The role of satellite data in census: Case study of an Indian State

Koel Roychowdhury*, Simon Jones, Colin Arrowsmith, Karin Reinke, Anthony Bedford
School of Mathematics and Geospatial Sciences, RMIT University, Melbourne

E-Mails: koel.roychowdhury@student.rmit.edu.au; simon.jones@rmit.edu.au;
colin.arrowsmith@rmit.edu.au; karin.reinke@rmit.edu.au; anthony.bedford@rmit.edu.au

* Author to whom correspondence should be addressed; Tel.: +61-3-9925-2472; Fax: +61-3-9925-2517

Abstract: Countries, such as India, conduct a census collection every ten years. Currently census in India is carried out manually, therefore suffering from a number of shortcomings including inconsistency issues, the Modifiable Areal Unit Problem (MAUP) and large temporal acquisition timeframes. This paper proposes a surrogate census method using satellite images captured at night by DMSP-OLS satellites to overcome some of these drawbacks. The lights on the earth surface captured by this satellite represent areas of human habitation. Correlations between stable lights and brightness information with available census metrics from the last Indian census (2001) were calculated using bootstrapping techniques. Linear regression and multivariate analyses were subsequently performed and models proposed for each of the selected census metrics (e.g. population density, number of households per square Kilometre, percentage of households with cars, jeeps and vans, Per Capita District Domestic Product (PCDDP) and urban population density) with results ranging from $r^2$ of 0.8 to 0.9 at the 95% confidence interval. Census metrics unavailable at spatial scales lower than districts were also predicted using the proposed models and maps were derived showing the predicted measures. The results demonstrate that DMSP-OLS night-time images may be successfully used to estimate census variables in real time.

Keywords: census; DMSP-OLS; sub-national level.
1. Introduction

Census is commonly defined as counting the population of a country. The Office of the Registrar General and Census Commissioner of India (ORG I) describe population census as “the total process of collecting, compiling, analyzing or otherwise disseminating demographic, economic and social data pertaining, at a specific time, to all persons of a country or a well-defined part of a country” [1]. The first regular census of India was taken in 1881 under the then British governance. Since then census in India has been carried out regularly every ten years [1, 2].

Indian census depends heavily on enumerators and therefore suffers from a number of drawbacks. For example, the census is still carried out manually by enumerators visiting every household in the country to fill out the questionnaire. As a result it takes about 12 to 14 months to collect census data and a similar amount of time for the census to be processed and published. Being such a financially and administratively intensive process, the census is only conducted every 10 years. There is also no database on the census metrics for the inter-censal period. Furthermore, for administrative regions smaller than districts such as taluks and villages, some census metrics are not even collected. This paper proposes a surrogate method for collecting key census metrics using satellite images which helps to overcome some of these problems. The method proposed in this paper will enable the prediction of census metrics more frequently than available now. This in turn will help in population policy making and development planning. Another benefit of this method is that it enables the prediction of unavailable census metrics at the sub-district level as well as areas as small as approximately 1 square Kilometre (i.e. the area represented by one pixel in the satellite images used) and therefore helps in overcoming the error arising from MAUP.

2. Study Area

The study was conducted for the state of Maharashtra in western India (fig 1). This is second most populated state of the country with an area of more than 3 million square Kilometres. The capital city is Mumbai which is the largest urban centre of the country with a population of more than 18 million. The state is representative of a broad spectrum of urbanization and densities of population. It contains the largest urban conurbation in India and also some rural areas such as the district of Gadchiroli.

3. Data used:

The research uses satellite images captured at night by the Operational Linescan System (OLS) sensor onboard the Defense Meteorological Satellite Program (DMSP) group of satellites. Two types of images were used for the study: global composites of a) stable lights and b)
radiance calibrated (brightness) data for the year 2001 (both concurrent with the last census) (fig 1).

Figure 1: The state of Maharashtra as obtained from two DMSP-OLS images of 2001. (a) Maharashtra shown using the stable lights dataset. (b) Maharashtra shown using the radiance calibrated dataset (showing brightness values)

a) The stable light data was obtained from National Geophysical Data Centre (NGDC) website [3] using the latest average DN data series. It contains the lights from cities, towns, and other sites with persistent lighting, including gas flares. Ephemeral events, such as fires have been discarded from this dataset. Data values range from 1-63, with background noise data replaced with a zero. Areas with zero cloud-free observations are represented by the value 255.

b) The second DMSP-OLS image used in the study is the global composite of brightness data prepared for 2000 – 2001. This has been prepared from fixed gain images taken from satellites F12 to F15. However, this data was not calibrated to radiance [4, 5] and brightness values for this image ranged from 0 to 653. A radiance calibrated DMSP-OLS dataset was produced in 1999 to overcome the problem of relatively low light pixel saturation for stable lights datasets and to detect a greater range of settlements by varying the gain of the sensor. The radiance calibrated datasets showed more variation in the digital number (DN) values and therefore it proved more useful in modelling and finding appropriate relationships between parameters of interest [4].

