought to minimise a number of economic externalities including commuting time and costs (Flood & Barbarto, 2005), traffic congestion, and more recently, greenhouse gas emissions associated with climate change (Burgess, 2000). When combined with other complementary strategies, a healthy level of employment containment is thought to make a city a better place to live.

During at least the past decade, urban planning policies in Melbourne have been driven by this Compact Development philosophy, and have aimed to improve local employment opportunities and therefore employment containment in the city's suburbs. For example in 2009 the State Government of Victoria released their *Melbourne@5 million* planning strategy that was aimed at improving the environmental and social sustainability of the city's population growth trajectory over twenty years (Department of Planning and Community Development, 2009). This policy explicitly outlined an aim to change employment distribution in the city so that its resident population will have greater access to employment closer to home, with the designation of targeted activity centres where new employment growth would be concentrated and supported by public

transport and other infrastructure. The strategy was a reaffirmation of the Government's long term desire to move Melbourne from a 'Monocentric' to 'Polycentric' city so that people have more opportunities to work and use necessary services closer to their homes. With a political change of government at the end of 2010, this policy has since been shelved in favour of a new Metropolitan strategy, still in development, which is almost certain to be concerned with the same overarching urban planning issues. Therefore there is still a clear need for ongoing empirical work in this area to underpin the development of good public policy.

To assess the effectiveness and relevance of such urban planning policies now and into the future, it is useful to understand the current journey to work and employment containment patterns in Melbourne. While policies such as *Melbourne@5 Million* have been developed with the assumption of a problem of inadequate employment containment and an overconcentration of jobs in Melbourne's Central Business District (CBD), analyses of commuting patterns and employment containment in Melbourne's suburbs are somewhat incomplete (Davies, 2010). Indeed, Davies points to the fact that while there is a high concentration of professional jobs in the city's CBD, around 90% of the Melbourne metropolitan region's jobs are located in the suburbs, suggesting that there are already many jobs located close to people's homes. Knowledge about who travels where, however, is still underdeveloped.

Work commuting studies are made possible in Australia by Australian Bureau of Statistics (ABS) data that records both the residential and work locations (where applicable) of all people in Australia on the five-yearly census day. An Origin-Destination matrix provides a summary of the number of people travelling from residential locations to employment locations. Due to privacy restrictions, the matrix is only made available in aggregated form, and most commuting and containment analysis is conducted at the Statistical Local Area (SLA) level, an ABS designation (see for example Johnson, (2010), and more recently the Bureau of Infrastructure Transport and

Regional Economics, 2011), the larger municipal Local Government Area (LGA) level (e.g Moriarty & Mees, 2006), or even at larger aggregations of these. Traditionally, measures of containment are derived by counting the proportion of people who both live and work within the same administrative or statistical boundary (Yigitcanlar, Dodson, Gleeson, & Sipe, 2007). SLAs in the greater Metropolitan Melbourne area range in size from 1.9 km^2 in the CBD to 1137 km^2 on the fringe of the metropolitan region, thus comparisons of containment rates in SLAs are complicated by the varying size of the areal unit. And while data detailing worker's origin and destinations is available at smaller ABS defined aggregations (namely Census Districts and Destination Zones), analysts have appeared to shy away from using the data at this smaller scale because the boundaries of these two data sources are not concurrent- thus the traditional containment analysis method is not possible. At any rate, these smaller geographies are also not uniform in size and shape and hence they still face the same problem of varying size as do the SLAs.

While there has been some analysis of the rates at which people both live and work within these broadly defined 'local areas', little work has been done to investigate employment containment using smaller and more uniform catchment areas as the unit of analysis. A finer scale analysis assisted by land use classification may provide the opportunity for more meaningful employment containment comparisons, and provide a greater understanding of which features of the urban landscape contribute to local employment.

This thesis supposes that assessments of employment containment could be assisted by interpolation or downscaling methods that have been widely used in the general geographic literature. This could help to overcome the constraints of area-aggregated data issued at the level of administrative or statistical boundaries. The rest of this chapter reviews some relevant background and previous research in downscaling or interpolating from areal data, as well as reviewing some relevant journey to work and

employment containment analysis, in order to formulate the thesis aims that appear at the end of the chapter.

1.2 Related Work

1.2.1 Journey to Work and Employment Containment

Employment containment is one of a number of related employment metrics that aims to describe the relationship between the working population's place of residence and place of work. Other related measures include job/housing balance, minimum commuting distance or excess commuting (Boussauw, Derudder, & Witlox, 2011; Boussauw, Neutens, & Witlox, 2010), and the number of accessible jobs. I consider employment containment to be of particular interest because it is the metric that can tell us the most about the current situation of employment localisation in Melbourne.

The traditional approach to employment containment is to calculate the proportion of people living in an administrative or statistical boundary that also work within that same administrative designation. This is a simple and quick indicator of the degree of containment in a given area, but from a geographic analysis perspective it has some limitations. The Modifiable Areal Unit Problem (MAUP) is often discussed when analysing areal data (Páez & Scott, 2004), which is a generalisation of the Change of Support Problem (COSP) (see the review of these topics by Gotway & Young, 2002). These problems recognise that the changing size and shape of administrative boundaries impacts on the rates at which a given phenomenon will be measured in those areas, and hypothetically shifting those boundaries could lead to significantly different counts and even an apparently different trend in the phenomena. It seems that this problem is especially significant when dealing with the question of employment containment- after all, the measure is traditionally based around the proportion of movement either within or across administrative boundaries and so the position of the boundary or relative size of any administrative unit will greatly impact on the measure given. Larger administrative

boundaries would, by geography, seem more likely to contain workers within it compared to another administrative boundary of a smaller size. Furthermore, one can imagine a situation where people living close to the administrative boundary might travel a very short distance across the administrative boundary for work, or a person living in the extreme north of the administrative region travels to the southern extreme of the area without passing the boundary. Within the traditional analysis of containment, the former would be considered not to be contained while the latter would be, even though the former had travelled a shorter distance.

When analysing commuting and employment containment some authors have acknowledged this shortcoming in employing the traditional containment analysis methods (Horner & Murray, 2002; Boussauw et al., 2011). Therefore such studies are often taken to be only indicative and descriptive for a given area, rather allowing meaningful comparison between different cities, or even different zones within a city. Some authors have gone partway to overcoming the problem by aggregating up the administrative regions until uniform rates of containment are reached within the aggregations, thus creating 'commuting regions' that are controlled for containment rate if not for physical size (eg Bill, Mitchell, & Watts, 2007; Watts, 2009; Johnson, 2010). However, little work has yet been done in the opposite direction of disaggregating rather than aggregating to overcome this problem (LeSage & Fischer, 2010).

At these broader scales, studies of 2006 ABS data have found that self-containment rates in Melbourne are highest in the CBD (with a self-containment rate over 50%) and the larger SLAs in the outer reaches of Melbourne, which had containment rates of 30-40% in most cases (Bureau of Infrastructure Transport and Regional Economics, 2011). Some authors have looked at employment containment alongside other socio-economic variables, for example a study of employment containment by occupation that found that self-containment rates don't vary greatly by occupation in Melbourne, but in general people in management, professional and knowledge-industry

jobs are more likely to travel further for work, compared to people in low skilled occupations (Bill et al., 2007).

Elsewhere in Australia, Yigitcanlar, Dodson, Gleeson, & Sipe (2007) used a road network analysis between census collection districts to examine origin-destination work flows in master-planned estates in Australia. This was a rare use of smaller areal units to study employment containment, but the scope of the study was limited to a small number of recently-developed suburbs that all exhibited low containment rates.

1.2.2 Areal interpolation

Areal interpolation is the process of inferring the data value of some phenomenon in space where it has not been directly measured. The principles of areal interpolation are useful where data has been collected and aggregated at one geographic resolution, but are desired at a different resolution or aggregation. There are two related but distinct areal interpolation problems- that of downscaling to sub-areal units, and spatial misalignment of similarly sized by mismatched sets of administrative boundaries or other polygons (see for example Lin, Cromley, & Zhang, 2011). Early attempts at areal interpolation were later characterised as simple areal interpolation, using a uniform areal weighting to assume that the target area for which some variable of interest is to be calculated will take a proportion of that variable measured at the source area, proportionate to the fraction of the source area that the target area occupies in space (Flowerdew & Green, 1992). A variation on this areal weighting principle is Tobler's pycnophylatic constraint method (W. R. Tobler, 1979, and more recently employed by Kim & Yao, 2010 and Yoo, Kyriakidis, & Tobler, 2010) which replaces the uniform distribution assumption of areal weighting, with a smooth density function extending to adjacent source zones, while retaining the original count of the source zones. Another, though less sophisticated variation is the Kernel Smoothing technique (Bracken & Martin, 1989; Martin, 1996)

which collapses all the population data into a point and employs inverse distance weighting.

1.2.3 Kriging and Geostatistics

Related to these smoothing techniques is the field of geostatistics, including area-to-point kriging which smooths known rates of the phenomenon across space via correlation of a variable with itself through space (e.g Kyriakidis, 2004). Geostatistical methods produce error estimates that can indicate the accuracy of the interpolated surface (Yoo et al., 2010). Smoothing may not be appropriate to the phenomenon under investigation in this current study, since population and employment density are products of their human-built urban environment, which can often have sharp edges and jumps in values rather than smooth continuous surfaces. Intuitively this seems especially true for measures of employment distribution as opposed to (residential) population distribution, as nodes at which employment occurs tend to be much more clustered in the urban space, compared to places at which residential population occurs, which tend to be more spread out throughout the urban space. Therefore, kernel smoothing and area-to-point kriging, with its focus on smoothing, do not seem to lend themselves to interpolating employment density at a small scale. Further the underlying smoothness of the process must be known or assumed. Nagle (2010) produced an employment surface via area-to-point factorial kriging in order to study employment agglomerations in the Denver metropolitan area. However, the scale of the interpolation was large and general, and no ancillary data such as land use information was used to inform the kriging. Instead, a standard covariance function was used. The main advantage of this method is the smooth surface it produces, though in the case of the current study a smooth statistical surface is not a high priority as employment density is not assumed to be a smoothly varying process at the scale that this project investigates.

1.2.4 Dasymetric methods

Following from the simple areal interpolation methods, intelligent methods were developed that incorporate ancillary information about the likely distribution of a given variable. These methods are often termed dasymetric and work in this field has tended to focus on residential population as the variable to be interpolated (see for example Mennis, 2003).

Dasymetric methods are characterised by their use of knowledge about the likely distribution of some phenomena within a *source zone*. This additional knowledge is used to distribute the known quantities of a given phenomenon at the source zones over the study area, in order to infer quantities in *target zones* of a different scale or areal aggregation. Dasymetric methods can be classed in a number of ways, with one of the key distinctions being between binary and 3-class methods. In binary methods, the variable in question is distributed evenly across areas that are thought to be occupied by the phenomena, and not attributed at all to areas that are deemed to be unoccupied. In three-class methods, the variable is allowed to have different densities or rates across the occupied area, given knowledge or assumptions about the rates of that phenomenon in different land classes. Many more recent studies have used this principle to downscale population data based on land use or land cover, with population distributed at different densities depending on known or assumed populations densities in different land use classes (Langford, 2006; Mennis, 2003; Reibel & Agrawal, 2007). Mennis & Hultgren (2006) termed these methods 'intelligent' and derived their own procedure where analysts could use both sampling of population in various land cover classes, and their own professional judgement to assign population density to various land cover classes. This method, like most others variations of dasymetric methods, employed a mass-preserving method in attributing population density, where the downscaled population must sum to the known population from the source data (Gregory, 2002).

Various techniques have been used to derive the downscaled population estimates. Ordinary Least Squares (OLS) (Langford, 2006; Yuan, Smith, & Limp, 1997), Poisson (Flowerdew & Green, 1989), Bayesian (Mugglin, Carlin, & Gelfand, 2000) and other regression methods have been used, where the coefficients derived in the regression are treated as the relative population density of the various land cover classes. The technique developed by Mennis and Hultgren (2006) relies on sampling population densities from source zones that are entirely covered (or in some cases, mostly covered, say by a threshold of 80% or more) by one of the land cover or land use classes of interest. It can be noted here that while some authors distinguish between 'dasymetric' (binary or three-class sampling based methods) and statistical methods, Langford (2006) writes that the various methods employed can be seen as variations along a continuum of these two techniques. Indeed, density coefficients may be derived by global or regional regressions, then re-scaled for each source zone to recreate the known population of that zone (employing the mass-preservation constraint). In this way, the regression-derived density coefficients act as ratios of the *relative* density that each land class should take. Langford tested the binary dasymetric method against the various formulations of a three-class dasymetric method. He found that combining regression with dasymetric re-scaling of the density coefficients produced the most accurate results of the three-class methods; however, within his case study a binary method actually performed better than all the three-class methods, probably because it is less sensitive to fluctuations in different land class densities.

Re-scaling is only one of the ways to ensure mass-preservation of the known counts. Liu, Kyriakidis, & Goodchild (2008) provide a variation on the regression-based dasymetric approach, by deriving population densities for each land use class via linear regression, but then using area-to-point kriging to smoothly distribute the residual source zone population counts to the target zones. Paralleling the sampling approach to dasymetric downscaling, the kriging process relied on sampling population density in

districts of entirely one land use zone in order to create the kriging semivariogram. The advantage of using kriging for this process is that spatial information is incorporated to attribute the population- e.g. if there is a high density population area in the vicinity then this information is used to attribute the population, rather than attributing the population uniformly. The authors of this study found that incorporating area-to-point kriging reduced Root Mean Square Error (RMSE) compared to the regression alone. Similarly, Kim & Yao (2010) combined dasymetric methods with pycnophylactic smoothing, created a smoothed population surface that was superior to either a standard dasymetric or standard pycnophylatic interpolation.

Whichever method is used, the effectiveness of intelligent dasymetric methods relies most heavily on the usefulness of the ancillary data for modelling the particular phenomenon in question. Gallego (2010) compared a number of dasymetric variations: the Expectation-Maximisation Algorithm, based on Dempster, Laird, & and Rubin (1977); logit regression; and the Limiting Variable method (based on Eicher and Brewer, 2001), which attributes a minimum population density to all classes in the study area, then distributes the remainder across other classes, assigning particular density thresholds along the way. Gallego found that the choice of algorithm did not have a great impact on the accuracy of the downscaling. Langford (2006) also emphasised the point that the ancillary data are of greatest importance. A variety of ancillary supports can potentially be employed, such as land cover or land use classes (Gallego, 2010), remotely sensed images (Deng, Wu, & Wang, 2010; Harvey, 2002; Silván-Cárdenas et al., 2010), road network data (Li et al. 2010) attributing population to areas where road network is present, cadastral boundaries (Maantay & Maroko, 2009) and address point data (Zandbergen, 2011).

In either the sampling or regression based methods, the classical approach assumes that relative population densities are stationary across the study area. However, it is possible that differences in population and urban form across a study area can result

in different relative population densities across the different land classes present. Some analysts have attempted to account for geographic non-stationarity of population density across the study area. Mennis (2003) sampled population densities from sub-units within each county to derive local densities, in order to account for differences in the urbanisation of different counties. Langford (2006) found that using regional (at UK District level) rather than global regressions produced a more accurate prediction, as long as enough units were available within regions to perform a reliable regression. However, Langford rightly pointed out that using District or any other such boundary as a basis for forming regions is quite arbitrary, and therefore may not be representative of density variations across space. Lin et al. (2011) approached this problem by using geographically weighted regression (GWR) to derive population densities. While noting that the GWR was more successful in spatial misalignment problems than downscaling problems, overall they concluded that for their study area GWR did a better job of interpolating than OLS regression did. Another form of weighted regression, Quantile Regression, was recently used to perform areal interpolation of population (Cromley, Hanink & Bentley, 2011). Although it is not specifically a form of spatial regression, quantile regression can be used to derive a regression line and therefore unique regression coefficients for each observation (i.e., source zone) in a study area. The authors of this study also found that this method outperformed OLS regressions and binary dasymetric methods.

Brinegar & Popick (2010) recently compared a number of different population estimation methods, including land-use based 3-class dasymetric methods, road network-based dasymetric methods and statistical regression. Their comparison and discussion points out that different methods have differing strengths and weaknesses in estimating a population based on different conditions, such as heterogeneous land use and unusually high or unusually low population density. This highlights the lesson that a particular technique that works for one case study and for a specific purpose, may not necessarily

produce the best result under other conditions. This raises an important distinction for the current study. This review of the dasymetric literature has mainly focused on downscaling population data. The bulk of the dasymetric literature focuses on this problem, rather than downscaling other socio-economic variables (such as employment distribution as in this study). Since employment has a different spatial distribution to residential population, findings from this literature may not translate easily to the downscaling of employment data.

One related problem that has attained some attention, however, is inferring daytime as opposed to nighttime population distribution. In general, population data record a person's residential address. Most people spend evenings and nights at their residential address, but much fewer are there during the day- they are either at work, school, or involved in other activities. Researchers interested in emergency management and traffic planning, for example, are more interested in daytime than nighttime populations. For example Sleeter and Wood (2006) used dasymetric techniques based on a detailed business database, and Kobayashi, Medina, & Cova (2011) used the pycnophylactic method to produce a smooth population surfaces that was appealing and easy for policy makers to look at and understand, and displayed expected population distribution at different times of the day. This work is similar to the interests of the current study, though the purpose for looking at the daytime population may be different.

1.2.5 Areal interpolation and journey to work

Use of areal interpolation methods alongside commuting analysis is so far limited. Li, Corcoran, & Burke (2010) used areal interpolation to paint a more detailed picture of the origin-destination work flows at a sub-areal scale, using the road network to perform binary dasymetric downscaling of commuting flows in South East Queensland. In another example, Boussauw, Neutens, & Witlox (2010) produced a 4 km grid interpolated surface of distance travelled to work in Flanders, Belgium.

Jang & Yao (2011) produced an interpolation using 'flow lines' of an origin-destination matrix of traffic flow data, focusing on aligning mismatched traffic zone and census zone data sources, rather than downscaling the origin-destination flows. In this case, no ancillary data was employed to support the interpolation. Kaiser & Kanevski (2010) produced a dasymetric population mapping method explicitly to assist with traffic modelling, focusing on residential population.

1.3 Formulation of this study

The aim of this study is to explore a process for estimating employment containment in uniform catchment zones around a given employment centre. In particular, the key questions posed by this study are:

1. Can dasymetric processes (binary, sampling and regression approaches) assist in producing a useful employment containment measure that overcomes some of the spatial irregularity problems associated with traditional employment containment measures?
2. What do these derived measures say about employment containment in the study area of Melbourne, Australia?

In order to achieve this, a downscaled estimate of both employment and working population distribution is required. In particular, binary dasymetric and 3-class methods are explored for deriving these distribution estimates. A workflow is developed that includes a validation step to give an indication of the accuracy of the derived employment and population surfaces.

From the final employment surface, a number of employment centres are randomly selected and an employment containment catchment is derived from a 5 km^2 commuting distance catchment. Commuting flows from an origin-destination matrix are areally weighted to estimate flows into the employment centres from the 5 km^2 catchment.

A brief analysis of the employment containment of these centres is presented, as well as an assessment of the performance of the downscaling process for enabling a useful measure of employment containment.

