

Compatibilização de populações entre malhas censitárias diferentes com o uso de imagens de sensores orbitais

Ilka Afonso Reis¹

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

Abstract. Studying data attached to census tracts such as population along time often brings a difficulty: these areal units can change from a census to another one. A solution to this problem is to spatialize the tracts population of a census and reaggregate this spatialized data using the geographical definition for the tracts of the other census. Despite being simple, this solution needs to assume the spatial distribution of the population over the tract area is homogeneous. This paper proposes an alternative to make population data attached to different census tracts compatible. The idea is to spatialize the tracts population using orbital images by regression linear models. Since the involved pixels datasets are often big, a fraction of the pixels in each tract is sampled to be used in the regression model. We have studied the influence of this fraction on the quality of the population estimates attached to the new census tracts. Using the whole dataset to spatialize the population, the proposed procedure has overestimated the total population by 3.4% and the absolute error at the tracts level has been 1.9% (median value). For sample fractions lower than 75%, this approach tends to overestimate the total population (macro level) as well as the tract population (micro level). However, the greater the sample fraction, the smaller and more homogeneous the error at micro level. Therefore, when the most part of the dataset can be used, we consider the proposed procedure a good alternative to make population compatible over different tracts definitions.

Palavras-chave: linear regression, population data, remote sensing, regressão linear, dados populacionais, sensoriamento remoto.

1. Introdução

No Brasil, o recenseamento da população é responsabilidade do Instituto Brasileiro de Geografia e Estatística (IBGE) e acontece, geralmente, de dez em dez anos. Em cada recenseamento, o território brasileiro é dividido em setores censitários. Segundo o IBGE (2010), um *setor censitário* é “a unidade territorial estabelecida para fins de controle cadastral, formado por área contínua, situada em um único quadro urbano ou rural, com dimensão e número de domicílios que permitam o levantamento por um recenseador”. O conjunto de setores censitários é conhecido como *malha censitária*. Os dados coletados são divulgados de forma agregada por setor censitário.

Os dados censitários são a base para diversos estudos nas áreas de Epidemiologia, Sociologia, Economia, Geografia, Gestão Pública, Marketing, entre outras. Esses estudos frequentemente envolvem um componente temporal, para o qual a análise de dados coletados em diferentes censos é imprescindível. Em alguns casos, por exemplo, é importante saber qual é a população de um determinado setor em dois censos diferentes. Para isto, é preciso que os setores de malhas censitárias em diferentes censos sejam compatíveis. Essa compatibilização de dados seria trivial se a malha censitária não mudasse de um censo para outro.

No entanto, dependendo do crescimento do número de domicílios em uma região, a definição dos domicílios que compõem um setor censitário pode mudar de um levantamento para outro. Basicamente, essa mudança pode ser feita de três maneiras, ilustradas na Figura 1: situação (a), na qual um novo setor censitário é formado pela união de dois ou mais setores antigos; situação (b), quando um setor censitário é dividido em dois ou mais novos setores; ou, no caso (c), quando uma nova definição dos setores é construída sem necessariamente guardar vínculo com a definição anterior. No primeiro caso, a agregação de dados segundo os novos setores pode ser feita de maneira simples, por meio de somas ou médias ponderadas.

No entanto, no segundo e no terceiro casos, a compatibilização dos dados de malhas censitárias diferentes não é tão trivial.

