

Population Estimates for Non-standard Geographical Areas. The Small Area Pilot Project

Alan Jenner

Australian Bureau of Statistics
Economic and Population Statistics Unit
GPO Box 796
Sydney, NSW 1041
Phone: 02 9268 4113 Fax: 02 9268 4805
Email: a.jenner@abs.gov.au

Introduction

For a number of years the Australian Bureau of Statistics (ABS) has provided a significant amount of census data in Geographic Information System (GIS) ready format. However, the spatial units upon which the data is based has not always been conducive to user needs, due to its reliance on standard geography. With the increasing interest in census-based data in other geographic forms, the NSW office of the ABS instigated the Small Area Pilot Project. This project explored the feasibility of producing population and other statistics from census data for geographical areas other than those normally provided by the ABS.

Information about the spatial distribution of population is widely used for major decisions in areas such as health, transportation, retail and urban planning. The present provision of this data through the ABS uses standard geography whilst the needs of government and non-government agencies often digresses from these standards towards areal units. Whilst ABS geography is based on the Collection District (CD), a unit small enough for a census collector to reasonably distribute and gather census information, user data needs often require information below that level or transecting across CD boundaries. Even when the user's area of interest is much larger than a CD, a simple aggregation of CDs may not always meet the user's requirements. The resulting CD derived area may not represent the user's requirement accurately enough in terms of either population or area. The CD derived area may for example exclude some topographical feature which is included in the user's area or vice versa.

This project therefore examined GIS modeling techniques which could produce population estimates at sub-CD level and which were more conducive to the myriad of requested non-standard geographies. A number of techniques were initially considered. The project though found that the most effective method for producing the required outcome was a combination of point interpolation and areal weighting combined with extensive non-population shadowing. The basic premises of this form of modeling did not provide a complete solution to the problem of sub-spatial heterogeneity. Nor did it completely remove the problem of non-population areas affecting the population distribution. It did however indicate that the production of small area population statistic estimates from the ABS 2001 census was feasible if a final methodology and accuracy level could be defined and accepted.

To Areal or Not to Areal - Methodological Considerations

GIS systems have evolved rapidly over the last 20 years. From their basis as systems developed as tools for the storage, retrieval and display of geographic information, current GIS systems now commonly include a range of spatial summarisation and analysis tools. Despite this, official census population statistics in Australia has not fully utilised these tools. Instead it has maintained the standardised system of irregular vector polygon spatial units based on the Collection District (CD).

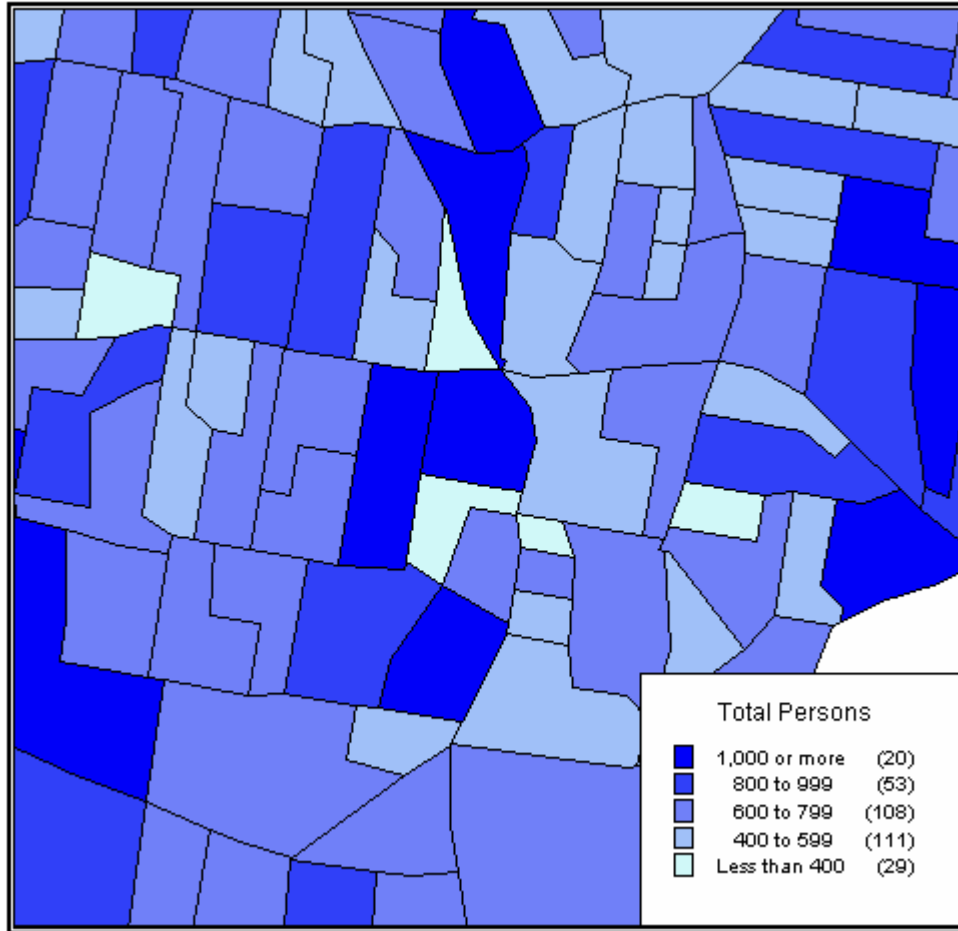
Today, GIS systems include a number of spatial modeling applications. Spatial models were one of the early 'add-ons' to basic GIS technology. Though they were initially created for the 'earth' sciences, more recent applications have seen a growth in use within the social sciences. Population modeling and the production of population-based statistics are one such application. As Bracken (1994 pp 249-50) suggested, "the representation of population and population-related phenomena can be best done using a structure which is independent of actual spatial enumeration. One way in which this can be achieved is to transform zone-based census data into a data structure that represents population more plausible as something approaching a continuous surface". This has already started to become a common practice overseas with examples such as the Surpop V2.0 program already operating within national statistical offices. This program produces small area estimates of population and other social statistics based on a 200x200 metre grid for the entire United Kingdom (<http://census.ac.uk/cdu/surpop/>). This project therefore approached its research task with a specific outcome in mind. Could the same type of statistics be produced using Australian data and if not, what were the problems encountered in producing similar estimates? As well as providing a solution to user demands, the project's outcomes would, as Fotheringham and Rogerson (1994 p 4) suggested, "lead to an improved understanding both of the attributes being examined and of the procedures used to examine it".

