Dasymetric Mapping of Census Data for Nepal towards Improved Disaster Risk Assessment Studies

Conference Paper · November 2015

Hari Krishna Dhonju
Pathway Technologies and Services Pvt. Ltd.

Sunita Duwal

Dasymetric Mapping of Census Data for Nepal towards Improved Disaster Risk Assessment Studies

Hari Krishna DHONJU, Manchiraju Sri Ramachandra MURTHY, Sunita DUWAL
(hari.dhonju@icimod.org, manchiraju.murthy@icimod.org, suniarchi@gmail.com)

Key words: Census data, choropleth, dasymetric, urban land cover, population grid

SUMMARY

Population census data is one of the most important variant for disaster risk assessment and also has numerous applications in different geospatial domain. Significant difficulties are encountered in the use of choropleth map using the census data especially for spatial analysis with the modifiable areal unit problem. Therefore it has been necessary for researchers to use ancillary information from census area and their boundaries in order to distribute population count into regular grid units for dasymetric mapping. This paper reviews the current population distribution models and presents an alternative population weighted centroid distribution algorithm. The gridded representation of population census data was done using urban land cover data as ancillary information. A comparative evaluation of the results at different grid resolutions in terms of accuracy, precision, scale and ease of use is presented. This work demonstrates several application of such gridded population data especially in disaster risk assessment where unique spatial overlay analysis is required and cannot be addressed using irregular area-based census data.

Dasymetric Mapping of Census Data for Nepal towards Improved Disaster Risk Assessment Studies

Hari Krishna DHONJU, Manchiraju Sri Ramachandra MURTHY, Sunita DUWAL
(hari.dhonju@icimod.org, manchiraju.murthy@icimod.org, suniarchi@gmail.com)

1. INTRODUCTION

Primary source of population is census data which can be readily transformed into choropleth map in rapidly growing geographic information systems (GIS). Although the choropleth map has been using extensively and continuously since decades, it has been recognized as a poor data structure for the representation of population based data. It is because of its uncertainty introduced by both the inappropriate location of zonal boundaries for the variable being mapped, and some potentially misleading nature of the choropleth mapping processes as it distributes population homogeneously over the zonal boundary regardless of settlements, land cover and landscapes. However, many researchers, planners and geographers readily use such data in computation of different population based indices in many application areas such as disaster risk assessment, health, migration, and retailing studies. Therefore, there is chance of carrying forward these problems to get compounded and propagate the errors easily by the unfamiliar users with the nature of the data. The recent development in GIS can offer an alternative representation of such data. In this paper, a review on the problems associated with the population mapping is discussed and followed by a method to represent population census data into raster grid and its implication in disaster risk assessment studies (DRAS).

The following four sections of this paper discuss the census data, problem associated with choropleth, method adopted for dasymetric raster representation of population data and its application towards DRAS.

2. NEPAL CENSUS AND ANCILLARY DATA

Central Bureau of Statistics, Nepal (Central Bureau of Statistics, 2015) publishes census data in the form of area-based aggregated population distribution for a range of geographical boundaries. The ward administrative boundary is the smallest population enumeration unit. Urban cover from land cover 2010 (Kabir, 2015), settlements location and geographic village development committee(vdc) boundary (Survey Department, 2015) have been used as the ancillary datasets for Nepal. Combining Nepal population census for 2011 and geographic vdc boundary, population has been relocated into regular raster grid using the ancillary datasets.

3. PROBLEMS ASSOCIATED WITH CHOROPLETH

In normal practices, use and analysis of population census data frequently involves the mapping of these data with vector geographies in choropleth form. A number of problems co-exist with this type of mapping because this type of data is essentially immature association between aspatial and highly irregular spatial areal units like administrative vdc boundaries. This uncertainty is provoked due to aggregation of individual data into imposed zonal units (Unwin, 1981), whose boundaries are not data-derived, but associated for ease of enumeration. As the consequences, it neglects underlying physical distribution such as settlements, urban and land cover of the population enumerable areal unit. Therefore, there is a need to define a function that encompasses underlying physical distribution to incorporate their natural boundaries, and try to represent the census data near to the existing real world.

Additionally, it is common to face difficulty with the application of choropleth map to small area studies and least appropriate for area displaying the most extreme socio-economic characteristics resulting to give inappropriate and misleading visual information in the map. For instance, choropleth assumes even distribution of population in Himalayan vdc though there is no settlement. Typically vdc boundaries encompass larger areas with open water bodies, agriculture, forest, snow areas with less populated area. In other words, population concentration should be greater in settlement and urban area rather than in other areas.