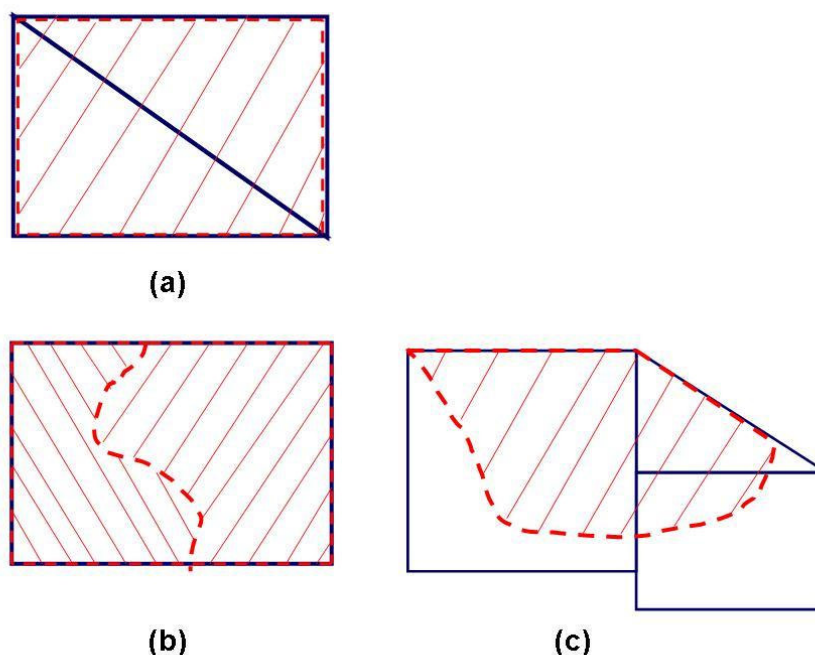


Figura 1. Definição de nova malha censitária (linhas pontilhadas em vermelho) e sua relação com a malha censitária antiga (linhas sólidas em azul). Caso (a): novos setores são definidos a partir da união de setores antigos. Caso (b): novos setores são definidos a partir da divisão de setores antigos. Caso (c): novos setores são definidos sem guardar relação com setores antigos.

Nos casos (b) e (c) da Figura 1, se pudermos supor que a distribuição espacial da população dentro um setor censitário é homogênea, isto é, sem concentrações em determinadas partes do setor, uma solução para a compatibilização da população de setores de duas malhas censitárias diferentes, digamos malha A e malha B, é espacializar a população dos setores da malha A em uma grade numérica de escala apropriada e reagregar essa população espacializada usando os setores da malha B. Essa foi a solução adotada em Reis *et al.* (2011) para criar os dados da variável “população no ano de 1996 por setor da malha censitária do ano de 2000”, que foi utilizada para prever a população por setor censitário no ano de 2000 para a cidade de Belo Horizonte (MG), juntamente com a utilização de imagens orbitais.

No entanto, essa suposição de homogeneidade na distribuição espacial da população dentro um setor censitário pode não traduzir adequadamente a realidade da maior parte dos setores de uma malha censitária. Uma alternativa à solução da espacialização simples da população vem do trabalho de Harvey (2002), que utilizou imagens do sensor TM/LANDSAT para estimar a população dos *collect districts*, equivalentes aos setores censitários brasileiros, em duas regiões da Austrália. Para isso, o autor distribuiu a população dos setores entre os pixels pertencentes a esses setores por meio de um modelo de regressão iterada, no qual as reflectâncias nas várias bandas do sensor TM foram utilizadas como variáveis explicativas (independentes). A solução de Harvey pode ser entendida como uma alternativa de espacialização mais informativa do que distribuir homogeneamente a população entre os pixels pertencentes a um setor. Nessa alternativa, as características de ocupação do setor captadas pelas imagens ajudam na distribuição da população ao longo do setor.

1.1 Objetivo

O objetivo deste trabalho é avaliar a utilização da espacialização de atributos via regressão com imagens orbitais (EI) na compatibilização da população dos setores no caso de mudanças na definição geográfica da malha censitária.

2. Materiais e Método

2.1 O banco de dados

Neste trabalho, foram utilizadas as malhas cartográficas do município de Belo Horizonte com as divisões em setores censitários utilizadas na contagem populacional de 1996 e no censo demográfico de 2000; a contagem populacional de 1996, por setor censitário; e as imagens das bandas 1 a 5 e 7 do sensor TM a bordo do LANDSAT5 (órbita/ponto 218/74 de 31/01/1996). As imagens sofreram correção atmosférica por meio do modelo 6S (*Second Simulation of Satellite Signal in the Solar Spectrum*, Vermote *et al.*, 1997).

Somente setores classificados como urbanos em 1996 (99.4%) foram utilizados. Depois de retirados setores considerados especiais (13) ou com menos de um (01) habitante por 10 m² (22), restaram 2060 setores censitários, que estavam classificados segundo o tipo de ocupação, de acordo com os critérios do IBGE: normal e aglomerado sub-normal (favelas).

O banco de dados de pixels foi construído com o uso do *software* SPRING (Câmara *et al.*, 1996), versão 5.2, e a regressão iterada foi realizada no ambiente de programação estatística R, versão 2.15.0 (R Development Core Team, 2012). Os pixels urbanos foram identificados por meio de classificação das imagens, do tipo supervisionada, por pixel, método da máxima verossimilhança, utilizando-se as bandas 3, 4 e 5. Ao final, 228957 pixels foram classificados como urbanos e cada um deles foi associado a um setor censitário, tanto na malha censitária de 1996 quanto na malha censitária de 2000.

