Adjusting population estimates using satellite imagery and regression models

Ilka Afonso Reis [1]
Vanessa Loureiro Silva [1]
Edna Afonso Reis [1]

[1] Universidade Federal de Minas Gerais - UFMG/ICEx/DEST/LESTE
Caixa Postal 702 – 31270-901 – Belo Horizonte - MG, Brasil
ilka@ufmg.br, vanessaloureirosilva@yahoo.com.br, edna@est.ufmg.br

Abstract. In this work, we describe a method to improve population estimates based on demographic data using satellite images. Our approach uses regression models whose explanatory variables represent demographic and remote sensing data. Furthermore, we assess the models performance taking account the variability in the estimation process via Monte Carlo simulation. To evaluate our models, we use the census districts of Belo Horizonte city (Brazil) and the images of sensor TM/LandSat5. Our final model has improved the estimates based only on demographic data. Furthermore, it has performed similar to the proposals in the literature at the micro level (census tract estimates). The advantage of our approach is to require low treatment of the explanatory variables, if compared to other proposals.

Palavras-chave: remote sensing, statistics, sensoriamento remoto, estatística.

1. Introduction

The technological and scientific development of the Remote Sensing brought an alternative to help solving the problem of estimating human populations: to use satellite images.

Regression models are the statistical method adopted in these cases. The dependent variable is the population counts (or density) and the explanatory variables are related to the satellite images. Harvey (2002a), for example, has used the reflectance in five bands of the sensor Thematic Mapper (TM), LANDSAT5, to estimate the population in the census tracts of a city in Australia. Lu et al. (2006) have used the images of the sensor ETM+, LANDSAT7, and the spectral mixture model to build an image of the impervious surfaces of a city in the United States of America. Using only this explanatory variable, they have estimated the population in the census tracts of the city.

Several levels of spatial aggregation can be used in the modeling. In general, the smallest collect unit of population counts is the census tract. However, these units can be aggregated to form larger ones. Usually, the area of the census tracts is larger than the area of the pixels of images with high or medium spatial resolution, which means that a census tract comprises several pixels. Therefore, it is necessary to summarize the values of the spectral radiance (or reflectance), for example, associated to the pixels in each tract.

Satellite images are routinely acquired for several reasons and some programs as LANDSAT and CBERS make them available at the internet for free. Then, having a satellite image of a city it is easier than counting people in its tracts. Even if one uses demographic projections to estimate tracts population, these predictions could not capture changes in areas that suffered a fast development. However, if this development is associated to changes in the land coverage, it can be captured by satellite images. These are probably the reasons why using satellite images to estimate population has gained attention in recent decades (Iisaka e Hegedus, 1982; Lo, 1995; Souza et al., 2002; Harvey, 2002a and 2002b; Reis, 2005; Liu et al., 2006; Lu et al., 2006).

Reis (2005) has adopted an approach similar to Harvey (2002b) to estimate the population counts in the census tracts of Belo Horizonte, the capital city of Minas Gerais state, Brazil. The author’s best model has performed better than Harvey’s model at the macro level (estimate of entire city population), but it has got a poor performance at the micro level (estimate of census tracts population).

The main goal of this work is to improve the performance of the model in Reis (2005). To do this, we add demographic data related to a previous period. In addition, we also improve the performance assessement using different datasets in modeling and in validatation phases, as well as calculating interval estimates for the performance measures and model parameters.

2. Materials and Methods

2.1 The dataset

In this work, we have used the population data of Belo Horizonte in two years: in 1996 and 2000, population has been counted by the Brazilian census bureau (IBGE). The city has got about 2500 census districts, which were occupied by about two millions people in 2000.

To build the dataset, we have used the bands 1 to 5 and 7 of LANDSAT-5/TM scene path/row 218/74, 2000/09/20 (Figure 1), the population counts of Belo Horizonte census tracts in 1996 and 2000 as well as the respective digitalized census tracts boundaries. The TM scenes have been co-registered to the digitalized census tracts boundaries.

