Applied GIS

an international, refereed, open source journal

ISSN: 1832-5505

URL:
http://www.appliedgis.net

MANAGING EDITORS:

Jim Peterson – Jim.Peterson@arts.monash.edu.au

Ray Wyatt – ray.wyatt@unimelb.edu.au

Volume 3, Number 9

September, 2007

CONTENTS:

All papers published during 2007 are part of *Volume* 3.
Each paper constitutes one *Number.*

Hence this paper should be cited as:

Li, T., Pullar, D., Corcoran, J. & Stimson, R. (2007) – A comparison of spatial disaggregation techniques as applied to population estimation for South East Queensland (SEQ), Australia, *Applied GIS*, 3(9), 1-16

A comparison of spatial disaggregation techniques as applied to population estimation for South East Queensland (SEQ), Australia

Tiebei Li
The University of Queensland Social Research Centre (UQSRC)
St Lucia, 4072, Australia
t.li@uq.edu.au

David Pullar
Department of Geography, Planning and Architecture (GPA)
University of Queensland, St Lucia, 4072, Australia
d.pullar@uq.edu.au

Jonathan Corcoran
The University of Queensland Social Research Centre (UQSRC)
St Lucia, 4072, Australia
Ji.corcoran@uq.edu.au

Robert Stimson
The University of Queensland Social Research Centre (UQSRC)
St Lucia, 4072, Australia
r.stimson@uq.edu.au

Abstract: The accuracy of spatial disaggregation techniques largely depends on their underlying density assumptions and the quality of the data applied. This paper presents the results of a comparative investigation of four spatial disaggregation methodologies to determine their relative accuracies. These methodologies include binary dasymetric, a regression model, a locally fitted regression model and three-class dasymetric, each of which provides different solutions for explaining spatially heterogeneous density when population data is spatially disaggregated. In contrast to previous studies, we apply the spatial disaggregation techniques to a comparably larger and more varied geographical area which allows the spatial disaggregation techniques to be more rigorously tested. Results indicate that the three-class dasymetric technique generates higher levels of accuracy compared to the other spatial disaggregation techniques and this result is more conclusive than previous findings.

Key words: MAUP, spatial disaggregation, dasymetric mapping, regression model

1. Introduction

Socio-economic data are typically collected and reported at a spatially aggregated level (e.g. census zones). This inevitably causes bias where the masked distributions of disaggregate, spatial data are aggregated to a larger spatial unit. In some cases it can lead to very misleading results where the aggregate unit represents an arbitrary regionalisation of space that is not well related to the disaggregate data. This is known in the literature as the Modifiable Area Unit Problem (MAUP) (Fotheringham & Wong 1991; Fotheringham & Rogerson 1993; Dennis & Wu 1996; Moon & Farmer 2001).

Policy issues often need to focus on variations across geographical space or regional details, and so portraying areas with a uniform distribution is distinctly uninformative

(Bracken 1989; Landis 1994; Rosenbaum & Koenig 1997; Hunt et al., 2005). This raises the need to break up the homogeneous aggregated areas and so make the internal variations observable, and this concept is called spatial disaggregation.

Yet unlike spatial aggregation, which only involves a straight forward statistical summary, there is no definitive way to carry out spatial disaggregation. The transfer of data in the spatial disaggregation process is complex because of the nature of mismatched boundaries between source areas and the boundaries for the target areas, and their heterogeneity in terms of density. Different techniques have been developed to solve the problem based upon underlying assumptions and the availability of other ancillary data to inform the process.

Inevitably however, all spatial disaggregation techniques generate error, because there are always limitations associated with assumptions they use. Some errors are caused by assumptions about the spatial distributions of the objects (e.g. homogeneity in density), and some errors are caused by the spatial relationship imposed within the spatial disaggregation process (e.g. size of target zones) (Lam 1983; Cockings et al. 1997; Sadahiro 1999). Basically, the accuracy of estimation primarily depends on the appropriateness of the assumptions applied, as well as on the geography of the areas. A summary of the assumptions made under different spatial disaggregation techniques is as follows.

Technique	**Method**	**Assumption**	**Control Surface (ancillary data)**	**Complexity (1—5)**
Simple Area Weighting	Cartographic	Homogeneous source zones	None	5
Regression Model	Statistical	Source zone composed of land classes with global uniform density	Discrete or Continuous	3
Binary Dasymetric Mapping	Cartographic	Source zone composed of populated and unpopulated areas	Discrete (binary)	2
Three-Class Dasymetric Mapping	Cartographic	Homogeneity at different land class (at each source zone)	Discrete	1- 2
EM Algorithm	Statistical	Source zone composed of land classes with global uniform density that conserve aggregate value	Discrete or Continuous	1- 2

Table 1 - A comparison of different spatial disaggregation techniques in terms of their assumptions, methods and data demand.

The simplest spatial disaggregation technique, known as Simple Area Weighting, assumes homogeneity within source zones. Clearly, this is far from the reality of expected spatial distributions for socioeconomic data (for example, population counts). Simple cartographic processing methods, such as overlay, are used to disaggregate the source zones. Other more advanced techniques deal with the more realistic expectation that source zones are heterogeneous but with an unknown structure. Note that it is possible to spatially overlay land use data, such as mapping from remote sensed data (Langford et al. 1991) or road density (Reibel 2005; Reibel and Aditya 2006) over the source zones to provide ancillary information to indicate variation in data distributions of the aggregated source data.