As imagens orbitais, a malha censitária, bem como os dados associados a ela, todos referentes ao ano de 1996, foram os mesmos utilizados em Reis (2005).

2.2 A regressão iterada

Os modelos de regressão utilizados por Harvey (2002) e Reis (2005) têm a mesma forma de um modelo de regressão linear usual (Draper e Smith, 1998), apresentada pela equação

$$p_i = \beta_0 + \sum_{j=1}^k \beta_j r_{ij} + \varepsilon_i \quad (1)$$

na qual p_i representa a população do pixel i , r_{ij} é a reflectância do pixel i na j -ésima banda do sensor, $j = 1, 2, \dots, k$, β_0 e β_j são os coeficientes a serem estimados e ε_i representa a parcela da variabilidade da população do pixel i que não é explicada pelo modelo de regressão (erro do modelo).

Como estimativa inicial para a população do pixel p_i , foi utilizada a população do setor dividida pelo número de pixels englobados por ele. A partir dos seus valores iniciais, as estimativas de população por pixel foram refinadas iterativamente. A cada iteração, os coeficientes da equação de regressão em (1) eram estimados e, a partir deles, os valores estimados para os p_i eram obtidos. Para que o total da população do setor censitário fosse mantido como o total populacional conhecido do setor, os valores estimados para os p_i eram ajustados. A população ajustada do pixel i era dada pela soma da população estimada para o pixel i (\hat{p}_i) com a média dos resíduos ($\bar{\varepsilon}$) do setor censitário ao qual pertence o pixel i , que é

definido como $\bar{e} = \sum_{i=1}^n (p_i - \hat{p}_i) / n$, onde n é o número de pixels do setor ao qual pertence o pixel i . Essa população ajustada substituí o valor de p_i da iteração anterior e a regressão era ajustada novamente. As iterações se repetiram até que a qualidade do ajuste do modelo, medida pelo coeficiente de determinação (R^2) ou pela variância dos erros, não se alterasse significativamente (incremento relativo menor do que 0.01%).

Estimativas negativas para a população dos pixels podem acontecer. A solução adotada aqui foi aquela proposta por Harvey (2002), que consiste em transformar as estimativas negativas em zero e ajustar a população dos outros pixels para que o total da população do setor fosse mantido em seu valor conhecido.

Devido ao grande esforço computacional de se trabalhar com todo o banco de pixels, Harvey (2002) sugere o uso de uma amostra deles, no seu caso, 2% dos pixels de cada unidade de coleta. Trabalhando com unidades de coleta menores do que as de Harvey, Reis (2005) utilizou uma amostra de 25% dos pixels de cada setor censitário. Como o valor dessa fração amostral é arbitrário, o presente trabalho avaliou sua possível influência nos resultados finais, trabalhando com frações amostrais de 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 75% e 90%. Depois de amostrados os pixels de cada setor, uma fração da população do setor igual a que foi usada para selecionar os pixels era distribuída homogeneamente entre os pixels selecionados. Essa foi a população inicial a ser redistribuída entre os pixels do setor por meio da regressão iterada.

Devido ao problema de multicolinearidade entre as bandas da imagem, principalmente entre as bandas 2 e 3 (Harvey, 2002; Reis, 2005; Reis et al., 2011), o modelo de regressão foi ajustado sem essas duas bandas. Além disso, na tentativa de melhorar a capacidade preditiva do modelo, a variável que indicava o tipo de ocupação do setor ao qual pertencia o pixel (normal ou aglomerado sub-normal) também foi incorporada ao modelo.

2.3 Avaliação da qualidade das estimativas da população distribuída

A qualidade das estimativas da população de 1996, distribuída nos setores de 2000, foi avaliada por meio de duas medidas de erro: o Erro Relativo Total (ERT) e o Erro Relativo Mediano (ERM).

O ERT representa a variação do total estimado em relação ao total observado, sendo chamado *erro no nível macro* (área urbana) e é calculado por meio da expressão

$$ERT = \frac{\text{soma dos valores } \textit{preditos} \text{ para 1996 nos setores de 2000}}{\text{soma dos valores } \textit{observados} \text{ para 1996 nos setores de 1996}} - 1. \quad (2)$$

Valores positivos e negativos para o ERT indicam, respectivamente, superestimação e subestimação da população total na área urbana do município.