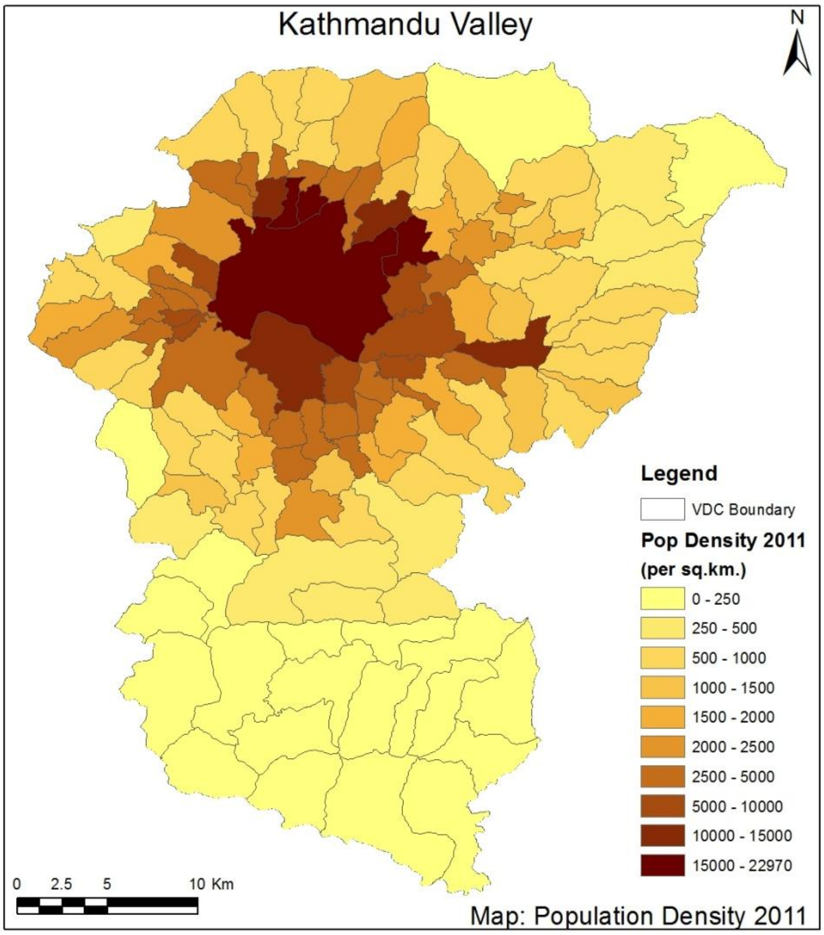

Figure 1: A Choropleth map showing VDC Population Density in the Kathmandu Valley (The VDC boundaries are population census enumeration zones)

Figure1 shows the distribution of population per square kilometers at vdc level based on population census 2011 in Kathmandu valley. The vdc boundaries are considered as population census enumeration zones and population density is represented in the form of choropleth map with evenly distribution of the population enumerated count. Figure 2 shows location of built-up area and settlements along with the other major land cover areas. The map clearly shows unevenly distribution of settlements in forest areas in comparison to built-up and agriculture areas. Here, it is noteworthy that though there is no settlement in certain forest area, population is evenly distributed over the enumerated vdc boundary, which is unreliable and this kind of boundary mapping frequently dominates the visual interpretation of the map. These types of misleading information are more likely to be encountered in the rural areas of Nepal, where the settlements occupy a very low percentage of the land area than other areas.

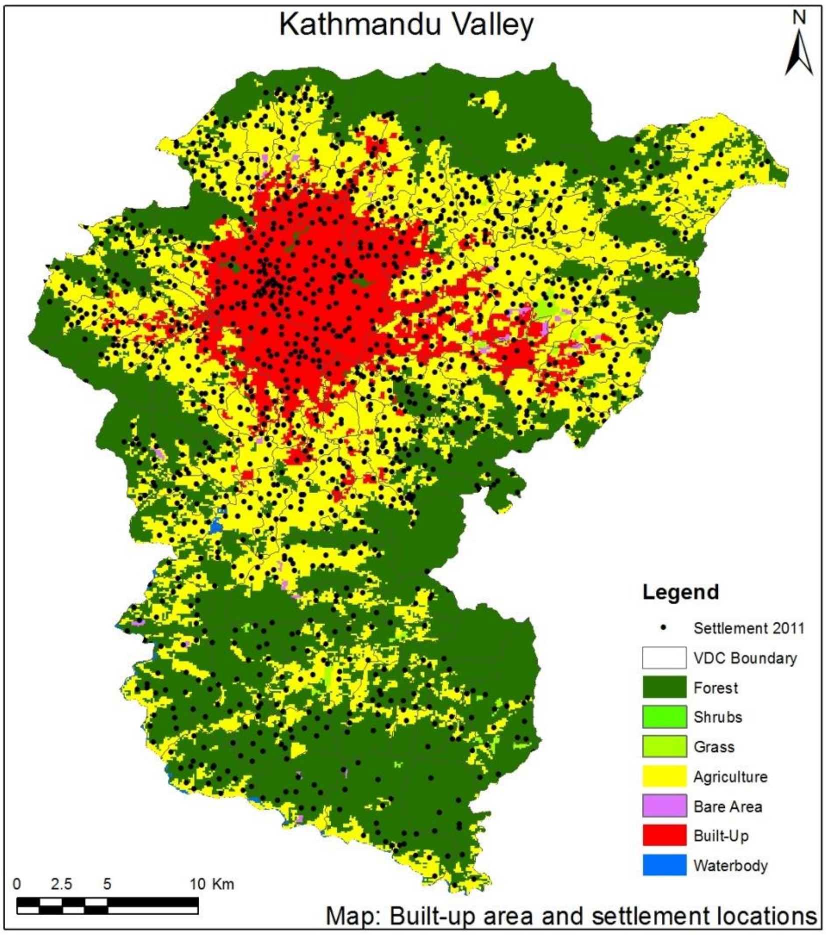

Figure 2: Location of built-up area and settlement along with other land cover areas

4. GRID-BASED DASYMETRIC MAPPING FROM CENTROID DATA

Though the problems exist in representing population and visual impact in choropleth map, it will never be possible to reconstruct and all aspect of spatial data structure based on aggregated census data. However, there is possibility of reconstructing population spatial data structure by representing the census data in raster cells using ancillary data by interpolation from the population-weighted settlement locations and built-up area.