The Base Data

The area chosen for the initial analysis within this project was the combined Statistical Local Areas (SLAs) of Auburn and Bankstown within the Sydney Statistical Division. Census population counts and geographic metadata for 1996 was extracted for the 729 Collection Districts (CDs) within and surrounding these SLAs using CData96. CData96¹ is a commercial sub-system which nestles within MapInfo and provides access to and display of 1996 census data. From CData, both macrodata and GIS metadata can be extracted for vector polygons within any of the geographies available through the Australian Standard Geographic Classification (ASGC). This includes data at the CD level. This data was in the form of a "mappable" MapInfo table and was able to be displayed as a thematic style representation of population distribution (Figure 1).

¹ This 1996 Census based product will be updated to the newer CData2001 with the release of 2001 census data in late 2002.

Figure 1: CD Population Vector Map for Auburn and Bankstown SLAs



CDs are the smallest available units within ABS standard geography. They are enumeration areas based on criteria revolving around the ability of a census collector to visit all dwellings within a 10 day period. Though these criteria have limits in the number of dwellings and persons present within any individual CD, the application of these criteria in pre-census design produces a heterogeneous rather than homogenous matrix of data. This can be further complicated by the time lag between CD design and census collection where large changes can occur in the built environment and population of many areas. Within the analysis area these factors produced a range of population values from 9 through to 1269 persons (excluding the 11 zero population CDs in the data set).

A number of ancillary sources of spatial data were also obtained to use as identification data and "masks" for non-population areas. These included a cadastre layer, road and railway, parks and reserves, national park, waterway, special land use data, and business and industry land layers. Some of this data was extracted from the CData96 Detailed Base Map or the Master Spatial Database provided by PSMA Australia. Other data was provided to the ABS by the NSW Department of Information Technology and Management, and Planning NSW. Though some of the CData96 map data was in polygon form, other data (e.g. airports) were in point form and had to be used to manually produce polygon layers based on cadastral parcel limits. Sydney Water also provided geocoded address points which were used for benchmarking purposes in the final evaluation phase of the project.

A second data set was created encompassing CDs within the entire Sydney Statistical Division. This data set, which included the test area, was used to examine overall trends within the data and was tested post analysis to evaluate procedural performance in larger data sets and to evaluate other possible modeling problems.

Methodology

Population modeling using GIS technology has become a major research area in recent years. So too has the growth of methods applicable to this social parameter. Nina Siu-Ngan Lam (1983 pp 129-49), in a 1983 review on spatial interpolation models, indicated a general split between two broad areas of these models, point and areal interpolation, with the selection of model strongly dependent on the type of data and required accuracy. One stream of areal models, those that are "volume-preserving", appear to have made significant inroads into population modeling.

One methodological solution considered was to utilise an existing, or produce a variant, of these available areal models. A number of these variants have been forwarded within GIS literature (e.g. Flowerdew et al 1991, Flowerdew and Green 1994, Ungerer and Goodchild, 2002). These methods employ specific algorithms which take into account the Boolean relationship between source and target polygons. Data within this process is reconfigured using weighting based on volume overlap to provide an estimate within the target zone/s. However, the results of this type of process have not always been considered accurate without the application of intelligent methods using ancillary data or by using small zone source data (Sadahiro, 2000, pp 25-26). This idea of "more intelligent" areal weighting includes the use of ancillary data such as non-population areas. Additionally, the data type and distributions within the data points also indicates model variations. Flowerdew and Green (1994, pp 126-139) proposed variants of their basic model for Poisson, bi-nomial and normal distributions, each relating to different population and housing statistics.