2. DATA AND METHODS

There are three parts to the analysis: Estimating employment distribution, estimating working population (residential) distribution, and then combining these for the employment containment estimate in fixed catchments around selected employment centres. A description of this process follows the details of the study area and the data used. The R statistical program was used to perform the regressions, ArcGIS analysis tools are primarily used to perform GIS tasks and Excel 2010 was used for data manipulation.

2.1 Study Area

The focus of this study is the greater metropolitan region of Melbourne, Australia (see Figure 1). At the time of the 2006 Census (from which all data for this study is sourced), the total population of the greater Metropolitan region was 3 599 644 people. There were 1 741 193 people living in Melbourne and adjacent regions (the population of focus in this study) who were employed (termed 'working population' from hereon). Furthermore, 1 507 060 people identified their workplace as within one of Melbourne's Destination Zones (as described further below).

2.2 Data

Following the formulation of Gregory (2002), Langford (2006) and Li and Corcoran (2010), the data used to perform the employment and working population downscaling in this study is designated at four nested levels, shown in Figure 2 and outlined in the text below. All data is derived from the Australian Bureau of Statistics (ABS) Census of Population and Housing, 2006.

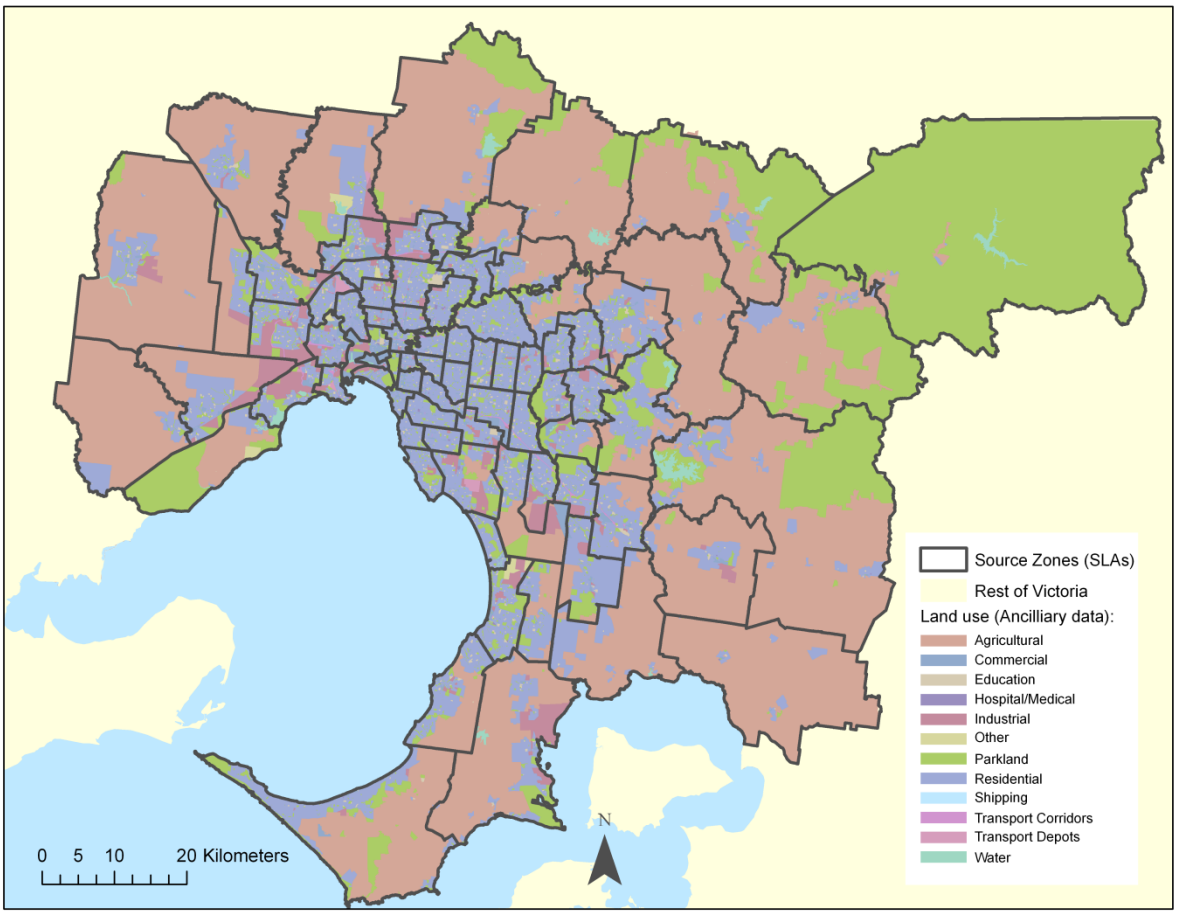


Figure 1: Map of the study area, showing source zones (the SLAs) and the ancillary data (land use classification) used in the study.

2.2.1 Source zones

Source zones are administrative boundaries with counts of either the number of people working in that zone (for the employment surface downscaling) or the number of employed people that live in that zone (for the working population downscaling). These are the known counts from which the downscaled estimates will be derived. In this study, the ABS SLA designation is used as the source zone in both the employment and working population downscaling. In the Melbourne study area there are 80 SLAs, which are displayed in Figure 1.

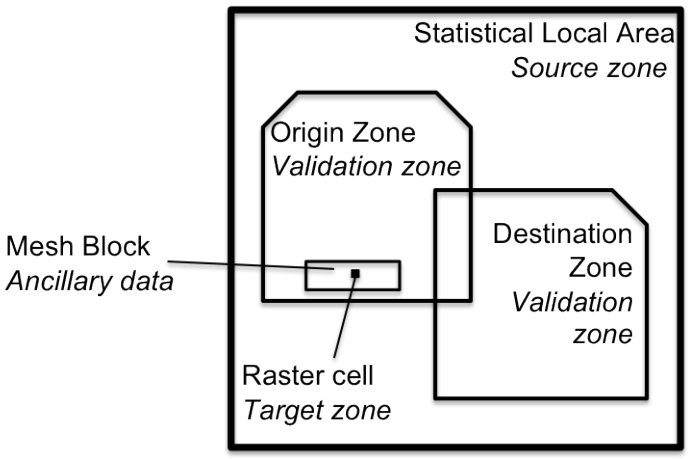


Figure 2: Schematic diagram of relevant nested Australian Bureau of Statistics data aggregations.

2.2.2 Target zones

Target zones are the unit at which the downscaled data is estimated. The zone may be another administrative designation for which the count data of interest is not available, some other user-defined designation, or a uniform raster surface. In this study the latter is chosen, at a 10m resolution, as this can more easily be integrated with the employment catchment study in which it will be deployed.

2.2.3 Ancillary data

Ancillary data supports the downscaling by providing information about how the variable of interest is distributed in the source zones (and therefore how it should be attributed to the target zones). Where dasymetric downscaling has been performed on population counts (by far the most common variable studied in the dasymetric literature), the ancillary data is usually in the form of land cover categories or remotely sensed imagery that may be used to categorise urban areas into residential density classes. However, little work has been done to perform downscaled employment estimates via these methods, and intuitively one suspects that estimating employment either by land cover or remotely sensed urban form may be a more difficult task. However, a land *use*

(as opposed to land cover) classification is available from the ABS, in the form of Mesh Blocks.

Mesh Blocks are the smallest designation at which ABS population data are available at, and includes a population count and a land use classification for each area. The land classes are relevant for determining employment and non-employment land uses. The land use classification is described in detail in Table 1, and their distribution with the study area is displayed in Figure 1. This classification is used as the basis for redistributing the employment and working population counts, and the population count is also used in the working population estimate. Note that in the original ABS classification the classes ‘Transport Depots’ and ‘Transport Corridors’ are a single class called ‘Transport’. They are split for the purposes of this study because they clearly represent different land uses. There are 47 725 Mesh Blocks in the study area. The distribution of the Mesh Block classes amongst the different classes, and the total area covered by each class, is summarised in Figure 3.

Table 1: Land use classification derived from Australian Bureau of Statistics Mesh Block data, with notes about features

Land class	Description	Number of parcels	Total area (km^2)
Commercial	Business and shopping zones, office blocks, strip shopping zones, shopping malls	1800	92.4
Education	Schools (primary and secondary), university campuses	1262	56.4
Industrial	Manufacturing, warehousing	814	251.4
Hospital/Medical	Hospitals or other large medical centres	78	3.9
Shipping	Mostly offshore regions designated as shipping zones	2	0.7
Residential	Areas primarily occupied by housing	37259	1676.2
Agricultural	Farming land	836	5190.3
Parkland	Regional and local parks of varying size	5082	5220.3
Other	Miscellaneous land uses such as water treatment plants and military accommodation	23	16.7
Transport Corridors	Land along railway lines	438	14.6
Transport Depots	Minor suburban airfields, public transport depots	5	8.2
Water	Lakes, dams and other bodies of water	126	49.6

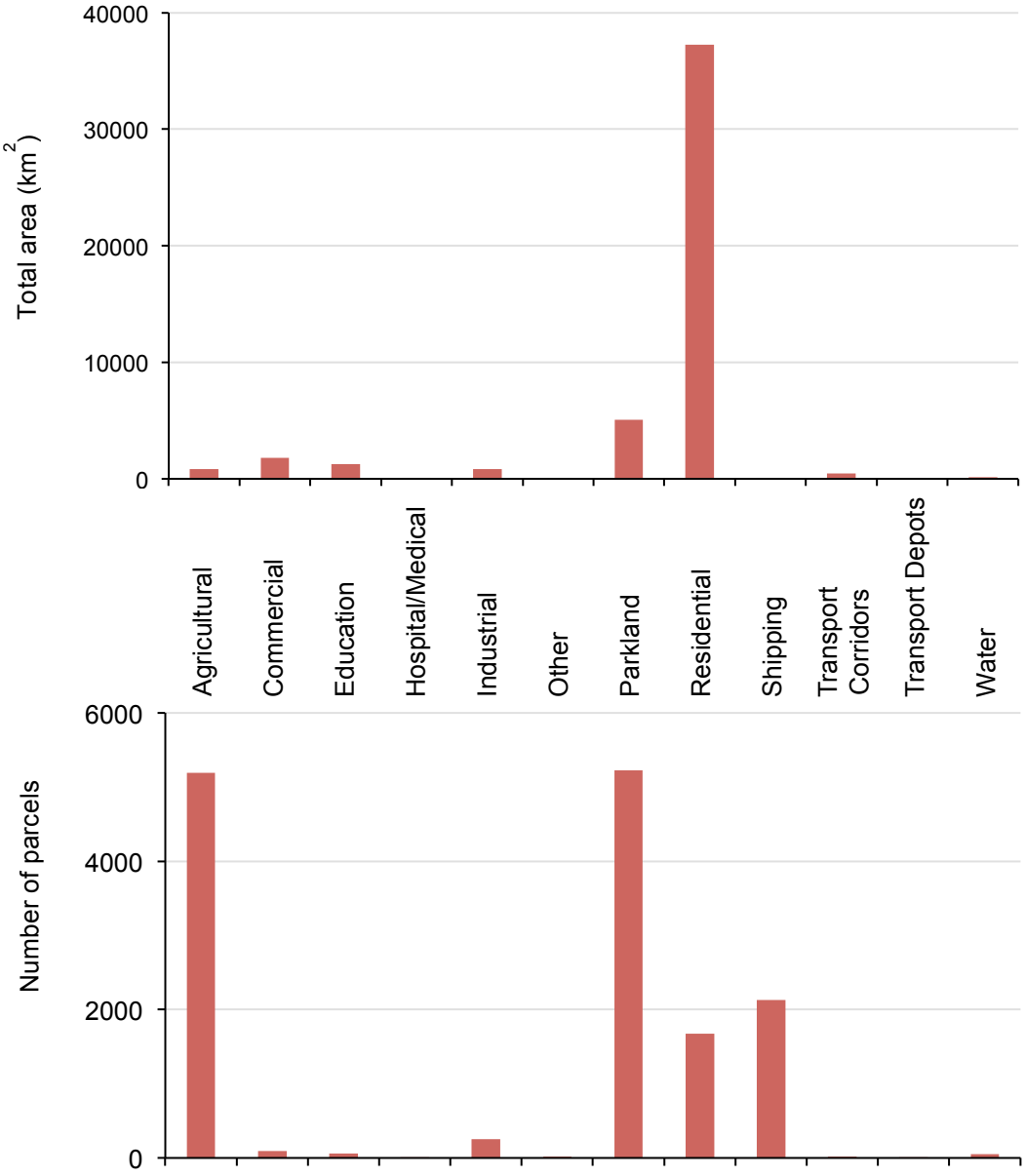


Figure 3: Graph showing the area occupied by the different land use classes within the study area (top), and the count of parcels of each land use type within the study area (bottom).

2.2.4 Validation zones

Validation zones are used to re-aggregate the downscaled data to test how accurately the downscaling process distributed the variable in question. Since ABS data aggregations are for the most part nested as smaller and smaller aggregations that can be aggregated up to a larger designation (e.g, Mesh Blocks can be combined until they are coincident with the SLA boundary), the known employment or working population counts at a designation smaller than the source zone must be used to validate the employment and working population estimates. The SLAs are further subdivided into Destination Zones associated with employment counts, and Origin Zones associated with residential counts. Origin Zones and Destination Zones are of similar size but generally have misaligned boundaries. In this study I use an aggregation that the Victorian Department of Planning and Community Development specifically developed for studying work and home locations of commuters- therefore the geography is different and slightly more coarse than that normally available from the ABS. There are 1301 Origin Zones and 652 Destination Zones in the geography used in this study.

2.2.5 Transport Network

To produce the commuting catchments around employment centres as part of the employment containment estimate, I use a vector layer of the Major Roads and Railway lines of Melbourne.

2.2.6 Commuting Data

To produce the employment containment estimate, information about the employment distribution must be linked to information about the working population distribution. An Origin-Destination matrix tallies the number of people residing in an Origin Zone travelling to a work in an SLA.

2.3 Downscaling employment data

The literature on areal interpolation provides a number of variations on the process of downscaling. Dasymetric approaches can roughly be divided into sampling-based approaches and regression-based approaches (Langford, 2006). The sampling-based method was discussed by Mennis (2003). The method relies on sampling the population (or in this case, employment) density of the various land classes from zones that are entirely covered or close to entirely covered (80% or more) by one land class. This is in order to derive an average employment density for that land class. Where the study area size allows it, Mennis recommends dividing the study area into sub-regions, such as municipalities, to better account for spatial differences in the variable of interest. Early data analysis for this study indicated that the SLAs are too large to provide sample zones that are covered or even 80% covered by each land class. Li & Corcoran (2010) had a similar finding when using dasymetric methods to perform downscaling based on SLAs in a South East Queensland study area. The authors instead use a regression method.

2.3.1 Regression-based approach to deriving employment densities

A number of authors use linear regression to produce relative population densities of classes within their study area. OLS regression, which minimises the sum of squares of the residuals, is commonly used in this context including by Harvey (2002), Langford (2006), Reibel & Agrawal (2007), Yuan et al., (1997), and Li & Corcoran (2010). Under this model, the variable of interest *E* (employment distribution) at the source zone is equal to the sum of the variable's density of each class *C* multiplied by the area *A* of that land class within the source zone.

$$E_s = \alpha + \left(\sum_{c=1}^{C} \beta_c A_{sc} \right) + \varepsilon_s$$

Therefore, given that *E* and *A* are known, the density in the land class equates to the regression coefficient and can be derived from the OLS regression. The coefficient corresponds to a global relative population density of each land class. In this case, the employment count for each SLA is regressed against the amount of each land use class in that SLA. A number of different configurations are tested in order to find a best fitting model.

An alternative to the OLS regression is Poisson regression. A Poisson distribution is often considered to be the best model for population counts (Flowerdew & Green, 1989, 1992), and a Poisson regression helps to avoid producing negative population totals in the final estimate (Langford, 2006). However a Poisson model has the assumption that the conditional variance of the outcome variable equals the conditional mean.

Regardless of whether OLS or Poisson regression is used, a regression without an intercept is recommended, since a theoretical source zone with zero area should have a zero employment count (Langford, 2006; Yuan et al., 1997; Harvey, 2002). In this study I test both OLS and Poisson regression models to compare their accuracy in describing relative employment densities in the study area.

A number of variations on the models are tested, altering the variables in the model to find the most fitting descriptors of employment distribution. These variations seek to distribute employment amongst the land classes in a way that best mimics the known distribution of employment in the source and validation zones. Should the model include information about all the land classes present in the source zone, or should it only incorporate those land classes that are presumed to be most closely associated with employment (Agricultural, Commercial, Industrial, Education, Hospital/Medical land

classes; from here on referred to as 'Employment land uses')? Or perhaps something between the two, excluding only land classes that are clearly *not* associated with employment (Parkland, Water, Shipping, and Transport corridors), while including some that might support some employment (Residential, Transport depots, 'Other', these along with the Employment land uses are from here on referred to as 'Urban land uses'). These variations are run in different versions of the model.

Local regressions that break the study area into some smaller regions for analysis have been used successfully in some cases (Langford, 2006; Yuan et al., 1997). As discussed by Langford (2006), using local regression raises the question of precisely how to break up the study area into 'local' areas and demands that the local areas have enough sub-areas for a robust statistical sample. Using a larger administrative designation is a commonly suggested solution, but is quite an arbitrary way to divide the study area, and at any rate in the context of this study there is no obvious administrative designation that could break the study area into smaller regions. An alternative is to break the study area into regions based on the relative representation of land use classes in the source zones. The Melbourne study area has a concentric form where the inner urban/suburban core, covered by relatively small SLAs dominated by urban covers, is surrounded by a ring of larger SLAs characterized by lower overall urban cover and a high proportion of agricultural land or parkland (Figure 4). Since the global density estimates derived by the regressions in fact represent the *relative* employment densities that each land use class should take, it is possible that globally derived employment densities may not adequately describe these differing forms. Therefore the study area is split into two regions, one where the Agricultural and Parkland classes represent 50% or more of the land in the source zone, and the other where all the other classes represent 50% or more of the source zone.

A final variation is introduced in an attempt to control for the difference in the size of the source zones. The dataset is altered so that the employment *density* of the

source zone is regressed against the fraction of the source zone that each land class represents. In this case the coefficient derived is a density fraction.

To assess the suitability of the resulting regression model, a number of statistics are used. The R^2 value is a measure between 0 and 1 of how well the model describes the

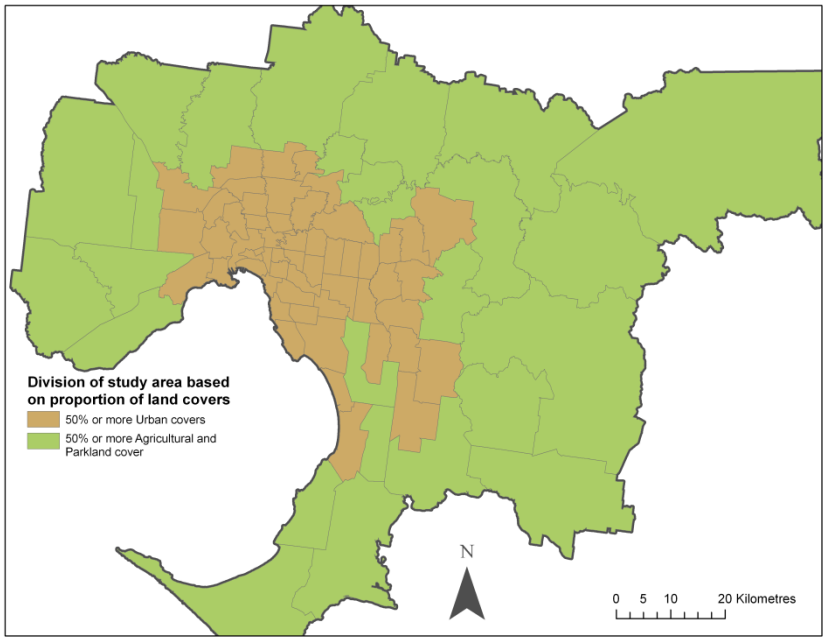


Figure 4: Division of study area based on the proportion of Urban or non-Urban (Agricultural and Parkland) covers.

regressed data, with values closer to 1 indicating better fitting models. While there is some controversy around best way to measure model fit for OLS regressions with no intercept, Eisenhauer (2003) notes that the R^2 value is valid as long as it is used to compare no-intercept models to each other.