The main concern is distribution of aggregated population data into regular raster cells, which can control on constraints over allowing the existence of cells with zero population to represent areas where no population is present. A precise technique can be developed for the production of raster maps from geometric zone centroids(Martin, 1989, 2007; Tobler, 1979) but these involve no more information about the underlying distribution than the boundaries themselves, and do not offer any solution to the issue of identifying unpopulated areas.

One of the prominent methods of interpolation is dasymetric mapping, defined as a method of setting disaggregated coarse resolution population data to a finer resolution by using ancillary data(Eicher & Brewer, 2001). Recent study suggests that dasymetric mapping can provide more accurate small-area population estimates than many areal interpolation techniques that do not use ancillary data(Gregory, 2002; Mrozinski Jr & Cromley, 1999). However, the method still needs to be refined for an optimal methodology for specifying the functional relationship of the ancillary data with the population density.

A robust methodology is required which can define a relationship between population density and ancillary data, and distribute population based on zonal boundary and population centroid locations. Here, a centroid data distribution algorithm has been adopted which can construct a detailed, higher-resolution population surfaces(Martin, 1989, 2007) with consideration of underlying ancillary data, built-up area and settlement locations. Built-up area has been taken as most populated area and is converted to vector points as the additional settlement points for a target raster resolution. Both settlement locations and additional settlement points from built-up area have been merged. Then, a vdc population has been redistributed equally among the merged settlement locations within the boundary. The following basic assumptions have been made for the algorithm development.

1. A centroid is settlement point with above average population density of the local area, of which it is a population summary point.
2. A centroid's population is distributed in the surrounding area using a distance decay function which has a finite extent.
3. No population may exist in water and snow area. So these areas have been excluded for population estimate.

Population count associated with the settlement points are distributed into the cells of a regular spatial grid. The algorithm proceeds moving of a kernel window over each centroid/cell. The size of the window determines the maximum possible extent of the centroid/cell influence in the assumption (2) above and it is a fundamental importance of the model operation. For each window, numbers of settlement points are counted, weight of population is calculated for these points according to a distance decay function and estimates population for the centroid of the window. Weight ($w_{ij)}$ of population for each settlement point is determined by equation 1(Martin, 2007).

$$w_{ij} = \left(\frac{k^2 - d_{ij}^2}{k^2 + d_{ij}^2}\right)^{\propto} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

Where, k is the initial window size and d_{ij} is the distance between the centre of the cell *i* and centroid *j*. The exponent α is the control factor over the shape of the distance decay function. Population for the centroid of the window is estimated using equation 2 i.e. the sum of contributed population weights from all settlement locations underlying the window.

$$\widehat{P_i} = \frac{\beta}{n} \sum_{j=1}^{N} w_{ij} P_j \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

Where, $\widehat{P_i}$ is a new estimated population for the centroid of the window *i*, P_j is the settlement population at *j* location, *N* is the number of settlement locations, n is the influencing population cells within the window and β is the zonal population factor. In ideal case, β equals to 1.0 and n equals to *N*.

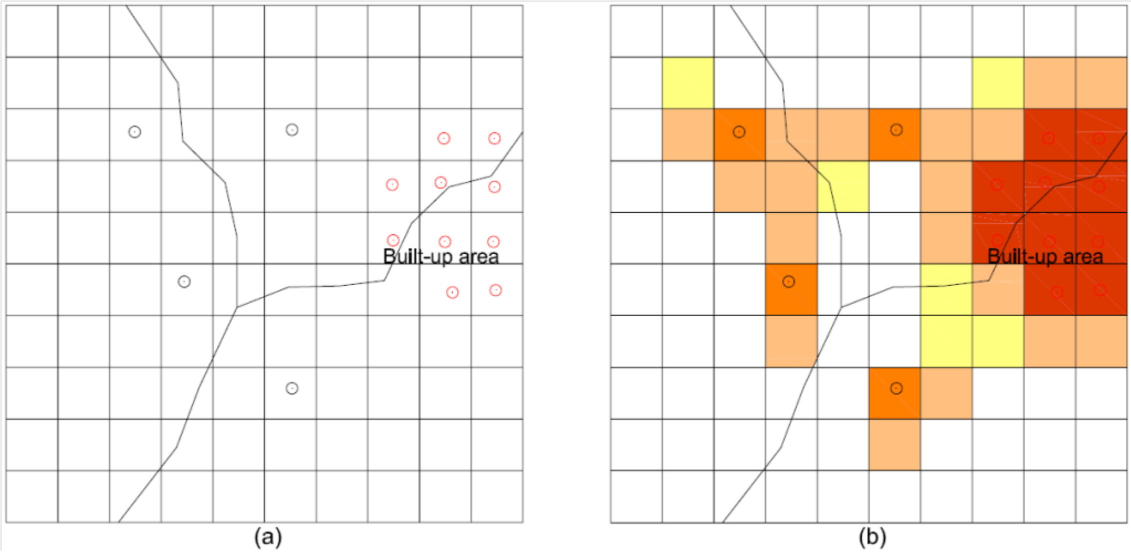

Figure 3: Raster cell based population estimates (a) vdc boundaries, existing settlement locations (black) and temporarily created settlement locations(red) in built-up area and (b) Estimated population raster grid.