Alternate methods to this form of areal interpolation do exist including those which approach Bracken's suggestion of a population surface independent of standard spatial units. Point interpolation models are already available in most GIS systems. Based on early geological modeling, these methods take sample points and convert them to continuous surfaces using distance decay algorithms. The main point of contention within these models is their applicability to population based parameters and statistics with critique of the application of these models to social statistics another theme in GIS literature (e.g. Batty & Xie, 1994 & Bracken 1994). Notwithstanding these points and critiques, a number of models based on point interpolation have been used in population statistics modeling. Most do not solely rely on point interpolation but include both the volume-preserving character and weighting process of areal interpolation. David Martin (1989 pp 90-97) suggested one such point method, before extending his research to produce the SurPop system for UK census small area estimates.

Data Considerations and Assumptions

There were a number of data problems which needed to be addressed prior to the process initiation. These relate to the form of the data, the types of systems being used and the quality of the data being used. Additionally, prior to modeling, the appropriateness of the modeling algorithm to the data needed to be tested.

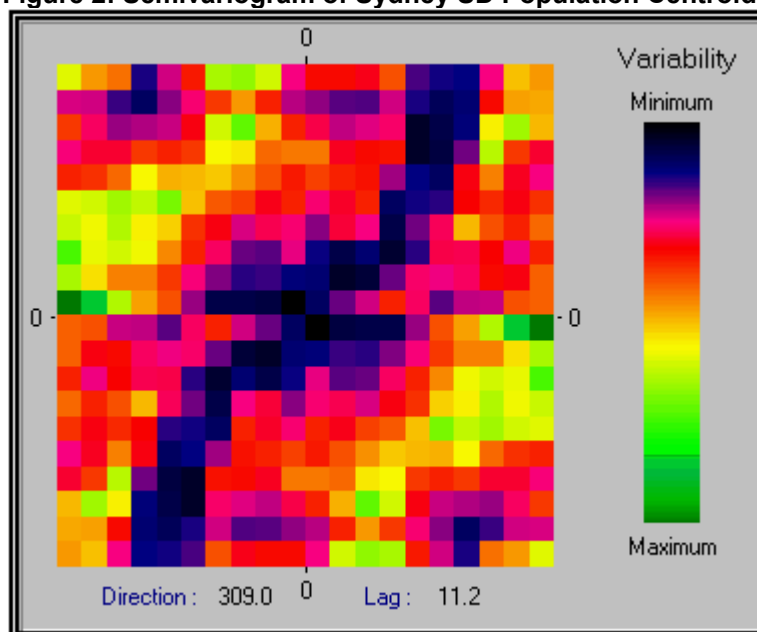
As previously stated, data was obtained in the form of CD based vector polygons as a MapInfo table. Though this form of data is conducive to target/zone areal interpolation methods, surface modeling requires data in point form. Therefore, to create a population surface model, the data needed to be converted from vector to point format. This was achieved within the MapInfo program using SQL standard operations and commands. This creates another form of the table where population values were assigned to generated centroid points. MapInfo creates these centroid points to reflect the mid-point of the vertical and horizontal extremes of any polygon. Where that point lies outside the initial polygon, as can occur in horseshoe shaped polygons, the centroid is moved to the nearest point on the polygon boundary. It should be noted that this process does not necessarily indicate the major population node within the polygon simply an arbitrary central point.

MapInfo has limited capacity to develop surface models or utilise three dimensional surface overlay procedures, but does have a readily accessible file exchange format (MIF) allowing between system data transfer. As MapInfo was to be the main platform for the desired output, a fully functional GIS system with a capacity to read MapInfo data was required. This secondary system was the Idrisi32 program which can read data from 'mif' files, the standard MapInfo import/export format. The choice of this second system was again arbitrary and reflects only the necessity of the research as it is considered that any GIS system with similar modeling capacities could be used in this process.

With data in formats which could be used in analysis, the next problem was to determine whether the data was suitable for the method being proposed. This is particularly important when looking at which algorithm/method is appropriate to interpret a generated surface. Central place theory is a well excepted tenet which indicates the effect of distance decay on population density across urban areas. Using this basic tenet, distance decay algorithms have been used to model population data. However, the modeling process here was based on population magnitude rather than density values and therefore the applicability needed to be examined. Point interpolation models use proximity and distance decay functions to generate a continuous surface from sample points. The point estimations created rely on a relationship between local pairs of observations as well as between local and distant pairs with the level of effect of distance mediated within the modeling process.

To investigate the relationship between data points, Idrisi32's Spatial Dependence Modeller was used. It produced a semivariogram of least squared residuals for distances out to 10 lags from all observations for all CD observations in the Sydney SD. The variogram cloud displayed in Figure 2 is the mapped outcome of the residual values between all pairs of observations taking into account direction and distance. For the Sydney SD, it suggested some consistency to the distance decay hypothesis among local pairs of observations with minimum variability in the centre and increased variability extending outward. There was though a strong suggestion of a pattern of variability not consistent with the standard distance decay hypothesis along the NE/SW axis and towards the NW/SE extremities. From this it was concluded that the use of a standard distance decay algorithm was plausible as long as that function was restricted by distance to local pairs (1-2 lags) within the modeling process.