Aikake's Information Criterion (AIC) can also be used to compare model fit. Models with a smaller AIC score are considered to be better models because the score penalizes models that have a large number of explanatory variables (Rosenhein, Scott, & Pratt, 2011). The statistical significance of the model coefficients is also seen as an

indicator of the model fit (Rosenhein, Scott, & Pratt, 2011). These three measures are used as a general indicator of model fit to select the best fitting models that are then tested by generating employment estimates based on the regression coefficients.

The detailed R script of the regression modelling is shown in Appendix A.

2.3.2 Generating employment estimates

There are two possible approaches for producing the downscaled employment surface: assigning the downscaled population to the areas that were used as classified ancillary data, or creating a raster surface. Liu et al. (2008) argue that a surface is the preferred output, as this is easier to integrate with other data sources. Therefore in order to generate the population estimates, I first produce a 10m-resolution raster with the values in the raster being the regression-derived density coefficients for the relevant land use class at that location (multiplied by a factor of 100 to equal the density of a $10m^2$ raster cell).

Applying the regression derived density coefficients to the study area results in over and underestimation of counts at the source zone level (literally the residuals of the regression performed to derive the density coefficients). There are at least three ways that these residuals are dealt with in the literature: 1) not at all; 2) by using area-to-point kriging to smooth the residual across the source zone (Liu et al., 2008); and 3) re-scaling the global density estimates within each source zone so that they reproduce the known count within the zone, a condition known as mass preservation or the pycnophylactic constraint (W. R. Tobler, 1979). Option 1) is disregarded as unsuitable for the current study; Option 2) is investigated but requires sampling of employment density in source zones of entirely one land use zone in order to create the kriging semivariogram. As was discussed in the introduction to this section in relation to employing the sampling-based dasymetric methods, the source zones are too large and heterogenous in land cover to

allow such sampling. Option 3) is therefore adopted as it is both possible and easy to apply in the context of this study.

To apply the re-scaling, the initial estimate in each source zone is summed using the Zonal Statistics tool in ArcGIS, and a re-scaling factor for each land class within each source zone is derived by the following equation (Gregory, 2002; Langford, 2006):

$$d_{cs} = \frac{E_s}{E_{is}} \cdot d_c$$

Where d_{cs} is the density estimate for land class c in source zone s, E_s is the actual employment count in the source zone, E_{is} is the estimated employment count in the source zone produced by the initial global density estimate, and d_c is the initial global density estimate in class c. The new locally-scaled densities are then applied to the study area using a Raster Calculator operation, multiplying the rescaling factor by the original global densities as in the above equation. This derives a final fine-scaled employment estimate for each $10m^2$ cell in the target raster.

In addition to the regression-based estimates, a binary estimate of employment distribution is produced, following the method discussed in Langford (2006) and Mennis (2003). This is a simplified version of the above process, where employment land uses are given a raster value of '1', and non-employment land uses a value of '0'. Zonal statistics are summed to find the area of employment-occupied land in the source zone. The employment count is divided by the employment-occupied area to derive an employment density for the source zone. The employment density for each source zone is then applied to the employment-occupied area via the Raster Calculator, producing a final binary estimate of employment distribution with an equal employment density for all employment land covers in each source zone.

To assess the accuracy of the downscaling, the final estimate is again summed, this time to produce an estimated employment count for the validation zones (the

Destination Zones). I compare a number of measures of the residuals of the estimate. The RMSE as is calculated, and the Adjusted RMSE which follows Gregory, (2002) and Lin et al. (2011), adjusting the RMSE to take account of the original observed population. These are measures of the average variance of the estimate compared to the known employment counts in each validation zone, and thus the accuracy of the downscaled estimate. A similar measure is the coefficient of variation following Fisher & Langford (1995), which divides the RMSE by the mean of the source zone populations. The Mean Error is the average of the difference between the estimated and observed employment count values. Note that given the mass-preserving constraint applied during the downscaling process, where the initial estimate was re-scaled to match the known employment counts in the source zone, the Mean Error should theoretically always be zero. However, some error is added when translating values between the areal-based employment counts or estimates, and the density values stored in raster form. Therefore the Mean Error can be seen as a measure of the error arising from reaggregating raster values to an areal-based count at the validation zones.

The abovementioned metrics are compared for each of the five employment distribution estimates to identify the most accurate estimate. A detailed script of the regression-based employment estimate procedure is given in Appendix B; the script for the binary estimate is given in Appendix C.

2.4 Downscaling working population data

For the working population estimate, the study area is expanded to include Source zones adjacent to Melbourne, in order to include any working population that may be within a 5 km commuting distance of employment centres in the Melbourne study area. Two estimates of the residential working population are produced. The first is a binary estimate, following a similar procedure as discussed in Section 2.3.2 for the binary

employment estimate. However, for the working population estimate, the residential land class takes the value '1' and all other classes take '0'.

The second estimate is weighted by the known total population density of each Mesh Block zone. A raster surface is produced of the known total population density of the study area, again at 10m resolution. Zonal Statistics in ArcGIS is used to calculate the total population of the source zone. Then, the known working population of the source zone is divided by the total population of the source zone, to derive a working population weighting for each source zone. The total population raster is multiplied by the working population weighting in the Raster Calculator, to derive the final working population estimate.

Once again the RMSE, Adjusted-RMSE, Coefficient of Variation and Mean Error are used to assess the accuracy of the estimates once re-aggregated to the validation zones. A detailed script of the working population estimate procedures are given in Appendices C and D.

2.5 Employment containment assessment

In order to make an estimate of employment containment, the employment and working population distribution estimates need to be linked to the known flows of people from the Origin Zones to a destination SLA; and then, these flows must be downscaled accordingly. The flows are stored as an origin-destination matrix. The matrix is imported to ArcGIS with the destinations represented by individual records and the origins represented by database fields.

From the best estimate of employment distribution, the final employment raster dataset is converted to a polygon feature class. A subset of 40 'employment centres' (any parcels that are estimated to have some employment associated with them) is selected to perform the containment assessment. The subset is randomly selected from those occurring in an employment validation zone that had been estimated to within ±10%

accuracy. The random selection was performed using the 'Sampler' toolbox in ArcGIS (Harold, 2011). The centres were visually inspected on Google Maps to identify their location within the local urban landscape, and the classification of the land use was identified from the Mesh Block data.

To generate the $5km^2$ commuting distance catchment around each employment centre, the centroids of each of these parcels was first derived. The centroid point is used to represent the employment centre as part of a Service Area Analysis, a feature of the Network Analyst extension in ArcGIS. Line coverage of the major road and railway for Melbourne is included in the analysis to create the $5km^2$ commuting catchments around each of the centres. While theoretically a number of different distances could be chosen for the study, 5km is chosen in order to investigate rates of highly localised employment. Travel surveys from the Victorian government suggest that about 35% workers travel less than 10 km to work each day, while around 15% travel 5 kms or less (Bureau of Infrastructure Transport and Regional Economics, 2011).

A number of spatial intersections and table joins were performed to derive the employment containment estimate. The aim is to calculate the number of workers in this catchment that work in the employment centre, as well as the total working population of the catchment in order to calculate a percentage of containment.

Each catchment polygon was associated with the employment centre point so that information about the location of the employment centre and the employment estimate was retained. The catchment polygons were intersected with Origin Zones (the working population validation zones) to identify the part of these zones that fall inside the employment catchment. Running Zonal Statistics calculates the estimated working population in each of the Origin-catchment intersection zones. This is related to the estimated working population in the entire Origin Zone (calculated during the validation step of the working population estimate). The proportion of the working population of the Origin Zone that falls inside the employment catchment is calculated.

The data is spatially joined with the employment source zones (the SLAs) so that the proportion of the source zone's employment that occurs in the employment centre, can be calculated.

To estimate the number of workers from the catchment that work in the employment centre, the following equation is used:

$$E_l = \sum_{o=1}^{n} \frac{E_t}{E_S} \times E_{so} \times \frac{P_{oc}}{P_o}$$

Where E_l is the number of people from the local catchment area working in the employment centre, E_t is the employment estimate for the employment centre, E_s is the known employment count from the source zone, and E_{so} is the count of people working in the source zone that travelled from origin o. P_{oc} is the estimated working population living in the part of the Origin Zone covered by the employment catchment, and P_o is the estimated total working population of the Origin Zone. The contribution of each Origin Zone in the catchment area is summed to reach E_l. The calculations were performed in Excel. The Employment containment of the employment centre is calculated as the proportion of the working population in the catchment, that works in the employment centre. Additionally, the percentage of people working in the employment centre that came from the local catchment area, is calculated.

3. RESULTS

3.1 Deriving employment density from regressions

During the analysis it became apparent that the one SLA covering the CBD of Melbourne could be considered an outlier as it has the highest employment count but the smallest overall area. Therefore models were run excluding this outlier from the analysis.

It is difficult to compare the Poisson and OLS models as an R^2 score is not produced for the Poisson models. Therefore, the Poisson and OLS models are considered separately. The results of the various OLS models are shown in Table 2. A simple ranking for the OLS models was produced by ranking all the models by their R^2 score, with the highest R^2 ranked as 1, second highest as 2, and so on. The models were then ranked again by AIC score, with the lowest AIC ranked as 1, and so on. The two ranks were then summed and the models ordered lowest to highest based on this score, to find the best fitting model.

The model that covered the agriculture-and-parkland dominated region only was the best fitting model. This does not necessarily tell us anything special as the high R^2 and low AIC may be a direct result of the small sample size of this regression compared to the other regressions (Langford, 2006): There were 26 source zones for the agricultural-parkland-dominated region and 53 source zones for the urban-dominated region, compared to 79 source zones for the whole study area.

Of the models covering the whole study area, the model with all land uses and using density fractions appears to have the best fit, followed by the employment model that used density fractions.

The OLS models of the whole study area that used the raw data were the lowest ranking of the OLS models. In general, however, these produce a significant estimate of the density coefficients of the various land covers, where other better-ranking models did not.

The results of the Poisson models is shown in Table 3. For the Poisson models, no R^2 score is generated for the models, so the Poisson models are compared and ranked based on their AIC score only. Additionally, as density fractions cannot be used in the Poisson model (as the Poisson model is designed specifically for count data), there is a smaller number of Poisson models than OLS models.

Of the Poisson models, the model with the full land covers was the best fitting, followed by the urban model and the employment model. All the Poisson models tested were found to produce significant estimates for all the land use classes included in that model.

It is difficult to select which are clearly the best fitting regression models, as there is inconclusive information provided by the two or three tests applied. In any case, these regression models only produce a preliminary density estimate that will be scaled during process of producing the employment estimates. Therefore I decided to generate employment estimates from all of the models, rather than eliminate possible best fitting models at this stage.

3.2 Producing employment estimates

A summary of the accuracy of the employment estimates produced by the various models is shown in Table 4. Comparing the various fit metrics produces the unexpected finding that the Poisson model with employment land classes and covering the entire study area produced the employment estimate with the lowest overall error, despite being one of the poorest fitting models based on the initial regression. Conversely, the estimate that incorporated different densities for the urban-dominated and agricultural-and-parkland-dominated regions produced more error than the estimate based on a universal model, even though the results of the initial regression models suggest that these were the best fitting models for the data. The Poisson models uniformly produced more accurate

Table 2: Results of Ordinary Least Squares regression models to determine relative employment densities of land use classes. Co=Commercial, Ed=Education, In=Industrial, Ot=Other, Sh=Shipping, HM=Hospital/Medical, Re=Residential, TD=Transport Depot

Model	R^2	AIC	Significant estimates	R^2 ranking	AIC ranking	Overall score
All land covers, Agriculture/park dominated region, raw data	0.971	495.3	Co, In, Ot, Sh	1	1	2
All land covers, All study area, density fraction	0.915	982.5	Co	2	4	6
Employment land covers, All study area, density fraction	0.878	720.2	Co, Ed	4	3	7
All land covers, Urban dominated region, raw data	0.900	1135.0	Co, HM	3	5	8
Urban land covers, All study area, raw data	0.876	1667.9	Co, HM, In, TD	6	2	8
All land covers, All study area, raw data	0.877	1672.7	Co, HM, TD	5	8	13
Urban land covers, All study area, density fraction	0.721	576.1	Co, In, Re	8	6	14
Employment land covers, All study areas, raw data	0.862	1669.9	Co, Ed, HM, In	7	7	14

Table 3: Results of Poisson regression models to determine relative employment densities of land use classes.

Model	AIC	Significant estimates	Ranking
All land covers, All study area, raw data	7650406	All	1
Urban land covers, All study area, raw data	8465964	All	2
Employment land covers, All study area, raw data	11411011	All	3

34

employment estimates than the OLS models. The best employment estimate is displayed on a map in Figure 5.

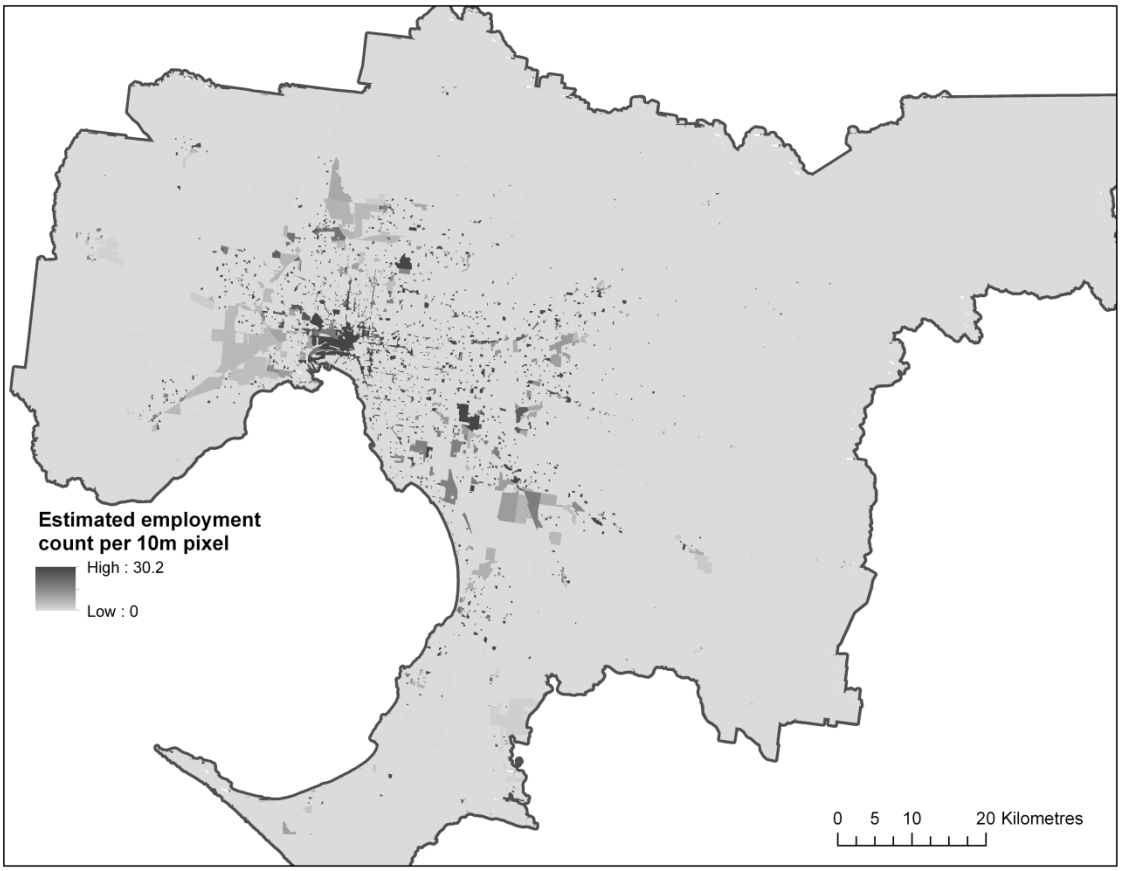


Figure 5: Best estimate of employment distribution, based on Poisson model with employment attributed to employment-related land uses.

Table 4: Comparison of error in the employment estimates

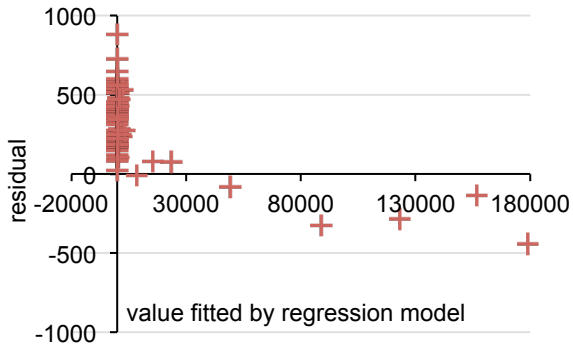
Model name	RMSE	Adjusted RMSE	Coefficient of Variation	Mean Error
Poisson, employment land uses, all study area, raw data	2060.6	1.55	0.90	-0.00009
Poisson, urban land uses, all study area, raw data	2164.6	2.04	0.94	-0.00002
Poisson, all land uses, all study area, raw data	2332.8	2.41	1.02	-0.012
OLS, all land uses, all study area, raw data	2489.0	1.93	1.08	-0.001
Binary estimate, employment land uses	2531.4	3.42	1.10	-0.000004
OLS, employment land uses, all study area, raw data	2625.9	3.68	1.14	-0.00003
OLS, all land uses, regions split, raw data	2633.7	3.41	1.15	58.1
OLS, urban land uses, all study area, raw data	2694.5	2.36	1.17	-0.00007
OLS, all land uses, all study area, density fraction	10193.5	8.04	4.44	-0.09
OLS, urban land uses, all study area, density fraction	11062.6	8.15	4.82	-0.0001
OLS employment land uses, all study area, density fraction	15596.7	10.09	6.79	0.0002

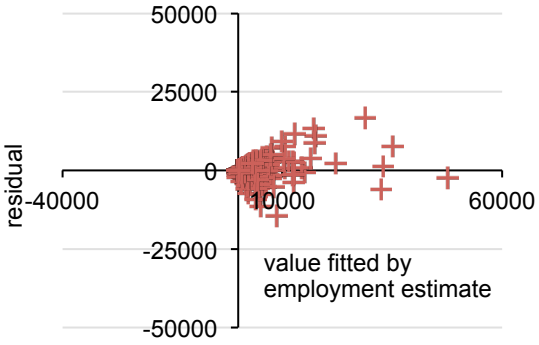
Figure 6 shows residual plots for each model tested, comparing the predicted values and residuals from the regression modelling stage (left) to the predicted values and residuals from the employment estimate stage (right). Note that that since the predictions at the modelling stage are based on source zones while the predictions at the employment estimate stage are based on validation zones, the number of predictions and their scale are different between each pair of plots. Examining these plots helps to understand the models and how the error (residual) in the estimates changes between the model stage and the employment estimate stage (where the estimated densities are rescaled to fit the know number of workers in the source zones).