Figure 3 (a) shows the relationship between the inputs- vdc boundary, existing settlement locations (black) and temporarily created settlement points in built-up area (red), and Figure 3(b) shows the resulting estimated population raster based on inputs provided. The algorithm can be used by varying different parameters of the model. However, it is important note here is that the key assumption of settlement locations with primary populations is taken as high information points to represent the local population distribution. Therefore, settlement locations and built-up area are the main influencing inputs to get more realistic population estimates.

The complexity of algorithm can be represented as O(n*m*l) in big O notation (Black, 2007). Where, n is the number of rows, m is the number of columns of the raster and l is the number of population centroids or settlement locations within a kernel window for population estimator. In other words, the raster grid space in terms of disk space will increase with the increase in the raster resolution (number of rows and columns) and in number of settlement locations within the increased window size.

5. APPLICAION AND USES IN DRAS

Essentially, raster maps are data models of the variable depicted rather than the fixed geography concept of vector data structures, which is integral to the data layer as the locational information. Therefore, each variable may have a unique characteristic data-derived geography, without fixed boundaries within the limit of the raster resolution. A more common approach to acquire vector data is to collect, point, line or area for a given variable to derive a surface model by areal interpolation techniques(Lam, 1983). Remotely sensed airborne and satellite imageries are also captured in raster form. Once the data are in raster formats, there are possibilities of a wide range of spatial overlay, operation and analysis among the rasters. In addition to this, it opens the gate of pixel based spatial data processing and modeling by using combination of mathematical functions, boolean, logical, and neighborhood functions within the rasters producing new map coverage at each stage of analysis(Berry, 1987). As a raster can be represented as one dimensional array, it is generally easy to handle and program for simple matrix operations. Rasters data can be represented for varying scales and compatibility of data over the time from different data sources, and also can be transformed, enhanced and degrade into a common raster format. These types of operations are notoriously complex while computing in vector-based formats(Martin, 1989).

In the event of any disaster and even for disaster preparedness and mitigation, information about the spatial distribution of population is prerequisite. Hazard, vulnerability and risk are three fundamentals and central pillars of the disaster risk assessment and management especially in the vicinity of disaster area. Population census is not frequently updated in most of the developing and under developed countries. Moreover, available data do not have information on the micro level spatial distribution of population. In contrast, information provided by risk assessment may contribute for decision making and mitigation processes towards management and planning in advance. Therefore, a real world representation of population distribution is the most important to calculate the effect of hazard and vulnerability with respect to the population exposure over the affected area, then to risk assessment. Thus, the raster representation of population distribution closer to real world justifies the use and applicability of such model towards DRAS and can be used more significantly.

6. RESULTS AND DISCUSSION

According to the methodology discussed in the section 4, shape factor (α) of value 1.0 and zonal population factor (β) that varies from zone to zone, has been used. The algorithm has been tested for 100m and 250m resolutions for whole Nepal. The complexity of computational and time expensiveness increases with the increase in resulting raster resolution and size of kernel window. Estimation of population for given raster grid cell converges at certain kernel size for a specified raster resolutions. Figure 4 illustrates the relationship between estimated population and kernel size. At certain level of increasing kernel size, the estimated population remains constant as it is the function of spatial distribution of settlement locations and raster resolution. The estimated population is almost constant when the kernel size is 13x13 for both the population grid with spatial resolution of 100m and 250m.

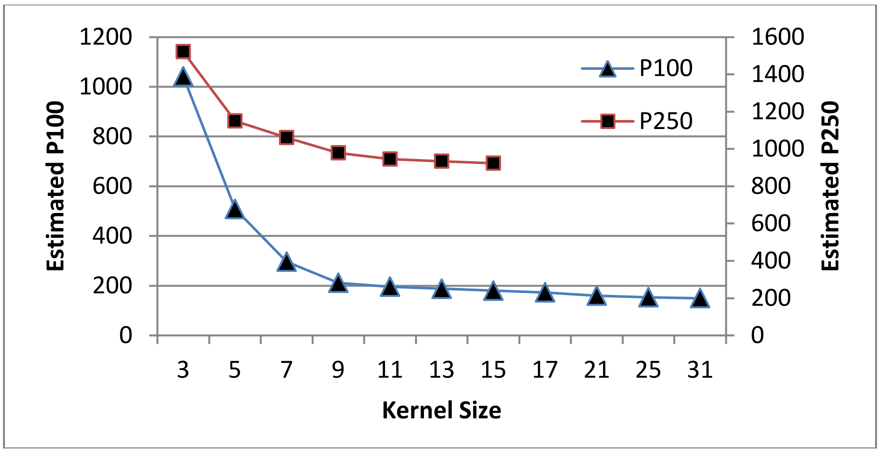

Figure 4: Relationship between estimated population and kernel size

The estimated population density grids with cell size of 100m and 250m are presented in Figure 5 for Kathmandu valley (see appendix for the whole Nepal). Figure 5(a) and 5(b) show population density distribution for raster resolution of 100m and 250m respectively. It is interesting to observe the difference between choropleth mapping and raster representation of population density developed by adopted methodology. In choropleth mapping of Figure 1, population density is evenly distributed all over the zones. In contrast, areas with no population can be distinctly observed in Figure 5. High population density can be found in the vicinity of settlement locations. As the cell size of population raster increases, area with no population also decreases and vice versa i.e. area with no population in Figure 5(b) is greater than in Figure 5(a). It shows that the population distribution and the range of population density are higher as the raster resolution. The reason behind might be the distribution of population in smaller area for higher resolution and larger area for lower resolution cell size.

