Correlating Night-time Satellite Images with Poverty and other Census Data of India and Estimating Future Trends*

Nischal K N
Student
M.S. Ramaiah Institute of Technology
Bangalore, INDIA
nischalkn@gmail.com

Radhika Radhakrishnan
Student
M.S. Ramaiah Institute of Technology
Bangalore, INDIA
radhika.radhakrishnan5@gmail.com

Sanket Mehta
Student
IIT Roorkee
Roorkee, INDIA
svanmuec@iitr.ac.in

Sumit Chandani
Student
IIT Kharagpur
Kharagpur, INDIA
sumit.chandani@iitkgp.ac.in

ABSTRACT

Given India's night-time satellite images and census data, this paper proposes a method to correlate light intensity from images with state-wise poverty, population, GDP, and forest cover, and forecast future values of the same for each state. We use the predictive model based method for imputation of missing data, multivariate regression analysis for correlating light intensity and census data, and the ARIMA model for forecasting census data. For forecasting, we compare results from the ARIMA model and the regression model to validate the authenticity of available census statistics. We outline a technique to obtain economical and timely information about poverty, which can help formulate monetary policy, foreign aid, or channelize other forms of support.

Keywords

Poverty, Satellite Images, Light Intensity, Census, Prediction

1. INTRODUCTION

Cities reveal themselves at night, while underdeveloped areas stay hidden. The presence of night-time lights captured from satellites across the globe is almost entirely due to some form of human activity. In this respect, such imagery provides a uniquely human view of the Earth's surface, as opposed to other remote sensing methods. These images are a rich source for analyzing developmental metrics for nations, and their application domain is vast [6].

*All authors contributed equally to this work.

CODS '15, March 18 - 21, 2015, Bangalore, India

http://dx.doi.org/10.1145/2732587.2732597

Poverty has historically been a pressing social concern for India with an estimated one-third of the world's poorest people living in the country [4]. Demographical data such as state-wise poverty, population, GDP, and forest cover is available through census. However, there are problematic issues with such census data in the context of India's population of over a billion.

- Data is not collected frequently, as the census process takes place only once in three to five years.
- Manual data collection is not an economical process for a large population.
- Data is often measured poorly and rarely measured at all (especially for sub-national regions), owing to widespread corruption in the country.

These problems lead to large chunks of data being unavailable for analysis, making it difficult to formulate solutions for poorer sections, such as monetary policy and foreign aid.

More importantly, a closer examination of the available census data reveals unnaturally optimistic drops in poverty levels in just a period of three years for some states, such as (but not limited to) Bihar and Jharkhand. This raises questions about the authenticity of the census collection and dissemination process in these states. However, absence of census data in the intermediate years hinders further examination of the matter. This calls for scientific evidence of tampering for it to be accepted as a valid concern. Our paper does that - it attempts to mathematically prove that the available census data of some Indian states has been tampered with. We do this by comparing the results of two methods - multivariate regression analysis and time-series analysis (the ARIMA model) - to validate the authenticity of available census data.

The need for an economical solution to the timely collection of data for India, and the investigation of specious fluctuations in available census data, were hence our motivations to conduct this research.

The majority of current research that relates night-time light sensing to developmental factors focuses on regions that commonly have complete statistical datasets [5]. However, the unavailability of such a dataset in many developing nations is an entry-point barrier to conducting such research in these demographics. This paper is different from others in the aspect that it focuses on imputing *missing* census data, and proposes a method to estimate future trends of the same, with primary focus on India, a developing country.

The rest of this paper is organized as follows. Section 2 describes the methods employed in conducting our research - namely, data collection, image processing, imputation of missing census data, correlation of light data with developmental statistics, and forecasting future values of census. Section 3 presents the results of our research. The paper then concludes in Section 4, followed by acknowledgments and references.

2. METHOD

This section describes the process of data collection and the calculation of light intensity from it. It then goes on to discuss the techniques employed for imputing missing demographic data, correlation using multivariate regression modeling, and future census forecasts.

2.1 Collection of Data

The Defense Meteorological Satellite Program Optical Line scan System (DMSP-OLS) Night-time Lights Time Series data used in this paper is available from the National Oceanic and Atmospheric Administration's (NOAA) National Geophysical Data Center (NGDC) [1]. State-wise GDP and population data for India from 2000-2012 is available from the Ministry of Statistics and programme Implementation, Govt. of India [7]. Forest Area and poverty (with missing values) statistics for India are available from the Planning Commission, Govt. of India [8].