Different approaches have been proposed based upon the assumptions made about the spatial structure imposed on the source zones resulting from the overlayed spatial data. Regression models (Langford et al. 1991; Yuan et al. 1997) assume that the ancillary land use classes define areas of global uniform density. That is, the land classes have a uniform area density that is related to the parameter of interest over the whole of the area, but it is unknown. Using a combination of the aggregate source values and the ancillary data with

unknown densities it is possible to developed regression equations to numerically resolve this relationship.

A drawback of this approach is that the global densities it computes allow for small errors between the estimated and the actual source-zone values. The quality of resolved densities maintaining the volume of the aggregate data value is called the pycnophylactic property (Tobler 1979; Goodchild et al. 1993).

hence there is another statistical technique for estimating the globally uniform density for each land class while satisfying the pycnophylactic property - the EM algorithm (Flowerdew & Green 1991; Flowerdew & Green 1992; Gregory & Paul 2005). However, the assumption of uniform area density for each land class might be problematic when dealing with many areas over a large region where relationships between population and land class are not spatially uniform. Langford (2006) argued that global fitted density can be estimated at local level by dasymetric mapping which allows for some global variability in density for each land class.

A simple example is binary dasymetric mapping (Eicher & Brewer 2001) which takes a binary land classification to control the population allocation. It assumes a non-zero density in the populated areas within each source zone and a zero density elsewhere. Hence varying assumptions can be made about the density in a functional way. A further refinement to this is three-class dasymetric mapping (Mennis 2003), which incorporates a functional relationship with area densities so that densities are uniform within a source zone even though they may vary across the larger region.

Overall, the density assumptions of different spatial disaggregation techniques can be illustrated by Figure 1, where the vertical bars represent density for each land use class and the parallel bar represents the density of the source zones. Comparably, the most relaxed assumption of homogeneity used by three-class dasymetric mapping is close to the complexity of real world.

a **b** **c** **d**

Figure 1 - Density assumptions of different spatial disaggregation techniques

The three-class dasymetric mapping is theoretically more appropriate to accommodate the spatial heterogeneity of a large geographical area. Langford (2006) evaluated spatial disaggregation techniques using UK Census data for the county of Leicestershire. The results show that the three-class dasymetric method largely outperforms other spatial disaggregation techniques, apart from the comparatively simpler binary dasymetric method.

One possible reason of this inconclusive result is the more complex three-class dasymetric technique is more sensitive to the land classification errors. On the other hand, as Fisher and Langford (1995) pointed out, the significance of comparative results is always limited by simplicity in the spatial structure of the study area, and a more conclusive result could be experimentally validated by broadening the study area to include more spatial heterogeneous density.

Therefore, the task in this study is to fully evaluate the accuracy of three-class dasymetric mapping using a larger and more complex geographical area, namely South East Queensland (SEQ), Australia, to verify previous findings. In other words, we will conduct a comparative investigation to determine the accuracies of four spatial disaggregation methodologies, namely binary dasymetric, a regression model, a locally fitted regression

model and a three-class dasymetric technique to spatially disaggregate population data. The aim is to obtain error estimates on spatial variables that will be used in other predictive models. Specifically, models will be developed to predict the future urban growth in SEQ.

The structure of remainder of this paper is as follows. The study area of SEQ and datasets used in the study are introduced in the next section – section 2. We demonstrate that the region of SEQ has more variability than areas used in other comparative studies of spatial disaggregation methods. In section 3, we describe the spatial disaggregation methods to be tested. In section 4, we discuss the results from the spatial disaggregation techniques, by examining visualised outputs of spatial disaggregation based on different assumptions, and by evaluating the absolute errors and root-mean-square-errors (RMSE) of the results. In the final section we summarise our findings and indicate some directions for future work.

2. The Study Area and Data

The SEQ region covers a relatively large geographical area (2,279,903 hectares) and houses a population of 2,479,295. Figure 2 depicts the settlement pattern of SEQ, which varies greatly, from the city for Brisbane and growing populations in nearby coastal settlements of the Sunshine and Gold Coast. The latter forms a metropolitan region which is colloquially described as “the 200km city” (Brisbane institute, 2004). By contrast, the population drops off dramatically away from the coast, with the exception of the two cities of Ipswich and Toowoomba.

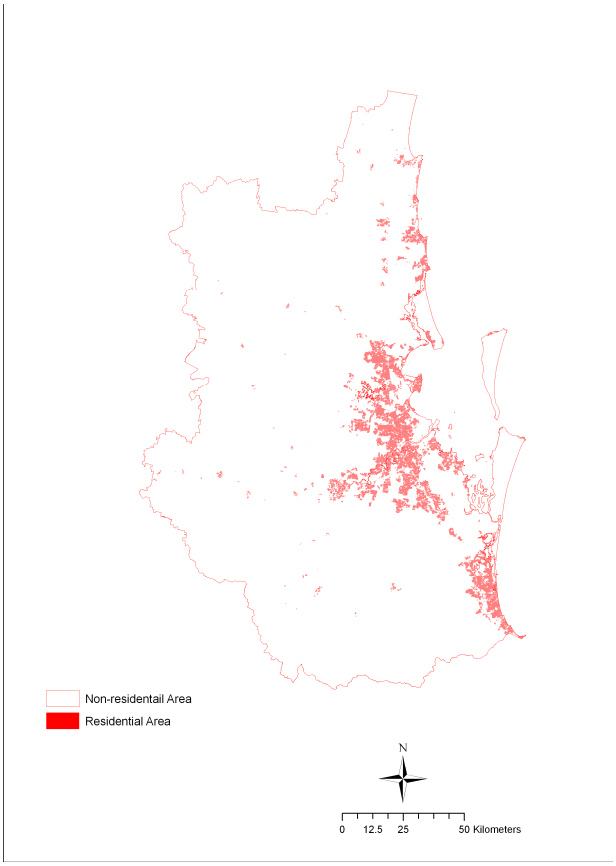


Figure 2 - Spatial characteristics of SEQ

Although, the eastern part of the region is heavily populated and urbanized, it is still mixed with other land covers in many small areas (see Figure 3). This makes he region substantially spatially unbalanced in terms of land use variations, with the population density varying either globally or locally.