For the Poisson employment model, the values fitted by the original model were biased, with many small predictions with positive residuals, but a number of much larger predictions with negative residuals balancing this out. The values fitted by the employment estimate, once re-scaling is taken into account, take the distribution more classically expected from such predictions, with residuals increasing away from zero in both a positive and negative direction and in a parabolic shape as the estimate value becomes larger. However the distribution shows that the estimate is somewhat biased, with the larger values tending to be over rather than under predicted. The other Poisson models follow a similar distribution at both the model fitting and employment estimate stage, with larger values tending to be overestimated. Additionally, these other Poisson-based estimates occasionally produce negative estimates. This may seem counter intuitive as the models are set to have no intercept and the estimates at the model stage are all positive. However, in fitting the model to the variables in the first stage, some negative coefficients are produced in many of the models, most commonly for the Agricultural land. In the model stage these negative coefficients are balanced by positive ones to produce overall positive prediction values, but when these coefficients are applied as global density estimates in the source zones where not all the different land covers are present, and are then re-scaled, some overall negative employment estimates result.

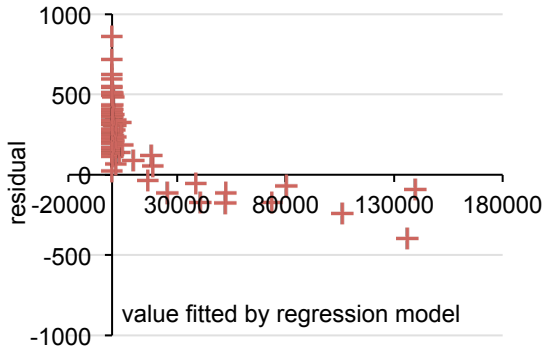
Figure 6: Residual plots comparing the predicted values and residuals from the regression modelling stage (left) to the predicted values and residuals from the employment estimate stage (right).

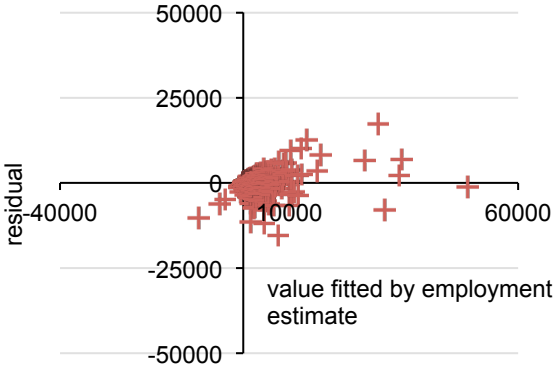
Poisson model with employment land uses



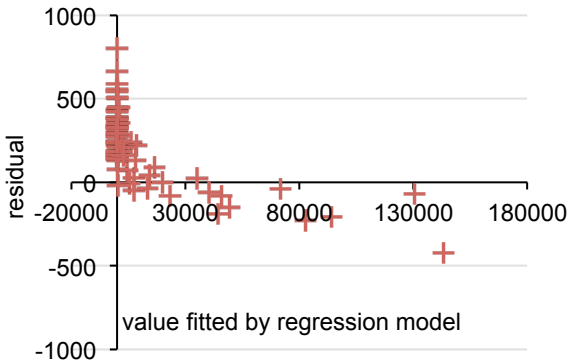


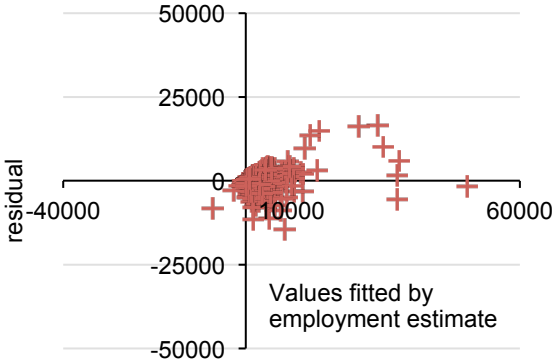
Poisson model with urban land uses



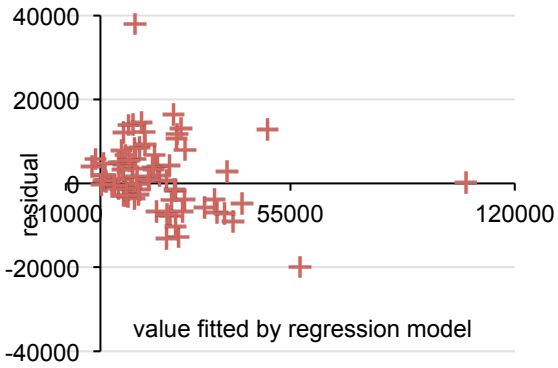


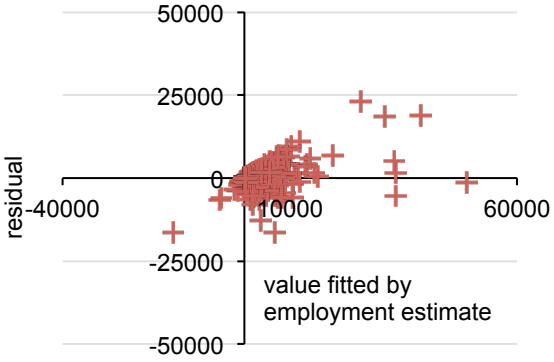
Poisson model with all land uses



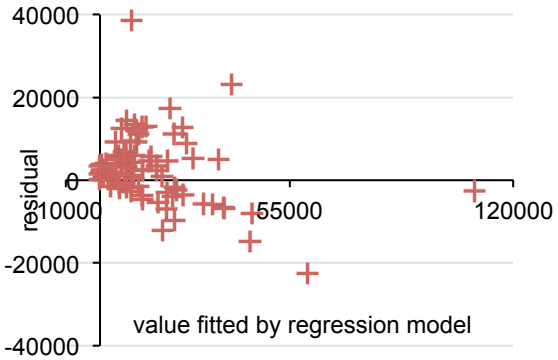


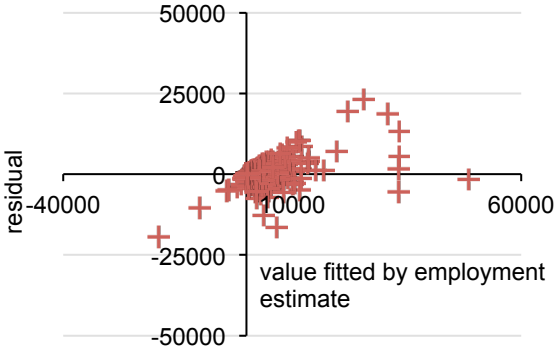
OLS model with all land uses



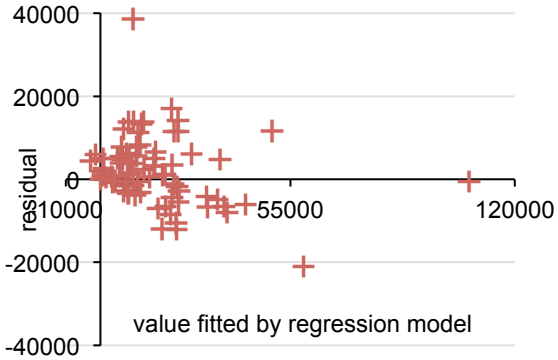


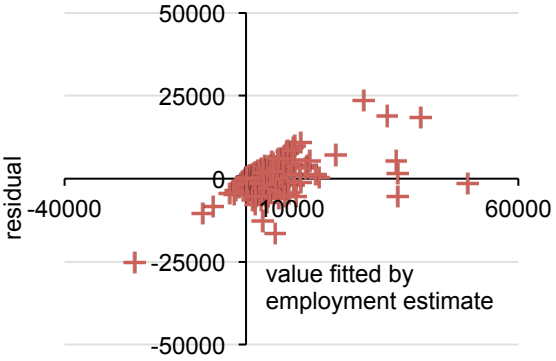
OLS model with employment land uses



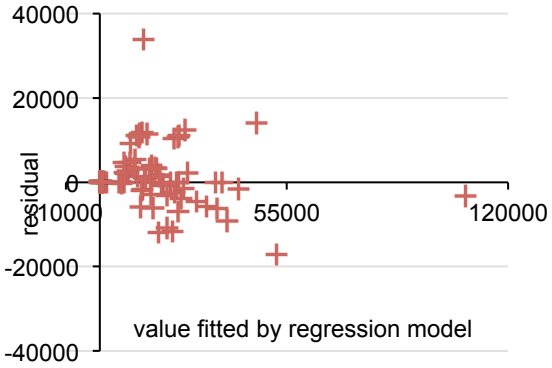


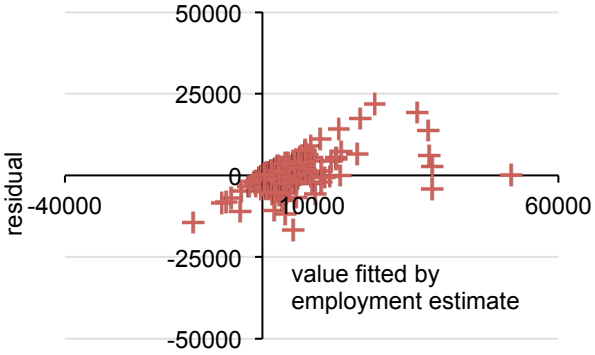
OLS model with urban land uses





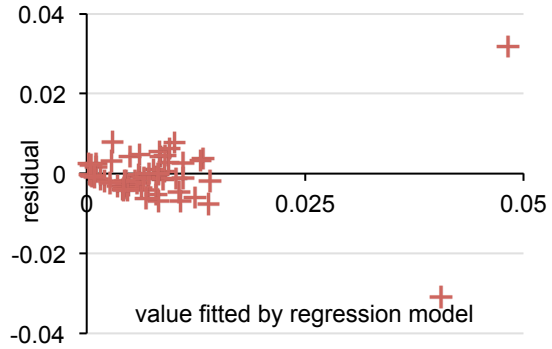
OLS model with all land uses, split by regions

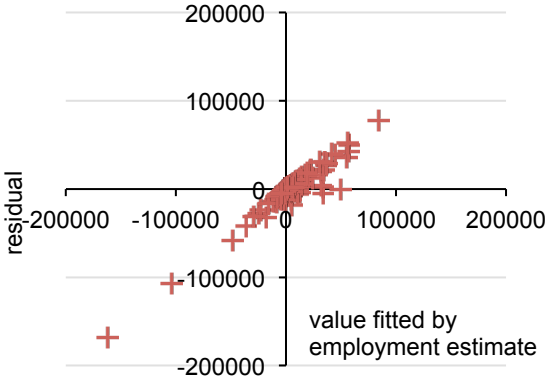




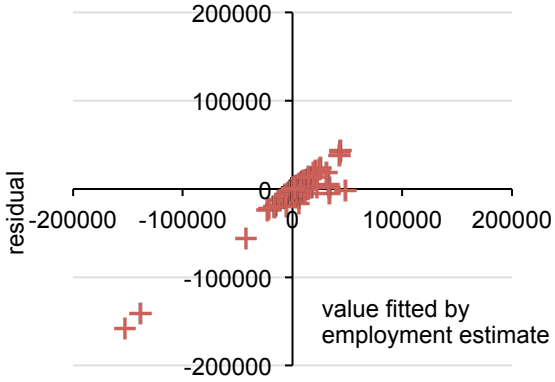
OLS model with all land uses and density

OLS model with urban land uses and density

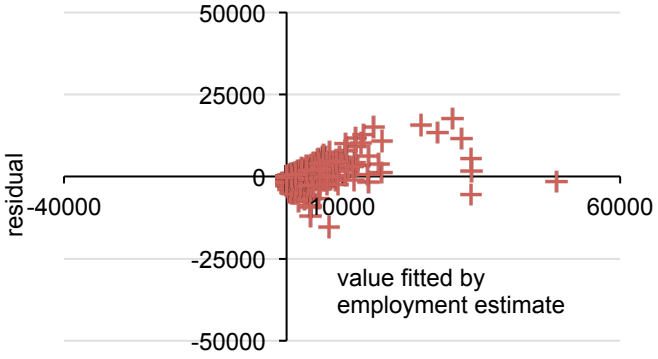




OLS model with employment land uses and density



Binary estimate with employment land uses



The negative estimates are, of course, all under predictions and show in the distribution as such.

The OLS models demonstrate a more or less expected distribution at the model fitting stage, though with some slightly negative estimates. The distributions appear slightly biased towards small overestimates, though not strongly so. At the employment estimate stage, however, this translates to some negative estimates and a tendency towards large overestimates rather than underestimates for larger predictions. In general, we can see that the residuals are somewhat larger than in the Poisson models.

The OLS models that used a density fraction rather than the raw data have uniformly negative prediction values for any underestimates, with all positive values being overestimated.

Overall, the plots appear to support the conclusions found when comparing the metrics in Table 4. Closer inspection of the geographic distribution of the residuals produced by the best employment estimate (Figure 7 and Figure 8) indicates that most estimates are within a count of ±2000 or ±20% of the known population of the Destination Zone. The relatively small number of large over or underestimates inflates the adjusted RMSE. The larger validation zones at the outer edge of the study area tend to have large overestimates, although there are some similarly large overestimates in the inner urban portion of the study area. Employment underestimates tend to be in the smaller zones and are distributed throughout the study area.

The observation that employment in Agricultural areas is generally overestimated in Poisson employment land uses estimate, raises the question of whether this situation is the same in the models where agriculture-and-parkland-dominated source zones were regressed separately from urban-dominated source zones. Indeed Figure 9 demonstrates that, in fact, under this regionally split model employment in Agricultural areas is *under*estimated. Figure 10 shows that these Agricultural areas have the greatest underestimates in terms of percentage error, whereas in the Poisson employment model these areas had the greatest overestimates in terms of percentage error. Evidently, comparing these models based on the measures described above does obscure some of the geographic variability associated with these different estimates.

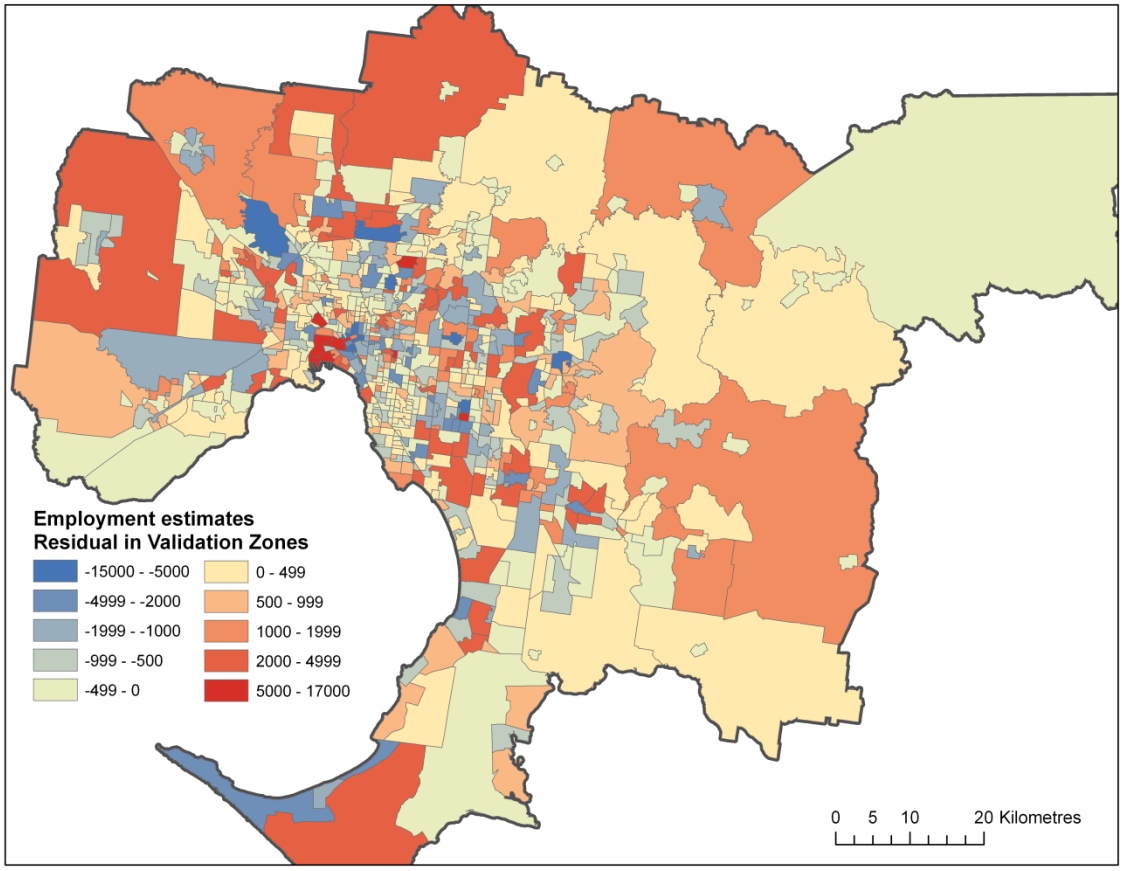


Figure 7: Residual count in validation zones from the employment estimate based on the Poisson employment land uses model. The values represent counts, i.e., workers. Positive values represent overestimates, negative values are underestimates.

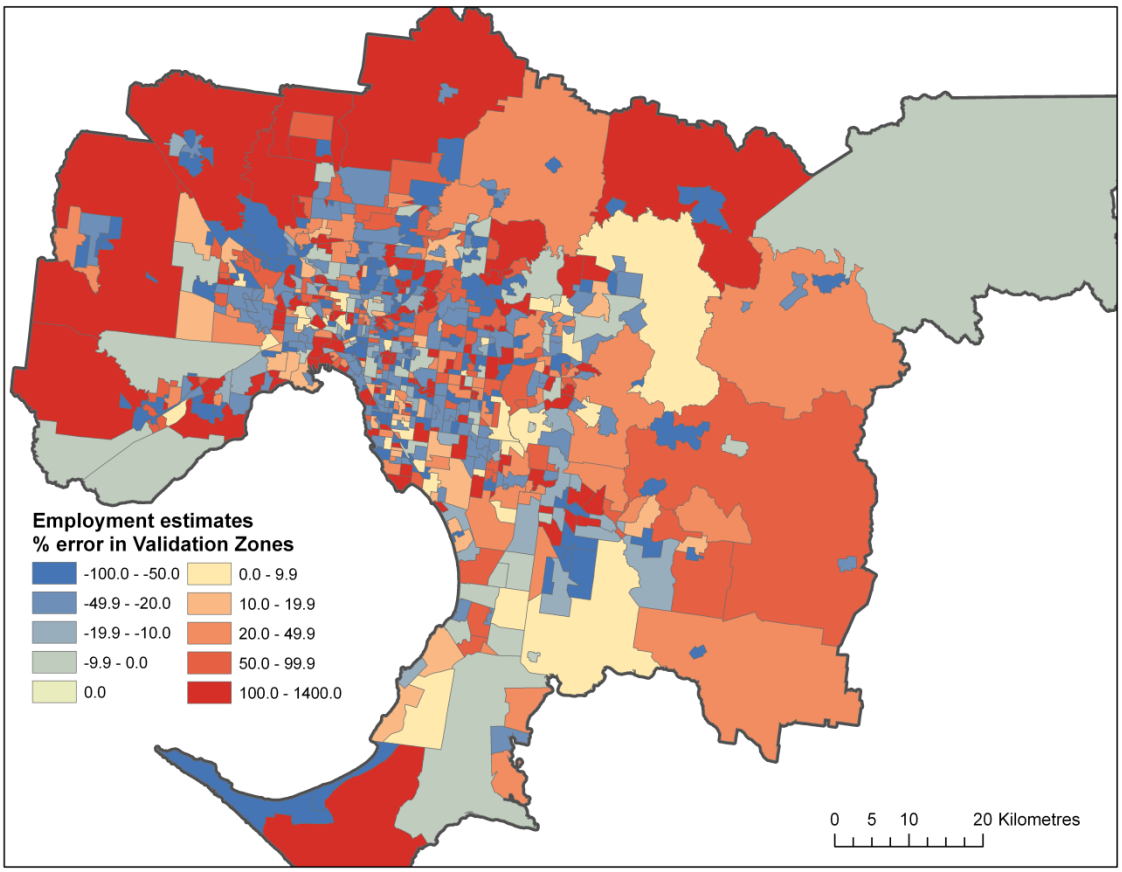


Figure 8: Percentage error in validation zones from the employment estimate based on the Poisson employment land uses model. The percentage error is calculated as the residual (difference between the estimated employment count and the known count) as a percentage of the known count. Positive values represent overestimates, negative estimates represent underestimates.