2.2 Computation of Light Intensity

To measure the amount of illumination in different states of the country, raw satellite night images for the years 2000-2012 were used [1]. These global satellite images were cropped to scale to obtain images of India, as shown in Figure 1(b). To separate various states from these satellite images, the images were masked by a political map of India showing its state boundaries (Figure 1(a)). On the satellite images, the most prominent cities (Bangalore, Mumbai, Chennai, Delhi, and Kolkata) were identified by their bright spots and marked. Similarly these cities were also marked on the outline map and used as reference points. These two images were scaled and rotated so that the reference points in both the images coincide when superposed. This helped to align both images for extraction of states. From the outline map of India, the masks (binary images) of different states were obtained (Figure 1(c)). The individual masks of the required states were multiplied with the cropped satellite images to obtain the satellite image of the respective state for the respective year, as shown in Figure 1(d). Finally, the light intensity information from these processed images was obtained by using Eq. 1 [3].

$$\mu \ln (L_n) = \frac{\sum_j \ln (I_j)}{A} \quad (1)$$

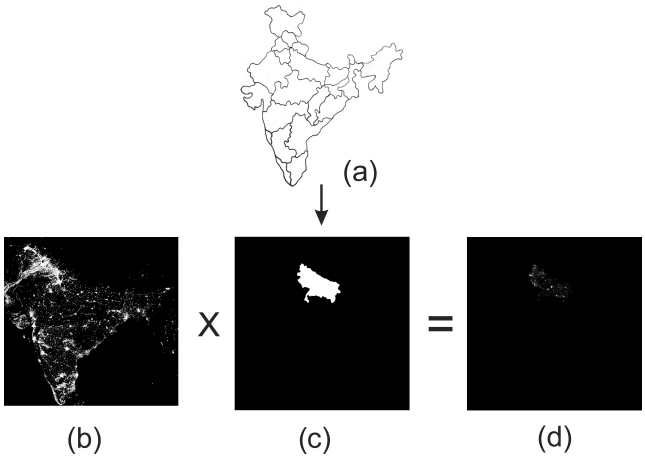


Figure 1: Light intensity extraction process: (a) Political map of India, (b) Night-time cropped satellite image of India, (c) Mask for the state of Uttar Pradesh, (d) Night-time satellite image of Uttar Pradesh.

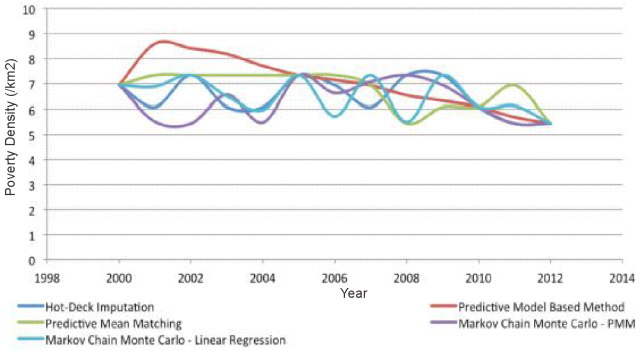


Figure 2: Comparison of various imputation techniques for poverty density data of Uttar Pradesh.

where I_j is the intensity for pixel j, A is the total count of pixels in the region, $\mu \ln (L_n)$ is the luminous intensity for state n.

2.3 Imputation of Missing Census Data

We compared several models for missing census data imputation. Figure 2 compares these models, with points of convergence representing data points from available census statistics. It may be observed that the predictive model-based method fitted the smoothest curve, reflecting the nature of variations in census data most accurately. It was computed using SOLAS software [2]. Table 1 shows the statstical data for Uttar Pradesh which was available from census, with NA indicating missing data. Table 2 shows the dataset of Uttar Pradesh after the predictive model-based method was applied to the available data.

Predictive model-based method [9] was used as follows-

1. We consider the multivariate regression expression, $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$, where $\varepsilon \sim N\left(0, \sigma^2\right)$

Table 1: Census data for Uttar Pradesh (UP) with NA indicating missing data

Year	GDP*	Population Density (km^{-2})	Poverty (km^{-2})	Forest Cover (km^2)
2000	181512	NA	219.9371	NA
2001	190269	689.8243	NA	NA
2002	206855	703.7248	NA	NA
2003	226972	717.8244	NA	14127
2004	260841	732.0610	NA	14127
2005	293172	746.3807	303.2856	14127
2006	336317	760.7335	NA	NA
2007	383026	775.1486	NA	NA
2008	444685	789.6716	NA	NA
2009	523394	804.2361	NA	14341
2010	600164	818.7965	7.409932	14339.5
2011	679007	833.2946	NA	14338
2012	769729	847.7636	248.2858	NA