No caso de malhas censitárias que foram redefinidas segundo os casos (b) e (c) da Figura 1, não há como saber qual seria a verdadeira população dos setores dessa nova malha censitária em um ano anterior a ela. Sendo assim, para o cálculo do ERM, foram escolhidos 20 setores dentre aqueles que não sofreram mudanças da malha censitária de 1996 para a de 2000, ou que foram simplesmente agregados em 2000, como ilustra a Figura 1(a). Os setores foram escolhidos manualmente considerando-se a qualidade da sobreposição das duas malhas censitárias. O erro relativo para cada setor i escolhido foi calculado pela diferença entre o valor observado para a sua população (P_i) e o valor estimado pelo modelo (\hat{P}_i), dividida pelo valor observado, como explicitado na expressão

$$ER_i = \frac{\hat{P}_i - P_i}{P_i}, \quad i = 1, 2, \dots, n_{SC}, \quad (3)$$

onde n_{SC} é o número de setores utilizados na avaliação. O ERM é a mediana dos valores absolutos dos erros relativos ER_i e representa o *erro no nível micro* (setores).

Utilizando os setores que foram redefinidos por meio da simples divisão, como ilustrado na Figura 1(b), foi possível avaliar a diferença entre as estimativas da população pelo procedimento de espacialização simples (\hat{P}_i^{ES}) e aquelas obtidas pelo uso das imagens orbitais (\hat{P}_i^{EI}), definida pela expressão a seguir

$$Dif_i = \frac{\hat{P}_i^{EI} - \hat{P}_i^{ES}}{\hat{P}_i^{ES}}, \quad i = 1, 2, \dots, n_{SC}^* \quad (4)$$

onde n_{SC}^* é o número de setores utilizados na avaliação. Para o cálculo das diferenças, foram escolhidos 20 setores da malha censitária do ano de 1996 que haviam sido divididos na malha do ano de 2000 de modo a formar dois novos setores. Sendo assim, a avaliação da diferença foi feita com os dados de 40 setores da malha de 2000. Para resumir os valores das diferenças, utilizou-se o valor mediano. Nesse caso, não se pode falar em erro, visto que não se conhece a verdadeira população dos novos setores no ano de 1996.

3. Resultados e Discussão

A Tabela 1 apresenta os resultados do ERT, do ERM nos setores sem mudança e do intervalo interquartil (percentis 25 e 75) das diferenças nos setores divididos para cada fração amostral utilizada na regressão iterada, assim como os valores para o modelo que utilizou todos os pixels e o valor do coeficiente de determinação (R^2) da última iteração do modelo de regressão.

A análise dos resultados usando todo o banco de pixels na regressão iterada mostra que a redistribuição da população via imagens leva a uma pequena superestimação do total populacional (ERT=3.4%), com magnitude comparável ao valor obtido por Harvey (2002) na estimação da população de uma zona urbana sem o problema de trabalhar com setores diferentes (ERT=-4.8%). Quanto à magnitude do erro em um dado setor, o valor mediano dos erros absolutos foi de 1.9%, contra 14.5% de Harvey (2002) e 31.57% de Reis (2005).

A utilização de apenas uma fração dos pixels deteriora bastante os resultados, tanto no nível macro quando no nível micro, até o valor de 75% dos pixels, a partir do qual o erro na estimação da população total (ERT) e o erro no nível dos setores (ERM) melhoram consideravelmente e se aproximam bastante dos resultados obtidos com todo o conjunto de pixels.

A Figura 2 mostra a distribuição dos erros relativos segundo a fração amostral utilizada na regressão iterada. Pode-se notar que o efeito do aumento da fração amostral é o de estabilizar a distribuição dos erros relativos em torno do valor mediano, tornando-os mais homogêneos. No entanto, até o valor de 60% de pixels amostrados em cada setor, os valores dos erros são bastante altos e com tendência à superestimação, como já evidenciado pelos resultados apresentados na Tabela 1.

A diferença entre as estimativas utilizando a espacialização simples (ES) e a espacialização via regressão com imagens (EI), avaliada nos setores que sofreram divisão na redefinição da malha censitária, segue a mesma tendência dos erros relativos e se torna menor com o aumento da fração amostral. O exame dos intervalos interquartílicos mostra que não há uma tendência clara de sub ou superestimação de um método em relação ao outro, podendo o

método EI gerar estimativas maiores ou menores do que as do método EI (diferenças positivas ou negativas, respectivamente).