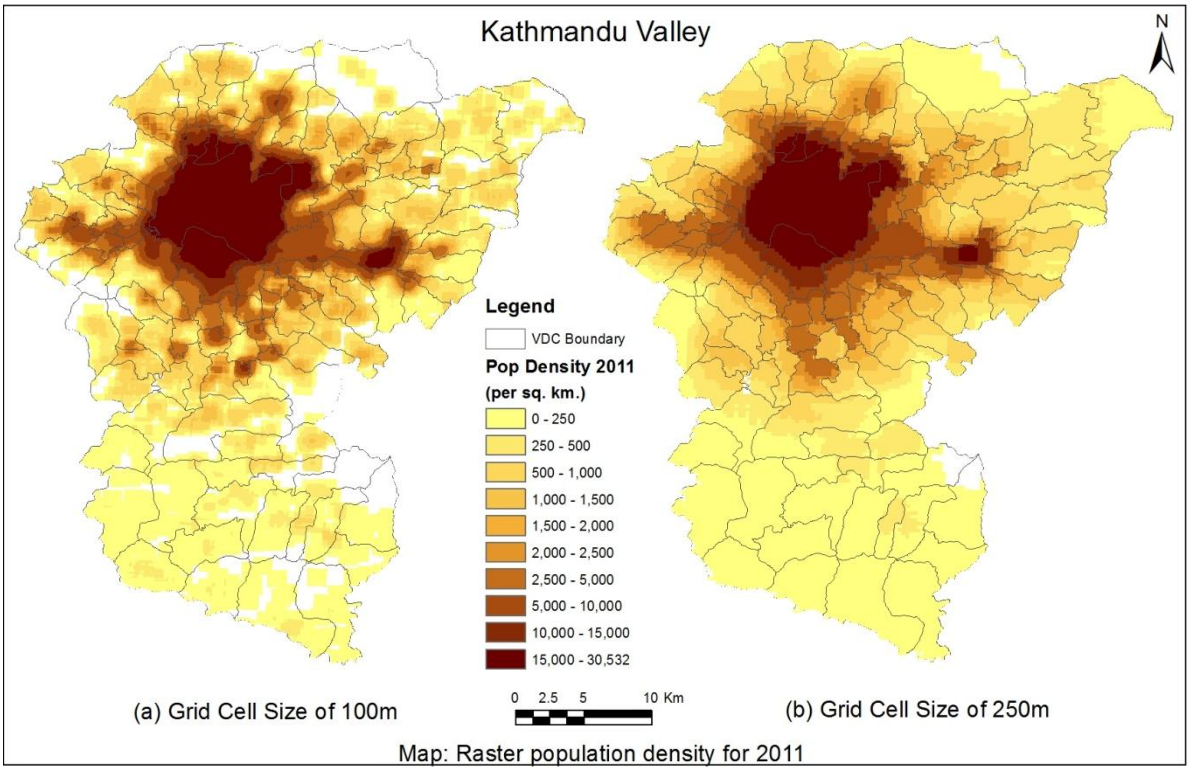

Figure 5: Estimated population density distribution in raster grids

7. CONCLUSION

In the earlier practices, population census data are represented in the form of choropleth in vector format. This paper has presented an alternative raster grid based population dasymetric mapping for Nepal discussing about its closer real world representation of population grid maps in the ground realities. This can be taken as new development and opens a number of research opportunities and challenges. Particularly, it can enable spatial modeling to enhance the new products in the frequent use and application of this product in DRAS. Further, it may be possible to use the new national-level model to calibrate population distribution so that they provide estimates of the spatially detailed pattern while preserving official counts at larger scales. This raster grid model can provide a new advantage of the data comparability and compatibility between the data sources over the time. There will be a bright future in the potential use of this standard raster grid representation of population as a publication framework for future census and other socioeconomic data.

ACKNOWLEDGEMENTS
The authors gratefully acknowledge International Centre for Integrated Mountain Development (ICIMOD) for providing financial support and essential datasets for this research. We would also like to acknowledge the contributions from many colleagues in ICIMOD to allow us to conduct this research and present in the wider audience.

REFERENCES

Berry, J. K. (1987). Fundamental operations in computer-assisted map analysis. *International Journal of Geographical Information System, 1*(2), 119-136.

Black, P. E. (2007). big-O notation. *Dictionary of Algorithms and Data Structures, 2007*.

Central Bureau of Statistics, N. (2015). http://cbs.gov.np/. Retrieved September 30

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

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

Kabir, U. (2015). Nepal Land Cover 2010, ICIMOD, Kathmandu, Nepal. Retrieved September 30

Lam, N. S. (1983). Spatial interpolation methods: a review. *The American Cartographer, 10*(2), 129-150.

Martin, D. (1989). Mapping population data from zone centroid locations. *Transactions of the Institute of British Geographers*, 90-97.

Martin, D. (2007). Gridded population data for the UK - redistribution models and applications. *Transactions of the Institute of British Geographers*, 1-10.

Mrozinski Jr, R. D., & Cromley, R. G. (1999). Singly±and Doubly±Constrained Methods of Areal Interpolation for Vector-based GIS. *Transactions in GIS, 3*(3), 285-301.