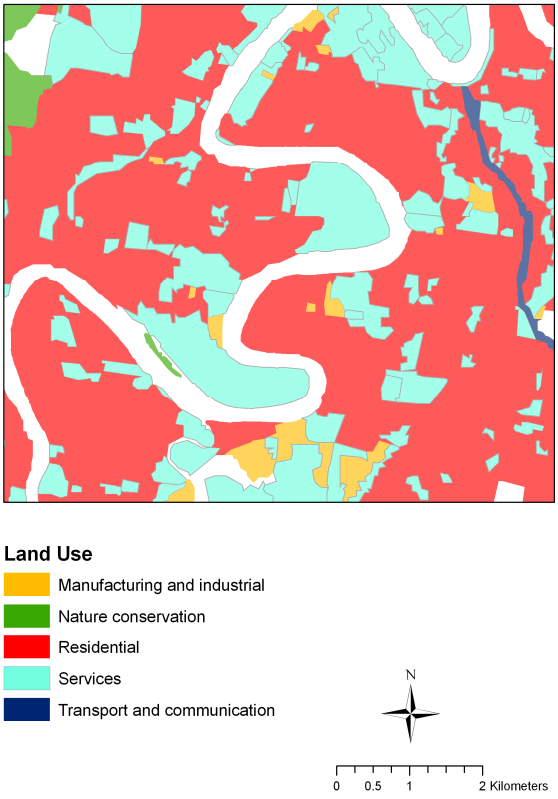


Figure 3 - Mixed land use in the local areas

In short, the degree to which population density varies throughout SEQ is much greater than in areas used for previous studies undertaken by Langford & Fisher (1996) and by Langford (2006), who used the county of Leicestershire, UK. Table 2 describes the degree of spatially heterogeneous density for the two study areas.

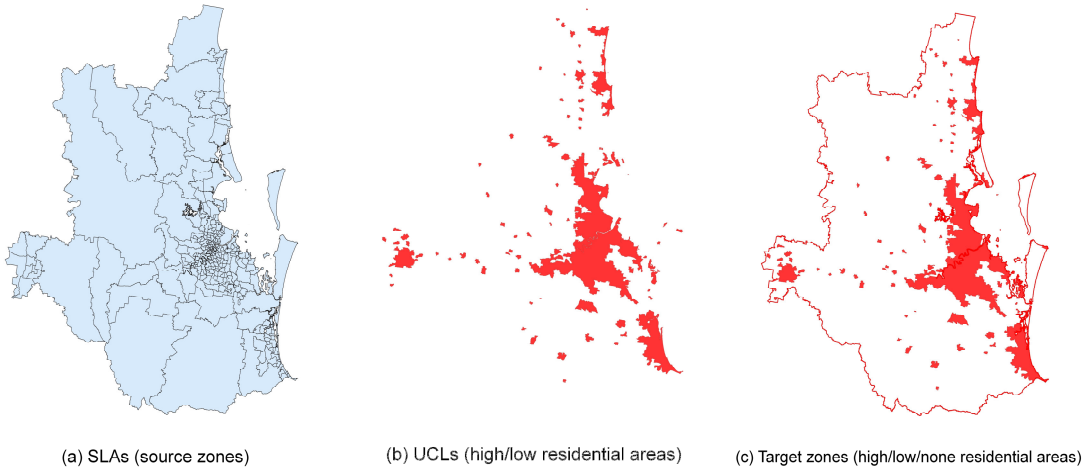
Leicestershire has a relatively uniformly population of 459,772 dispersed across high-, medium- and low-density residential areas, and it covers an area of 81,700 hectares. By contrast, SEQ exhibits a greater degree of heterogeneity. There are much greater extremes of residential density, which is highlighted in the final column (density ratio) of Table 2. The density variation for SEQ is due to population concentrations that are distributed unevenly across the region and so SEQ is considered a more suitable study area to more rigorously evaluate the relative performances of spatial disaggregation techniques.

Study Area	Total Pop. ('000)	Total Size ('000 hect.)	Ave. Density	Ave. high density	Ave. medium density	Ave. low density	Density Ratio (high/low)
Leicestershire	459	82	5.628	37.839	21.114	4.746	7.97
SEQ	2,479	2,280	1.087	6.85	2.99	0.0379	180.8

Table 2 - Population density variations for SEQ in comparison to Leicestershire.

2.1 Data

The Australian Bureau of Statistics (ABS) provides census data for statistical collection zones and urban areas. We test a single case of error for each spatial disaggregation technique using the 2001 census data. We obtained population counts for 298 statistical local areas (SLAs) that are used as source zones (see Figure 4. a). The census data at urban centre localities (UCL), classified as low-density urban and high-density urban areas, were used as both control zones (79 polygons) and target zones (80 polygons). In this case, the target zones are spatially non-contiguous with source zones and congruent with control zones (see Figures 4.b and 4.c).