DMSP-OLS images have been used for many applications such as population prediction, studying greenhouse gas and Carbon dioxide emissions, global maps of GDP, light pollution and its impact on species [4, 6-9]. Radiance calibrated datasets are produced from fixed gain DMSP-OLS images and show variation of radiance within urban areas in addition to diffused lights from populated rural settlements [10]. This data has been used for mapping urban extents [11, 12] in the United States. At the sub-national level, radiance calibrated global composite data has also been used to predict population density [13] and map urban areas [14]. Other studies [15] have
investigated the utility of separate fixed gain radiance calibrated datasets for predicting census metrics at different spatial scales.

4. Method:

4.1: Satellite Image Processing Methods:

The two types of images used in the study were each processed separately.

Due to the absence of on-board calibration, the stable lights product was calibrated using an empirical method [16]. After intercalibration, the pixel values were truncated to range from 0 – 63. Stable light data obtained from F15 satellite was chosen for further analyses. The brightness values obtained from the second DMSP-OLS image were converted from floating numbers to integers and were used in the study. Finally, a subset of the study area was taken from both the images (fig 1) and the mean and standard deviation of stable lights and brightness were calculated for the districts and used in the analyses.

4.2: Census data processing:

The first step of census data processing was to select a sample of districts. There are 35 districts in the state of Maharashtra. Of these, the districts of Mumbai, Greater Mumbai and Thane were not included for sample selection as they had very high values of both mean and standard deviation of brightness and stable lights compared to others. From the remaining 32 districts, 24 were randomly selected and 8 districts were withheld for model validation.

The analysis began using 144 census variables (demographic and some socio-economic). The distribution of these variables was tested by plotting histograms and calculating skewness and kurtosis [17]. From these tests, 48 variables such as sex ratio, female literacy rate, work participation rate, percentage of households with television, education facilities per square Kilometre and percentage of households with bicycles were shortlisted all of which had confirmed normal distribution over the selected districts.

Correlation coefficients were calculated for this subset list of variables using bootstrapping techniques to help overcome the limitations of the study’s small sample size [18, 19, 20]. Coefficients were calculated from 1000 bootstrap samples at 95% confidence interval. Bias and standard error of all the correlations was noted. Variables with a bias of less than 0.05 and standard error of less than 0.2 were chosen.
On the basis of these tests, the final list of census metrics consisted of ten variables: number of households per square kilometre, total population density, urban population density, female literates per square kilometre, total number of workers per square kilometre, percentage of households with car, jeep and van, percentage of households with access to electricity as power source, Percentage of households with television, Percentage of permanent houses and Per Capita District Domestic Product (PCDDP).

### 4.3 Model Development:

For building the surrogate census method, two types of models were tested: linear regression models with single independent variables or predictors and multiple regression models where two or more independent variables are used (fig. 2) [17, 21].

![Figure 2: An example multiple regression model of percentage of households using electricity as power source](image)

(a) model using mean and standard deviation of brightness (b) model using mean and standard deviation of brightness and stable lights (c) model using mean and standard deviation of stable lights. (n = 24)

#### 5. Results

All the chosen census metrics exhibited a positive correlation with the mean and standard deviation of stable lights and brightness. The highest correlations were observed between the census metrics and mean brightness followed by that between census metrics and mean of stable lights. The results from linear regression models and multiple regression models are presented below.
The simple linear regression models were created with no intercepts. The adjusted $r^2$ of these models ranged from 0.8 to 0.97 at 95% confidence interval for the models with average brightness and average stable lights. Models with standard deviation of brightness and stable lights showed moderate correlations ranging from 0.6 to 0.8 for models with standard deviation of stable lights and 0.4 to 0.8 for models with standard deviation of brightness at same confidence interval. Table 1 shows the adjusted $r^2$ values for the simple linear regression models.

