Developing a geo-data frame to facilitate data integration (Part 1)

by Gerbrand Mans, CSIR

This article discusses the development of a disaggregation procedure for socio-economic data based on the principles of dasymetric mapping and areal interpolation in order to develop a flexible geo-data frame which allows the data to be assigned to different demarcations seamlessly. The geo-data frame is based on the Spot Building Count points dataset from Eskom.

The need to work with data in an integrated fashion is a topic that has enjoyed much attention in the past two decades, especially with the rise in popularity in the use of geographical information systems (GIS). Walker and Young [1] made some of the earliest and most compelling arguments regarding this saying that:

"Area-based analysis of social and environmental factors taken together should aid understanding of the range and variety of needs for policy intervention. It would enable a more customised approach to be taken to resource allocation for environmental protection, social development and well-being, and the maintenance or improvement of economies... in the end: ensuring that economists and ecologists are more aware of the spatial and environmental consequences of their recommendations."

People with possibly the greatest need to work with data in an integrated fashion are policy and decision makers within government. Also, integration of data is necessary on primarily the social, economic and ecological level in order to ensure sustainability [2, 3]. An area-based analysis of social and environmental factors taken together, aids in understanding the range of needs on which to base policy intervention. It enables a more customised approach to be taken to resource allocation for environmental protection, social development and well-being, and the continuance or improvement of economies.

Despite many calls for the integration of socio-economic and ecological data for decision making, there has been little real progress. In the South African government context economic and ecological advice arrives from different departments or sections and is then integrated subjectively, for example in the integrated development plans (IDPs)1 of municipalities, and thus does not pertain address the question of data integration.

The first solutions pertaining to data integration centered on the use of geographic information systems (GIS). This was however in the early days, and soon people realised that using GIS brought with it many issues of its own. The first of these was that the construction of geographical information system databases and the licensing of proprietary software are expensive and are criticised by some government administrators responsible for strategic policy development as investments incapable of providing adequate returns [4, 5]. Recently this has started to change as the initial development of databases is usually the most expensive aspect and the databases are now established and cheaper to maintain. In terms of the software, the market for GIS has become competitive, extensive and settled enough such that licensing and maintenance is not such an issue any more and there are also freeware and open source GIS-packages available.

Obstacles to data integration: boundaries and scale

The biggest issue with respect to the integration of socio-economic and environmental data arises from the fact that often one has datasets with differing scales and differing demarcations. These differences are commonly dictated by academic discipline, nature of the study, survey methodology, administrative / service considerations or physiographic characteristics [6].

Typically, environmental and socio-economic studies adopt different types of spatial areas as their basic units. Data structures usually reflect the specific interests of data collectors and reporters [7] and different units are selected and tailored to meet specific needs, often for organisational as well as for technical reasons.

Socio-economic studies, for example, tend to be based on administrative units such as local authority districts, wards or primary care trusts [8]. The boundaries of such units are abstract creations with no particular relation to physical realities on the ground and may be subject to periodic change [9]. In contrast, environmental data is

---

Note 1: An IDP is a participatory approach to integrate economic, sectoral, spatial, social, institutional, environmental and fiscal strategies in order to support the optimal allocation of scarce resources between sectors and geographical areas and across the population in a manner that provides sustainable growth, equity and the empowerment of the poor and the marginalised. An IDP is therefore a plan that guides the activities and decisions of a municipality for the next five years in terms of Chapter 5 of the Municipal Structures Act, 2000.
mainly collected on the basis of units related to either physical attributes of the environment such as watersheds, or regular sampling grids related to the technology or methodology used to acquire or store the data. In attempts to overcome the problems related to scale and demarcation, basic spatial units have been created to which the attribute data of the different layers is then assigned. An example of a study where basic spatial units (BSUs) were used is the SECRA-study by Huby et al. [6]. In this study they were concerned with the integration of spatial data sets designed to characterise rural England in terms of what is there, what it is like, the living and working conditions, and the political and economic context from a sustainable development point of view. The CSIR developed the mesozones demarcation for a similar purpose in mind, i.e. integration of socio-economic and environmental data [10].

There are a host of considerations when defining a BSU. The problem one ends up with is that the demarcation only suits the purpose of the specific project at hand and therefore it is not generic enough, i.e. just another arbitrary demarcation. The other issue is – with the fact that socio-economic studies’ boundaries are dictated by administrative demarcations (e.g. South African census demarcations) – that when a BSU takes these administrative boundaries into consideration it become superfluous once the administrative demarcation changes, which is the case in point for the CSIR mesozones (BSUs). Especially in a South African context this is a major issue with local municipality boundaries changing frequently.