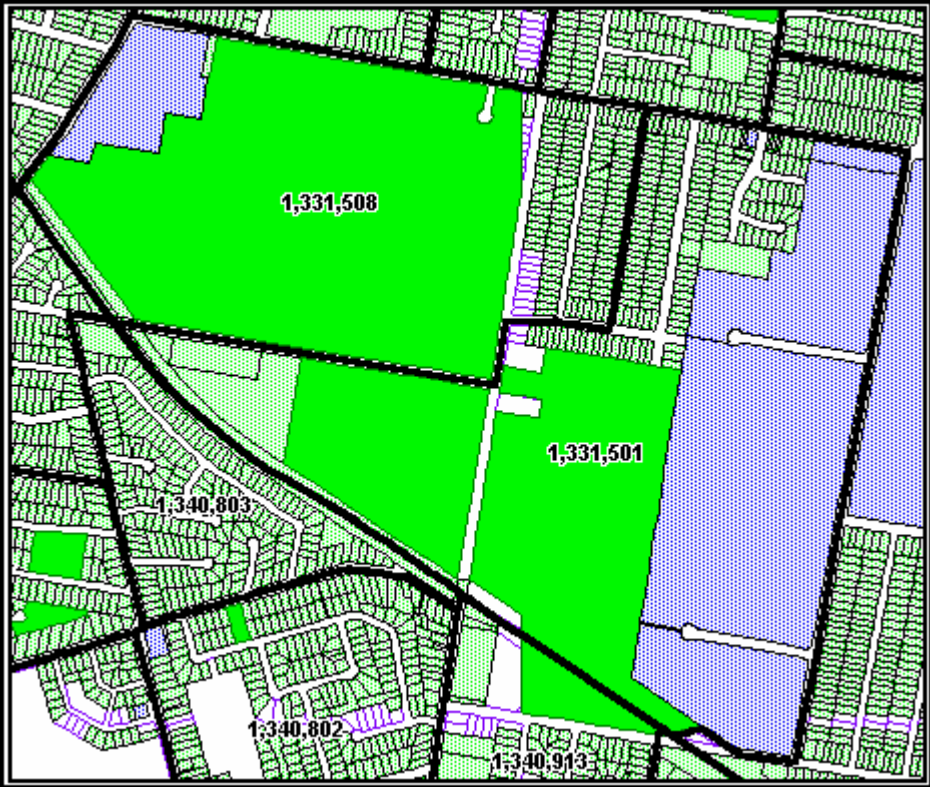
Figure 2: Semivariogram of Sydney SD Population Centroids.



Interrelationship between CD populations do not necessarily guarantee accurate estimates within the CD boundary. This is due to the heterogeneous nature of the population distribution within individual CDs. Though some CDs have relatively homogenous housing characteristics and therefore population densities across their area, the majority are extremely heterogeneous. Within their boundaries they can include a myriad of land uses which are non-population or shared population zones. These include open space, business areas, industrial areas, farms, education institutions and many other special use areas such as cemeteries.

As indicated in Figure 3, CD number 1,331,501 has extensive internal non-population areas including business/industry land (dotted) and park land (shaded) with the actual population confined mainly to the western side of the northern part of the CD. CDs can also include mixes of low, medium and high-density housing. It is possible to partially exclude known non-population areas if spatial data is available or can be reasonable produced. However, to account for business/residential and low/medium/high density housing mixes through weighting is not feasible at this time due to limitations within data such as the NSW cadastre. Thus, although much can be done to remove the population from known non-population zones, any estimate within a source CD boundary is apt to have some level of error associated with it.

Figure 3: Sub CD Land Use



The final consideration revolves around the optimum output spatial units. The output from the research needed to be in the form of a continuous layer of values across the entire area. This form of gridding needs to represent spatial units small enough to be both useful and reliable. Work in the UK indicated a grid of 200x200 metre squares as the optimum and this was used for the output from the SurPop program. Bracken (1994 p250) used a 100 meter approximation when he worked on population related social indicators. However, both of these considered grid sizes were related to UK based statistics and their Enumeration Districts which provide a different original surface based on the small landmass in the United Kingdom. Within Australia, urban centres and localities contain a similar structure. This is not necessarily true so outside these areas where CDs can encompass large spatial tracts to which these smaller grids could be incompatible.

Selected Methodology

Taking into account the modeling problems and the data considerations, the selection between either point or areal types of modeling could be considered problematic. Therefore, it was decided to focus first on producing an output similar to that in David Martin's SurPop method. This selection was not made purely on the basis of suggested model accuracy, as both areal and point methods can have similar problems in this area. This method selection was weighted towards which methodological output would maximise utility. Thus the method chosen involved producing a population based surface model then applying areal weighting and volume-preserving measures to the surface to produce a point layer of population estimates. This layer could then be utilised to provide estimates for any defined spatial unit.

