The Brazilian Population Grid: a hybrid approach

Maria do Carmo BUENO
IBGE, Brazil, mbueno@ibge.gov.br
David MARTIN
University of Southampton, UK
Alvaro D'ANTONA
University of Campinas, Brazil

ABSTRACT

There is a growing interest in the generation of statistical population grids due to their stability through time and ease of integration with different spatial data sources. The conventional means of producing these estimates may be divided into bottom-up (aggregation) and top-down (disaggregation) approaches, depending on the national data environment. This paper introduces a hybrid model proposed for creation of a population grid for Brazil by the national statistical agency using data from Census 2010. It has been necessary to develop this novel hybrid methodology due to the diverse data environments found in urban and rural settings. Two regions in the state of Para were selected as study areas to evaluate several methods in order to find the most suitable combination. Results of both the aggregation and disaggregation methods are available for the study region, making possible comparison of the results obtained using different methods. The analysis suggests that each set of conditions requires a different treatment and highlights the importance of good metadata. The insights gained from this analysis have potential application in many countries facing similar data challenges.

KEYWORDS population grid, Brazil, aggregation, disaggregation

INTRODUCTION

A Census produces essential information for national, state and municipal policy-making, including service planning (educational, health, social and utilities), emergency services related to natural disasters and numerous analyses (poverty, labour force, marketing, epidemiological). The growth of geospatial technologies has led to much wider use of this statistical information, increasing the demand for detailed and geographically disaggregated information. Although these types of data are traditionally produced for

irregular geographical units, there are many benefits to the production of data on a regular geographical grid. The latter offers particular advantages of stability over time and relative ease of integration with spatial data from other sources.

In Nordic countries the population grid is a regular product offered by statistical agencies for example, since 1970 in Finland and 1980 in Sweden. This is only possible due to the underlying point-based statistical system, permitting aggregation from a georeferenced building-code system to the cells of the grid (UN 2007). This approach to generation of grid-based data is known as the aggregation approach. In many other countries where this approach is not adopted – most often due to the absence of a suitably detailed georeferencing basis, researchers need to use some alternative spatial and/or statistical method to reallocate census data from irregular units into a population grid, termed the disaggregation approach. There are a variety of disaggregation methods, some of which use ancillary data. Examples without ancillary data include areal weighting, pycnophylactic interpolation (Tobler 1979) and kernel estimation (Bracken and Martin 1989). Examples using ancillary data include dasymetric mapping utilising land use classification derived from remotely sensed images (Eicher and Brewer 2001; Mennis 2003; Holt et al. 2004; Langford 2006), a network vector layer (Reibel and Bufalino 2005) or point addresses (Zandbergen 2011).

This paper addresses the challenge of producing a population grid for Brazil, in a context in which the data available to the national statistical agency varies greatly between urban and rural areas. It is proposed that generation of the population grid requires development of a novel hybrid model which combines both aggregation and disaggregation approaches. The following section describes the Brazilian context, proposed methods and the study area. The third section presents the study results and evaluation of the performance of the different methods. The final section presents conclusions and recommendations regarding population grid generation in this challenging context.

METHODS AND DATA

The Brazilian 2010 Census had two significant advances that deserve mention: integrated digital census mapping covering both urban and rural areas, and an address list combined with use of handheld computers with GPS. The first advance has allowed the georeferencing of dwellings in urban areas via addresses, and the second has allowed the capture of point locations of buildings in rural areas via GIS (IBGE 2010). These two technological advances permit for the first time aggregation of census data into grid cells. However, analysis of an initial sample revealed a significant number of enumeration areas with missing locational data, preventing direct aggregation. In urban areas and rural conglomerates (villages and small settlements) the spatial location is based on street block face codification. This presents two potential sources of missing locational data: there may

be no street network map, or the network may be missing block face codes. In the state of Para we have around 10% of enumeration areas that are possibly in this situation. In rural areas with a sparse settlement pattern, enumerators captured the building GPS points at the time of enumeration, but these may be missing due to operational and technical failures. In the state of Para around 3% of enumeration areas has locational data missing whilst about 3% has incomplete locational data. Under these circumstances it is not possible to simply apply an aggregation approach to the generation of a national population grid. Rather, it will be necessary to develop a hybrid approach which combines both aggregation and disaggregation approaches according to the local data context. It remains, however, to determine the most appropriate disaggregation method to be used in these circumstances.

Two study areas have been chosen in the state of Para in northern Brazil. The results of both aggregation and disaggregation are available for these areas, permitting a comparative analysis. The two areas have similar features, with a large rural portion and a small urban zone. Area 1 encompasses part of the municipality of Santarem (300,000 inhabitants) and Area 2 encompasses part of the municipality of Altamira (100,000 inhabitants). The rural part of both study areas is a mix of forest and agro-pasture. Area 1 presents a settlement pattern strongly related to the road network, while in Area 2 settlement pattern is more diverse and sparse. The urban part of Area 1 is more densely populated than that of Area 2.