(a) SLAs (source zones) (b) UCLs (high/low residential areas) (c) Target zones (high/low/none residential areas)

a b c

Figure 4 - Data for the test

For the purpose of accuracy evaluation, we only spatially disaggregate population data for the urban areas in the SEQ region. Hence overall there are less target zones than source zones. Most target zones are smaller than source zones except in the central urban areas where the reverse is the case (i.e. the target zones are larger than the source zones). However, this only represents a small part of the data and such a limitation was deemed acceptable for testing each of the spatial disaggregation techniques.

3. Methodology

The principle of spatial disaggregation is to transfer data from one zonation to another in a spatially disaggregate manner. The original spatial units, with known data, are called source zones, and the final spatial units that accept the refined data are called the target zones which describe the same region. There are many techniques to disaggregate different types

of spatial variables, but the main focus in this paper is on count measurements, that is, disaggregating population or census counts.

The primary aim of this paper is to compare the relative accuracies of four heterogeneous density solutions represented by a binary dasymetric approach, a regression model, a locally fitted regression model and a three-class dasymetric technique. It is recognised that more spatial disaggregation techniques exist, but this paper includes the most representative classes of techniques based on the different density assumptions, as detailed in the literature.

3.1 Simple Area Weighting

Simple area weighting is based on proportioning the source attribute by area, given the geometric intersection of the source zones and target zones. It assumes homogeneity within each source zone, and based on this assumption the data for each target zone can be estimated as:

$$P_t = \sum_s \frac{A_{st} \times P_s}{A_s} \tag{1}$$

where: P_t is the estimated population count at target zone *t*;

P_s is the observed population for the source zone *s*;

A_s is the area size of source zone s and

A_{st} is the area size of intersection of source and target zones.

The technique is rather easy to implement without any ancillary data demand.

The problem with this method is that it is incorrect to assume that density of population within the source zones is uniform. There have been numerous studies that have shown the overall low accuracy of simple area weighting in comparison to other techniques (see for example, Langford 2006; Gregory 2005; Reibel & Aditya 2006), we exclude the simple area weighting method from the our comparison set in this study.

3.2 Binary Dasymetric Mapping

Researchers have attempted to relax the restrictive assumption of homogeneous density by adding supplementary knowledge about locality onto the spatial structure of source zones. Binary dasymetric mapping (Langford & Unwin 1994; Langford & Fisher 1996) uses a binary land use classification (either populated or empty areas) as ancillary data to aid area-based data interpolation.

The purpose of the ancillary information is to allow the internal structure of population distribution within source zones to be inferred. Binary dasymetric mapping assumes that the population is concentrated,with fixed population density, inside the urban areas that are within each source zone. The method is different to simple area weighting as it only considers the populated areas in the target zones for allocating population to:

$$P_t = \sum_{s=1}^{S} \frac{A_{tsp} \times P_s}{A_{sp}} \tag{2}$$

where: P_t is estimated population at target zone *t*;

A_{tsp} is the area of overlap between target zone *t* and source zone *s* having land cover identified as populated;

A_{sp} is the source zone area identified as populated and

P_s is the total population in source zone *s*.

However, this method is unable to address more complex land uses that have a variety of population concentrations. Also, areas with low populations are always arbitrarily ignored and this may be a significant problem.

3.3 The regression model

The regression model exploits ancillary spatial information to improve estimation accuracy. Detailed ancillary information is typically derived from remotely sensed data, or land use data, to classify the levels of urbanization on the land which will take higher or lower population concentrations. That is, regression methods obtain global estimates for density for each land class over the entire study area.

In other words, they assume that the given source zone population may be expressed in terms of a set of densities related to the areas assigned to the different land classes. Other ancillary variables may be included for these area densities, but the basic model is:

$$P_s = \sum_{c=1}^{C} (d_c \cdot A_{sc} + \varepsilon_s) \quad (3)$$

where: P_s is the total population count for each source zone *s*;

c is the land cover class;

A_{sc} is the area size for each land class within each source zone;

d_c is the coefficient of the regression model and

ε_s is the random error.

The intercept is always set to zero, as an area with size zero should have zero population, and although a linear regression equation Is usually used Langford et al. (1991) found that increasing the complexity of the equation will improve the accuracy of the interpolation results. Accordingly, in this research we take three land classifications (high density urban,; low density urban and non-urban) as independent variables for establishing regression relationships with population counts.

However, regression analysis is not supported by current GIS products and so it requires additional statistical software (e.g. SPSS) to do the analysis. Another disadvantage is that because the densities are derived from a global context, they remain spatially stable within each land class throughout the study area. Hence the nature of spatial variation between different census reporting zones that have the same land class cannot be addressed properly. This is why Langford (2006) argued that the locally fitted approach used by dasymetric method will always outperform the global fitting approach used by regression models.

Another statistical approach in the same density-solution class as the regression model is the EM algorithm (Flowerdew and Green 1992). Rather than using a regression approach, the EM algorithm incorporates an iterative, best-fitting approach to derive the density for each land class that satisfies the pycnophylactic property.

Although the EM algorithm is complex, it is still based on the same assumption that the densities for each land class are constant across the space (see Figure 1). The method is presumed to have same level of ability to address spatially heterogeneous density as the regression model. Therefore, in this study we only implement the regression model . It is deemed unnecessary to duplicate of the same type of density solution.

