Estimação em Pequenas Áreas para o Acesso às Tecnologias de Informação e Comunicação na Pesquisa TIC Domicílios

Carolina Valani Cavalcante
Dra. Denise Britz do Nascimento Silva

10 de novembro de 2016

Instituto Brasileiro de Geografia e Estatística
Escola Nacional de Ciências Estatísticas
Programa de Pós-Graduação em População, Território e Estatísticas Públicas

Sumário

Introdução

Introdução

- Presença cada vez mais intensa da internet na vida da população colocou o acesso à rede como requisito primordial nas atividades cotidianas.

- Ocorreram impactos na forma de comunicação entre as pessoas e nas relações econômicas entre empresas e países ao redor do mundo, criando uma complexa rede de interação.

- Propagação das novas tecnologias não aconteceram da maneira uniforme no planeta, existem pessoas que ainda estão à margem da nova forma de conexão.

- É importante conhecer e levantar dados que ilustrem a distribuição digital e auxiliem na criação de políticas voltadas para o acesso e utilização das Tecnologias de Informação e Comunicação (TIC).

Motivação e objetivo

Conforme a última pesquisa TIC Domicílios divulgada pelo CETIC.br (CGI,2015):

- 50% dos domicílios de todo o país têm acesso a computador/ internet;
- 57% dos domicílios da região Sul e 59% na região Sudeste possuem computador;
- 33% dos domicílios da região Norte e 37% na região Nordeste possuem computador;
- 98% dos domicílios da classe A e 14% nas classes D e E possuem acesso a internet.

Objetivo geral

Produzir estimativas para o número de domicílios com acesso ao computador e número de domicílios com acesso à internet em todas as UFs entre 2011 e 2014.

Tecnologia de informação e comunicação

Tecnologia de informação e comunicação

- É essencial que os dados referentes às TICs sejam de fácil compreensão, confiáveis e comparáveis, de maneira que ajude o governo na tomada de decisões para políticas públicas (UIT,2014).
- As estatísticas sobre este assunto ainda são limitadas, tanto na qualidade como na disponibilidade.

Duas as principais fontes brasileiras de informação sobre a utilização e acesso às TICs:

- Suplementos da PNAD, realizada pelo IBGE.
- Pesquisa TIC Domicílios, realizada pelo CETIC.br.

Dados utilizados

Importante ressaltar que os dados utilizados na dissertação foram cedidos pelo CETIC.br como parte do acordo de cooperação entre a ENCE e o NIC.br.

PNAD

- A PNAD é realizada anualmente pelo IBGE e coleta informações sobre os indivíduos e domicílios, além de suplementos que abordam temas específicos.
- Utilização de internet e posse de telefone móvel foram investigadas como suplemento em 2005, 2008, 2011, 2013 e 2014.
- Em 2009 as variáveis mais usuais sobre TIC foram incluídas no questionário básico da PNAD, entre elas a posse de telefone móvel, a posse de computador e o acesso à internet no domicílio.
- As variáveis coletadas não são suficientes para as análises necessárias sobre o desenvolvimento do tema no país.
- As informações levantadas a partir dos suplementos não possuem periodicidade definida.

TIC Domicílios

- A pesquisa TIC Domicílios é realizada desde 2005 pelo CETIC.br e divulga seus resultados anualmente com o objetivo de retratar a distribuição e acesso das TICs nos domicílios.

- Os dados obtidos e os indicadores calculados por esta pesquisa abrangem mais aspectos sobre a utilização e acesso às TICs do que a PNAD. Alguns exemplos são: motivos para a ausência de computador ou/e conexão à internet no domicílio, gasto com a conexão e tipos de atividades realizadas na internet.

- Os indicadores coletados pela TIC Domicílios permitem ao pesquisador realizar análises profundas sobre a distribuição das TICs no país.

- As estimativas só possuem boa precisão para o nível geográfico das grandes regiões.

Estimação em pequenas áreas

Estimação em pequenas áreas

- A Estimação em Pequenas Áreas, do inglês *Small Area Estimation* (SAE), abrange métodos e modelos para a produção de estimativas amostrais confiáveis para áreas geográficas, ou subgrupos, que apresentam limitação nos dados e para os quais não é possível produzir estimativas diretas com qualidade.

- Uma área, ou domínio, é considerada(o) "pequena(o)" quando não apresenta uma amostra com tamanho suficiente para possibilitar o cálculo de estimativas diretas com precisão adequada.

- As pequenas áreas não estão relacionadas somente a uma delimitação geográfica menor, também podem se referir a diferentes domínios da população como idade, sexo, entre outros (RAO,2003).

- A principal abordagem dos métodos de SAE é "pegar força emprestado" de informações da área de interesse, ou adjacentes e, a partir dessas variáveis auxiliares, ajustar um modelo para estimar as quantidades de interesse.