Aggregated microdata from Census 2010 is here used as reference data he aggregation method differs between urban and rural areas. In rural areas, the grid cell result is the simple summation of the population count at each GPS point inside each grid cell. In urban areas, the block face is the smallest geographical unit and a linear weighting method is used to reallocate the population count from each block face into grid cells. For the grid cells that are partially urban and partially rural both results are summed. From now on this combined method will be referred as aggregation (AGG). The grids used here are based on a geographic projection with approximately square cells with sides around 1 km in rural areas and 250 m in urban areas. Four disaggregation methods are evaluated, each based on population count by enumeration area from Census 2010.

1. WEIGHT. Areal weighting based on 2010 Census data. It assumes that the distribution is homogenous within source (enumeration area) and target (grid cell) areas.

2. IMAGE. Dasymetric method using binary land use classification derived from 2009-10 Landsat 30m image data. Land use class "impervious surface" has been considered as populated and classes related to vegetation and water as unpopulated. Some known non-residential impervious features (e.g. airports) have been deleted from the information layer.

3. STREET. Dasymetric method using edited 1:5000 road network from 2010 IBGE Census Mapping. It is only available in urban areas.

4. POINT. Dasymetric method using 2007 IBGE Population Count residential building points. It is only available in rural areas.

The first evaluation assesses the populated and unpopulated cells correctly and incorrectly identified by disaggregation, compared to aggregation. These are termed omission and commission errors and are tabulated for each study area and disaggregation method..

The second evaluation concerns the population values estimated by each method. Linear regression is employed and selected goodness of fit statistics reported, although there is not space for these to be fully tabulated here.

The third evaluation is related to the difference between population estimated by the disaggregation models and the population count resulting from AGG. Cell values were grouped into classes and then the difference was computed. The formula used is:

$$D = (POP_{model} - POP_{AGG}) \div (POP_{model} + POP_{AGG})$$

POP_{AGG} is the population value obtained from the aggregation method and POP_{model} is the population value estimated by the disaggregation models. This measure has been chosen because it is able to describe the direction as well as the magnitude of the error. Negative values occur when predictions are smaller than observations.

RESULTS AND EVALUATION

Results are considered separately for urban and rural areas due to the different methods available in each context. As noted above, some areas are missing the information required for the aggregation approach and these are therefore excluded from the analysis. The total number of urban cells in the analysis is 1,347 in Area 1, and 542 in Area 2; the total number of rural cells is 4,424 in Area 1 and 2,656 in Area 2.

The tabulation of omission and commission errors for urban and rural areas respectively is presented in Tables 1 and 2. Omission errors correspond to cells not recognized as belonging to a class whilst commission errors are related to the incorrect identification of the class. Map accuracy concerns the probability that the classification is correct. In relation to urban areas (Table 1), all the methods perform better in more densely populated areas. The areal weighting presents the worst results and performs particularly poorly in Area 2, as it is not able to identify unpopulated places. The two dasymetric methods have similar overall accuracy in both areas, but the IMAGE model is less accurate in identifying unpopulated places in Area 1. Inspection of the mapped results (not shown here) suggests that this may relate to poorer performance in urban areas with plenty of open spaces and lower population density, but with a significant built street network.

Table 1 – Omission and commission errors (%) and map accuracy (%) in urban areas

		WEIGHT			IMAGE			STREET		
		Om	Com	Acc	Om	Com	Acc	Om	Com	Acc
AREA 1	Populated	0.00	28.86	100.00	3.34	13.61	96.66	0.21	8.16	99.79
	Unpopulated	100.00	0.00	0.00	37.63	11.68	62.37	21.91	0.66	78.09
	Overall Accuracy			**71.14**			**86.79**			**93.54**
AREA 2	Populated	0.00	54.80	100.00	11.43	21.66	88.57	4.90	17.67	95.10
	Unpopulated	100.00	0.00	0.00	20.20	10.57	79.80	16.84	4.63	83.16
	Overall Accuracy			**45.20**			**83.76**			**88.56**

In rural areas (Table 2) the WEIGHT model performs poorly in identifying populated places, and therefore has a low overall mapping accuracy. This is due to the very low population density and scattered pattern of human settlements in rural areas. The other methods produced good overall results, but IMAGE shows a map accuracy less than in urban areas. The POINT model shows a very low accuracy in identifying populated places in Area 1 due to the poor quality of the point layer, previously noted. All dasymetric methods identify unpopulated places better than populated places – reflecting the far greater number of unpopulated cells.