3.4 The locally fitted regression model

The locally fitting approach was introduced by Yuan et al. (1997) and Langford (2006) to improve the reliability of estimated densities derived from the regression model. It was developed initially to ensure that the populations reported within target zones were constrained to match the overall sum of the source zones (the pycnophylactic property). That is, the globally estimated density for each land class is locally adjusted within each source zone by the ratio of the predicted population and census counts. In this way a variation of the absolute value of population densities is achieved by reflecting the differences in terms of local population density between source zones.

The mathematical expression of the density-adjusting approach is:

$$d_{cs} = \frac{P_s}{P_{es}} \times d_c \tag{4}$$

where: d_{cs} is the specific density estimates for class *c* in zone *s*;

p_s is the actual population of source zones *s*;

P_{es} is the estimated population of source zone *s* and

d_c is the initial global density estimate of land class *c*.

The use of locally fitted regression has modified the assumption of the regression model by objectively allowing spatially inconsistent density values within each land class. This approach is comparably simple, but based on the relaxed homogeneity assumptions regarding density. It is quite desirable for SEQ and well worth being compared with the more advanced, three-class dasymetric mapping method.

3.5 Three-class dasymetric mapping

The three-class dasymetric mapping (Mennis 2003; Langford 2006) takes advantage of binary dasymetric mapping and the regression model with a limited number of ancillary class variables (i.e. non-urban, low-density residential and high-density residential) to present a range of residential densities within each source zone. The technique is based on the most relaxed assumption about homogeneous density for each land class within each source zone (same homogeneity level within the locally fitted regression approach, see Figure 1).

Different variables have been used to define density variability. Langford (2006) requires the relative densities for each land class within each source zone to implement three-class dasymetic mapping. The equation is given as:

$$P_t = \sum_{s=1}^{S} \sum_{c=1}^{C} \frac{A_{stc} P_s}{A_{sc}} = \sum_{s=1}^{S} \sum_{c=1}^{C} A_{tsc} d_{sc} \tag{5}$$

where: P_t is the estimated population of target zone *t*;

A_{stc} is the area of intersection between target zone *t* and source zone *s*, and identified as land class *c*;

P_s is the population of source zone *s* and

A_{sc} is the area of source zone *s* identified as land class *c*.

Alternatively, Mennis (2003) established the relative ratios of density values (density fraction given by equation 6) assigned to each land cover to accept certain proportions of the total population from each source zone (see equations 6 – 9). Mennis's work is a modified version of the previous simplistic method of dasymetric mapping which applied a fixed proportion of

population for each land use type (Eicher & Brewer 2001). Similarly to the regression model, Mennis (2003) used an area-based, locally fitting approach to adjust the global density fractions within each source zone while allowing spatial inconsistency to exist (see equations 7 and 8).

$$D_c = d_c \Big/ \sum_{c=1}^{n} d_c \tag{6}$$

$$A_{sc} = (a_{sc} / a_s) \tag{7}$$

$$f_{sc} = (D_c \times A_{sc}) / \left[\sum_{c=1}^{n} (D_c \times A_{sc}) \right] \tag{8}$$

$$P_t = \sum_{s=1}^{S} \left(f_{sc} \times P_s \times a_{tsc} \right) / a_{sc} \tag{9}$$

where: in eqn. (7), A_{sc} is the area ratio;

a_{sc} is the area size for each land class within each source zone and

a_s is the size of each source zone.

In eqn. (8), f_{sc} is the adjusted density fraction which will be an overall score used by the target zone when taking population counts.

The geographic pattern of SEQ is similar to that of the Southeast Pennsylvania (Mennis 2003). Therefore we adapted the three-class dasymetric technique proposed by Mennis (2003) using a density fraction. The technique is expected to well accommodate the spatially skewed land use variation. Compared with simpler techniques, this method is complex, and modification is required due to different data conditions.

3.6 Modification of three-class dasymetric mapping - a hybrid model

Mennis (2003) suggested a sampling approach to assess the relative density fractions for each land class. This assumes the original spatial units of population are small enough to be contained entirely within each ancillary land class. However, the sizes of source zones in SEQ are fairly large, especially in the non-urban areas, and this makes the sampling approach unfeasible.

Instead, we estimated the initial density fractions using a regression model (see equation 10) rather than selective sampling. This approach integrates the elements of the regression model and the dasymetric method, and so it can be described as a hybrid model (Langford 2006).

$$P_s = \sum_{c=1}^{n} (d_c \cdot a_{sc}) \tag{10}$$

Overall, the conceptual working process of three-class dasymetric mapping is illustrated as follows:

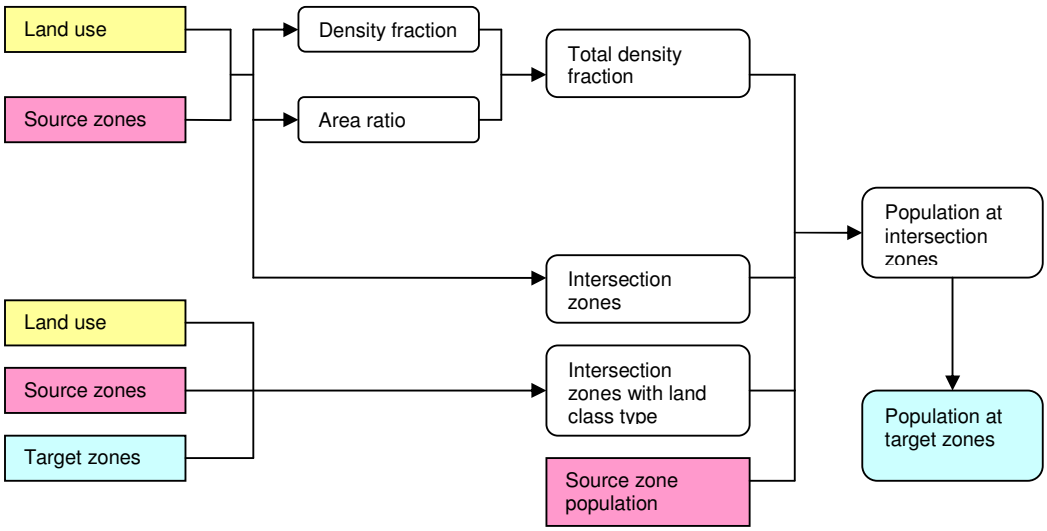


Figure 5 - A flowchart of the dasymetric mapping process

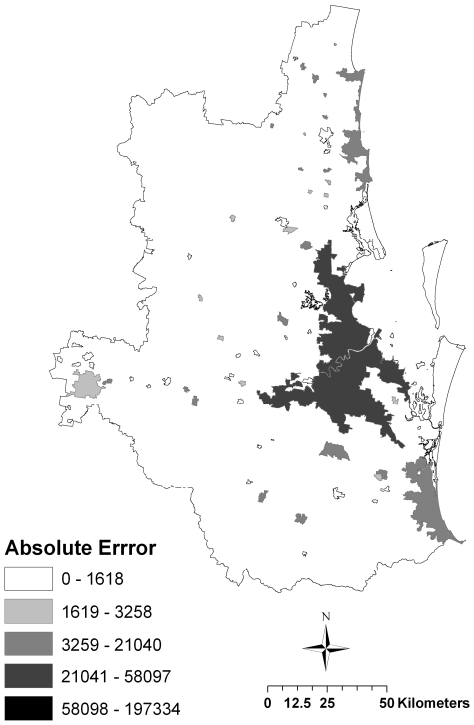
4. Results and discussion

As with previous studies, we evaluated errors for spatial disaggregation using actual values from independent data sources at the most refined spatial unit of population, namely the CCD (census collection district). However, without simulation of target zones the original CCDs are spatially contiguous with SLAs, which are not appropriate for testing the selected spatial disaggregation techniques.

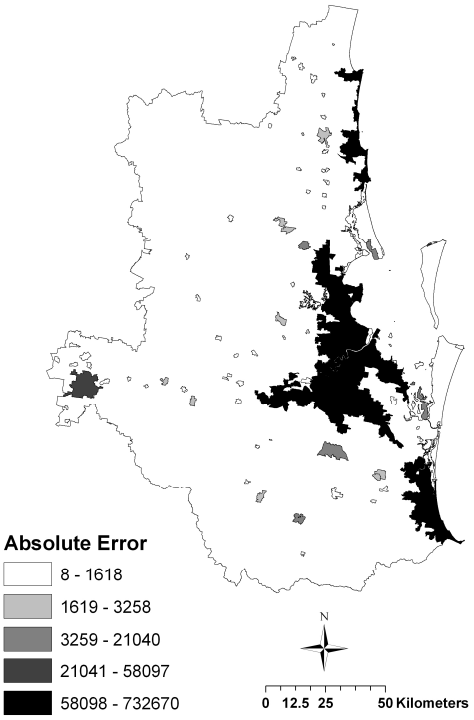
As an alternative, we make single comparisons using urban footprints (UCL) as target zones. The latter contain true population data at a smaller spatial unit than SLAs and so they may be used to check the performance of the different disaggregation methods. As the population count for urban footprints are reported at two density levels (high-density urban and low-density urban), the errors of each spatial disaggregation technique are only identified within urban areas, and error assessment within non-urban areas is not included.

In Figure 6 (a - d), the absolute errors of the disaggregated values generated by each technique are visualized at the target zones. Examination of this visualized error has been an efficient method for analysing the results (Eicher & Brewer 2001; Reibel 2005). A visual presentation of results provides a comparison against our knowledge of existing development and patterns of settlement change (Langford 1991).

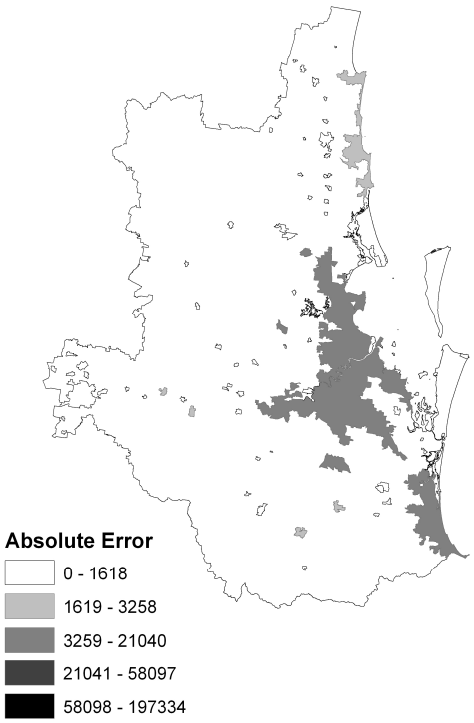
In each case the majority of errors are concentrated in the higher density urban areas, the rural areas generally having lower errors due to smaller populations within these zones. By comparing the error maps, we can distinguish the degree of errors between each technique for the same target locations. The three-class dasymetric mapping presents the lowest overall degree of error across the region. The regression model produced the highest error, especially in the high density urban area. The binary dasymetric mapping and the locally fitted regression approach gave a similar distribution of error.