Survey Department, N. (2015). http://dos.gov.np/language/en/. Retrieved September 30

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

Unwin, D. (1981). Introductory spatial analysis: Methuen. *London and New York*.

BIOGRAPHICAL NOTES

Hari Krishna Dhonju is currently working in International Centre for Integrated Mountain Development (ICIMOD), Khumaltar, Lalitpur, Nepal as an Associate GIS-DSS Development Specialist and he has more than 12 years of experience in engineering software/application development with major contribution in development of SW_DTM, SW_Road, HKKH DST (GIS Viewer and 3D Viewer) and GIS based web science applications. His recent interests are focused in remote sensing and GIS based desktop and web science application and spatial decision support system developments. Computer programming is his passion through which he aspires to make geospatial technology access to everybody. Beside his works, he enjoys hiking and travelling.

CONTACTS

Name: Mr. Hari Krishna Dhonju
Institution: International Centre for Integrated Mountain Development (ICIMOD)
Address: Khumaltar, Lalitpur
City: Lalitpur
COUNTRY: Nepal
Tel. +977-1-5003222
Fax +977-1-5003277
Email: hari.dhonju@icimod.org
Web site: www.icimod.org

APPENDIX

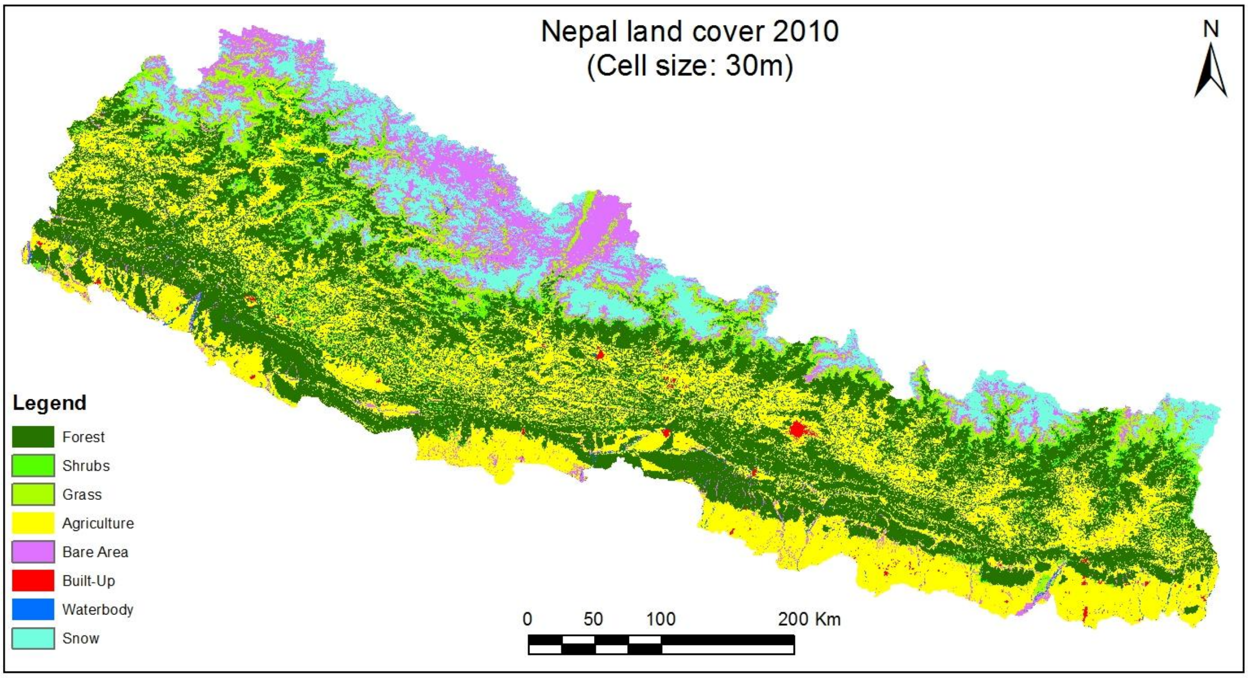

Figure 6: Nepal land cover 2010

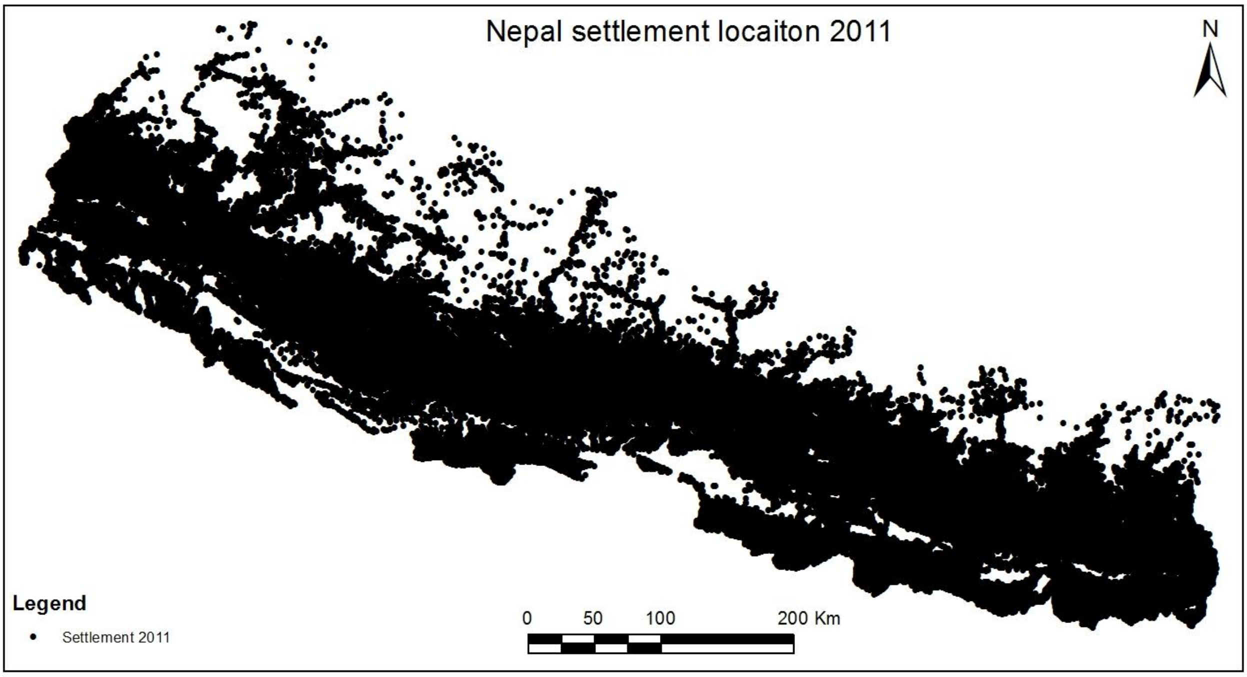

Figure 7: Nepal settlement location 2011

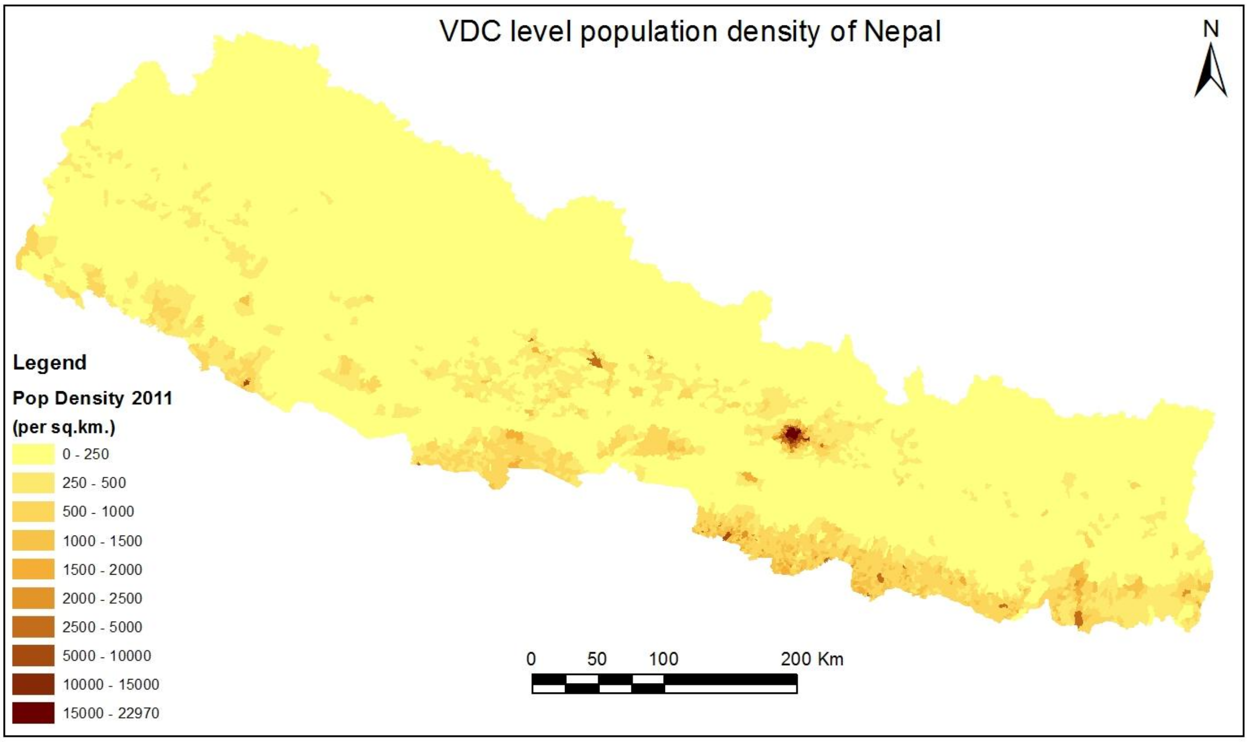

Figure 8: VDC level population density of Nepal

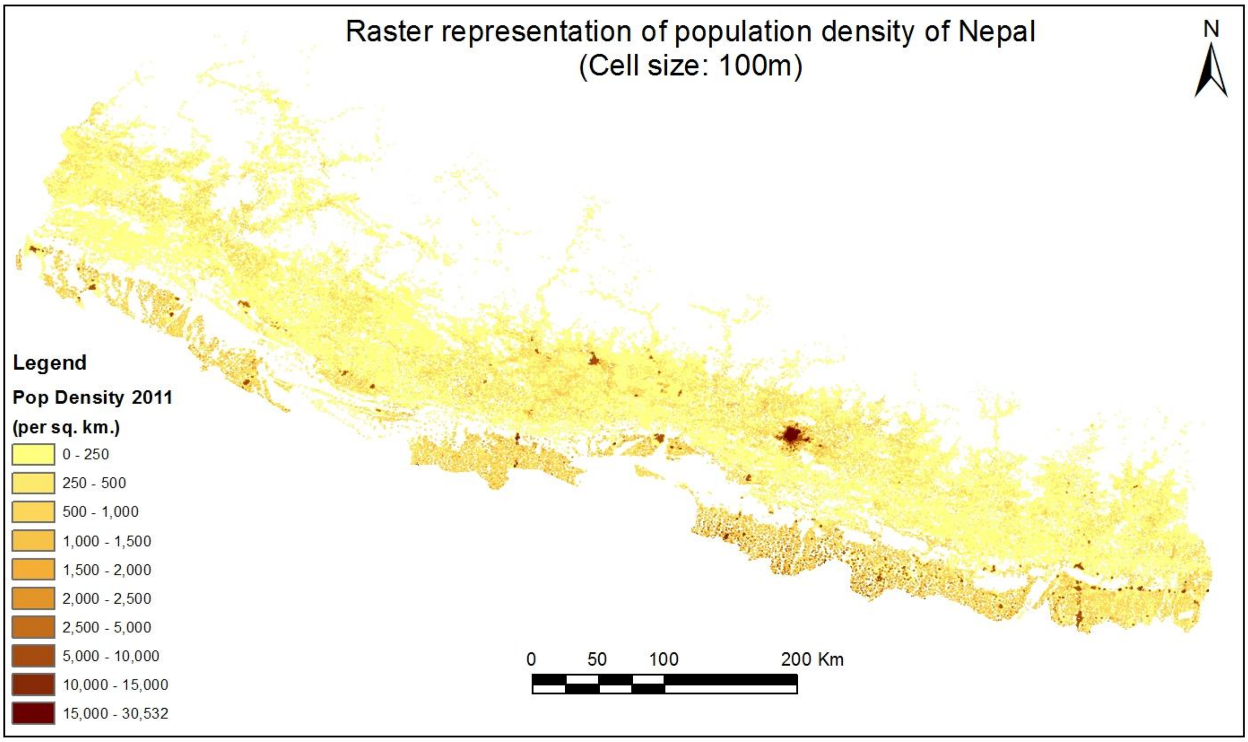

Figure 9: Raster representation of population density of Nepal (Cell size: 100m)

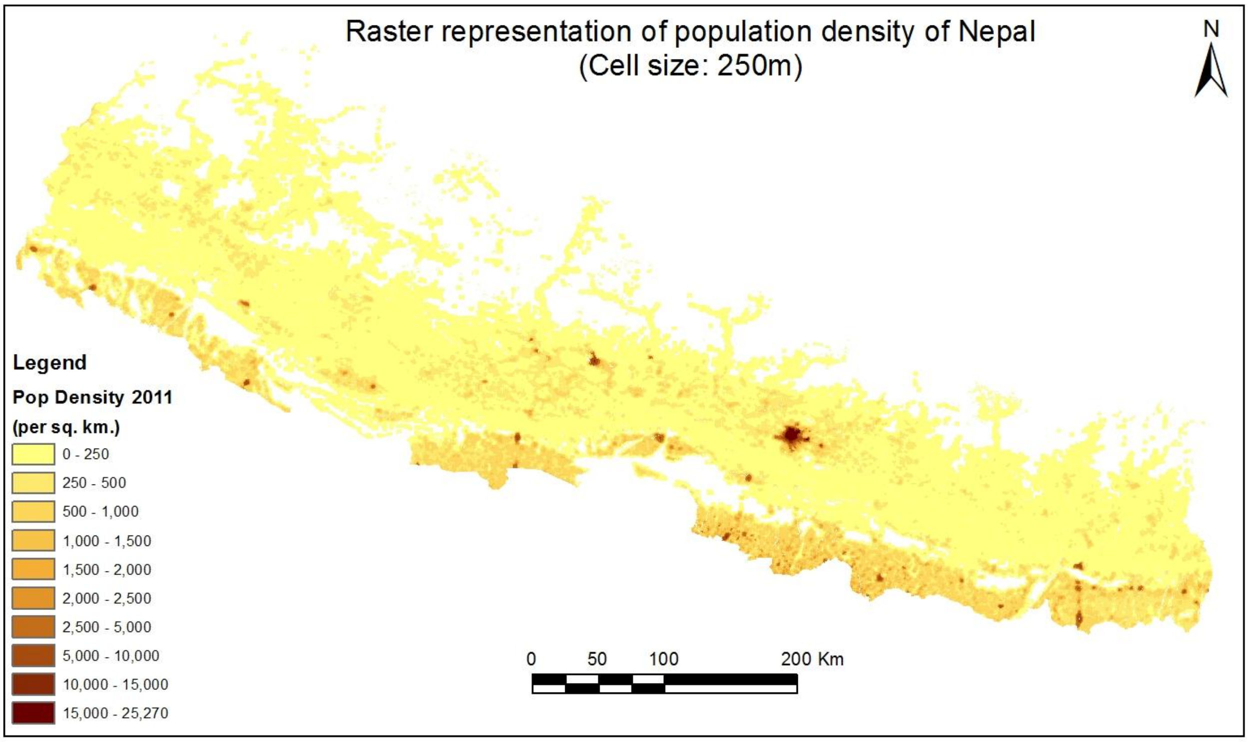

Figure 10: Raster representation of population density of Nepal (Cell size: 250m)