The method employed involved a number of steps. The first step was to model a surface based on the population count data. This was achieved using the point interpolation module in Idrisi32. All points, including zero CDs were used in the model. Due to their effect on slope within the surface model, removal of zero CDs was initially considered. This would remove the tendency for the surface model to decline towards features such as waterways, which may not necessarily be a true representation of the population distribution. This assumption however is not consistent. Thus, without individual consideration for every zero CD point based on local knowledge, simple removal could corrupt the model. Therefore, all zero population points were considered valid and used in this analysis. This modeling produced a continuous surface of values covering the research area and surrounds. This model was not an end in itself as it represented the population based on totals for CD areas without reference to the CD area. This macro level slope surface was the used as a conduit for the areal interpolation of estimated values by providing intermediate value between centroid points which could be scaled and distributed across original CD areas.

The second step involved the production of a blank grid surface onto which the values were calculated. Initially, the 200x200 metre grid was chosen and this was relatively effective for the research area. However, post testing on the entire Sydney SD indicated a flaw with this grid where a large number of small CDs were removed entirely from the analysis. This occurred when a cell area encompassed more than one original centroid point. In creating the scale factor, primacy was given to the larger CD and the small CD was thus dropped out of the grid. At the 100x100 metre grid level, the loss of CDs within the Sydney SD was reduced to less than ten. From this it was concluded that the 200x200 metre grid may need to be replaced by a 100x100 metre grid for urban centres, with both grids produced and tested for the project. The small number of "lost CDs" could then feasibly be added back into the final product as single points at the end of the estimation process. At the other end of the CD scale another problem occurred when apply the grids to non-urban areas. Within these areas, as indicated by large CDs on the urban fringe, the 200x200 metre grid produced estimates of extremely low values. As land use data was restrictive in that, for example, large farm tracts could not be identified, a coarser grid became an option and thus a 1x1 kilometre grid was suggested as more appropriate but at this stage had not been tested.

The third step involved removing all non-population areas from the population surface produced. This was achieved by setting all grid values representing non-population areas to a value of zero and overlaying the surface and non-population maps. The combined use of these two layers and the original CD map could then be applied to produce a scaling factor to apply to the original surface map in order to spread the population values across all cells within the CD where population could exist. The scaling factor was based on the modified surface map which reduces its dispersion to only those areas within a CD where population can exist. This also produces a scaling factor of zero for any indicated non-population area. The scaling factor for each cell was therefore defined as:

$$\text{Scale Factor cell } x = \frac{N_{CD}}{\sum CD_{PS}}$$

Where N_{CD} is the original population of the CD containing cell x

and $\sum CD_{PS}$ is the summed population surface values of the CD containing cell x .

The scaling factor can then be applied to the original population surface values to distribute these values across the available cells within each CD. This produces a new matrix of cell values representing estimated population in each cell for the entire area under investigation with the exception of non-population areas. The initial estimated population for each cell is thus defined as:

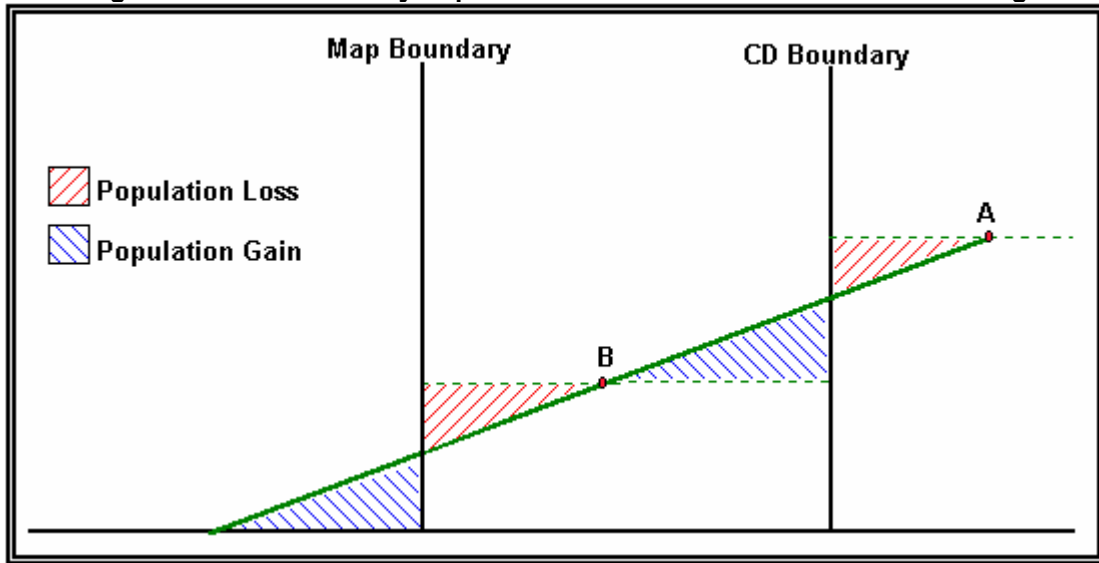
$$\text{Initial Population Estimate cell } x = X_{SF} * X_{PS}$$

Where X_{SF} is the scale factor for cell x

and X_{PS} is the population surface values of cell x .