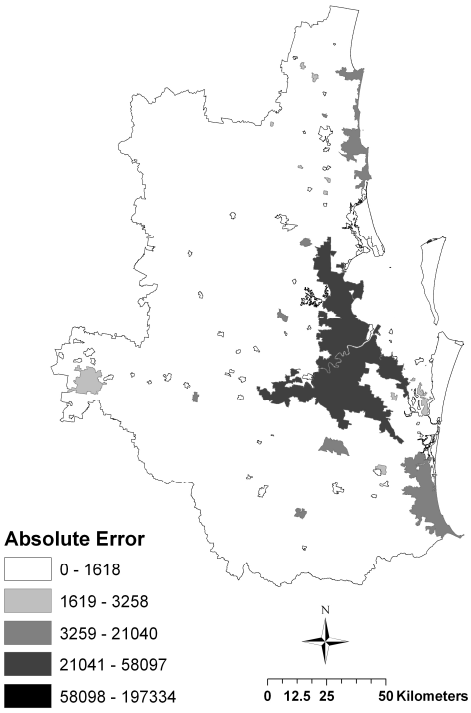
a. Error distribution of binary dasymetric mapping



b. Error distribution of the regression model



c. Error distribution of three-class dasymetric mapping



d. Error distribution of the locally fitted regression model

Figure 6 - The visualized absolute errors of four techniques that disaggregate population from SLAs to urban footprint polygons.

Table 3 summarizes the outputs and errors for each technique in detail. The errors of the results are measured by absolute error for each density category and root mean square errors (see equation 11) (Gregory 2002). The italicised figures in the table show the value of overestimation and the value of underestimation.

$$E^{RMS} = \left[\frac{1}{t} \sum_{t} \left(Y_t - \hat{Y}_t \right)^2 \right]^{1/2} \tag{11}$$

Techniques	Low density urban	High density urban	Non-urban	Sum of population	RMSE
Binary dasymetric mapping	92824 *+ 66017*	2376345 *+ 142400*	--	2469619	5775.1
Three-class dasymetric	27152 *+ 345*	2283716 *- 49771*	168290	2479158	2956.3
Global Regression	37187 *+ 10380*	1721474 *- 512471*	76415	1835076	49744.4
Locally fitted Regression	40567 *+ 13760*	2353053 *+ 119108*	85639	2479259	5029.5
UCL (True Value)	26807	2233945	--	2260752 (Target Zones)	
SLA				2479295 (Source Zones)	

Table 3 - A comparison table of the accuracies of different spatial disaggregation techniques (by absolute error and RMSE). The italicised figures represent the value of overestimation or underestimation.

From Table 3 we can see that the technique using constant, globally fitted parameters (density for each land class) is extremely unreliable for disaggregating population data for SEQ with its diverse densities. The overall RMSE (49774.4) of the regression model is surprisingly much higher than other techniques. This has not been revealed from previous studies based on simpler datasets.

The result showed significant underestimation of population in the high density urban area (-512,471), which means the natural bias of population concentration is basically not well addressed by the regression model. This result suggests that the regression model, being based on the constant density assumption, is an unreliable method for disaggregating spatial data within a complex geographical area.

The binary dasymetric mapping technique was also found unsatisfactory. This is evidenced by an overestimation for both low-density residential areas (+66,017) and high-density residential areas (+142,400), and there is no population left on the non-urban land. The main problem is the limitation of data for testing binary dasymetric mapping, because target zones are congruent with control zones and the non-urban polygons do not cross the boundaries of any populated areas. Based upon the method of binary dasymetric mapping, there was not any population in the populated area that got a chance to be allocated to the non-urban area.

From this case we can see that the application of binary dasymetric mapping will be constrained whenever binary land classifications are defined at target zones. Nevertheless, the overall accuracy of binary dasymetric mapping at target zones is better than that of the regression model that uses absolute classified densities (5775.1 is smaller than 49744.4 RMSE). These results are consistent with previous findings: a globally fitted regression model gives poorer accuracy than the locally fitted dasymetric method (Fisher & Langford 1995).

Also, the locally fitted regression model with variable densities has better accuracy than the regression model. Based on the improvement of the high-density population estimation, we conclud that calibrating the global density at each source zone is necessary when dealing with large areas that have complex density distributions. Also, this technique slightly outperforms binary dasymetric mapping (5029 is smaller than 5775 RMSE). This is plausible because the technique takes advantages of both a locally fitting approach and a multiple population density classification. However, its accuracy did not seem to be competitive with three-class dasymetric mapping that uses a more complex rescaling approach.

Three-class dasymetric mapping produced the least error in all density categories and overall RMSE (2956.3) compared to the other techniques. These improvements are because:

(a) SEQ is more complex than previous study areas;

(b) the estimate of initial density fractions, using a regression model, is better than a traditional sampling approach;

(c) the area-based scaling approach improved the ability of three-class dasymetric mapping.

Note that unlike the three-class dasymetric mapping test by Langford (2006), which used constant density fractions, we applied variable density fractions. This is more appropriate for addressing the spatial non-stationarity for each land use class over the space. The result is reasonable because a technique based on complex density assumptions will always be expected to be more accurate.

On the other hand, we can see that three-class dasymetric mapping produced more error in the high-density urban area than in the low-density urban area. This is possibly caused by the quality of control zones. The coarse high-density urban polygons did not subdivide the urban area into multiple densities, and the latter actually need to be specified because errors in land classification might mistakenly inform the disaggregation process. Improvement might be possible through refinement of ancillary data.