Tabela 1. Avaliação das estimativas da população de 1996 nos setores censitários de 2000 segundo a fração amostral utilizada na regressão iterada, considerando o nível macro (ERT) e o nível micro (ERM)

Fração Amostral	Erro Relativo Mediano (ERM)	Erro Relativo Total (ERT)	Intervalo interquartilício* das diferenças Dif_i	R ²
10%	0.396	0.229	-0.470 a 0.463	0.134
15%	0.336	0.232	-0.560 a 0.381	0.141
20%	0.380	0.231	-0.588 a 0.388	0.147
25%	0.286	0.238	-0.463 a 0.351	0.151
30%	0.284	0.250	-0.484 a 0.310	0.143
40%	0.280	0.259	-0.644 a 0.336	0.157
50%	0.268	0.259	-0.360 a 0.299	0.142
60%	0.232	0.268	-0.631 a 0.333	0.173
75%	0.028	0.028	-0.044 a 0.046	0.413
90%	0.018	0.031	-0.057 a 0.052	0.416
100%	0.019	0.034	-0.032 a 0.045	0.409

* Percentis 25 e 75

O valor do coeficiente de determinação (R²) mede o percentual da variabilidade da população do pixel que é explicada pelo modelo de regressão. Ele é baixo para os modelos que utilizaram até 60% dos pixels de cada setor, tornando-se razoável quando frações amostrais maiores são utilizadas. Vale notar que não são esperados valores do R² muito grandes em problemas desse tipo, visto que há muitos outros fatores influenciando a variabilidade da população do setor que não podem ser captados pelas imagens orbitais.

4. Conclusões

A compatibilização de dados entre malhas censitárias diferentes é importante para a análise de dados em diversos estudos. Neste trabalho, é proposto um procedimento para compatibilizar dados de setores definidos em malhas censitárias diferentes.

Para dados censitários que estejam correlacionados com o comportamento espectral de alvos captados por imagens orbitais, como é o caso da população, as reflectâncias dos alvos nas diversas bandas de um sensor podem ajudar na distribuição desses dados nas áreas dos setores antigos, utilizando-se a grade definida pelos pixels das imagens.

Neste trabalho, a espacialização da população nos pixels pertencentes aos setores é feita por meio de iterações de um modelo de regressão linear que tem as imagens orbitais como variáveis explicativas. Posteriormente, os valores assim distribuídos espacialmente são re-agregados utilizando-se as novas definições geográficas para os setores.