At this stage the produced population estimates maps (both 100 and 200 metre versions) were converted into point form and transferred to MapInfo for further analysis and interpretation. A number of quality checks were instigated to see how well the model compared with the original data. Cross checking indicated that during this initial estimation process some of the original CDs had gained or lost parts of their total population. This was caused when the population surface creates a slope between centroid points with large differences or unequal boundary distances, or where surface slopes increase or decrease external of the aggregate CD boundaries (Figure 4). At the individual point level, the amount of gain/loss was also mediated by the number of boundary points the CD had. Thus a large amount of population could be leached across a boundary by small losses/gains from a large number of points as well as the possibility of a single point losing a large part of its population when it represented the only cell within a CD.

Figure 4: Cross Boundary Population Gain/Loss in Point Method Modeling



Thus, in order to maintain the volume-preserving component of the required areal method a final step was thus instigated. This involved calculating the residual difference between the original CD population value and the aggregated population estimates within the CD. This residual value was then re-distributed across the cells within that CD to produce the final estimate values which still maintained the slope function of the original estimates. The final population estimate produced for each cell was therefore defined as:

$$\text{Final Population Estimate cell } x = X_{EST} + \left(\left(\frac{X_{EST}}{\sum CD_{PS}} \right) * CD_{RES} \right)$$

Where X_{EST} is the initial population estimate for cell x ,

$\sum CD_{PS}$ is the summed population surface values of the CD containing cell x

and CD_{RES} is the residual difference between $\sum CD_{PS}$ and N_{CD} .

Application within the Small Area Pilot Project

The method used in this project was able to deal with a number of the problems of areal interpolation to produce a reasonable population estimate at sub-CD levels. By using the best available spatial data to enhance its "intelligent" areal capacity it was able to successfully remove most parks and major non-population areas such as industry land from the final estimates layer. The method was applied to produce two versions of the estimated population layer based on 100 and 200 metre grids. Both layers were transferred to MapInfo in the form of a point layer with estimated values attached to the centre point of each grid. Grid cells containing values of zero were not included in the point file. The initial grid layers were then adjusted for cross boundary gain/loss to produce the final estimated population grid. Thus, from the original data containing 721 CD centroid values for 442,502 persons in the analysis area, the 200 metre grid process initially returned 6,075 points representing 437,881 persons, whilst the 100 metre grid returned 24,374 points representing 442,502 persons (Table 1).

Table 1: Data Set Parameters.

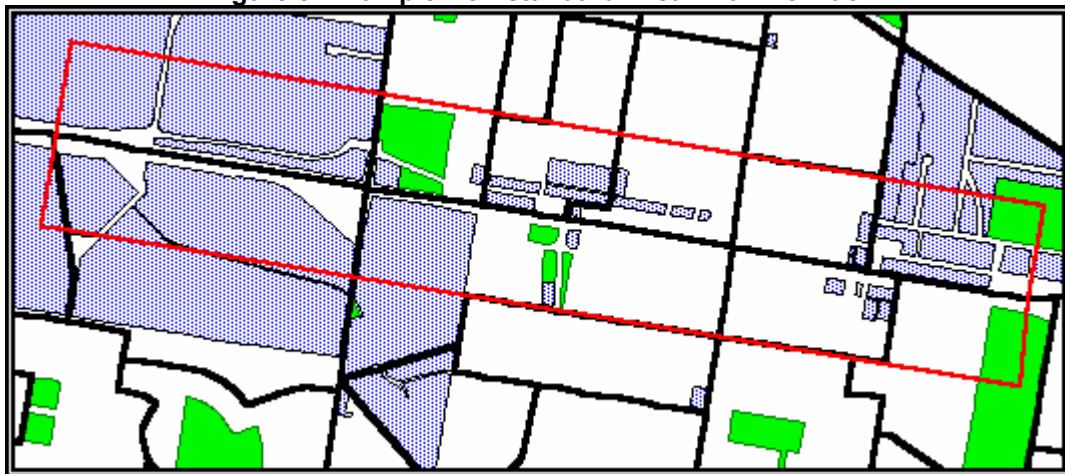
Data Set	Cases	Persons	Minimum Value	Maximum Value	Notes
Original CDs	729 CDs	442502	0	1269	11 Zero CDs
Auburn/Bankstown SLA CDs	321 CDs	208003	0	1269	1 Zero CD
200m Grid Initial	6075 Points	437881	0.007	814	
100m Grid Initial	24374 Points	442502	0.001	475	
200m Grid Adjusted	6086 Points	442502	0.007	814	11 CDs points added
100m Grid Adjusted	24374 Points	442502	0	475	2 Points set to Zero

At first glance, the loss of 4,621 persons within the 200 metre grid could have been considered a problem. However, investigation of the grid indicated that the lost population was represented by 11 CDs which had not been included in the grid analysis due to their small size and shape. Slight movements in the grid coordinates may have forced the inclusion of these 11 CD points but could have excluded other points thus not solving the problem. Therefore, it was decided to add the 11 CD points to the final 200 metre point file as extra data points after the adjustment for population gain/loss.

The initial estimates were compared with the initial CD populations for gains/losses. Within the 100 metre grid, which included data for all initial CDs, the majority of CDs lost or gained less than 1 person (676 CDs) with the maximum gain being 108.3 persons and a corresponding maximum loss of 108.3 persons. However, the 200 metre grid had a maximum gain of 147.3 persons with a maximum loss of 123.1 persons even though more CDs lost/gained less than 1 person (709). The major gains/losses were between adjoining CDs where the population gradient was high. Smaller levels of leakage also occurred across all other boundaries and across external limits of the aggregate CDs.