A number of researchers have inverted the problem of the BSU identification and rather advocate the development of base units derived from the distribution of the underlying dataset [11]. Openshaw and Rao [12] argue that census users should abandon official zones and re-engineer zoning systems dependent on the data and relationship they are investigating. Whilst this would result in the optimal solution with regard to data representation and integration, from a more pragmatic policy-making point of view it is not the optimal solution either. This approach assumes full competency in working with GIS-based data in different formats (vector and grid based) and integration of the data in a platform suitable for the question at hand. A person tasked with decision making based on integrated datasets will not have the time and, most probably, not the skill to do this.

Methodology

Purpose of the study
What we wish to address in this article is the need to integrate data in an appropriate way in order to allow for two things: firstly, integrated decision making on socio-economic and ecological or environmental data; and, secondly, to ensure that the data is portrayed in such a way that it overcomes the MAUP effect. This will be done by developing a flexible socio-economic frame using the Spot Building Count (SBC) dataset (point dataset) and assigning socio-economic characteristics to each point using the principles of dasymetric mapping. The SBC was produced by the CSIR and Eskom in 2008 and is a geo-referenced building frame developed using Spot 5 satellite imagery. The inventory concerned contains all classifiable building structures within the borders of South Africa [13]. The accuracy of this process will be tested and discussed herein.

The reasons why we suggest putting the socio-economic data set in a flexible format are as follows:

- Environmental data is dictated by physiographic boundaries (for example plant biomes, river catchments, mountainous areas, etc.) which do not adhere to any administrative or human activity related boundaries.
- Administrative boundaries are dictated by many influences and subject to constant change. It is neither practical nor implementable to adjust these boundaries according to environmental based demarcations, or to try and make it fit socio-economic trends accurately.

The SBC-points dataset is therefore considered as the static spatial units to which one will assign the socio-economic data.

Dasymetric mapping
A dasymetric map is the result of a procedure applied to a spatial dataset for which the underlying statistical surface is unknown, but for which the aggregate data already exists. The aggregate dataset’s demarcation is however not based on variation in the underlying statistical surface, but rather the result of convenience of enumeration [14, 15]. Thus, the process of a dasymetric map involves transforming data from the arbitrary zones of the aggregate dataset to recover (or try to recover) and depict the underlying statistical surface. This transformation process incorporates the use of an ancillary dataset that is separate from, but related to, the variation in the statistical surface [14]. Dasymetric mapping has a close relationship with areal interpolation – the transformation of data from a set of source zones to a set of target zones with different geometry [16, 17, 18]. Areal interpolation is mostly an areal weighting procedure and does not take ancillary sources into consideration when the spatial distribution of data is refined. Many areal interpolation methods can be incorporated into dasymetric mapping methods to improve the detail of a choropleth map below the level of the enumeration unit [17, 19].

In the examples taken from the literature, a dasymetric map is the result of intersecting polygon layers which predicts where the actual concentration of variability would be within the data source layer [14]. We propose to move away from a polygon based dataset which represents the underlying statistical surface to using a point dataset by using the Spot Building Count (SBC) data as an ancillary source. This makes it a novel approach from a dasymetric principle point of view. The argument is that the SBC-points are an accurate ancillary source for human activity and therefore for all socio-economic related activities. The inverse of the argument is that one would not find any socio-economic activity where there is not any type of building present, whether formal or informal.

Establishing a flexible geo-data frame
In order to test the viability of using the SBC data set as a flexible frame to which socio-economic characteristics can be assigned, we only focused on classifying the dataset in order to do an accurate population distribution classification. The SBC is only a building count and no other information is available regarding the size or type of structure it represents. It was therefore necessary to develop a method by which the points (buildings) could be classified.
To use the SBC as a proxy layer for population distribution, a weight had to be assigned to each point indicating the relative contribution of that point to the total population. In other words the weight of the point would represent the probable household size of the building (household) in question. It was, however, not possible to do a regression analysis of the 2001 StatsSA census data and the SBC data regarding the ratio between SBC points and the number of households in, for instance, a census sub-place as the SBC was only done from 2006 to 2008. If the SBC data was available for 2001, the growth from 2001 to 2006 could be used to predict what the growth or expansion of a certain area was as well as to predict new growth areas beyond the boundaries of more densely populated areas. The characteristics of the new growth area could then be related to the 2001 area of which it is a natural extension of and then regression analysis used to determine population growth. The incompatibility of the datasets arising from the different time lines of the datasets, however, rendered this impossible.