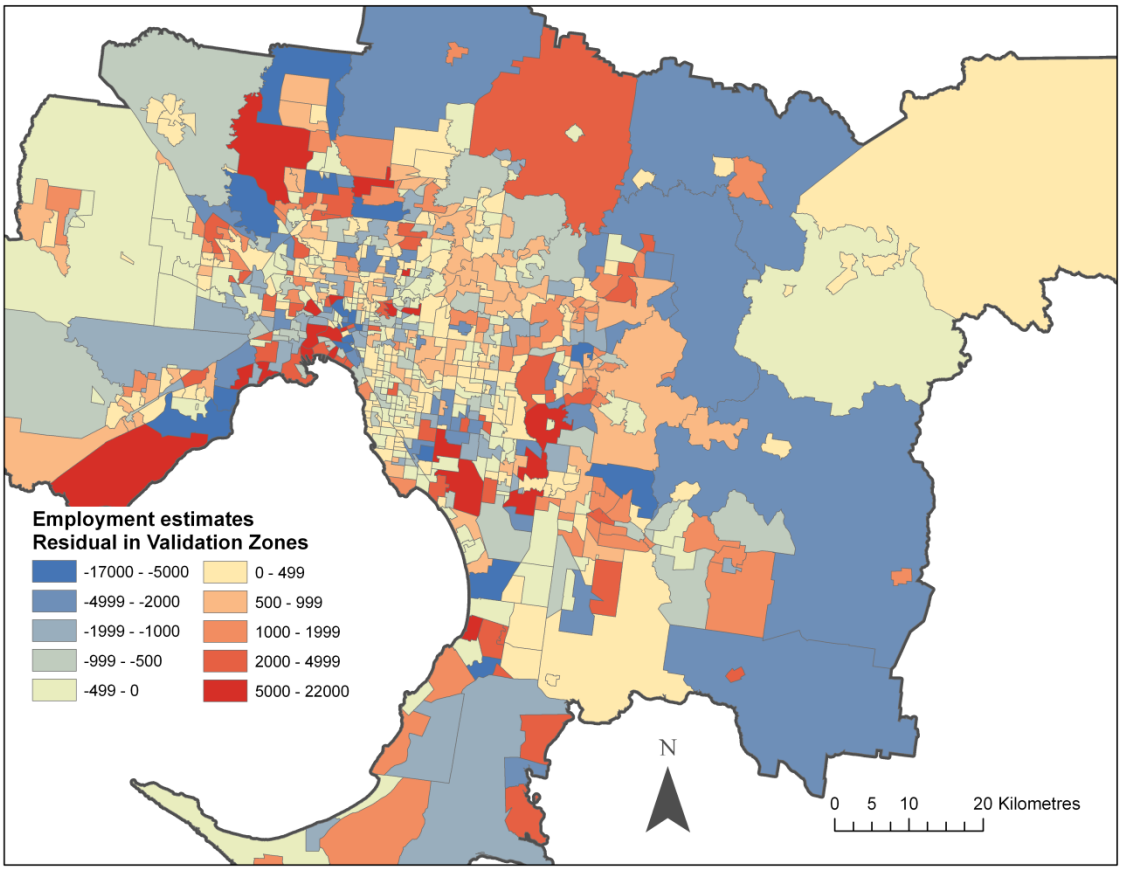


Figure 9: Residual count in validation zones from the employment estimate based on the OLS all land uses model that was split by region, i.e., agricultural-and-parkland dominated source zones were regressed separately to urban-dominated source zones. The values represent counts, i.e., workers. Positive values represent overestimates, negative values are underestimates.

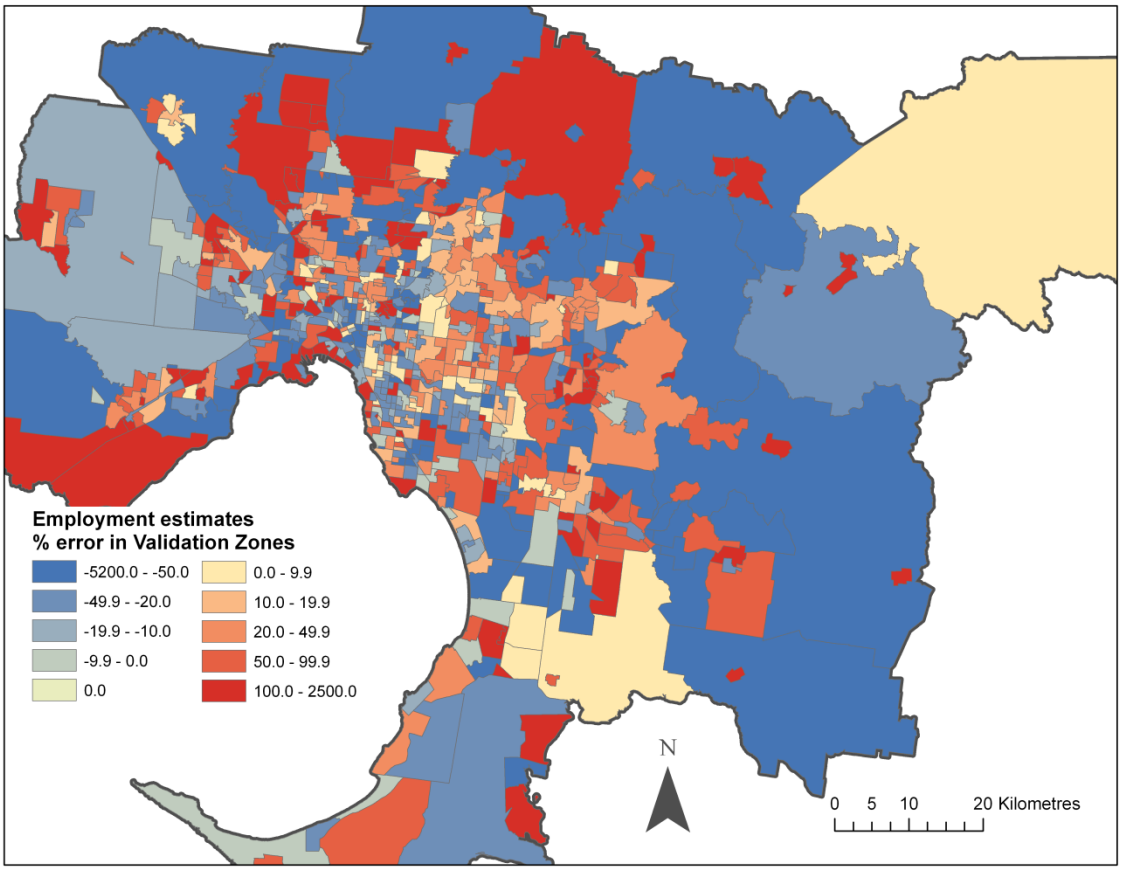


Figure 10: Percentage error in validation zones from the employment estimate based on the OLS all land uses model that was split by region, i.e, agricultural-and-parkland dominated source zones were regressed separately to urban-dominated source zones. The percentage error is calculated as the residual (difference between the estimated employment count and the known count) as a percentage of the known count. Positive values represent overestimates, negative estimates represent underestimates

3.3 Downscaling residential data

A summary of the accuracy of the working population estimates produced by the two different methods is shown in Table 5.

Table 5: Comparison of mean square error, etc. for the population estimates produced by different models

Model name	Root Mean Square Error	Adjusted Root Mean Square Error	Coefficient of Variation	Mean Error
Binary residential	1039.742	2.581	0.565	-29.779
Using total population density weight	262.637	0.355	0.143	-46.826

As can be seen from the results, the estimate of working residents is greatly improved by incorporating data about overall population density at the Mesh Block level, compared to a binary dasymetric estimate. It should be noted here that even the coarser binary estimate for residential working population scored better in terms of RMSE and Coefficient of Variation, than any of the regression-based estimates for employment. Figure 11 plots these residuals on a similar scale as the employment estimate residual plots in Figure 6. The smaller magnitude of the working population residuals is a reminder that employment distribution is a rather more difficult phenomenon to model, when compared with residential population, given the ancillary data available. The best estimate of the working population is displayed on a map in Figure 12. The residual map in Figure 13 shows that residuals were small in general compared to the employment estimate, and the largest residuals were not in the largest validation zones, indeed they tended to be in smaller zones. The map of percentage errors (Figure 14) follows this trend of larger errors occurring in relatively smaller zones. These smaller zones are likely associated with the high population counts, being inner urban areas where population is

most densely concentrated. Therefore it appears reasonable that the residuals and percent error are high in those zones.

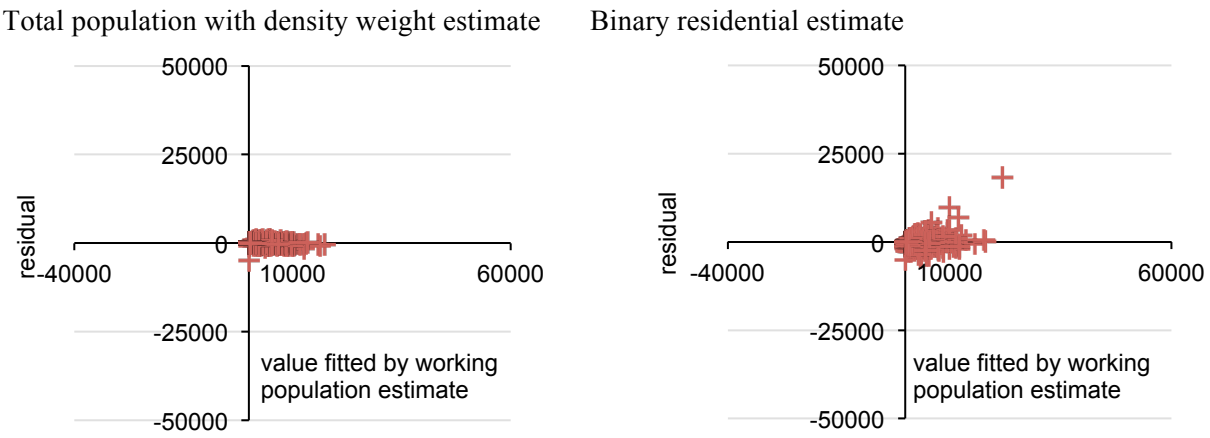


Figure 11: Residual plots comparing the predicted values and residuals from two working population estimates

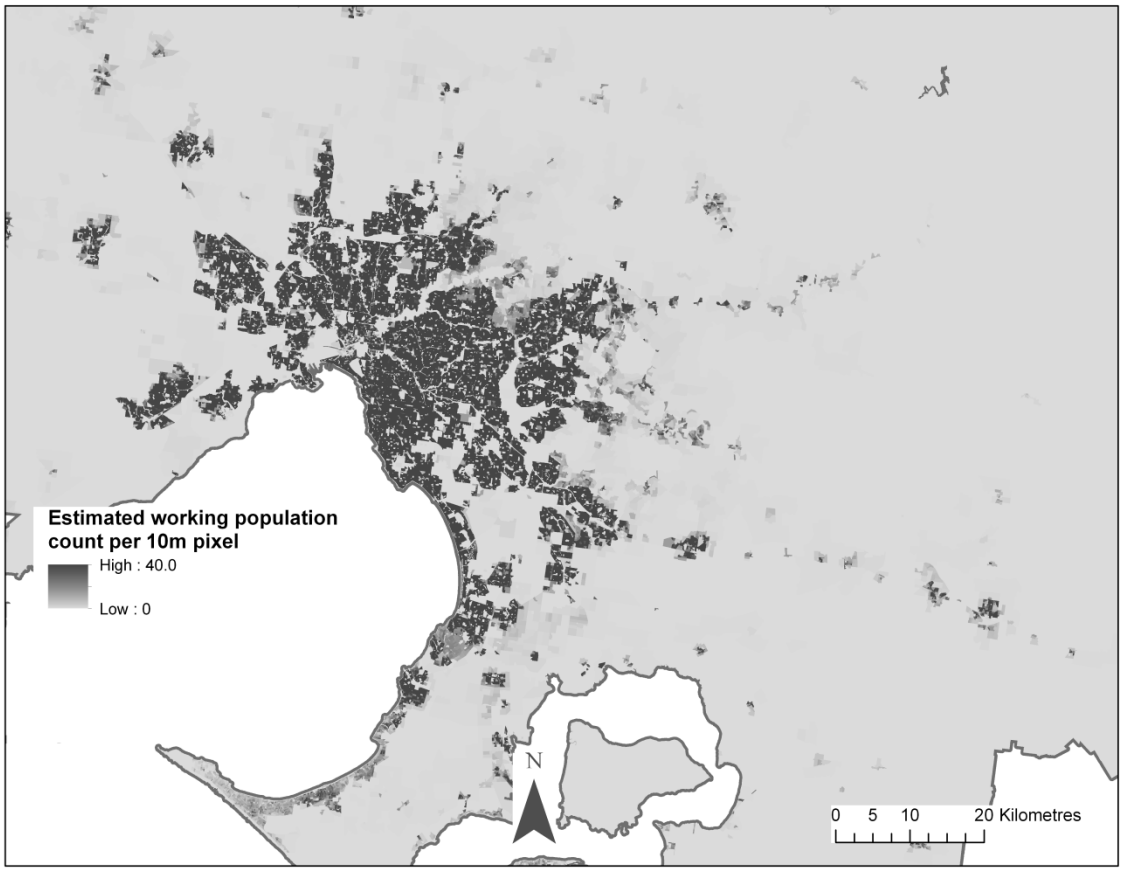


Figure 12: Best estimate of working population distribution, using the density-weighted distribution method.

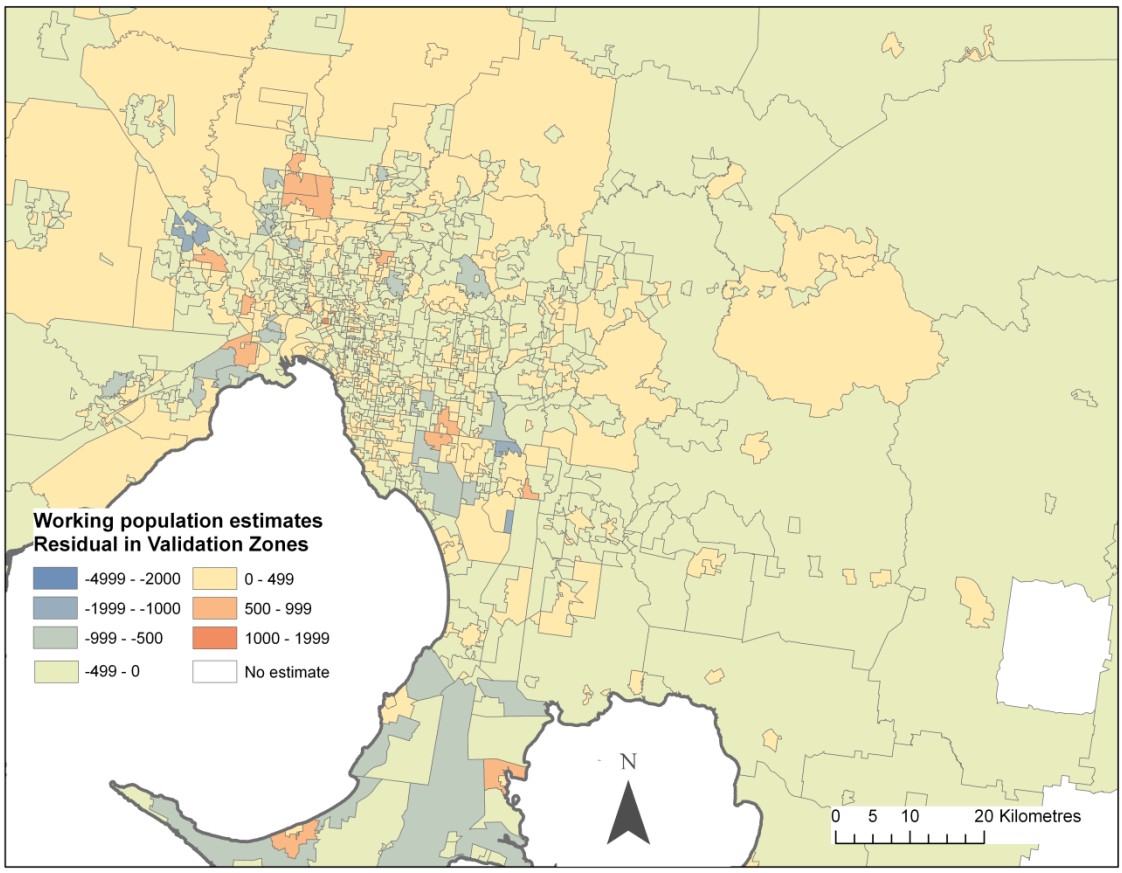


Figure 13: Residual count in validation zones from the working population estimate using the density-weighted distribution method. The values represent counts, i.e., workers. Positive values represent overestimates, negative values are underestimates.