*Crores of Rupees

Table 2: Census data for UP with highlighted values indicating missing data imputed using the predictive model-based method

Year	GDP*	Population Density (km^{-2})	Poverty (km^{-2})	Forest Cover (km^2)
2000	181512	675.6000	219.9371	14018.5600
2001	190269	689.8243	180.2109	14087.4827
2002	206855	703.7248	264.5859	14052.5660
2003	226972	717.8244	191.1042	14127
2004	260841	732.0610	275.4265	14127
2005	293172	746.3807	303.2856	14127
2006	336317	760.7335	194.2408	14185.4358
2007	383026	775.1486	98.0965	14205.1123
2008	444685	789.6716	146.1680	14216.6020
2009	523394	804.2361	194.1743	14341
2010	600164	818.7965	7.4099	14339.5000
2011	679007	833.2946	149.1640	14338
2012	769729	847.7636	248.2858	14372.2544

2. We modify co-variance σ^2 using $\sigma^2 = \sigma^2 (\eta_{abs} - q)/g$ and further modify to get a new value of co-variance β^* using $\beta^* = \beta + \sigma^* Z\sqrt{V}$ where Z and q are random Gaussian and chi-squared distributed.

3. We hence compute the missing values using the expression $Y^*_{mis} = X_{mis}\beta^* + e^*$ where $e^* = \sigma^* z$.

2.4 Finding of Correlation between Light Intensity and Census Data

Multivariate regression modeling was applied to find the correlation between intensity and other variables as shown in Eq.2.

$$\begin{aligned}\mu \ln (L_n) = &\beta_0 + \beta_1 Population Density \\ &+ \beta_2 Poverty Density + \beta_3 GDP \\ &+ \beta_4 Forest Cover + \varepsilon\end{aligned} \quad (2)$$

Where, β_0,β_1,... are coefficients of independent variables, $\mu \ln (L_n)$ is the luminous intensity of state n. Based on the R^2 value, we can determine how good the correlation between the above variables is for a particular state.

2.5 Forecast of Future Census Data

The ARIMA (Auto Regressive Integrated Moving Average) model was used to forecast future values for GDP, population density, forest cover, and poverty density of 16 Indian states for the year 2013 using Statgraphics Centurion software. The method was used as follows-

1. We use the standard ARIMA expression as given in Eq.3,

$$y_t(k) = \theta_t y_{t-1} + \ldots + \theta_p y_{t-p} + \varepsilon_t + \alpha\varepsilon_{t-1} + \ldots + \alpha_q \varepsilon_{t-q} \quad (3)$$

where y_t is the current value, y_{t-n} is the known census value for the previous n^{th} year, and θ is the moving average parameter.

2. The predictor is well defined when the constants are known. To evaluate the value of the constants, we minimize $\varepsilon_t(k) = y_{t+k} - y_t(k)$ which is the error between the predicted and current values found in step 1.

Variable	Coefficient	Value
Population Density	β_1	0.1789
Poverty Density	β_2	-0.1735
GDP	β_3	0.1609
Forest Cover	β_4	-0.001962
R square		0.7895
Adjusted R Square		0.6843

Table 3: Coefficients indicating the degree of correlation for the state of Uttar Pradesh. A negative sign indicates an inverse relation.

Variation	DoF	SS	MS	F	Significance (P)
Regression	4	2180.453	545.113	7.504	0.008153
Residual	8	581.114	72.639		
Total	12	2761.567			

Legend: (DoF) Degrees of Freedom, (SS) Sum of Squares, (MS) Mean Square, (F) F ratio.

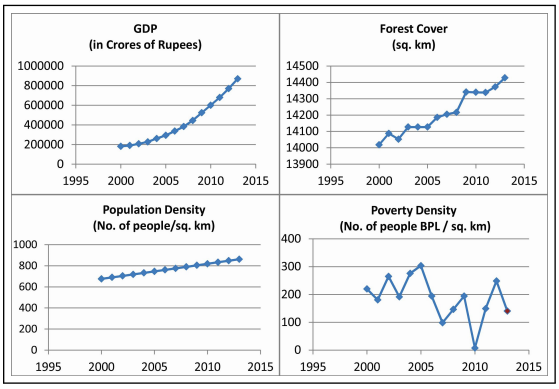
Table 4: ANOVA table for the state of Uttar Pradesh.