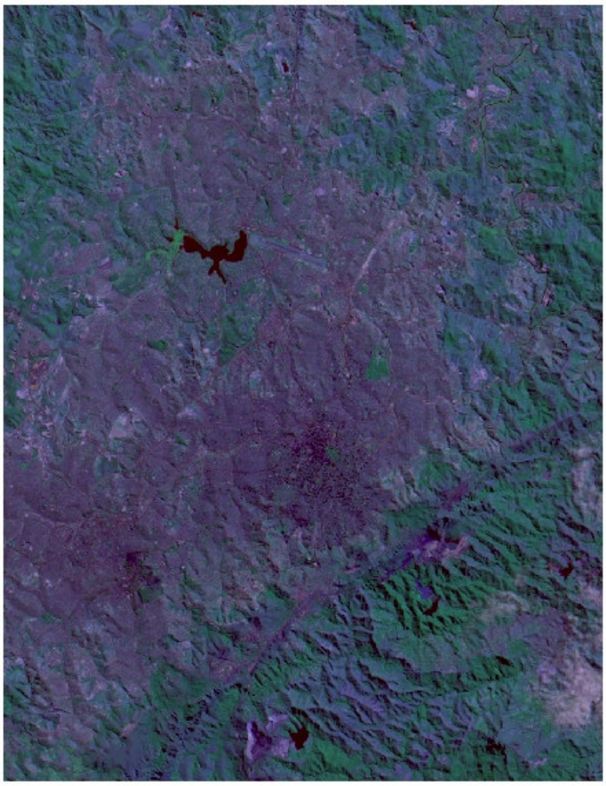


Figure 1 – Belo Horizonte city in the composition R(3)G(4)B(5) of part of LANDSAT-5/TM scene path/row 218/74, 2000/09/20.

Since some of the census tracts in 2000 were different from the census tracts in the former period, we had to make them compatible. To get the estimated population counts of 1996 period using the census tracts of 2000 period, we have spatially distributed the 1996 population over a grid of pixels (30m x 30m). Then, this grid was superposed the layer with the digitalized census tracts boundaries in 2000 and most often population value in a tract was associated to that tract as its estimated population in 1996.

To get the atmospheric correction and produce the surface reflectance images, we have used the 6S model (Vermote et al., 1997). To each band, the average reflectance of the pixels associated to a census tract has been associated to that tract as its measure of reflectance in that band.

The operations with images and census tracts described early have been performed using the GIS SPRING (Camara et al., 1996). Table 1 presents the variables in the dataset, which have been exported to be read in the statistical environment R (R Development Core Team, 2009).

Table 1 – Description of the variables in the dataset.

Name	Description	Source*
ID	The identification number of the census tract (CT)	IBGE[1]
Type	0, if the occupation of CT is normal; 1, if the occupation of CT is special (slum or "*favela*")	IBGE
Pop00	Population of CT in 2000	IBGE
Pop96	Population of CT in 1996	IBGE
TM1	Average reflectance in band 1 of LANDSAT-5/TM	INPE[2]
TM2	Average reflectance in band 1 of LANDSAT-5/TM	INPE
TM3	Average reflectance in band 1 of LANDSAT-5/TM	INPE
TM4	Average reflectance in band 1 of LANDSAT-5/TM	INPE
TM5	Average reflectance in band 1 of LANDSAT-5/TM	INPE
TM7	Average reflectance in band 1 of LANDSAT-5/TM	INPE
Area00	Area of the census tract	IBGE
UrbArea00	Percentage of urban area of the census tract in 2000 (after image classification)	INPE, IBGE

* This is the original source of the data, which have been processed to generate the dataset for analysis, except by the data of variable Pop00.