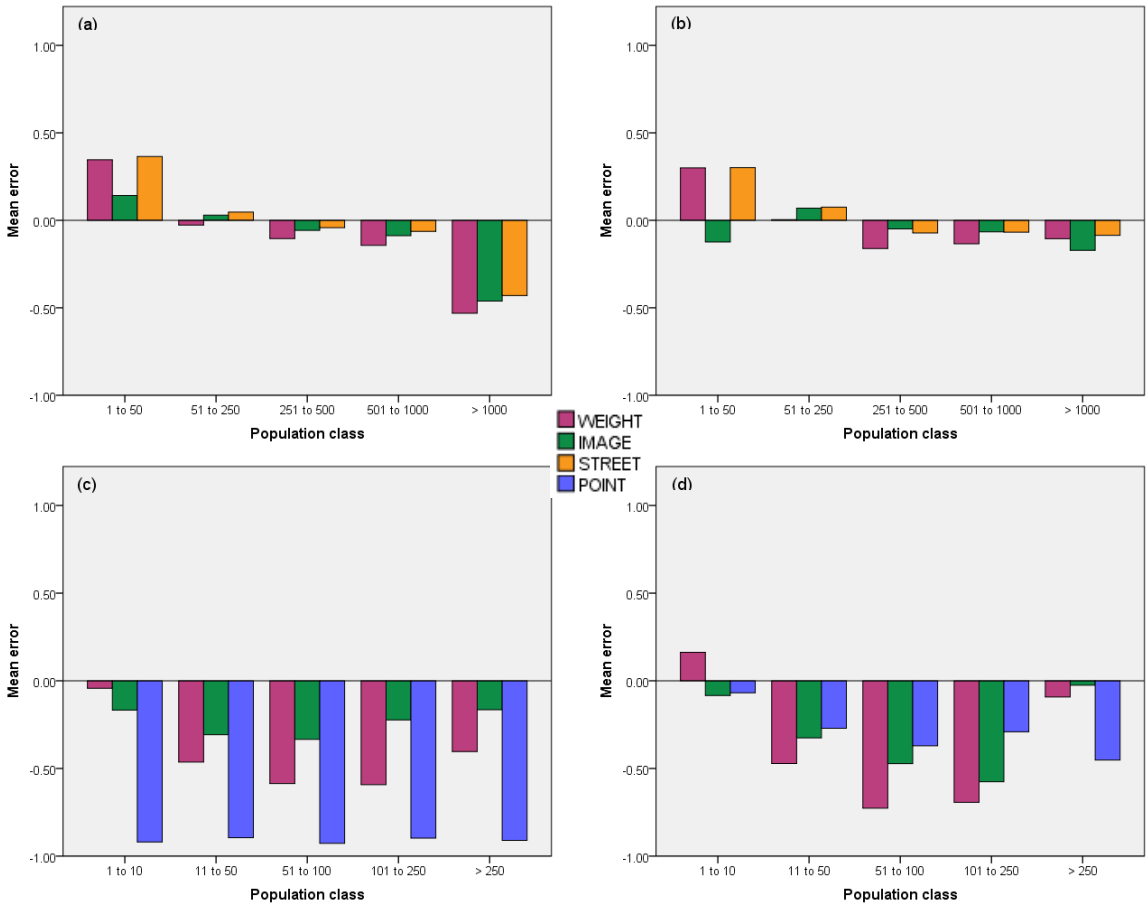
Table 2 – Omission and commission errors (%) and map accuracy (%) in rural areas

		WEIGHT			IMAGE			POINT		
		Om	Com	Acc	Om	Com	Acc	Om	Com	Acc
AREA 1	Populated	0.00	82.62	100.00	22.76	56.23	77.24	92.85	36.05	7.15
	Unpopulated	100.00	0.00	0.00	20.88	5.71	79.12	0.85	16.46	99.15
	Overall Accuracy			**17.38**			**78.80**			**83.16**
AREA 2	Populated	0.00	70.07	100.00	37.99	38.22	62.01	34.47	25.25	65.53
	Unpopulated	100.00	0.00	0.00	16.39	16.25	83.61	9.46	13.99	90.54
	Overall Accuracy			**29.93**			**77.15**			**83.06**

Turning to the regression models for urban areas, there are not great differences between the two study areas or methods analysed. R^2 measures the proportion of the variability in the dependent variable explained by regression. In Area 1, the WEIGHT model explains 85.90% of the variance of the values of population count, IMAGE explains 90.80%, and STREET explains 94.10%. In Area 2, the WEIGHT model explains 86.00%, IMAGE explains 92.30%, and STREET explains 94.50%. The models can be ordered by ascending accuracy: WEIGHT, IMAGE, and STREET. All F ratios are statistically significant at the 0.01 level. We speculate that the strong performance of the STREET model might be due to a circularity effect, as both aggregation and disaggregation methods in urban areas use

the network vector as georeferencing layer and ancillary data respectively. However, the F ratio values for STREET are notably greater than for IMAGE and WEIGHT.