Once the final estimate layer was produced an example non-standard geographic area was constructed representing a 400 metre wide rail corridor within the research area (Figure 5). The railway line on which this area was based run down the centre of the area along the CD boundaries (black solid line). This area included residential (non-shaded) and commercial areas (blue dotted) as well as parks and reserves (green solid) within its boundaries. The corridor cut across 9 CDs and enclosed 2 CDs. One of these 11 CDs contained only a minor proportion within the corridor and thus was excluded from the results to reduce external bias. The total population from the CD based data (10 CDs) was therefore 5,267 people which would have been a reasonable figure extracted using standard geography for a request of this nature. By comparison, the 200 metre grid produced an estimate of 3289 people whilst the 100 metre grid produced an estimate of 3019 people.

Figure 5: Example Non-standard Area - Rail Corridor



Though both the estimates could be considered reasonable based on the spatial data presented, without extensive fieldwork, the possibility of testing their accuracy is limited. As this type of fieldwork was not viable, other data sources were investigated to see if any concordance could be viewed. Sydney Water generously provided the ABS with a geocoded address point data set which, in the absence of extensive fieldwork, provided some scope for a comparative analysis. The data included all connections within the Sydney region but these could not be easily separated between residential and non-residential connections. Additionally, the data was for 2001 and, in comparison to 1996 dwellings data, would include a number of significant urban developments within the research area. These would tend to under-count the 1996 figures when comparisons were made. Even after considering the data limitations, it was concluded that the Sydney Water data was the only available source which was fit for the purpose of comparison. Table 2 lists the comparative tests conducted at both the CD and Grid level.

Table 2: Comparisons of Data Sets with Sydney Water Connection Data.

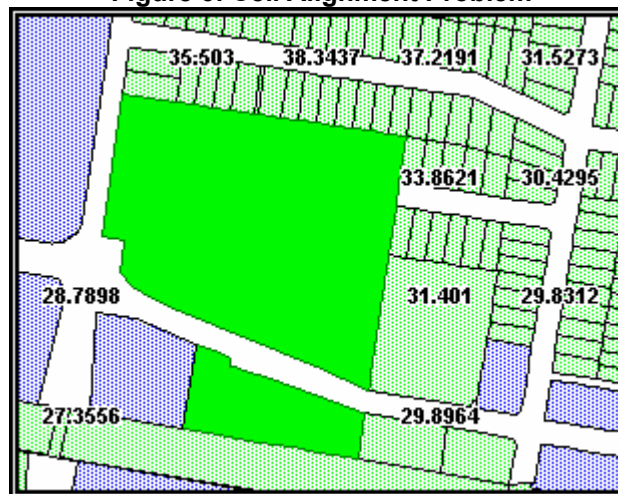
Data Set	Cases	Min	Max	Mean	Variation	Standard Deviation	T Value CD Base Data
Base Data (Persons Per Private Dwelling)							
CDs	321	0.0	176.3	3.7	115.1	10.7	
Grid Data Original (Estimated Persons Per Water Board Connection)							
100m	26581	0.0	475.0	1.8	22.6	4.8	-13.73
200m	7112	0.0	196.2	2.1	43.9	6.6	-3.07
Grid Data Outliers and Zero CDs Removed (Estimated Persons Per Water Board Connection)							
100m	5651	0.0	475.0	2.9	53.5	7.3	-1.14
200m	1417	0.0	166.6	3.0	37.2	6.1	-0.75
Grid Data Outliers and Zero CDs Removed, Zero Grid Cells Removed (Estimated Persons Per Water Board Connection)							
100m	4259	0.2	475.0	3.8	67.3	8.2	0.10
200m	1114	0.5	166.6	3.8	44.3	6.7	0.05

In order to compare the different data, the number of persons per private dwelling was calculated within the original CDs for the 2 SLAs of Auburn and Bankstown. The figure of 3.7 persons per private dwelling was then used as a population mean which could be compared to a similar figure within the grid data. To calculate a similar figure, a comparative figure for private dwellings was needed. This figure was obtained by counting the number of connections within each grid cell. Due to the difference in time and the over-count caused by business and industry connections, investigations of outlier changes in the original CD pattern were instigated. From this investigation, a number of CDs were identified as having large differences between the number of private dwellings in 1996 and the number of connections in 2001. The largest difference was 1,205 more connections, which was associated with development in the Sydney Olympic Site. It was therefore decided to compare only those areas where these differences were not greater than 50. Thus, within the grids, points which were within 83 of the 321 initial CDs were separated from further analysis. Additionally, the number of cells containing zero values also needed to be removed to ensure that data was only calculated on valid cells. This altered the original 100 and 200 metre grid means for estimated persons per connection from 1.8 and 2.1 respectively to 3.8 for both. The concordance with the original data set was calculated using a T test and this indicated that the means were similar. However, the variance of the distribution of values had decreased from 115.1 at the CD level to 67.3 and 44.3 for the 100 and 200 metre grids respectively.