[1] Instituto Brasileiro de Geografia e Estatística. URL: http://www.ibge.gov.br

[2] Instituto Nacional de Pesquisas Espaciais. URL: http://www.inpe.br

Census tracts which have very low population density (less than 1.5 habitant per 100m^2) probably do not represent residential areas. After the elimination of these tracts, the dataset had 2520 census tracts.

2.2 The statistical modelling

The basic regression model used in the analysis is described as following

$$\begin{aligned} \text{Pop00}_i = \beta_0 + \beta_{96}\text{Pop96}_i + \beta_T\text{Type}_i + \\ \beta_1\text{TM1}_i + \beta_2\text{TM2}_i + \ldots + \beta_7\text{TM7}_i + \beta_U\,\text{UrbArea00}_i + \\ \varepsilon_i, \qquad\qquad i = 1, 2, \ldots., n_a, \end{aligned} \tag{1}$$

where n_a is the amount of census tracts used in the analysis and the variables are described in Table 1. The exceptions are the response variable (*Pop00*) and the explanatory variable *Pop96*, which can represent the population counts, the population density or the logarithm of the population density in the respective time periods.

The two first explanatory variables represent the demographic part of the model. The remaining ones (from TM_1 to *UrbArea00*) represent the contribution of the remote sensing data to the model and ε is the error term. Therefore, the role of the remote sensing part of the model is to adjust the predictions made by the demographic part.

Since the reflectance in the bands of a sensor can be highly correlated to each other, the model described by Equation (1) can suffer of the *multicolinearity* problem. The presence of multicolinearity in a regression model inflates the variance estimates, which makes more difficult to consider statistically significant the relationship between the response variable and the explanatory variables (Draper and Smith, 1998). To evaluate the correlation among explanatory variables, we have used the Variance Inflaction Factors (VIF), defined as following

$$VIF_j = 1 \big/ \left(1 - R_j^2\right), \qquad j = 1, 2, ...k, \tag{2}$$

where R^2_j is the determination coefficient of a regression linear model that has the explanatory variable j as the response and the other (k-1) variables as the explanatory ones. Variables having a value for VIF greater than 10 must leave the model.

The census tracts have very different sizes. Then, we have decided to use the population density as the response variable. Furthermore, since population density usually has a skewed distribution, we have adopted the neperian logarithm of the density population as the response variable. Then, *Pop00* represents the neperian logarithm of the density population in 2000, and, to be coherent, *Pop96* represents the neperian logarithm of the density population in 1996.

2.3 Assessing the predictions performance

To assess the performance of the model predictions, we have used the strategy in (Lu et al., 2006) to separate a part of the dataset to be a *validation dataset*. Then, one third of the census tracts have been used to evaluate the model that has been fit using the remaining tracts (the *modeling dataset*).

The performance of the model prediction for the census tract *i* has been evaluated by the relative error (RE_i), defined as following

$$RE_i = \frac{\hat{d}_i - d_i}{d_i}, \qquad i = 1, 2, ..., n_T, \tag{3}$$

where $\hat{d}_i$ and d_i are, respectively, the exponential of the predicted value and the observed value for the population density in the census tract *i* of the validation dataset and n_T is the size of the validation dataset (n_T =842).

To summarize the relative error values, we have calculated MdRE, which is the median of the absolute values of RE_i. Since there are census tracts with outlying values for population density, the median is a better choice than the usual mean.

MdRE assesses the performance at the micro level. To evaluate the performance at the macro level, we have calculated

$$TRE = \frac{\text{sum of predicted values for the } n_T \text{ tracts}}{\text{sum of observed values for the } n_T \text{ tracts}} - 1 \tag{4}$$

There are many possible ways to select tracts to compose the fitness and validation datasets. Each selection leads to a different estimate for MdRE and TRE. To be able to account for this variability, we have performed Monte Carlo simulation experiments. In each one of 10000 experiments, for example, we have selected a modeling dataset and, consequently, a validation dataset. The model has been fitted and the parameters estimates as well as performance measures have been calculated. Then, we have got 10000 estimates for the model parameters, MdRE and TRE. They have been used to calculate confidence intervals for the parameters.