Figure 1 - Mean error difference between observed and estimated populations



Turning to rural areas, the R^2 are much lower. For the POINT model in Area 1 this reflects known poor data quality. R^2 values for WEIGHT in Area 1 are 38.5% and for POINT in Areas 1 and 2 are 1.5% and 36.3% respectively. All other models explain more than 60% of the variability in the AGG population values. An analysis of the F ratios suggests that IMAGE is the best model in Areas 1 and 2, although in Area 2 the WEIGHT model has also a good fit. There are potentially errors in the POINT model due to the distribution of 2010 population count on a 2007 point layer.

Figure 1(a) and (b) show that overall in urban areas the models underestimate the population count when population is greater than 250 and overestimate it when it is lower than this. An exception is population class 1 to 50, where the IMAGE model overestimates

population in Area 1 and underestimates in Area 2. In general we can say that all the methods underestimate population in rural areas (Figures 1(c) and 1(d)). The models can be ordered by descending errors: WEIGHT, IMAGE, and STREET. The first population class (1 to 10) in Area 2 has small positive errors for WEIGHT and small negative errors for IMAGE and POINT, which probably reflects the very large number of truly vacant cells. The POINT method has an atypical behaviour in Area 1 due to the missing data. The last population class (> 250) in rural areas refers to very small numbers of cells.

CONCLUSION

The analysis presented here clearly shows that disaggregation methods can perform well in places where aggregation is not possible. However, no one method is best suited for use in all contexts. The choice will need to be determined by the characteristics of the application region, data availability and quality and the purpose of the analysis. With regard to the latter, it is important to consider whether the location (presence/absence) of population or the overall accuracy of the estimated counts is most important. For a national statistical agency, it is important to adopt a strategy that meets the analytical needs of many different users. Different models are likely to perform better in urban and rural areas but, more particularly, model performance is sensitive to density and settlement pattern. The best performance is achieved in dense urban areas and dasymetric methods consistently perform better than simple areal weighting. The choice of dasymetric method needs to take account of completeness, date, resolution/scale, format and availability of ancillary data. The quality of the metadata available on potential ancillary data sources can be critical in helping to inform these decisions. Further the output population grid should contain as much metadata as possible to inform the user about the method used and the uncertainties involved. We conclude that where countries face internal diversity in collected census data, a hybrid approach presents a viable means of generating a national population grid but that further research is needed on the optimal way of performing the choice of dasymetric disaggregation method based on the ancillary data available.

ACKNOWLEDGEMENTS

This study was performed as part of PhD research by Maria do Carmo Dias Bueno, funded by Instituto Brasileiro de Geografia e Estatística and by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (process number 17235-12-0). 2010 Census microdata and mapping were provided by IBGE solely for use in this research.

REFERENCES

Bracken I, Martin D (1989). The Generation of Spatial Population Distributions from Census Centroid Data Source. *Environment and Planning A* 21(4): 537-543.

Eicher C L, Brewer C A (2001). Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation. *Cartography and Geographic Information Science* 28(2): 125-138.

Holt, J B, Lo, C P, Hodler, T W (2004). Dasymetric Estimation of Population Density and Areal Interpolation of Census Data. *Cartography and Geographic Information Science* 31(2): 103-121.

IBGE – Brazilian Institute of Geography and Statistics (2010). 2010 Census – Summary of Survey Steps. <http://censo2010.ibge.gov.br/images/pdf/censo2010/sintese/sintese_censo2010_ingles.pdf>. Accessed 25 July 2013.

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: 161–180

Mennis, J (2003). Generating Surface Models of Population Using Dasymetric Mapping. *Professional Geographer* 55(1): 31-42.

Reibel, M, Bufalino, M E (2005). A test of street weighted areal interpolation using geographic information systems. *Environment and Planning A* 37: 127–139.

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

UNITED NATIONS, United Nations Economic Commission for Europe (2007). Register-based statistics in the Nordic countries - Review of best practices with focus on population and social statistics. New York and Geneva: United Nations. <http://www.unece.org/fileadmin/DAM/stats/publications/Register_based_statistics_in_Nordic_countries.pdf>. Accessed 25 July 2013.

Zandbergen, P A (2011). Dasymetric Mapping Using High Resolution Address Point Datasets. *Transactions in GIS* 15(s1): 5–27.