**Table 1: Adjusted $r^2$ values of simple linear regression models (p < 0.05)**

<table>
<thead>
<tr>
<th>Census Metrics</th>
<th>Mean brightness</th>
<th>SD brightness</th>
<th>Mean Stable Lights</th>
<th>SD Stable Lights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of households per square Kilometre</td>
<td>0.95</td>
<td>0.44</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td>Total population per square Kilometre</td>
<td>0.96</td>
<td>0.44</td>
<td>0.96</td>
<td>0.92</td>
</tr>
<tr>
<td>Female literates per square Kilometre</td>
<td>0.94</td>
<td>0.48</td>
<td>0.93</td>
<td>0.89</td>
</tr>
<tr>
<td>Total workers per square Kilometre</td>
<td>0.94</td>
<td>0.37</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td>Percentage of households with cars, jeeps and vans</td>
<td>0.86</td>
<td>0.54</td>
<td>0.81</td>
<td>0.74</td>
</tr>
<tr>
<td>Percentage of households with television</td>
<td>0.97</td>
<td>0.50</td>
<td>0.95</td>
<td>0.92</td>
</tr>
<tr>
<td>Percentage of permanent houses</td>
<td>0.92</td>
<td>0.38</td>
<td>0.93</td>
<td>0.87</td>
</tr>
<tr>
<td>Percentage of households using electricity as power source</td>
<td>0.92</td>
<td>0.31</td>
<td>0.94</td>
<td>0.91</td>
</tr>
<tr>
<td>Urban population per square Kilometre</td>
<td>0.81</td>
<td>0.80</td>
<td>0.71</td>
<td>0.68</td>
</tr>
<tr>
<td>PCDDP (Rs 1998 - 99)</td>
<td>0.96</td>
<td>0.45</td>
<td>0.94</td>
<td>0.89</td>
</tr>
</tbody>
</table>

In order to test the impact of the independent variables together with the census metrics, multiple regression models were tested. These models were also built without intercepts. The models were constructed using a) mean and standard deviation of brightness; b) mean and standard deviation of stable lights and c) models with both mean and SD of brightness and stable lights. All models were predicted at 95% confidence interval. The adjusted $r^2$ values of these models are presented in table 2.

**Table 2: Adjusted $r^2$ values of multiple Regression Models (p < 0.05)**

<table>
<thead>
<tr>
<th>Census Metrics</th>
<th>Mean and SD brightness and stable lights</th>
<th>Mean and SD of brightness</th>
<th>Mean and SD stable lights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of households per square Kilometre</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td>Total population per square Kilometre</td>
<td>0.97</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td>Female literates per square Kilometre</td>
<td>0.95</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>Total workers per square Kilometre</td>
<td>0.97</td>
<td>0.96</td>
<td>0.97</td>
</tr>
</tbody>
</table>
The models were validated over the withheld districts and the most appropriate models selected. The model with the least error in predicted values was chosen.

5.1 Results after validation:

The mean and standard deviation of brightness and stable lights from the withheld districts were used to validate the proposed models. The predicted values were compared with the actual census metrics. The models which most accurately predicted the census metrics within a margin of error of 25% were chosen. For some census metrics such as percentage of households with televisions, percentage of permanent houses percentage of households with access to electricity and PCDDP, there was more than one model predicting values within 25% margin of error. In those cases the model with the highest $r^2$ values was selected.

The proposed models for the chosen census metrics are shown in table 3.

**Table 3: Proposed models for the selected census metrics at district level**

<table>
<thead>
<tr>
<th>Census Metrics</th>
<th>Chosen Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of households per square Kilometre</td>
<td>4.08 x ‘mean brightness’</td>
</tr>
<tr>
<td>Total population per square Kilometre</td>
<td>28.85 x ‘SD stable lights’</td>
</tr>
<tr>
<td>Female literates per square Kilometre</td>
<td>5.84 x ‘mean brightness’ – 0.01 x ‘SD brightness’</td>
</tr>
<tr>
<td>Total workers per square Kilometre</td>
<td>12.48 x ‘SD stable lights’</td>
</tr>
<tr>
<td>Percentage of households with cars, jeeps and vans</td>
<td>0.25 x ‘mean brightness’ + 0.17 x ‘mean stable lights’ – 0.32 x ‘SD stable lights’</td>
</tr>
<tr>
<td>Percentage of households with television</td>
<td>2.16 x ‘mean brightness’ + 1.56 x ‘mean stable lights’ + 1.09 x ‘SD stable lights’</td>
</tr>
<tr>
<td>Percentage of permanent houses</td>
<td>3.82 x ‘mean brightness’ – 0.01 x ‘SD brightness’</td>
</tr>
<tr>
<td>Percentage of households using electricity as power source</td>
<td>6.85 x ‘mean stable lights’</td>
</tr>
<tr>
<td>Urban population per square Kilometre</td>
<td>20.98 x ‘mean brightness’ – 22.04 x ‘mean stable lights’ + 4.51 x ‘SD stable lights’</td>
</tr>
</tbody>
</table>
6. Discussion and prediction of metrics at finer spatial scales:

The results obtained from the models in this research suggest there are several potential benefits of using remotely sensed measurements and statistical models for predicting census metrics. One of the major advantages of the method proposed here is that, it is possible to overcome the long (10 years) time lag between two censuses. DMSP-OLS captures data every night over the earth surface. The monthly composites of these datasets may be used to predict the census metrics using the proposed models at much higher temporal frequencies than was previously possible. Although there is observed error, the models are nevertheless useful in estimating trends of change of the selected census metrics consistently at various spatial scales.

Another advantage of these models is that, in addition to predicting metrics at district level, they enable prediction of census metrics unavailable for administrative regions smaller than districts. The Indian census does not account for some metrics at spatial scales smaller than that of districts. Some of these metrics include percentage of households with access to electricity, percentage of households with cars, jeeps and vans and per capita domestic productivity at sub-district level. The models were applied to produce maps of these metrics for administrative regions smaller than districts. Figure 4 shows the maps produced for percentage of households with access to electricity at subdistrict level for Pune. The values of these metrics can be predicted for administrative regions smaller than districts defined by the census (such as taluks) as well as for areas approximately 1 square Kilometre in size (i.e. the equivalent to the area covered by a single pixel in the satellite image). As a result the method proposed in this paper helps to overcome the MAUP that usually arises out of aggregation of smaller units into larger areas [22, 23]. “The areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating” [22]. These biases can be overcome by using the proposed models for predicting the census metrics.
Figure 3: Percentage of households with access to electricity at sub-district level for the district of Pune in the state of Maharashtra

Figure 3 (a) shows map at taluk level (administrative area below districts in the Indian census). Taluks such as Khed, Haveli and Pune city have more than 90% of households with access to electricity. This area surrounds the Pune city and forms part of the major urban centre of the district. Taluks such as Bhor and Ambegaon has 60 – 70 % of their households using electricity as their power source.

Figure 3 (b) shows map of percentage of households with access to electricity at approximately ground sample distance (pixel size). This map gives an idea of the smaller patches within each taluk which have more than 80% of households with access to electricity. Areas around Pune city have the highest percentage of households using electricity. Some smaller patches with high electricity usage can be found around smaller cities such as Purandhar, Baramati, Indapur and Lonavala spread over the district.

7. Conclusions:
This paper demonstrates that annual composites of brightness and stable lights obtained from DMSP-OLS satellite are useful in predicting census metrics and develops a surrogate census approach. The results provide an exciting opportunity to map variables at the subdistrict level. The method proposed in this study helps to overcome some of the major shortcomings of manually conducting a census for a big country like India such as MAUP and large temporal
acquisition time. Models proposed using brightness data are more effective in predicting the census metrics compared to those using stable lights. The metrics can be predicted at a far higher temporal frequency and at a greater level of detail than is available from the census. Out of the ten models proposed in the study, seven models use information on brightness. Some of the variables such as percentage of households with access to electricity had moderate correlations with both brightness and stable lights. This shows that distribution of electricity is one of the key metrics which can be predicted using DMSP-OLS night time images. Although the results obtained contain some level of error in them arising due to data conversions and scaling issues, the method described in the paper helps to provide an approach for mapping census metrics at spatial scales that were previously unavailable.

8. Acknowledgements:

The authors kindly acknowledge AUSAID for funding this research. The authors like to thank Ben Tuttle of National Geophysical Data Centre, USA for providing the DMSP-OLS datasets and clarifying any related questions; the Space Application Centre, Ahmedabad, India for supplying the vector datasets and the anonymous reviewers for their detailed comments and feedbacks.


5. Tuttle, B.T., Personal communication, K. Roychowdhury, Editor. 2008.


Accepted October 14, 2010
Published December 31, 2010

© 2010 by the authors; licensee Asia Pacific Advanced Network. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).