3. Results and Discussion

The model described by Equation (1) has been the first one we have fitted. The explanatory variables *TM₂* , *TM₃* and *TM₅* have presented very large values for VIF. We have fitted a second model, without the mentioned bands, which has decreased the problem of multicolinearity (VIF < 5).

To investigate the possibility of different models for different types of occupation, we have added the terms of interaction between the variable *Type* and the other explanatory variables to the second model. Only the interactions between *Type* and the variables *TM₄* and *Pop96* have been statistically significant.

The final version of the model is described by the following equation

$$\begin{aligned}\text{Pop00}_i = \beta_0 + \beta_{96}\text{Pop96}_i + \beta_T\text{Type}_i + \\ \beta_1\text{TM1}_i + \beta_2\text{TM4}_i + \beta_7\text{TM7}_i + \beta_U\,\text{UrbArea00}_i + \\ \beta_{96T}\left(\text{Pop96}_i \times \text{Type}_i\right) + \varepsilon_i, \qquad i = 1,2,3,\ldots,n_a\end{aligned} \tag{5}$$

Table 2 presents the summaries for the results of the Monte Carlo experiments using 5000 simulations. We have calculated 95% confidence intervals for parameters and performance measures as well as the median values.

Table 2 – 95% confidence intervals for parameters and performance measures of the model described in Equation (5) (based on Monte Carlo experiments with 5000 simulations).

		2.5 Percentile	**Median**	**97.5 Percentile**
Estimates for the model parameters (based on the modeling dataset)	β_0	-763.24	-532.44	-238.02
	β_{96}	0.567	0.607	0.653
	β_T	-1.84	-1.49	-1.15
	β_1	0.051	0.063	0.075
	β_4	-0.024	-0.021	-0.017
	β_7	-0.011	-0.007	-0.002
	β_U	201.04	484.90	709.53
	β_{96T}	-0.526	-0.457	-0.387
Performance measures (based on the validation dataset)	**$\mathbf{R^2_{back}}$**	0.437	0.477	0.524
	MdRE	0.182	0.198	0.216
	TRE	-0.143	-0.102	-0.058

Since we have used the logarithm scale to transform the response variable, it is not trivial to interpret the values of the parameters estimates. Then, we are going to discuss the signal of the estimates. As expected, the correlations between population density in 2000 and the variables population density in 1996, percentage of urban area and reflectance in band 1 have been positive. Asphaltic surfaces and some roof materials have high reflectivity in band 1 (Forster, 1980). Moreover, band 1 is probably capturing the buildings shadows, since we have excluded tracts with large water bodies. As expected too, the correlation between population density in 2000 and the reflectance in band 4 has been negative, since the reflectance in this band is positively correlated with vegetation coverage.

Because of the interaction between population density in 1996 and tract type, we have to consider the tract type to interpret the effect of changing the population density in 1996 on the population density in 2000. Since the signal of interaction term is negative, the effect of changing the population density in 1996 is lower if the census tract is classified as "*favela*".

Having in mind that the reflectance in band 7 is positively related to built surfaces, we have expected a positive coefficient for reflectance in band 7. Then, we consider the negative value for the coefficient of band 7 quite intriguing. A possible reason for this negative correlation is that band 7 may be capturing the reflectance of bare ground and clay tiles (Harrison and Jupp, 1989). Since clay tiles are an indicator of the presence of houses, which have a lower habitant density compared to buildings, a rise in the reflectance in band 7 would have the effect of decreasing the density population of the census tract.

On the performance measures, R^2_{back} is calculated as the coefficient of linear correlation between the observed and the predicted values for the response, using the original scale. Considering the complexity of this prediction problem, the value for R^2_{back} of the final model is relatively good, especially if compared to the value for the model using only the population density in 1996 as explanatory variable (about 0.36).