3. RESULTS

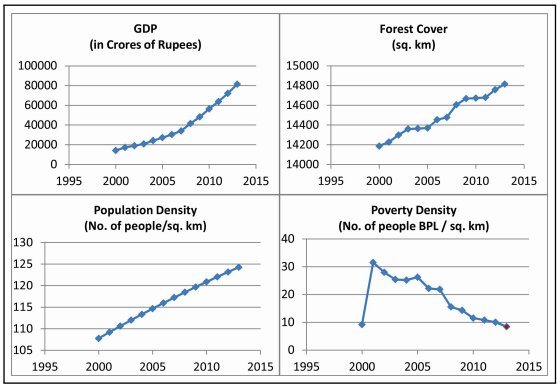
3.1 Correlation

The values of the multivariate regression co-efficients from Eq.2 for Uttar Pradesh are given in Table 3, along with the R^2 and adjusted R^2 values. The R^2 value assumes that every independent variable in the model helps to explain variation in the dependent variable. The adjusted R^2 value denotes the percentage of variation explained by only those independent variables that truly affect the dependent variable, and penalizes for adding independent variable(s) that do not belong to the model.

The values of β_n, R^2, and adjusted R^2 jointly indicate the degree of correlation between night-time light intensity and

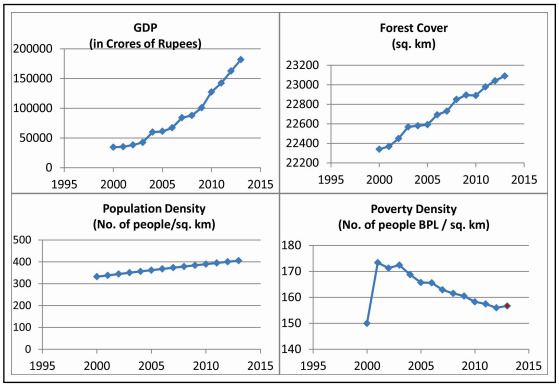


(a) Uttar Pradesh, Adjusted $R^2 = 0.6843$

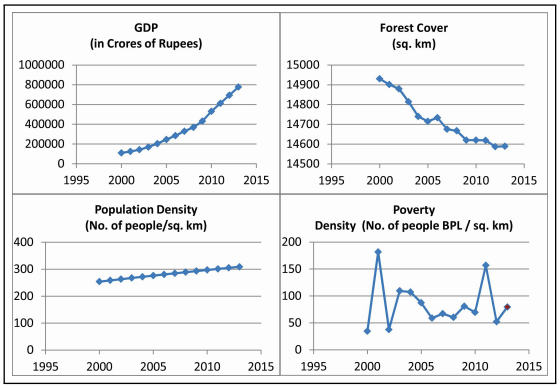


(b) Himachal Pradesh, Adjusted $R^2 = 0.6885$

Figure 3: Factor-wise trend models depicting most favourable results



(a) Jharkhand, Adjusted $R^2 = 0.3790$



(b) Gujarat, Adjusted $R^2 = 0.2513$

Figure 4: Factor-wise trend models depicting least favourable results

poverty density, population density, forest cover, and GDP. The high value of the co-efficients of poverty density, population density, and GDP indicate a strong correlation with light intensity. However, the low value of the coefficient for forest cover β_4 indicates that forest cover doesn't have a significant contribution in the multivariate regression model. Thus, forest cover has no significant effect on the determination of poverty levels using light intensity, whereas poverty density, population density, and GDP do.

Table 4 shows the ANOVA table with standard conventions for Uttar Pradesh. Here, we use the standard F-test and P-value to test the overall significance of the multivariate regression model, with a significance level of 5%. The significant value of F in the table shows that the model fits the data well.

3.2 Prediction

Table 5 shows the poverty density values for 2013, forecasted using the ARIMA model (fitted in Section 2.5), and the multivariate regression model (developed in Section 2.4). we note that the degree of variation between values forecasted using the ARIMA model and the multivariate regression model differs from state to state. This calls for further investigaton of the authenticity of available census data for states with high variation, such as Bihar. State-wise comparison of

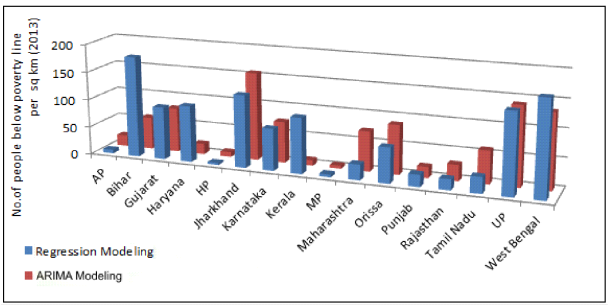


Figure 5: State-wise comparison of poverty values predicted for 2013 from Regression Modeling and ARIMA modeling.

forecasted poverty values for 2013 from the ARIMA model and the multivariate regression model is represented in Figure 5.