5. Conclusions and future research

This paper has examined the comparative accuracy of four methods for spatially disaggregating data in the context of modelling the population of SEQ. In this study, three-class dasymetric mapping based on the better assumption of homogeneous density in addition to incorporating detailed ancillary data was found to provide the most accurate result. The degree to which this technique was found to be superior is attributed to the fact that SEQ is a relatively large geographical area with a considerable range of population densities. Another finding is that, although based on the same level of density assumptions, the three-class dasymetric mapping using an area-based, locally fitted approach outperforms a locally fitted regression model that uses a density-based, locally fitting approach.

However, since the three-class dasymetric mapping is still based on the limited classification of residential densities, improvement could be made by incorporating multi-class or spatially continuous ancillary information. For example, the land class could be further divided into five or eight density categories. In addition, the application of a Monte Carlo simulation approach could be employed to randomly simulate different sets of target zones to generate multiple RMSEs based on various geographical situations. By evaluating the statistical distributions of errors, the hidden relationships between estimation accuracy and spatial factors (i.e. MAUP) could be further visualized and analysed.

Overall, the results indicate that the three-class dasymetric approach is the most appropriate technique for spatially disaggregating census-derived population data in SEQ. The ability to reliably generate spatially disaggregated estimates of population is of prime importance for inputting into urban modelling and policy development.

References

Brisbane institute. (2004) - The 200 km city, from Noosa to the Tweed. [Internet]. The 200km city exhibition resources. accessed 10 October 2005 - http//www.brisinst.org.au/resources/brisbane_institute_200kmcityindex.html

Bracken, I. (1989) - The generation of socioeconomic surfaces for public policymaking, *Environment & Planning B,* 16, 307-325

Cockings, S., Fisher, P. F., Langford, M. (1997) - Parameterization and visualisation of the errors in areal interpolation, *Geographical Analysis,* 29(4), 314-328

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

Dennis, E. J., Wu, J. (1996) - The Modifiable areal unit problem and implications for landscape ecology, *Landscape Ecology,* 11(30), 129-140

Eicher, C. L., Brewer, C. A. (2001) - Dasymetric mapping and areal interpolation, implementation and evaluation, *Cartography & Geographic Information Science,* 28(2), 125-138

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

Flowerdew, R., Green, M. (1991) - Data integration, statistical methods for transferring data between zonal systems, in I. Masser, I. & M. Blakemore (eds.) *Handling Geographical Information, Methodology & Potential Applications*, New York, Wiley, 38-54

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

Fotheringham, A. S., Rogerson, P. (1993) - GIS and spatial analytical problem, *International Journal of Geographical Information System*, 7(1), 3-19

Fotheringham, A. S., Wong, D.W. (1991) - The modifiable areal unit problem in multivariate statistical analysis, *Environment & Planning A,* 23, 1025-1044

Goodchild, M. F., Anselin, L., Deichmann, U. (1993) - A framework for the areal interpolation of socioeconomic data, *Environment & Planning A,* 25, 383-397

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

Gregory, I. N., Paul, S. EII. (2005) - Breaking the boundaries, geographical approaches to integrating 200 years of the census, *Journal of the Royal Statistic Society,* 168, 419-437

Hunt, J. D., Kriger, D.S., Miller, E.J. (2005) - Current operational urban land use - transportation modelling frameworks, a review, *Transport Reviews,* 25(3), 329-376

Lam, N. S. (1983) - Spatial interpolation methods, a review, *The American Cartographer,* 10, 129-149

Landis, G. (1994) - The California urban futures model, a new generation of metropolitan simulation models, *Environment & Planning B,* 21(4), 399-420

Langford, M., Maguire, D. J., Unwin, D. J. (1991) - The areal interpolation problem, estimating population using remote sensing in a GIS framework, in Masser, I. (ed.)

Handling Geographical Information, Methodology and Potential Applications, Michael Blakemore, New York, Wiley, 55-77

Langford, M., Unwin, D. J. (1994) - Generating and mapping population density surfaces within a GIS, *The Cartographic Journal,* 31, 21-26

Langford, M., Fisher, P. F. (1996) - Modelling sensitivity to accuracy in classification imagery, a study of areal interpolation by dasymetric mapping, *Professional Geographer,* 48(3), 299-309

Langford, M. (2006) - Obtaining population estimations in non-census reporting zones, An evaluation of the three-class dasymetric method, *Computers, Environment & Urban Systems,* 30, 161-180

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

Moon, Z. K., Farmer, F. L. (2001) - Population density surface, a new approach to an old problem, *Society & Natural Resources,* 14, 39-49

Reibel, M. (2005) - Street-weighted interpolation techniques for demographic counts estimation in incompatible zone systems, *Environment & Planning A,* 37, 127-139

Reibel, M., Aditya, A. (2006) - Land use weighted areal interpolation, *GIS Planet 2005 International Conference*, Estoril, Portugal

Rosenbaum, A. S., Koenig, B. E. (1997) - Evaluation of modelling tools for assessing land use policies and strategies, San Rafael, California, Systems Application International, Inc.

Sadahiro, Y. (1999) - Accuracy of areal interpolation, a comparison of alternative methods, *International Journal of Geographical Information Science*, 1, 323-346

Tobler, W. R. (1979) - Smooth pycnophylactic interpolation for geographical regions, *Journal of American Statistical Association,* 74(367), 519-530

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