Within the test non-standard area (rail corridor), of the 1388 connections, 358 were within the areas indicated as business or industry land. This suggests that around 1030 connections would be considered as mainly residential connections. With a population mean of 3.7 persons per dwelling this would suggest that there were 3811 people in the area. Considering the accuracy of the data layers being used and the difference in years, this value, though higher than the 3289 obtained from the 100 metre grid, provides some support for the estimates validity.

One final problem was noted which could effect estimates obtained from the proposed output. In a number of cases, data points were produced in areas where no population should have existed. These points were located on roads between areas designated as non-population zones. The reason this occurred relates to both the accuracy and inaccuracy of the non-population data. This occurred during the rasterisation of non-population vector data. As the data values are transferred onto a grided image, a grid cell receives the value if at least 50% of its area is covered by the vector to which the value relates. This means that for the 200 metre grid, some smaller non-population areas as well as areas where major roads/highways exist were made available for process data (Figure 6). On the 100 metre grid, more non-population areas are removed but, in this case, smaller roads can become areas where data points could exist. This occurred when the business and industry data was used as the vectors within this data set were extremely detailed rather than zonal. This meant that a vector often excluded the streetscape between/within areas allowing cells to be generated within these areas.

Figure 6: Cell Alignment Problem



Modeling Population Parameter and Statistics in Australia

As shown within the Small Area Pilot Project, population modeling of this type can readily be achieved using GIS procedures for data within Australia. However, it is also apparent that there are many problems which could cause errors in the output from such models. The data shown here at least supports the modeling done overseas and suggests that limited distance decay algorithms are applicable to Australian data. The use of such algorithms may not however be appropriate for other social statistics modeling and this requires further investigation. The algorithm must also be used in conjunction with a set grid size. Within urban areas, there is no clear benefit to using either a 200m or a 100m grid with both indicating some form of problem within the process or from the available data. The 200m grid used in the United Kingdom may not be as useful in Australia due to the size of urban periphery and rural CDs, a larger and more robust 1000 metre grid appears more plausible for rural Australia. Further research will need to be conducted to validate these points.

It should also be noted that without significant non-population data, particularly within urban centres, spatial estimation errors would occur. As shown in this research, even with accurate spatial data from multiple sources, inappropriate grid points can be generated. With a lack of available data in NSW, particularly zone based map data on business and industry land use as well as map data on residential zoning, accurate intelligent areal modeling would be restricted to areas where maximum data is available (e.g. Sydney) or by the ability of individuals to digitise data from non-electronic forms.

It is also suggested that alternative spatial units may provide better results. For example, the use of units which are smaller than a CD like mesh blocks (as occurs in New Zealand) or small postcode unit (as occurs in the United Kingdom) may provide one alternate method to future data needs. The possibility of such spatial changes is already under investigation within the ABS.

Thus although the method and project indicated a level of success in that reasonable estimates could be made at both sub and cross CD spatial areas, it is not conclusive that the errors contained within the provided estimates outweigh the advantages of this data form. It will therefore be necessary for extensive peer review and consultation to take place before any decision is made on whether the ABS should adopt sub-CD population modeling. This will include further investigations of appropriate algorithms, grids and non-population data to ensure that any final product maintains the high standards of all ABS data sources.

Note: Views expressed in this paper are those of the author and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted or used, they should be attributed clearly to the author.

References:

- Batty M. & Xie Y., *Modelling inside GIS: Part 1. Model structures, exploratory spatial data analysis and aggregation*, International Journal of Geographic Information Science, 8:3, 1994 pp 291-307.
- Bracken I. *A surface model approach to the representation of population-related social indicators*, in *Spatial Analysis and GIS*, Fotheringham S. and Rogerson P. ed., Taylor and Francis, London, 1994.
- Flowerdew R., Green M. & Kehris E., *Using Areal Interpolation Methods in Geographic Information Systems*, PAPERS IN REGIONAL SCIENCE: The Journal of the RSAI, 70:3, 1991 pp 303-15.
- Flowerdew R. & Green M., *Areal Interpolation and types of data*, in *Spatial Analysis and GIS*, Fotheringham S. and Rogerson P. ed., Taylor and Francis, London, 1994.
- Fotheringham S. & Rogerson P., *GIS and spatial analysis: introduction and overview*, in *Spatial Analysis and GIS*, Fotheringham S. and Rogerson P. ed., Taylor and Francis, London, 1994.
- Martin D., *Mapping population data from zone centroid locations*, Transactions of the Institute of British Geographers, 14, 1989, pp 90-97.
- Sadahiro Y., *Accuracy of count data transferred through the areal weighting interpolation method*, International Journal of Geographic Information Science, 14:1, 2000 pp 25-50.
- Siu-Ngan Lam N., *Spatial Interpolation Methods: A Review*, The American Cartographer, 10:2, 1983 pp129-49.
- Ungerer M.J. & Goodchild M.F., *Integrating spatial data analysis and GIS: a new implementation using the Component Object Model (COM)*, International Journal of Geographic Information Science, 16:1, 2002 pp 41-53.