A five-step hierarchical approach – incorporating a process of elimination – was followed in order to get a more detailed distribution of the current population figures for South Africa. This process is based on the principles of dasymetric mapping which is reliant on underlying ancillary data sources to determine the characteristics of the SBC-points. The process was as follows:

**Step 1 – Assign weights (potential household size) to the SBC-points**

The average household size per census small area layer (SAL) was calculated (population divided by households) and this figure was then assigned the SBC-points falling inside each SAL. The assumption underlying this process is that each SBC is a potential household, but we know that this is not always the case. Hence, a hierarchical exclusion process is followed to get a more accurate representation (Steps 2 to 5).

**Step 2 – Identify new growth areas**

The household sizes determined in Step 1 are dependent on the 2001 census data. From 2001 to 2008 there has been development in urban areas which extends into the SALs which have previously been classified as rural. These new growth areas will therefore have similar characteristics to that of the urban areas that they extend from and the SBC-points here must therefore inherit the characteristics of these urban areas.

**Step 3 – Identify the growth of informal areas**

In the SBC-dataset there are polygons which indicate the boundaries of informal areas for which it was not possible to identify individual houses in high density informal settlements. There is thus no indication of the number of dwellings, and hence households, situated in these areas. The approximate number of households was determined and the representative amount of points added to the existing SBC-points in order for the points to more accurately represent population distribution by the following process:

- Amalgamate informal area polygons with SALs
- Select the most dense SAL (2001) – based on density of households per square metre – overlapping the informal polygon (2008)
- Assign the same density to the rest of the informal area polygon
- Based on this density, calculate the approximate amount of households
- Create random points equal to the amount of households within the polygon
- Use the household size of the densest SAL overlapping the informal area as the average household size of the informal area.

The underlying assumption of the above mentioned process is that the density and household size of an informal area are the same as those of the densest SAL (2001) it overlaps.

**Step 4 – Identify SBC-points situated in commercial and industrial areas**

Census SALs are determined by the demarcation of enumerator areas (EAs). Enumerator areas are functional demarcations to ensure that a representative sample of the population is taken during the census survey; therefore EAs also include land-uses other than just residential land. In Step 1 the assumption was made that each SBC-point is a residential house and therefore represents a household but it may actually be a different type of building. Hence, all commercial and industrial areas had to be identified as these are usually high density built-up areas and the ratio between buildings and households may differ significantly for these two land-use types. Points which are part of the informal areas and which are situated in industrial or commercial areas were excluded from the process as the informal areas are structures used for housing and therefore the weights calculated in the

---

Note 2: Regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed [22].
previous process are appropriate for these cases. Step 4 thus entailed the following:

- Identify industrial and commercial areas from the 2001 land-use dataset
- Calculate the potential 2001 population of these areas using the area weighting method
- Calculate the ratio between the 2001 total population and the amount of SBC-points to obtain the relative weight of the point as a contributor to the total population.

This process therefore assumes that the same ratio between the amount of buildings and amount of people will apply as was the case in 2001.

**Step 5 – Calculate weights for SBC-points situated in rural areas**

Unlike urban residential areas where the amount of buildings strongly correlates with the amount of households in the area, rural areas have a lot of buildings which may either be uninhabited or used for a different purpose, for example storage facilities on farms. All SBC-points situated in SALs classified as rural were selected while SBC-points which had already been part of any of the above mentioned processes were omitted. The procedure followed was to:

- Identify SBC-points situated in 2001 census SALs classified as rural
- Exclude SBC-points which have been part of any of the previous processes (Steps 1 to 4)
- Calculate the ratio between total population and SBC for each SAL in order to assign a weight to the point for the dasymetric process.

**Assumption:** The ratio between the amount of buildings and the total population stayed the same from 2001 to 2008 (population divided by SBC).

**Exceptions:** All points which were part of Steps 2 to 4.

Part 2 of this article will be published in the June 2012 issue of PositionIT.

**Acknowledgement**

This paper was presented at AfricaGEO 2011 and is republished here with permission.

**References**