Analysing the performance at the micro level (MdRE values), our proposal has improved the model in Reis (2005), which has got a median relative error equal to 0.304. Moreover, this value may be underestimating the true value, since the author has not used a validation dataset to calculate the performance measures. On the other hand, our proposal performance at the macro level (TRE values) is quite inferior to the performance of the model in Reis (2005) (TRE=0.0047), even if we consider the underestimation problem.

The interval of values for MdRE are quite similar to the value obtained (0.23) in the best model using the impervious surface (Lu et al., 2006). However, since the authors have not adopted interval estimates to assess their model performance, we can not make secure conclusions about the inferiority of our MdRE values compared to theirs.

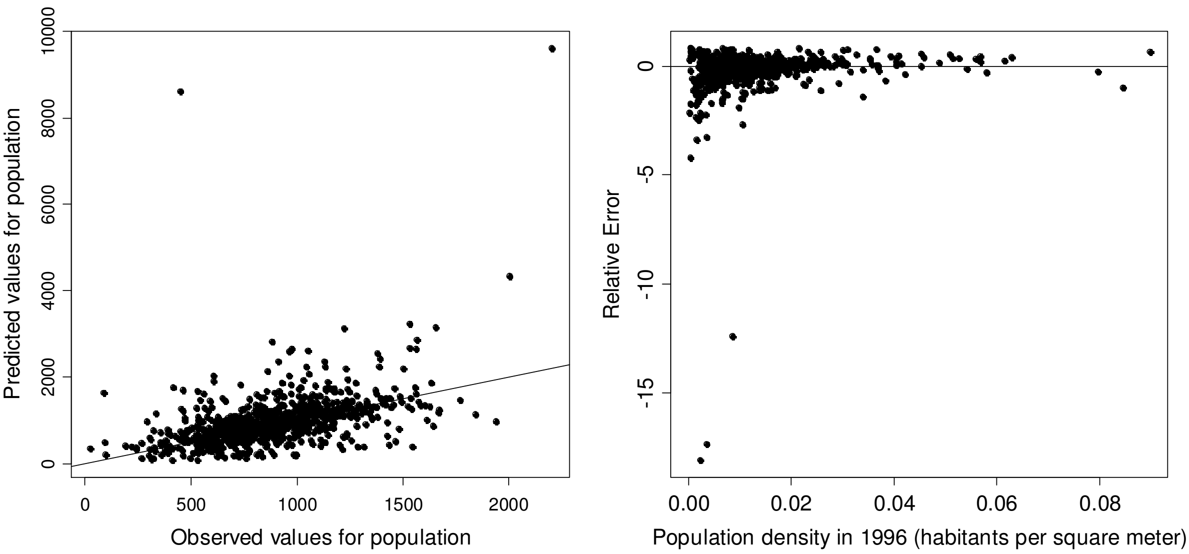
The estimate for TRE shows that the model tends to underestimate the overall population in the city. This is also observed by Lu et al. (2006), which have calculated TRE using the entire dataset and obtained a value equals to -0.0097.

Using more complex analysis (pixels of the image as analysis units instead of census tracts), Harvey (2002b) has obtained MdRE and TRE of 0.171 and -0.03, respectively. However, these values are not comparable to ours, since Harvey (2002b) have not used a validation dataset to calculate the performance measures of his models, which can produce underestimated values. Moreover, the authors have not taken account for variability in the estimation process and report only punctual estimates for the performance measures.

Figure 2(a) presents the relationship between the observed and predicted values for population of the census tracts in the validation dataset. There is a clear positive correlation between observed and estimated values, which has been also observed by Harvey (2002b) and Lu et al. (2006). Those tracts that are very distant of the points cloud can be identified in Figure 2(b). They are the ones with the highest relative errors and have low population density. This suggests that this kind of census tract should be treated in a special way in the modeling phase.