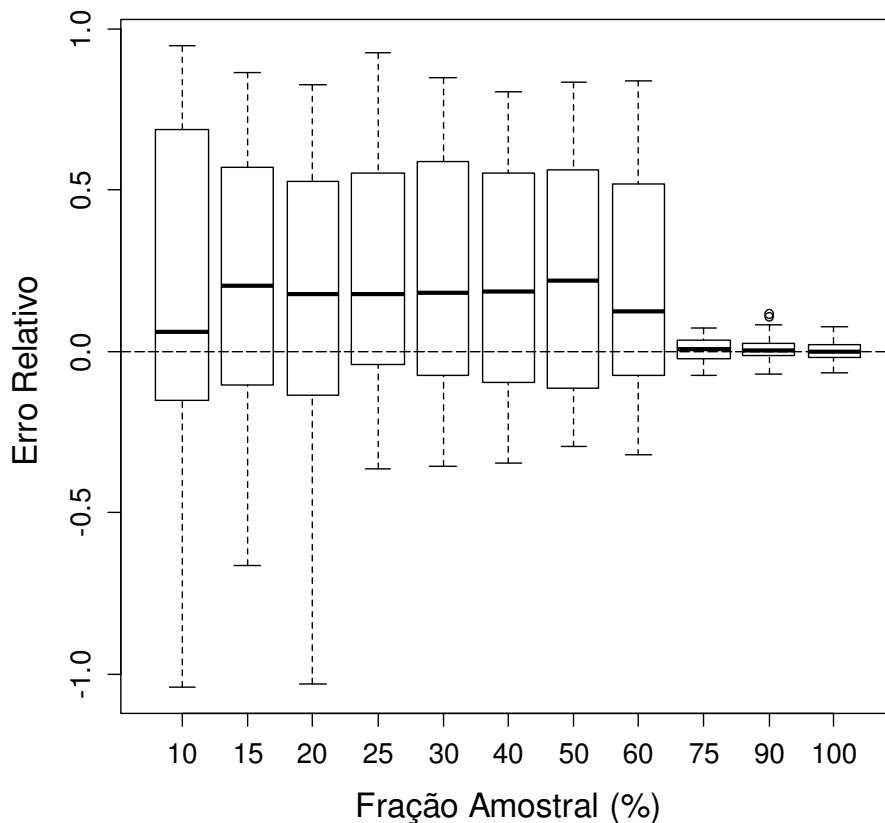


Figura 2. Distribuição dos erros relativos dos setores censitários que não sofreram mudanças segundo fração amostral. A linha sólida dentro das caixas marca a mediana dos erros relativos. A base e a tampa da caixa representam os percentis 25 e 75, respectivamente.

Utilizando-se uma fração de pixels maior ou igual a 75% no ajuste do modelo de regressão, os resultados podem ser considerados muito bons, tanto no nível macro (toda a população) quanto no nível micro (setores). Em ambos os níveis, houve uma pequena superestimação da população. Isso indica que o procedimento proposto consegue reproduzir bem os dados populacionais de setores nos quais não houve mudanças.

A diferença entre as estimativas populacionais utilizando-se procedimento da espacialização via regressão com imagens e o da espacialização simples foi avaliada nos setores que sofreram divisão na redefinição da malha censitária de um censo para outro. Essa diferença pode ser considerada pequena, de magnitude comparável à magnitude dos erros de estimação, indicando que os aspectos captados pelas imagens orbitais não causam grandes diferenças na espacialização da população quando ocorre a simples divisão dos setores. Isto pode acontecer porque os setores divididos acabam sendo muito parecidos como o setor original em termos das características que podem ser captadas pelas imagens.

Nos setores que são redefinidos sem nenhuma relação com os setores da malha antiga, não foi possível fazer nenhum tipo de avaliação, visto que o valor da população em um censo, caso os setores fossem os da malha de outro censo, não é conhecido. Esses valores só poderiam ser conhecidos se os dados populacionais fossem divulgados no nível do domicílio e os domicílios fossem georeferenciados, permitindo a agregação dos dados em qualquer definição da malha censitária.

Setores redefinidos sem guardar relação com os setores da malha antiga são os mais problemáticos para se aplicar a espacialização simples, pois a redefinição desses setores provavelmente está relacionada à nova configuração de ocupação das áreas. Os novos setores podem, por exemplo, conter áreas anteriormente não habitadas na configuração antiga. Assim, não seria razoável supor que parte da população do setor está nessas áreas, que é a suposição por trás do método da espacialização simples.

No caso desses setores e também para aqueles formados pela simples divisão dos setores antigos, a espacialização via imagens orbitais seria mais recomendada para viabilizar a compatibilização de dados populacionais entre malhas censitárias diferentes.

Agradecimentos

A autora agradece à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro na participação do XVI SBSR e também à Professora Edna Afonso Reis, do Departamento de Estatística da UFMG, pela revisão crítica do texto deste trabalho.

Referências

Câmara, G.; Souza, R.C.M.; Freitas, U. M.; Garrido, J. C. P. SPRING: Integrating Remote Sensing and GIS with Object-Oriented Data Modelling. **Computers and Graphics**, v.15, n.6, p.13-22, 1996.

Draper, N. R. ; Smith, H. **Applied Regression Analysis**, 3a. edição. John Wiley and Sons, EUA, 706 p, 1998.

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

IBGE. Guia do Censo 2010 para Jornalistas. Disponível em http://www.ibge.gov.br/home/presidencia/noticias/pdf/Guia_do_censo2010.pdf. Acesso em 15.out.2012.

R Development Core Team (2012). R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org/>.

Reis, I. A. Estimção 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. Disponível em: < <http://marte.dpi.inpe.br/col/ltid.inpe.br/sbsr/2004/11.18.18.39/doc/2741.pdf>>.

Reis, I. A.; Silva, V. L.; Reis, E. A. Adjusting population estimates using satellite imagery and regression models. In: Simpósio Brasileiro de Sensoriamento Remoto (SBSR), 15., 2011, Curitiba. **Anais...** São José dos Campos: INPE, 2011. Artigos, p. 830--837. CD-ROM, On-line. ISBN 978-85-17-00056-0. Disponível em : < <http://urlib.net/dpi.inpe.br/marte/2011/07.15.14.49>>.

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 Trans. Geosc. Remote Sens**, vol. 35, n. 3, p. 675-686, 1997.