Figure 3 and Figure 4 show factor-wise trend models of the best and least-favourable state-wise results respectively, plotted using the ARIMA model. The favorability of results was determined by the adjusted R^2 values for each state from the multivariate regression model.

Table 5: Predicted values by ARIMA and Regression model for the year 2013

State	Light Intensity	GDP (in Crores of Rupees)	Population Density*	Forest Cover (sq. km)	Poverty Density†- ARIMA Model	Poverty Density†- Regression Model
Andra Pradesh		843395	313.51	46130.3	24.0168	6.272904399
Bihar		377855	1065.04	7344.81	57.6547	180
Gujarat		776697	308.96	14589.3	79.5053	93.96792717
Haryana		394715	594.276	1613.85	18.6316	100.769115
Himachal Pradesh		81472.5	124.244	14814.3	8.44206	3.437471658
Jharkhand		181799	405.2	23089.3	156.711	129.1374253
Karnataka		588385	316.057	36600.6	73.5767	74.72593782
Kerala		405833	901.747	17768.5	9.46017	99.1430863
Madhya Pradesh		405209	17.2885	78400.3	5.05872	4.809140941
Maharashtra		1590000	376.208	51889.4	71.5921	27.8
Orissa		287053	266.237	49120.9	87.9061	62.85811239
Punjab		332421	561.609	1808.02	19.4284	21.29579416
Rajasthan		536984	204.4	16165.1	28.7071	19.17291105
Tamil Nadu		858080	524.883	23616.9	58.616	28.70232361
Uttar Pradesh		870351	862.214	14427.7	140.475	143.6822071
West Bengal		723393	1025.64	13248.6	133.565	169.8326862

*No. of people/sq. km, †No.of people BPL/ sq. km

4. CONCLUSION

We found the overall results to be satisfactory, though the level of accuracy varied from state to state. It may be concluded that poverty density levels can be successfully estimated using satellite images.

Moreover, it was observed that available census data of some states did not fit the model. This brings to light the possibilty of incorrect census data that may reveal tampering during the process of gathering or disemmination of census, or both. Given the widespread corruption in India, this is an important finding that must be taken into urgent and serious consideration.

This project has vast scope for expansion. Further relevant statistics such as possibly, road networks, traffic counts, climate patterns etc. may be included as factors in determining poverty levels. Forecasting values for one state can also take into account the statistical values of neighboring states. Given an expanded dataset, district level analysis can be carried out for each state, as smaller geographical areas may produce better results. The recent global trend towards procuring better-resolution and frequent satellite imagery may make this possibility a reality soon. Since we have only analyzed correlation, future work can also take into account causation of factors, which may provide a better insight into understanding poverty as a social phenomenon.

5. ACKNOWLEDGEMENTS

We are thankful to Dr. Bhiksha Raj and Dr. Rita Singh of Carnegie Mellon University, Pittsburgh for their guidance and encouragement. We would also like to thank CMU's Internship Program in Technology-Supported Education (IPTSE) for providing a supportive platform to carry out this research.

6. REFERENCES

[1] Ngdc satellite images. http://ngdc.noaa.gov/ngdc.html.

[2] Solas for missing data analysis. http://www.solasmissingdata.com.

[3] Illuminating poverty: Using satellite light detection for a poverty indicator. December 2013.

[4] W. Bank. The state of the poor: Where are the poor and where are they poorest?

[5] DataKind, editor. *DC Big Data Exploration, Final Report*, March 2013.

[6] C. N. Doll. Ciesin thematic guide to night-time light remote sensing and its applications. *Center for International Earth Science Information Network of Columbia University, Palisades, NY*, 2008.

[7] M. o. S. Government of India and P. Implementation. Selected socioâĂŘeconomic statistics. October 2011.

[8] P. C. Government of India. Press note on poverty estimates, 2011-12. July 2013.

[9] Y. C. Yuan. Multiple imputation for missing data: Concepts and new development (version 9.0). *SAS Institute Inc, Rockville, MD*, 2010.