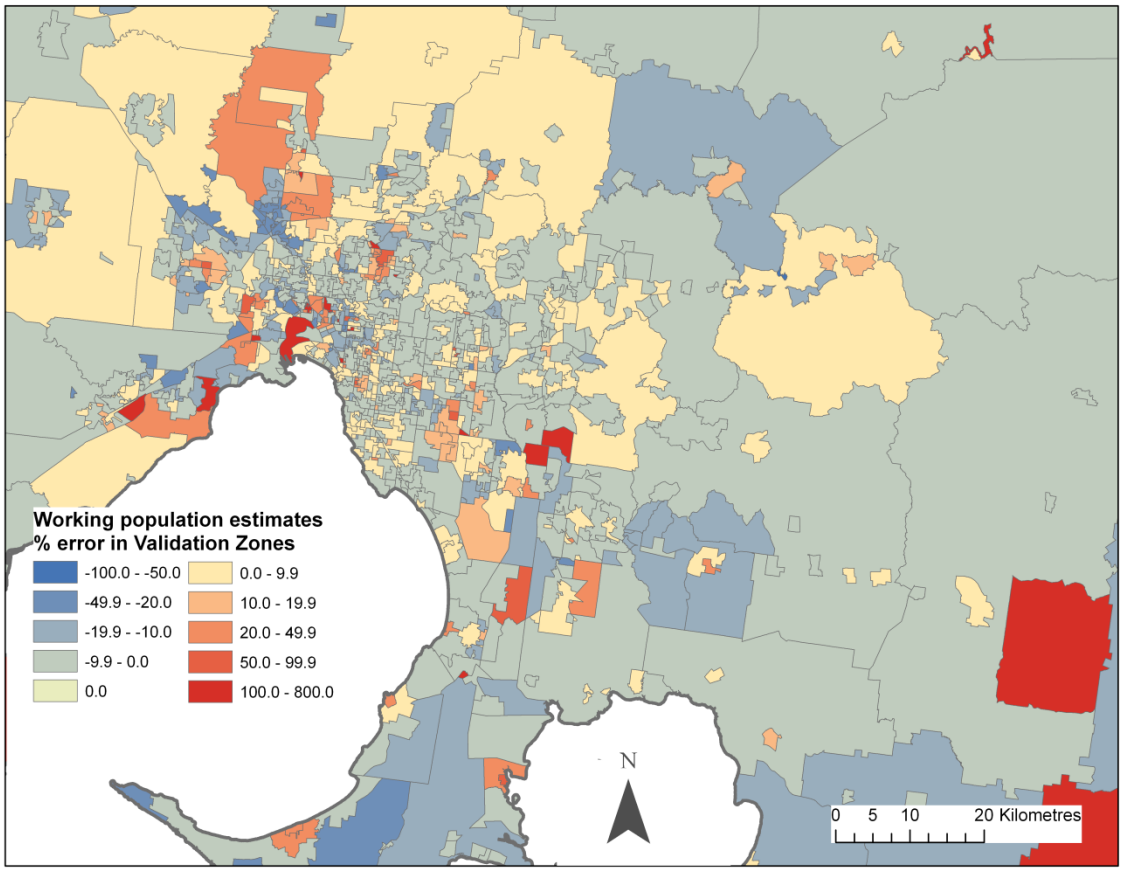


Figure 14: Percentage error in validation zones from the working population estimate using the density-weighted distribution method. The percentage error is calculated as the residual (difference between the estimated working population count and the known count) as a percentage of the known count. Positive values represent overestimates, negative estimates represent underestimates.

3.4 Employment Containment

The final employment raster dataset was converted to a polygon feature class of parcels with uniform employment density. This yielded a total of 8 187 parcels. Of these, 330 were within Destination Zones that had been estimated to within ±10% accuracy. The location of the 40 parcels randomly selected for employment containment analysis are displayed in Figure 15.

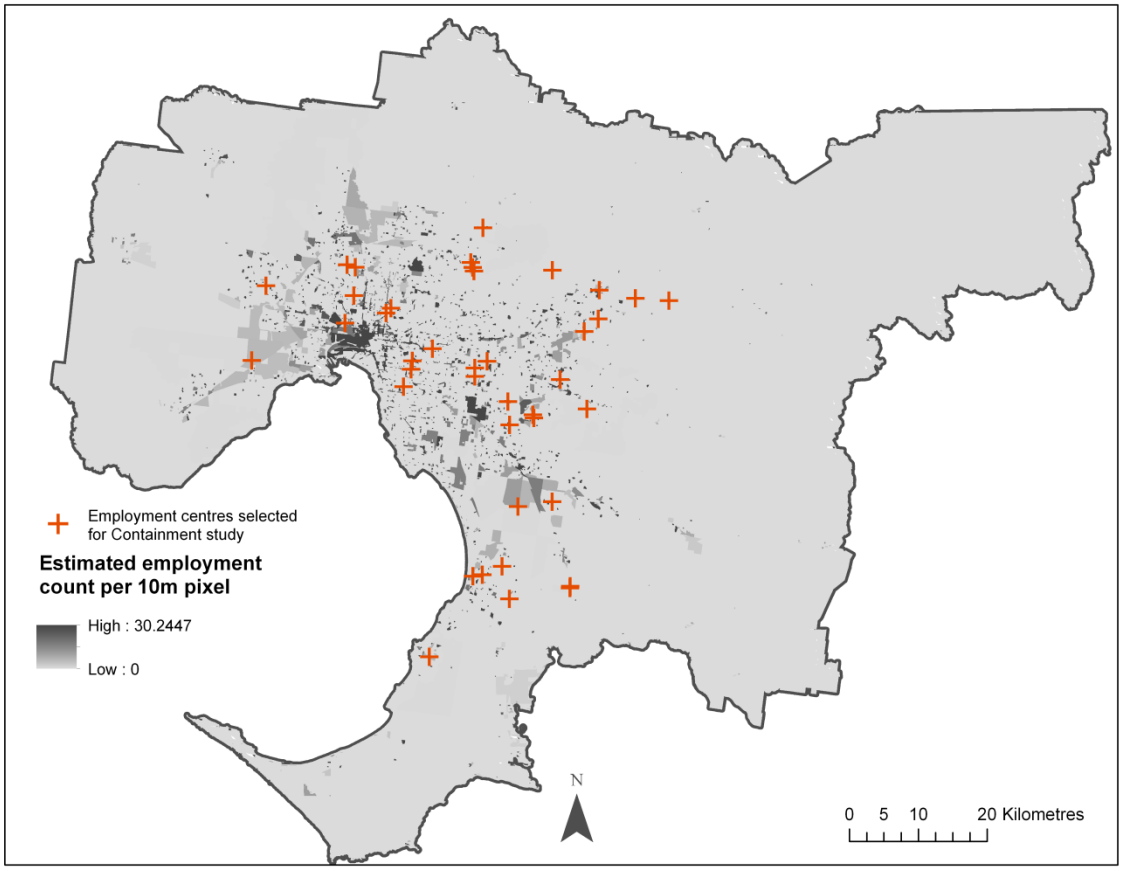


Figure 15: The 40 sites randomly selected for the employment containment study. They are overlaid on the best final employment estimate surface (based on the Poisson employment land uses model).

The full results of the containment analysis are shown in Table 6. The lowest containment rate was effectively zero in some agricultural parcels and one small section of an industrial estate where employment was estimated to be less than one person. The highest containment rate was 7.6% in a large industrial estate that had a large employment estimate and a relatively sparsely populated catchment. In general, the employment containment estimates are small, less than 1%, as could be expected for small employment centres that have a small number of people working there, compared to the working population of the catchment as a whole. The estimate of the percentage of workers in the employment centre that live in the catchment is lowest again in the industrial estate where employment was estimated to be less than one person, with only 0.1% local employment; it was highest at a car salesroom in Frankston (outer south-eastern suburbs) where local employment was estimated to be 84.3%. The average percentage of workers from the catchment was 20.5%.

Figure 16 shows the relationship between the estimated number of workers in the employment centre and 1) employment containment in the centre, and 2) number of employees in the centre from the catchment (rather than the percentage of employees from the catchment that was discussed above). In general, the results suggest that the rate of containment increases as the number of people employed in the centre increases. This is to be expected, both intuitively and because the estimate incorporates information about the estimated employment as a fraction of the total employment in the source zone (SLA). However, there is still not a perfect linear relationship between the estimated employment and the estimated containment, and this captures the information about the origins that workers travelled from to work in the SLA (and, by extension, the employment centre). With regard to the number of workers (rather than the percentage) that came from the catchment, the scatterplot shows that this also rises with employment,

with a strong relationship that is nonetheless not perfect for having taken account of the origin that workers travelled from.

Table 6: Summary of employment and employment containment in selected employment centres. EmEc= Employment estimate for the centre, WP= Working population estimate for the employment centre catchment, WpEc= the estimated number of workers in the centre originating from the catchment, %Co= Estimated percentage of employment containment for the centre, %LEm= Estimated percentage of people working in the centre that also live in the catchment.

ID	Type	Description	EmEc	Wp	WpEc	%Co	%LEm
1	Commercial	Diamond Creek Shopping centre, between Main Hurstbridge Rd and the Hurstbridge railway line, Diamond Creek	249	12024	86	0.71	34.4
2	Industrial	Warehousing area or business park, Bridge St Eltham.	221	23292	83	0.36	37.6
3	Education	Pascoe Vale Girls College, near Boundary Rd and Cumberland Rds, Pascoe Vale	705	25529	31	0.12	4.4
4	Education	Pascoe Vale North Primary School, Derby St, Pascoe Vale	237	32221	64	0.2	27
5	Education	Eltham High School, Wither Way, Eltham	646	22416	248	1.11	38.4
6	Education	Unidentified school, Wonga Park	145	3926	14	0.36	9.8
7	Commercial	Small shopping strip, Main Rd, Eltham	245	21420	81	0.38	33
8	Commercial	Shopping centre (part), East Esplanade and Main Rd, Keilor	180	10240	11	0.1	5.9
9	Commercial	Shopping strip, Main St, Lilydale.	683	16342	167	1.02	24.4
10	Commercial	Commercial area at the corner of Albion St and Melville Rd, Brunswick West	138	69333	44	0.06	31.8
11	Education	Wandin North Primary School, Wandin	190	1955	4	0.22	2.3
12	Education	Unidentified school, School Rd, Seville	84	2350	10	0.4	11.3
13	Education	Westgarth Primary School (part), Northcote	170	32448	37	0.11	21.9
14	Commercial	Commercial area, corner Cummingham St and High Sts, Northcote.	59	37543	14	0.04	24.6
15	Commercial	Shopping strip, Macaulay Rd, Kensington	60	57528	6	0.01	9.7

55

ID	Type	Description	EmEc	Wp	WpEc	%Co	%LEm
16	Agricultural	Agricultural land, corner of Swansea Rds and Cambridge Rd, Lilydale	86	14316	14	0.1	16.7
17	Education	Gladeville Primay School, Gladesville Drive, Kilsyth	157	14641	21	0.15	13.6
18	Education	Camberwell High School, near Riversdale Road, Camberwell	993	47139	314	0.67	31.6
19	Commercial	Shopping strip, Milton Parade, Toorak	41	40325	6	0.02	15.3
20	Education	Forest Hill College, corner of Hawthorn Rd and Mahoneys Rd, Burwood East	2174	43354	416	0.96	19.1
21	Education	Burwood East Special Development School, Mudgee St, Burwood East	218	44408	35	0.08	15.9
22	Education	Unidentified School, Eva St, Malvern	206	33579	31	0.09	15
23	Education	Mount Waverley North Primary School, Josephine Ave, Mount Waverley	389	44405	66	0.15	17
24	Industrial	Business park near the corner of Dorset Rd and Burwood Hwy, Ferntree Gully	100	30655	28	0.09	27.6
25	Commercial	Commercial shopping strip on Hawthorn Rd, near the corner of Glen Eira Rd.	91	31972	26	0.08	28
26	Industrial	Industrial area near the corner of Fitzgerald Rd and Doherty's Rd, Laverton North	15503	2689	204	7.6	1.3
27	Education	Caulfield Grammar School, Jells Park Primary School, near Jells Rd, Waverley East	2544	21203	494	2.33	19.4
28	Education	Upwey South Primary School, Morris Rd, Upwey	162	6915	31	0.45	19.1
29	Commercial	Rowville Shopping Centre, Stud Rd, Rowville	466	25343	97	0.38	20.8
30	Education	Rowville Secondary College, Turramurra Drive, Rowville	1003	24703	198	0.8	19.8
31	Education	Unidentified school, Gladeswood Drive, Mulgrave	476	30859	94	0.31	19.8
32	Commercial	Hampton Park Shopping Square, Corner Hallam Rd and Pound Rd, Hampton Park	234	20649	35	0.17	15.1
33	Industrial	Small part of a large industrial estate, corner Frankston Dandenong Rd and Bangholme Rd, Dandenong South	<1	288	0	0	0.1
34	Education	Flinders Christian Community College, Ballarto Rd, Carrum Downs	1092	15839	146	0.92	13.3
35	Education	Monterey Secondary College, Silvertop St, Frankston	715	20623	154	0.75	21.6
36	Commercial	Car salesroom, Wells Rd, Frankston	83	19130	70	0.36	84.3

56

ID	Type	Description	EmEc	Wp	WpEc	%Co	%LEm
37	Agricultural	Agricultural land, corner Craig Rd and South Gippsland Highway, Junction Village	<1	2812	0	0	11
38	Agricultural	Agricultural land near the corner of Craigs Rd and Browns Rd, Junction Village	<1	2381	0	0	10.6
39	Education	Elizabeth Murdoch College, Frankston- Warrandyte Rd, Frankston	715	12584	173	1.38	24.3
40	Industrial	Part of business park near Mornington-Tyabb Rd and Nepean Hwy, Mornington Peninsula	651	10327	171	1.65	26.2

57

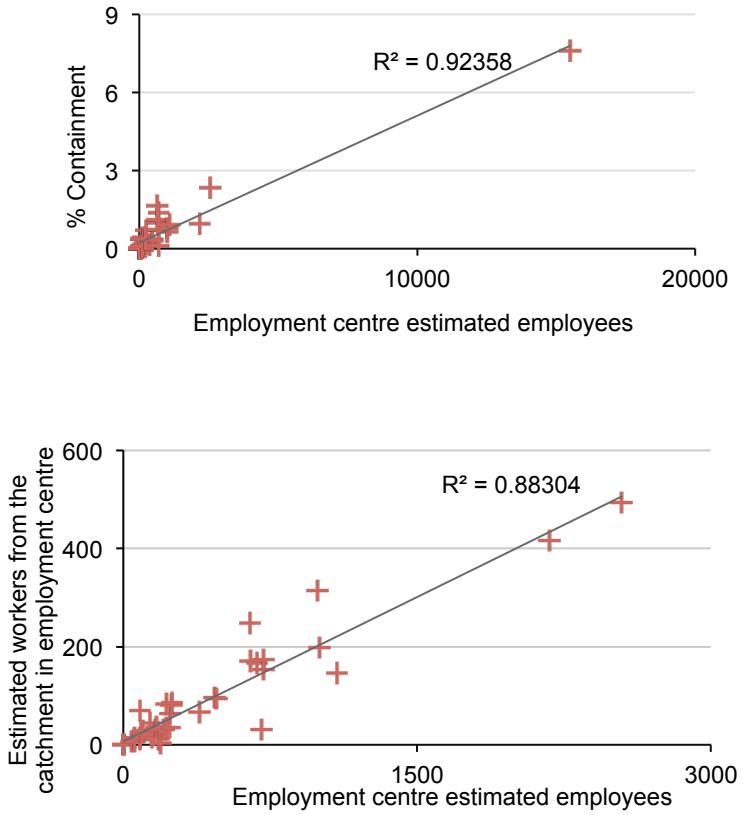


Figure 16: Relationship between the estimated number of employees in the selected employment centres, and 1) the estimated percentage of employment containment for the centre, and 2) The percentage of workers in the employment centre that live in the catchment. Note that for the second plot one outlying value with very high employment has been eliminated to display the overall trend

The limited analysis of this sample data suggests the possibilities for containment analysis that could be performed using this containment calculation method. However, I refrained from performing detailed analysis on this data here, partly for reasons of scope but also because the data shows some flaws that indicate that the employment estimate figures, and therefore the containment estimates, cannot be relied on. Although the employment centres were selected from the validation zones where the employment

estimate was found to be most accurate (at the validation zone scale), closer inspection of the estimated employment in the selected employment centres suggests that the estimates are not reliable at a sub-validation zone scale. In particular, the estimates for Education land uses (mostly primary or secondary schools) appear to be inflated. Comparing the estimates for the Educational employment centres to the known number of jobs in the Education and Training sector for the relevant validation zones (ABS, 2006) suggests that the estimated employment is at least twice that observed by the ABS data, and in some cases much greater. Given that, the other land classes are evidently under-estimated. It is also interesting that 50% of the selected employment centres have Education land use, when Education land uses represent around 26% of the employment-related parcels in the original Mesh Block data set. A possible explanation for this is that the Agricultural and Industrial land use parcels tended to be located in the larger source zones, and this is where percentage error tended to be higher, whereas Educational land uses are quite well distributed throughout the entire study area. Therefore the Educational land uses are overrepresented in the most accurately predicted validation zones.

4. DISCUSSION

In general, the Poisson models produced more accurate employment estimates than the other OLS-based models. At the estimate stage, size of these residuals was generally smaller and these models produced no or fewer negative count estimates. Of the Poisson models, the one that attributed employment counts to employment land use classes produced the best estimates. This may be because with employment counts being attributed to a smaller number of land classes and distributed over a smaller number of parcels overall, there is less opportunity for employment counts to be erroneously distributed.

While successive refinements to the employment surface model improved the overall prediction, the final 'best' model still produced RMSE, Adjusted RMSE, Coefficients of Variation and Mean Error that users of the resulting data might consider to be unacceptably high. Reaggregating the data from the best estimate of employment distribution showed that the downscaling incorrectly distributed the employment data to a large degree. The source of this error was hinted at when performing the employment containment analysis on selected employment centres. Comparison of some of the estimates in Education employment centres to known counts of Education and Training industry workers in the validation zones suggests that Education employment centre estimates were routinely over-estimated, and that therefore the employment density attributed to this land class was too high.

The large overestimates in the larger validation zones on the edge of the study area, where Agriculture is concentrated, suggests that in the best fitting model Agricultural land was attributed with too high an employment density in many cases. It should be noted that this is in contrast to the estimate where the urban-dominated and agriculture-and-parkland dominated regions were regressed separately. In that case, employment counts in Agricultural land tended to be under-estimated. This highlights the

sensitivity of the downscaling process to changes in the model and the data used, but also gives encouragement that with further refinements a useful downscaled employment estimate could be developed.

The working population surface was found to be more accurate, and information about overall population density in the ancillary data zones greatly improved the estimates. Even so, the working population estimate produced has room for improvement.

The finding that Education-related employment centres appear to have highly overestimated employment counts highlights a more general limitation for using these downscaled estimates. While the formulation of this study relies on detecting the residuals produced when we consider the estimates at validation zone scale, the misattribution of employment or working population to different locations and land uses *within* the validation zone is not detected. Using the derived data from the downscaled/small areas in isolation therefore has both known and unknown error associated with it.

There are certainly achievable options that could be explored to improve the employment surface model. In the current study, the total number of all workers in the source zone was downscaled via linear regression, using information about the amount of land in different use classes. Additional data is available from the ABS that provides the number of people employed in different industry sectors and occupations at both the source zone and validation zone level. Where a given industry of employment can be sensibly associated with a particular land use class (for example, education and training workers can be associated with Education land use, manufacturing and warehousing workers can be associated with Industrial land use), this additional information could be used to control how much employment is attributed to some land use classes. Another point for improvement is the negative density coefficients that many of the models produced for some land use classes. Future refinements to the model could constrain the

model to produce positive coefficients, for example using the non-negative least squares function in R (e.g. Mullen & van Stokkum, 2010)

The employment containment method developed in this project is a positive step towards the development of new technique for comparing employment containment across different areas and employment centres within Melbourne. By producing uniform catchment areas at which to assess employment containment, the method does control for spatial factors that traditional employment containment methods cannot. Thus the finding that some employment centres appear to contribute more to employment containment and local employment, than others. Of course, the method would be greatly improved if the input datasets (the employment and working population estimates) were able to be made more accurate. The employment containment estimate downscales flows from an origin-destination matrix, areally weighting the flows based on the employment and working population estimates- the proportion of the total workers in the source zone (SLA) that the employment centre is estimated to represent, and the proportion of workers living in the Origin Zones that are included in the employment centre's catchment. However, aside from this areal weighting, flows between different sub-areas of each origin and destination are assumed to be the same. An improved employment containment estimate could take account of the finding that people working in higher skilled occupations are more likely to travel long distances to work, than those in lower skilled occupations (Bill et al., 2007), and attribute flows based on the likely propensity of workers to travel longer distances to a given employment centre. Conceptualising this heterogeneous propensity to travel by using a Gravity model (such as in Trendle & Siu, 2005) may be possible.