4. Conclusions

In this work, we have proposed a model to use remote sensing data to adjust the population estimates based on demographic data. In addition to population data of a previous period and the type of occupation of the census tract, we have used the data of satellite images to build a linear regression model. Moreover, we have evaluated our proposal taking account the variability present in the modeling process.



(a) **(b)**

Figure 2 – (a) Observed values for population of the census tracts and the respective values predicted for them using the model in Equation (5); (b) Population density in 1996 of census tracts and the relative error in the estimation of their population. These results have been obtained for one of the many possible selections for the validation dataset.

Especially at the micro level, our proposal has achieved results that are quite similar to the results of models using more complex forms or explanatory variables that require more complex treatments.

We believe that our proposal can be improved if we give different treatments to tracts with different population density. That is what we intend to do in a future work.

5. Acknowledgments

The second author was supported by CNPq/Brazil under PIBIC program. The authors thank to the Professor Renato Assunção (DEST/UFMG) for his excellent suggestions, which have contributed to improve this work.

6. References

Draper, N. R. ; Smith, H. **Applied Regression Analysis**, 3^{th} edition. John Wiley and Sons, EUA, 706 p, 1998.

Forster, B.C. Urban residential ground cover using Landsat digital data**, Photogrammetric Engineering & Remote Sensing**, vol. 46, pp. 547-558, 1980.

Harrison, B.A. ; Jupp, D.B.L. **Introduction to Image Processing**, CSIRO, Canberra, Australia, 256 p, 1990**.**

Harvey, J. T. Estimating census district populations from satellite imagery: some approaches and limitations. **International Journal of Remote Sensing**, vol. 23, n. 10, p. 2071-2095, 2002a.

___________ . Population estimation models based on Individuals TM Pixels. **Photogrammetric Engineering and Remote Sensing**, vol. 68, n. 11, p. 1181-1192, 2002b.

Iisaka J.; Hegedus, E. Population estimation from Landsat Imagery. **Remote Sensing of Environment**, v. 12, p. 259- 272, 1982.

Liu, X.; Clarke, K. ; Herold, M. Population Density and Image Texture: A Comparison Study. **Photogrammetric Engineering & Remote Sensing,** v. 72, n. 2, pp. 187–196, 2006.

Lo, C. P. Automated population and dwelling unit estimation from high-resolution satellite images: a GIS approach. **International Journal of Remote Sensin**g, v. 16, n.1, p. 2071-2095, 1995.

Lu, D.; Weng, Q.; Li, G. Residential population estimation using a remote sensing derived impervious surface approach. **International Journal of Remote Sensing**, vol. 27, n. 16, 3553–3570, 2006

Souza, I. M; Pereira, M. N.; Kurkdjian, M. L. N. O. Evaluation of high resolution satellite images for urban population estimation. In: International Symposium Remote Sensing of Urban Areas, 3., 2002, Istanbul, Turkey. **Proceedings** … Instanbul, 2002. Papers, CD-ROM,

R Development Core Team. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, 2003, Vienna, Austria. ISBN 3-900051-00-3, URL http://www.R-project.org

Reis, I. A. Estimação da população dos setores censitários de Belo Horizonte usando imagens de satélite. In: Simpósio Brasileiro de Sensoriamento Remoto (SBSR), 12., 2005, Goiânia. **Anais...** São José dos Campos: INPE, 2005. Artigos, p. 765-773. CD-ROM, On-line. ISBN 85-17-00018-8. Available at:
< http://marte.dpi.inpe.br/col/ltid.inpe.br/sbsr/2004/11.18.18.39/doc/2741.pdf>.

Vermote, E.F. ; Tanre, D. ; Deuzé, J.L. ; Herman, M., and Morcrette, J.J. Second simulation of the satellite signal in the solar spectrum, 6S: an overview. **IEEE Transactions on Geosciences and Remote Sensing**, vol. 35, n. 3, p. 675-686, 1997.