When the employment centres for the employment containment analysis were selected, the land use classification or the type of activity at the centre was not controlled for, so it was not possible to look at the way that employment containment and the percentage of workers from the catchment varied in different types of employment

centres. However this could be an interesting piece of analysis for any further work, particularly once an improved estimate could be implemented.

While the caveats on the results of this method are many, the method can be viewed as being useful in producing a more refined picture of how employment is distributed throughout Melbourne, incorporating urban land use and therefore a picture of urban form. There are options available for improving the estimate of employment distribution, in particular. However, given the demonstrated difficulty in accurately modelling employment distribution at a very small scale, alternative formulations of this problem could be investigated, particularly those within the field of geostatistics. The regression-based interpolation method could be replaced with Poisson area-to-point or area-to-area kriging, which has been used widely in health statistics to map rates of disease incidence (and more recently, crime rates) over aggregated areas in a way that eliminates the visual bias associated with chloropleth maps (Goovaerts, 2008; Goovaerts 2006; Kerry, Goovaerts, Haining *et al*., 2010). This effectively produces a downscaled estimate of rates and numerically quantifies the uncertainty associated with the estimated spatial distribution of the variable in question. This process is used to identify likely clusters of the phenomena in space, so could be applied to identifying concentrations of employment or employment containment. Indeed, analogous approaches using area-to-point factorial kriging and pycnophylactic smoothing were recently undertaken by Nagle (2010) to identify employment clusters and by Kobayashi et al. (2011) to map daytime populations in a city, respectively. These studies produced distribution maps at a much coarser scale than that attempted in this study, which aimed to associate employment counts with individual blocks of a given land type, rather than producing a smoothed surface based on count rates across the whole study area. Such a geostatistical approach may have its merits for the current research topic: while removing the downscaling component of the process would change the spatial scale of the containment analysis, a more coarsely (but more accurately) described employment and population surface could

allow generic assessments of employment in a fixed area that is not constrained by administrative boundaries, but not necessarily around an identified ‘employment centre’.

5. CONCLUSION

This research began with the supposition that assessments of employment containment could be assisted by interpolation or downscaling methods in order to overcome the constraints of area-aggregated data issued at the level of administrative or statistical boundaries. I pursued an implementation of dasymetric downscaling methods that could describe employment and working population distributions at a fine scale, in order to provide a method to assess employment containment over uniform catchment zones in Melbourne. Supported by ABS Mesh Block land use classification data at a fine scale, a number of regression model variations as well as a binary dasymetric estimate were trialled to produce employment density estimates for different land use classes in the study area. The model that produced the smallest error was a Poisson model that distributed employment to employment-related land use classes. Unfortunately the error produced by this model was still high. Two approaches were compared when producing the working population estimate: a binary dasymetric method and a dasymetric method weighted by total population density data. The population-density weighted estimate was the more accurate of these two, and overall produced low error.

The employment and working population estimates were combined with areally-weighted commuting flows taken from an origin-destination matrix to calculate employment containment estimates in a small sample of employment centres. The method was found to be potentially useful; inspecting the results of this employment containment calculation highlighted flaws in the current estimates that should be addressed before the measures can be used to further analyse employment containment in Melbourne. Improvements to this method would support urban strategic and transport planning analyses at a metropolitan-wide scale.

5.1 Further work

Further work on this topic should focus on the following:

- Refining and improving the employment distribution estimate by incorporating additional ancillary information, such as the number of people in the source zones employed in different industries.
- Constraining the regression models so that they only produce positive density coefficients, for example using the non-negative least squares function in R (e.g. Mullen & van Stokkum, 2010)
- Incorporating some additional information into the containment calculation, taking account of the propensity of different people to travel less or more for work based on their occupation, industry of employment or other socio-economic characteristics.
- Investigating the effectiveness of a broader-scale estimate of employment and working population distribution using various geostatistical techniques such as area-to-point Poisson kriging (Goovaerts, 2006; Kerry *et al.*, 2010) area-to-area Poisson kriging (Goovaerts, 2008) area-to-point factorial kriging (Nagle, 2010) or pycnophylactic smoothing (Kobayashi *et al.*, 2011), for contributing to more accurate employment containment estimates.

6. REFERENCES

Australian Bureau of Statistics. Census of population and housing (2006) Customised Place of Work by Industry of Employment. Customised data cube for Victorian Department of Planning and Community Development, 2009.

Bill, A., Mitchell, B., & Watts, M. (2007). The occupational dimensions of local labour markets in Australian cities. *Proceedings of the 2007 State of Australian Cities Conference, Adelaide,* 172-187.

Boussauw, K., Derudder, B., & Witlox, F. (2011). Measuring spatial separation processes through the minimum commute: the case of Flanders. *European Journal of Transport and Infrastructure Research*, *11*(1), 42-60.

Boussauw, K., Neutens, T., & Witlox, F. (2010). Relationship between spatial proximity and travel-to-work distance: The effect of the compact city. *Regional Studies*, (907465083), 1-20.

Bracken, I., & Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A*, *21*, 537–543.

Bureau of Infrastructure Transport and Regional Economics. (2011). *Population growth, jobs growth and commuting flows in Melbourne*. *Transport* (p. 404). Canberra.

Burke, M., Li, T., & Dodson, J. (2010). The transport impacts of employment decentralisation in Brisbane. *Cities*, (October), 1-15.

Cromley, R., Hanink, D. M., & Bentley, G.C. (2011). A quantile regression approach for areal interpolation. *Annals of the Association of American Geographers, 102*(In press).

Debenham, J., Stillwell, J., & Clarke, G. (2003). The estimation of self-containment and catchment size indicators for use in small area classification. *International Journal of Population Geography*, *9*(3), 253-271.

Dempster, A. P., Laird, N. M., & and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39*, 1-38.

Deng, C., Wu, C., & Wang, L. (2010). Improving the housing-unit method for small-area population estimation using remote-sensing and GIS information. *International Journal of Remote Sensing*, *31*(21), 5673-5688.

Eicher, C. L., & Brewer, C. A. (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, *28*, 125-38.

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics*, *25*(3), 76-80.

Fisher, P. F., & Langford, M. (1995). Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation. *Environment and Planning A*, *27*(2), 211 – 224.

Flood, M., & Barbarto, C. (2005). Off to Work: Commuting in Australia. Canberra: The Australia Institute, Discussion Paper No. 77.

Flowerdew, R., & Green, M. (1989). Statistical methods for inference between incompatible zonal systems. *Transactions of the Institute of British Geographers* (pp. 239-247).

Flowerdew, R., & Green, M. (1992). Developments in areal interpolation methods and GIS. *The Annals of Regional Science*, *26*, 67-78.

Gallego, F. J. (2010). A population density grid of the European Union. *Population and Environment*, *31*(6), 460-473.

Goovaerts, P. (2006). Geostatistical analysis of disease data: accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. *International Journal of Health Geographics, 5*(52).

Goovaerts, P. (2008). Geostatistical analysis of health data: State of the art and perspectives. *geoENV VI - Geostatistics for Environmental Applications: Proceedings of the Sixth European Conference on Geostatistics for Environmental Applications*, 3-22. Netherlands: Springer Science+Business Media.

Gotway, C. a, & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, *97*(458), 632-648.

Gregory, I. (2002). The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, *26*(4), 293-314.

Harold, B. (2011). Sampler Toolbox for ArcGIS 10.0. ESRI.

Harvey, J. T. (2002). Estimating census district populations from satellite imagery: Some approaches and limitations. *International Journal of Remote Sensing*, *23*(10), 2071-2095.

Horner, M. W., & Murray, A. T. (2002). Excess commuting and the Modifiable Areal Unit Problem. *Urban Studies*, *39*(1), 131-139.

Jang, W., & Yao, X. (2011). Interpolating spatial interaction data. *Transactions in GIS*, *15*(4), 541-555.

Johnson, K. (2010). The geography of Melbourne's knowledge economy. *Proceedings of the Melbourne 2010 Knowledge Cities Summit* (pp. 1-14).

Kaiser, C., & Kanevski, M. (2010). Population distribution modelling for calibration of multi-agent traffic simulation. *13th AGILE International Conference on Geographic Information Science* (pp. 1-10). Guimarães.

Kerry, R., Goovaerts, P., Haining, R.P., & Ceccato, V. (2010). Applying geostatistical analysis to crime data: Car-related thefts in the Baltic States. *Geographic Analysis 42*, 53-77.

Kim, H., & Yao, X. (2010). Pycnophylactic interpolation revisited: integration with the dasymetric-mapping method. *International Journal of Remote Sensing*, *31*(21), 5657-5671.

Kobayashi, T., Medina, R. M., & Cova, T. J. (2011). Visualizing diurnal population change in urban areas for emergency management. *The Professional geographer : the journal of the Association of American Geographers*, *63*(1), 113-30.

Kyriakidis, Phaedon C. (2004). A geostatistical framework for area-to-point spatial interpolation. *Geographical Analysis*, *36*(3), 259-289.

Langford, M. (2006). Obtaining population estimates in non-census reporting zones: An evaluation of the 3-class dasymetric method. *Computers, Environment and Urban Systems*, *30*(2), 161-180.

LeSage, J. P., & Fischer, M. M. (2010). Spatial econometric methods for modelling origin-destination flows. In M. M. Fischer & A. Getis (Eds.), *Handbook of Applied Spatial Analysis* (p. 24). Heidelberg, Germany: Springer.

Li, T., & Corcoran, J. (2010). *Testing dasymetric techniques to spatially disaggregate regional population forecasts for South East Queensland*. *Social Research* (p. 36). Brisbane.

Li, T., Corcoran, J., & Burke, M. (2010). Investigating the changes in journey to work patterns for South East Queensland – a GIS based approach. *Australasian Transport Research Forum 2010 Proceedings* (pp. 1-19). Canberra: PATREC.

Lin, J., Cromley, R., & Zhang, C. (2011). Using geographically weighted regression to solve the areal interpolation problem. *Annals of GIS*, *17*(1), 1-14.

Liu, X. H., Kyriakidis, P. C., & Goodchild, M. F. (2008). Population-density estimation using regression and area-to-point residual kriging. *International Journal of Geographical Information Science*, *22*(4), 431-447.

Maantay, J., & Maroko, A. (2009). Mapping urban risk: Flood hazards, race, & environmental justice in New York. *Applied geography (Sevenoaks, England)*, *29*(1), 111-124.

Martin, D. (1996). An assessment of surface and zonal models of population. *International Journal of Geographical Information Systems*, *10*(8), 973–989.

Mennis, J. (2003). Generating surface models of population ssing dasymetric mapping. *The Professional Geographer*, *55*(1), 31-42.

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science*, *33*(3), 179-194.

Moriarty, P., & Mees, P. (2006). The journey to work in Melbourne. *29th Australasian Transport Research Forum* (pp. 1-13).

Mugglin, A. S., Carlin, B. P., & Gelfand, A. E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, *95*(451), 877.

Mullen, K. M., & van Stokkum, I. H. M. (2010). The Lawson-Hanson algorithm for non-negative least squares (NNLS). The R project for statistical computing.

Nagle, N. N. (2010). Geostatistical smoothing of areal data: Mapping employment density with factorial kriging. *Geographical Analysis*, *42*, 99-117.

Páez, A., & Scott, D. M. (2004). Spatial statistics for urban analysis: A review of techniques with examples. *GeoJournal*, *61*(1), 53-67.

Reibel, M., & Agrawal, A. (2007). Areal interpolation of population counts using pre-classified land cover data. *Population Research and Policy Review*, *26*(5-6), 619-633.

Silván-Cárdenas, J. L., Wang, L., Rogerson, P., Wu, C., Feng, T., & Kamphaus, B. D. (2010). Assessing fine-spatial-resolution remote sensing for small-area population estimation. *International Journal of Remote Sensing*, *31*(November 2010), 5605-5634.

Sleeter, R., & Wood, N. (2006). Estimating daytime and nighttime population density for coastal communites in Oregon. *Urban and Regional Information Systems Association Annual Conference Proceeding* (pp. 1-15). Vancouver.

APPENDIX A: R SCRIPTS FOR REGRESSION MODELLING

```
-----------------------------------------------------------------
#author    Christabel McCarthy
                #purpose   Modelling employment density of land use
                classes in Melbourne, Australia
#date      1 February 2012
-----------------------------------------------------------------
#required libraries

#set working directory
setwd("/Users/christabelmccarthy/Documents/Study/Semester 3/Data
analysis")

#read data for whole study area (total employment and amount of land in
each land use class in source zones)from csv file
SLAland<-read.csv("SLAland.csv")

#read data for agriculture-and-parkland dominated portion of study area
(total employment and amount of land in each land use class in source
zones)from csv file
SLAland<-read.csv("SLAParkAg.csv")

#read data for urban-dominated portion of study area (total employment
and amount of land in each land use class in source zones)from csv file
SLAland<-read.csv("SLAUrban.csv")



#set up Ordinary Least Squares models with raw data inputs

olsfull<-lm(Tot_work ~ Agricultural + Commercial + Education +
Hospital.Medical + Industrial + Residential + Other + Parkland +
Transport + Transport_a + Water + Shipping + 0, SLAland)

olsurban<-lm(Tot_work ~ Agricultural + Commercial + Education +
Hospital.Medical + Industrial + Residential + Other + Transport_a + 0,
SLAland)

olsemp<-lm(Tot_work ~ Agricultural + Commercial + Education +
Hospital.Medical + Industrial + 0, SLAland)

#set up Ordinary Least Squares models with the study area split into
agriculture-and-parkland dominated and urban dominated areas for
separate regression

pafull<-lm(Tot_work ~ Agricultural + Commercial + Education +
Hospital.Medical + Industrial + Residential + Other + Parkland +
Transport + Transport_a + Water + Shipping + 0, SLAParkAg)

urbanfull<-lm(Tot_work ~ Agricultural + Commercial + Education +
Hospital.Medical + Industrial + Residential + Other + Parkland +
Transport + Transport_a + Water + Shipping + 0, SLAUrban)
```

```
#set up Poisson models with raw data inputs

poissfull<-glm(formula = Tot_work ~ Agricultural + Commercial +
Education + Hospital.Medical + Industrial + Other + Parkland +
Residential + Transport + Transport_a + Water + Shipping + 0, family =
"poisson", data = SLAland)

poissurban<-glm(formula = Tot_work ~ Agricultural + Commercial +
Education + Hospital.Medical + Industrial + Other + Residential +
Transport_a + 0, family = "poisson", data = SLAland)

poissemp<-glm(formula = Tot_work ~ Agricultural + Commercial +
Education + Hospital.Medical + Industrial + 0, family = "poisson", data
= SLAland)

#set up land cover fraction and employment density calculations for
density-based full OLS model

attach(SLAland)
Agricultural.f<-Agricultural/Grand.Total
Commercial.f<-Commercial/Grand.Total
Education.f<-Education/Grand.Total
Hospital.Medical.f<-Hospital.Medical/Grand.Total
Industrial.f<-Industrial/Grand.Total
Other.f<-Other/Grand.Total
Parkland.f<-Parkland/Grand.Total
Residential.f<-Residential/Grand.Total
Transport.f<-Transport/Grand.Total
Transport_a.f<-Transport_a/Grand.Total
Water.f<-Water/Grand.Total
Shipping.f<_Shipping/Grand.Total
density<-Tot_work/Grand.Total

#set data frame for density-based full OLS model

density.df<-data.frame(density, Agricultural.f, Commercial.f,
Education.f, Hospital.Medical.f, Industrial.f, Other.f, Parkland.f,
Residential.f, Transport.f, Transport_a.f, Water.f)

attach(density.df)

#set up density-based full OLS model
olsfulldf<-lm(density ~ Agricultural.f + Commercial.f + Education.f +
Hospital.Medical.f + Industrial.f + Other.f + Parkland.f +
Residential.f + Transport.f + Transport_a.f + Water.f + Shipping.f + 0,
density.df)



#set up land cover fraction and employment density calculations for
density-based urban OLS model

Agricultural.f.u <- Agricultural / (Agricultural + Commercial +
Education + Hospital.Medical + Industrial + Residential + Transport_a +
Other)

Commercial.f.u <- Commercial / (Agricultural + Commercial + Education +
```

```
Hospital.Medical + Industrial + Residential + Transport_a + Other)

Education.f.u <- Education / (Agricultural + Commercial + Education +
Hospital.Medical + Industrial+ Residential + Transport_a + Other)

Hospital.Medical.f.u <- Hospital.Medical / (Agricultural + Commercial +
Education + Hospital.Medical + Industrial + Residential + Transport_a +
Other)

Industrial.f.u <- Industrial / (Agricultural + Commercial + Education +
Hospital.Medical + Industrial + Residential + Transport_a + Other)

Residential.f.u <- Residential / (Agricultural + Commercial + Education
+ Hospital.Medical + Industrial + Residential + Transport_a + Other)

Transport_a.f.u <- Transport_a / (Agricultural + Commercial + Education
+ Hospital.Medical + Industrial + Residential + Transport_a + Other)

Other.f.u <- Other / (Agricultural + Commercial + Education +
Hospital.Medical + Industrial + Residential + Transport_a + Other)

density.u <- Tot_work / (Agricultural + Commercial + Education +
Hospital.Medical + Industrial + Residential + Transport_a + Other)


#set data frame for density-based urban OLS model
density.u.df <- data.frame(density.u, Agricultural.f.u, Commercial.f.u,
Education.f.u, Hospital.Medical.f.u, Industrial.f.u, Residential.f.u,
Transport_a.f.u, Other.f.u)

#set up density-based urban OLS model

olsurbandf <-lm(density.u~Agricultural.f.u + Commercial.f.u +
Education.f.u + Hospital.Medical.f.u + Industrial.f.u + Residential.f.u
+ Other.f.u + Transport_a.f.u + 0, density.u.df)



#set up land cover fraction and employment density calculations for
density-based employment OLS model

Agricultural.f.e <- Agricultural / (Agricultural + Commercial +
Education + Hospital.Medical + Industrial)

Commercial.f.e <- Commercial / (Agricultural + Commercial + Education +
Hospital.Medical + Industrial)

Education.f.e <- Education / (Agricultural + Commercial + Education +
Hospital.Medical + Industrial)

Hospital.Medical.f.e <- Hospital.Medical / (Agricultural + Commercial +
Education + Hospital.Medical + Industrial)

Industrial.f.e <- Industrial / (Agricultural + Commercial + Education +
Hospital.Medical + Industrial)

density.e<-Tot_work / (Agricultural + Commercial + Education +
```

```
Hospital.Medical + Industrial)

#set data frame for density-based employment OLS model

density.e.df <- data.frame(density.e, Agricultural.f.e, Commercial.f.e,
Education.f.e, Hospital.Medical.f.e, Industrial.f.e)

attach(density.e.df)

#set up density-based employment OLS model

olsempdf <- lm(density.e~Agricultural.f.e + Commercial.f.e +
Education.f.e + Hospital.Medical.f.e + Industrial.f.e + 0,
density.e.df)

#inspect model results

summary(olsfull)
summary(olsurban)
summary(olsemp)
summary(pafull)
summary(urbanfull)
summary(poissfull)
summary(poissurban)
summary(poissemp)
summary(olsfulldf)
summary(olsurbandf)
summary(olsempdf)

#generate AIC models for values
AIC(olsfull)
AIC(olsurban)
AIC(olsemp)
AIC(pafull)
AIC(urbanfull)
AIC(poissfull)
AIC(poissurban)
AIC(poissemp)
AIC(olsfulldf)
AIC(olsurbandf)
AIC(olsempdf)

#export residuals for comparison to employment estimate residuals

resolsfull <- residuals(olsfull)
write.table(resolsfull, file= “resolsfull.csv", sep=",",row.names=F)

resolsurban <- residuals(olsurban)
write.table(resolsurban, file= “resolsurban.csv", sep=",”, row.names=F)

resolsemp <- residuals(olsemp)
write.table(resolsemp, file= “resolsemp.csv", sep=",”, row.names=F)

respafull <- residuals(pafull)
write.table(respafull, file= “respafull.csv", sep=",”, row.names=F)

resurbanfull <- residuals(urbanfull)
write.table(resurbanfull, file= “resurbanfull.csv", sep=",”,
```

```
row.names=F)

respoissfull <- residuals(poissfull)
write.table(respoissfull, file= “respoissfull.csv", sep=",”,
row.names=F)

respoissurban <- residuals(poissurban)
write.table(respoissurban, file= “respoissurban.csv", sep=",”,
row.names=F)

respoissemp <- residuals(poissemp)
write.table(respoissemp, file= “respoissemp.csv", sep=",”, row.names=F)

resolsfulldf <- residuals(olsfulldf)
write.table(resolsfulldf, file= “resolsfulldf.csv", sep=",”,
row.names=F)

resolsurbandf <- residuals(olsurbandf)
write.table(resolsurbandf, file= “resolsurbandf.csv", sep=",”,
row.names=F)

resolsempdf <- residuals(olsempdf)
write.table(resolsempdf, file= “resolsempdf.csv", sep=",”, row.names=F)
```

APPENDIX B: ARCPY SCRIPT FOR EMPLOYMENT DISTRIBUTION ESTIMATES FROM REGRESSION COEFFICIENTS

```
# ---------------------------------------------------------------------------
# employment estimate from regression.py
# By Christabel McCarthy
# Created on: 2011-12-23 15:45:30.00000
#   (generated by ArcGIS/ModelBuilder)
#  Usage: employment estimate from regression <Land_use_classifcation>
<Density_coefficient>        <Source_zone_raster>        <Source_zone_table>
<Validation_zone_raster> <Validation_zone_table>
# Description:
# Generates employment estimates that use the global density estimates (from
Poisson and OLS models) as the intial inputs
# ---------------------------------------------------------------------------

# Import arcpy module
import arcpy

# Check out any necessary licenses
arcpy.CheckOutExtension("spatial")

# Set Geoprocessing environments
arcpy.env.scratchWorkspace = "C:\\Thesis\\Employment_estimate.gdb"
arcpy.env.outputCoordinateSystem = ""
arcpy.env.snapRaster = "C:\\Thesis\\Data.gdb\\SLA_raster"
arcpy.env.extent = "MINOF"
arcpy.env.geographicTransformations = ""
arcpy.env.workspace = "C:\\Thesis\\Employment_estimate.gdb"


# Local variables:
Land_use_classification = "C:\\Thesis\\Data.gdb\\Mesh_blocks_Melbourne"
Density_coefficient                                                    =
"C:\\Thesis\\Employment_estimate.gdb\\Density_coefficient"
Initial_density_raster                                                 =
"C:\\Thesis\\Employment_estimate.gdb\\Initial_density_raster"
Source_zone_raster = "C:\\Thesis\\Data.gdb\\SLA_raster"
Inital_employment_estimate_table                                       =
"C:\\Thesis\\Employment_estimate.gdb\\Inital_employment_estimate_table"
Source_zone_table = "C:\\Thesis\\Data.gdb\\SLAs"
Rescaling_factor_raster                                                =
"C:\\Thesis\\Employment_estimate.gdb\\Rescaling_factor_raster"
Final_density_estimate                                                 =
"C:\\Thesis\\Employment_estimate.gdb\\Final_density_estimate"
Validation_zone_raster = "C:\\Thesis\\Data.gdb\\Destination_zone_raster"
Validation_zone_final_estimate                                         =
"C:\\Thesis\\Employment_estimate.gdb\\Validation_zone_final_estimate"
Validation_zone_table = "C:\\Thesis\\Data.gdb\\Destination_zones"


# Join the land use classification table to the table with density
coefficients for each land use category (derived from regression)
```

```
arcpy.JoinField_management(Land_use_classifcation,                "CATEGORY",
Density_coefficient, "CATEGORY", "")

# Create a raster layer with the density coefficient values at each cell
arcpy.PolygonToRaster_conversion(Land_use_classification,
"Density_coefficient", Initial_density_raster, "MAXIMUM_AREA", "NONE", "10")

# Sum the value of the raster within each source zone for comparison to the
true employment counts in each source zone
arcpy.gp.ZonalStatisticsAsTable_sa(Source_zone_raster,              "Value",
Initial_density_raster, Inital_employment_estimate_table, "DATA", "SUM")

# Join the summary table to sourc zone table containing the source zone
employment counts
arcpy.JoinField_management(Source_zone_table,                     "SLA_Code",
Inital_employment_estimate_table, "SUM", "")

# Add a field in which to calculate the ratio of the true count to the
estimated count (rescaling factor)
arcpy.AddField_management(Source_zone_table, "Ratio true est", "FLOAT", "",
"", "", "", "NULLABLE", "NON_REQUIRED", "")

# Calculate the rescaling factor
arcpy.CalculateField_management(Source_zone_table,    "Ratio    true    est",
"!Tot_work!/!Initial employment estimate.SUM!", "PYTHON 9.3", "")

# Produce a raster of the rescaling factor values
arcpy.PolygonToRaster_conversion(Source_zone_table,    "Ratio    true    est",
Rescaling_factor_raster, "MAXIMUM_AREA", "NONE", "10")

# Produce a raster of the original density coefficient multiplied by the
rescaling factor
arcpy.gp.RasterCalculator_sa("\"%Initial  density  raster%\"  *  \"%Rescaling
factor raster%\"", Final_density_estimate)

# Sum the value of the rasters within each validation zone for comparison to
the true employment counts in each validation zone
arcpy.gp.ZonalStatisticsAsTable_sa(Validation_zone_raster,          "Value",
Final_density_estimate, Validation_zone_final_estimate, "DATA", "SUM")

# Join the summary table to source zone table containing the source zone
employment count
arcpy.JoinField_management(Validation_zone_table,              "Destinat_1",
Validation_zone_final_estimate, "VALUE", "")

# Add a field in which to calculate the residual (difference) between the
estimated and the true employment counts in the validation zones
arcpy.AddField_management(Validation_zone_table, "Difference", "FLOAT", "",
"", "", "", "NULLABLE", "NON_REQUIRED", "")

# Calculate the residual
arcpy.CalculateField_management(Validation_zone_table,          "Difference",
"!Validation_zone_final_estimate.SUM!-!Tot_work!", "PYTHON 9.3", "")
```

APPENDIX C: ARCPY SCRIPT FOR BINARY EMPLOYMENT OR WORKING POPULATION DISTRIBUTION ESTIMATES

```
# ---------------------------------------------------------------------------
# binary estimate.py
# By Christabel McCarthy
# Created on: 2011-12-23 23:16:25.00000
#   (generated by ArcGIS/ModelBuilder)
# Usage: binary estimate<Validation_zone_raster> <Validation_zone_table>
# Description: Binary estimate of working population distribution. Can be
adapted to produce the binary estimate of employment
# ---------------------------------------------------------------------------

# Import arcpy module
import arcpy

# Check out any necessary licenses
arcpy.CheckOutExtension("spatial")

arcpy.env.scratchWorkspace = "C:\\Thesis\\Working_pop_estimate.gdb"
arcpy.env.outputCoordinateSystem = ""
arcpy.env.snapRaster = "C:\\Thesis\\Data.gdb\\SLA_raster"
arcpy.env.extent = "MINOF"
arcpy.env.geographicTransformations = ""
arcpy.env.workspace = "C:\\Thesis\\Working_pop_estimate.gdb"

# Local variables:
Land_use_classification = "C:\\Thesis\\Data.gdb\\Mesh_blocks_Melbourne"
Binary_classification = "C:\\Thesis\\Data.gdb\\Mesh_blocks_Melbourne"
Binary_classification_raster                                              =
"C:\\Thesis\\Working_pop_estimate.gdb\\Binary_classification_raster"
Source_zone_raster = "C:\\Thesis\\Data.gdb\\SLA_raster"
Populated_area_table                                                      =
"C:\\Thesis\\Working_pop_estimate.gdb\\Total_populated_area"
Source_zone_table = "C:\\Thesis\\Data.gdb\\SLAs"
Source_zone_working_pop_density                                           =
"C:\\Thesis\\Working_pop_estimate.gdb\\Source_zone_working_pop_density"
Final_binary_estimate                                                     =
"C:\\Thesis\\Working_pop_estimate.gdb\\final_binary_estimate"
Validation_zone_raster = "C:\\Thesis\\Data.gdb\\Destination_zone_raster"
Validation_zone_table = "C:\\Thesis\\Data.gdb\\Destination_zones"
Validation_zone_estimate_summary                                          =
"C:\\Thesis\\Working_pop_estimate.gdb\\Validation_zone_estimate_summary"

# Add a field to the land use classification table,  in which to assign the
binary class (1=has working population, 0=doesn't have working population)
arcpy.AddField_management(Land_use_classification,  "Binary_class",  "SHORT",
"", "", "", "", "NULLABLE", "NON_REQUIRED", "")

# Assign the value '1' to the land class that has counts
#(in this case, Residential is assigned the counts, for the working
population estimate, Agricultural, Commercial, Education, Industrial and
Hospital/Medical land classes take the counts
```

```
arcpy.CalculateField_management(Land_use_classification, "Binary_class",
"Lc", "VB", "dim Lc\\nif [CATEGORY]=\"Residential\" then\\nLc=1\\n\\nElse
\\nLc=0\\n\\n\\nEnd if")

# Convert the binary classification to a raster
arcpy.PolygonToRaster_conversion(Binary_classification, "Binary_class",
Binary_classification_raster, "MAXIMUM_AREA", "NONE", "10")

# Sum the value of the raster within each source zone to find the populated
area within the source zone
arcpy.gp.ZonalStatisticsAsTable_sa(Source_zone_raster, "Value",
Binary_classification_raster, Populated_area_table, "DATA", "SUM")

# Join the summary populated area table to the source zone table containing
the source zone working population counts
arcpy.JoinField_management(Source_zone_table, "SLA_Code",
Populated_area_table, "SUM", "")

# Add a field in which to calculate the working population density of the
populated area of the source zones
arcpy.AddField_management(Source_zone_table, "Working pop density", "FLOAT",
"", "", "", "", "NULLABLE", "NON_REQUIRED", "")

# Calculate the working population density of the populated area of the
source zones
arcpy.CalculateField_management(Source_zone_table, "Working pop density",
"!Res_workers!/!Populated area table.SUM!", "PYTHON_9.3", "")

# Create a raster of the working population density of the populated area of
the source zones
arcpy.PolygonToRaster_conversion(Source_zone_table, "Working pop density",
Source_zone_working_pop_density, "MAXIMUM_AREA", "NONE", "10")

# Calculate the downscaled working population estimate by multiplying the
working population density raster by the binary classification raster
# Population density will be assigned to the populated area only (those with
a value of '1')
arcpy.gp.RasterCalculator_sa("\"%Binary_classification_raster%\" * \"%Source
zone working pop density%\"", Final_binary_estimate)

#Process: Sum the value of the rasters within each validation zone for
comparison to the true working population counts in each validation zone
arcpy.gp.ZonalStatisticsAsTable_sa(Validation_zone_raster, "Value",
Final_binary_estimate, Validation_zone_estimate_summary, "DATA", "SUM")

# Join the summary table to the validation zone table containing the
validation zone working population count
arcpy.JoinField_management(Validation_zone_table, "Origin_Cod",
Validation_zone_estimate_summary, "VALUE", "")

# Add a field in which to calculate the residual (difference) between the
estimated and the true working population counts in the validation zones
arcpy.AddField_management(Validation_zone_table, "Difference", "FLOAT", "",
"", "", "", "NULLABLE", "NON_REQUIRED", "")

# Calculate the residual
arcpy.CalculateField_management(Validation_zone_table, "Difference",
"!Validation_zone_estimate_summary.SUM!-!Res_workers!", "PYTHON_9.3", ""
```

APPENDIX D: ARCPY SCRIPT FOR WORKING POPULATION DISTRIBUTION ESTIMATES FROM TOTAL POPULATION DENSITY

```
# ---------------------------------------------------------------------------
# Working population from total population density.py
# Created on: 2011-12-23 17:30:58.00000
#   (generated by ArcGIS/ModelBuilder)
#    Usage:    Working    population    from    total    population    density
<Land_use_classifcation>       <Source_Zone_Raster>       <Source_zone_table>
<Validation_zone_raster> <Validation_zone_table>
# Description:
# Generates working population estimates from the total population count at
mesh block level
# ---------------------------------------------------------------------------

# Import arcpy module
import arcpy

# Check out any necessary licenses
arcpy.CheckOutExtension("spatial")

# Set Geoprocessing environments
arcpy.env.scratchWorkspace = "C:\\Thesis\\Working_pop_estimate.gdb"
arcpy.env.outputCoordinateSystem = ""
arcpy.env.snapRaster = "C:\\Thesis\\Data.gdb\\SLA_raster"
arcpy.env.extent = "MINOF"
arcpy.env.geographicTransformations = ""
arcpy.env.workspace = "C:\\Thesis\\Working_pop_estimate.gdb"


# Local variables:
Land_use_classification = "C:\\Thesis\\Data.gdb\\Mesh_blocks_Melbourne"
Total_pop_density_raster                                                    =
"C:\\Thesis\\Working_pop_estimate.gdb\\Total_pop_density_raster"
Source_zone_raster = "C:\\Thesis\\Data.gdb\\SLA_raster"
Total_pop_count = "C:\\Thesis\\Working_pop_estimate.gdb\\Total_pop_count"
Source_zone_table = "C:\\Thesis\\Data.gdb\\SLAs"
Scaling_factor_raster                                                       =
"C:\\Thesis\\Working_pop_estimate.gdb\\Scaling_factor_raster"
Final_density_estimate                                                      =
"C:\\Thesis\\Working_pop_estimate.gdb\\Final_density_estimate"
Validation_zone_raster = "C:\\Thesis\\Data.gdb\\Destination_zone_raster"
Validation_zone_final_estimate                                              =
"C:\\Thesis\\Working_pop_estimate.gdb\\Validation_zone_final_estimate"
Validation_zone_table = "C:\\Thesis\\Data.gdb\\Destination_zones"

# Add a field to the land use classification table, in which to calculate
total population density at mesh block level
arcpy.AddField_management(Land_use_classifcation, "Popdensity", "FLOAT", "",
"", "", "", "NULLABLE", "NON_REQUIRED", "")

# Calculate the total population density
arcpy.CalculateField_management(Land_use_classifcation,         "Popdensity",
"!TURPOP2006!/ !Shape_Area!", "PYTHON_9.3", "")
```

```
# Create a raster layer with the total population density values at each cell
arcpy.PolygonToRaster_conversion(Land_use_classification,       "Popdensity",
Total_pop_density_raster, "MAXIMUM_AREA", "NONE", "10")

# Sum the value of the raster within each source zone for comparison to the
working population counts in each source zone
arcpy.gp.ZonalStatisticsAsTable_sa(Source_zone_raster,              "Value",
Total_pop_density_raster, Total_pop_count, "DATA", "SUM")

# Join the summary table to the source zone table containing the source zone
working population counts
arcpy.JoinField_management(Source_zone_table,  "SLA_Code",  Total_pop_count,
"SUM", "")

# Add a field in which to calculate the ratio of the working population count
to the total population count (rescaling factor)
arcpy.AddField_management(Source_zone_table, "Ratio working total", "FLOAT",
"", "", "", "", "NULLABLE", "NON_REQUIRED", "")

# Calculate the rescaling factor
arcpy.CalculateField_management(Source_zone_table,  "Ratio  working  total",
"!Res_workers!/!Total_pop_count.SUM!", "PYTHON_9.3", "")

# Produce a raster of the rescaling factor values
arcpy.PolygonToRaster_conversion(Source_zone_table,  "Ratio  working  total",
Scaling_factor_raster, "MAXIMUM_AREA", "NONE", "10")

#  Produce  a  raster  of  the  total  population  density  multiplied  by  the
rescaling factor
arcpy.gp.RasterCalculator_sa("\"%Total  pop  density  raster%\"  *  \"%Density
ratio raster%\"", Final_density_estimate)

# Sum the value of the rasters within each validation zone for comparison to
the true working population counts in each validation zone
arcpy.gp.ZonalStatisticsAsTable_sa(Validation_zone_raster,          "Value",
Final_density_estimate, Validation_zone_final_estimate, "DATA", "ALL")

#  Join  the  summary  table  to  the  validation  zone  table  containing  the
validation zone working population count
arcpy.JoinField_management(Validation_zone_table,                  "Obj_ID",
Validation_zone_final_estimate, "VALUE", "")

# Add a field in which to calculate the residual (difference) between the
estimated and the true working population counts in the validation zones
arcpy.AddField_management(Validation_zone_table,  "Difference",  "FLOAT",  "",
"", "", "", "NULLABLE", "NON_REQUIRED", "")

# Calculate the residual
arcpy.CalculateField_management(Validation_zone_table,          "Difference",
"!Validation_zone_final_estimate.SUM!-!Res_workers!", "PYTHON_9.3", "")
```

Tobler, W. R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, *74*(367), 519–30.

Trendle, B., & Siu, J. (2005). Commuting patterns of Sunshine Coast residents and the impact of education. Brisbane: Queensland Government Department of Education and Training Working Paper No. 37. (pp. 1-17).

Watts, M. J. (2009). The impact of spatial imbalance and socioeconomic characteristics on average distance commuted in the Sydney metropolitan area. *Urban Studies*, *46*(2), 317-339.

Yigitcanlar, T., Dodson, J., Gleeson, B., & Sipe, N. (2007). Travel self-containment in master planned estates: Analysis of recent Australian trends. *Urban Policy and Research*, *25*(1), 129-149.

Yoo, E.-H., Kyriakidis, P. C., & Tobler, W. (2010). Reconstructing population density surfaces from areal data: A comparison of Tobler's pycnophylactic interpolation method and area-to-point kriging. *Geographical Analysis*, *42*(1), 78-98.

Yuan, Y., Smith, R. M., & Limp, W. F. (1997). Remodelling census population with spatial information from LandSat TM imagery. *Comput., Environ. and Urban Systems*, *21*(3/4), 245-258.

Zandbergen, P. A. (2011). Dasymetric mapping using high resolution address point datasets. *Transactions in GIS*, *15*(s1), 5-27.