Estimação em pequenas áreas - Modelo Fay-Herriot

Modelos de área são amplamente utilizados, principalmente quando as informações existentes para as covariáveis são apenas em nível de área. Para um conjunto de covariáveis $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})'$ para a área i, o modelo é definido por:

Modelo Fay-Herriot

$$\begin{aligned} \hat{Y}_i &= Y_i + e_i, & e_i | Y_i &\sim N(0, \sigma_i^2), \\ Y_i &= \boldsymbol{x}_i^{'}\boldsymbol{\beta} + u_i, & u_i &\sim N(0, \sigma_u^2). \end{aligned}$$

Estimação em pequenas áreas - Fay-Herriot espaço-temporal

Diversas extensões do modelo de Fay-Herriot foram realizadas, em particular o modelo proposto por Marhuenda, Molina e Morales (2013) assume uma estrutura temporal e leva em consideração a correlação espacial entre os vizinhos. O modelo é expresso da seguinte forma:

Fay-Herriot espaço-temporal

$$\hat{Y}_{it} = \boldsymbol{x'}_{it}\boldsymbol{\beta} + u_{1i} + u_{2it} + e_{it}.$$

Estimação em pequenas áreas - Fay-Herriot espaço-temporal

O vetor de efeitos de área $(u_{11}, \cdots, u_{1D})'$ segue um processo SAR(1) com variância σ_1^2, de autocorrelação espacial ρ_1 e de matriz de proximidade padronizada $\boldsymbol{W} = (w_{i,l})$, e é dado por:

$$u_{1i} = \rho_1 \sum_{l \neq i} w_{i,l} u_{1l} + \epsilon_{1i}, \qquad |\rho_1| < 1, \epsilon_{1i} \sim N(0, \sigma_1^2).$$

O vetor de efeitos aleatórios de área-tempo $(u_{2i1}, \cdots, u_{2iT})'$ são independentes e identicamente distribuídos em cada área i. Estes efeitos seguem um processo AR(1) com parâmetro ρ_2 de autocorrelação e são definidos por:

$$u_{2it} = \rho_2 u_{2i,t-1} + \epsilon_{2it}, \qquad |\rho_2| < 1, \epsilon_{2it} \sim N(0, \sigma_2^2).$$

- A precisão das estimativas é calculada por meio do Erro Quadrático Médio (EQM), do inglês *Mean Squared Error* - (MSE). Dentre os métodos mais utilizados estão o de Prasad e Rao (1990) e Molina, Salvati e Pratesi (2009).

Análise descritiva

Análise descritiva - Variáveis de interesse

Os indicadores existentes permitem avaliar a distribuição das tecnologias entre os indivíduos e a presença delas nos domicílios e para isso são utilizados dois termos que ajudam a caracterizar essa investigação: uso e acesso.

Dentre os indicadores existentes, foram escolhidas as seguintes variáveis como as variáveis de interesse para produzir estimativas e assim avaliar o acesso de TICs nos domicílios. São elas:

- Número de domicílios com acesso ao computador;
- Número de domicílios com acesso a internet (exceto celular).

Análise descritiva - Variáveis de interesse

Distribuição das estimativas diretas (em milhares).

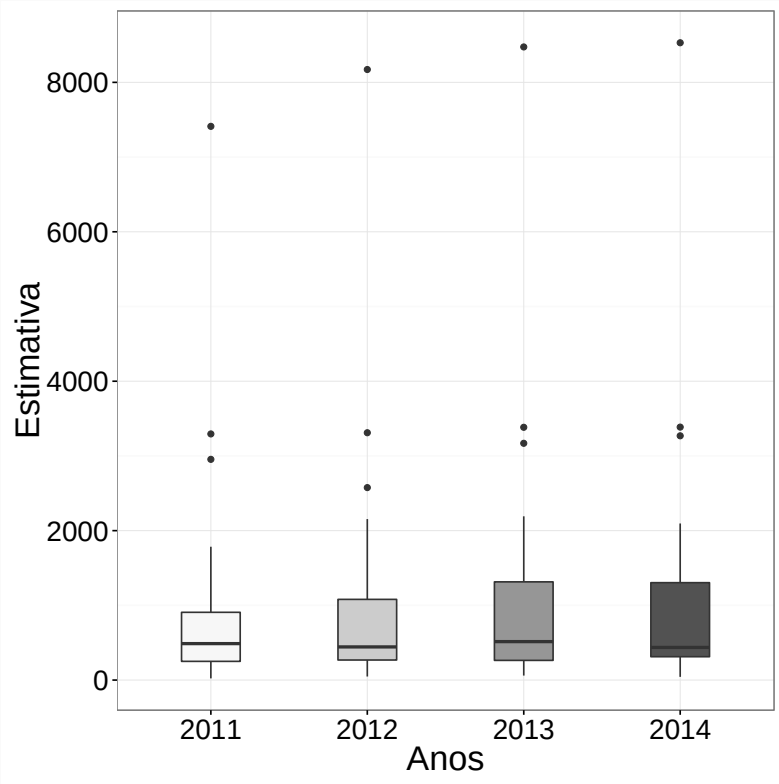


Figura 1: Acesso a computador.

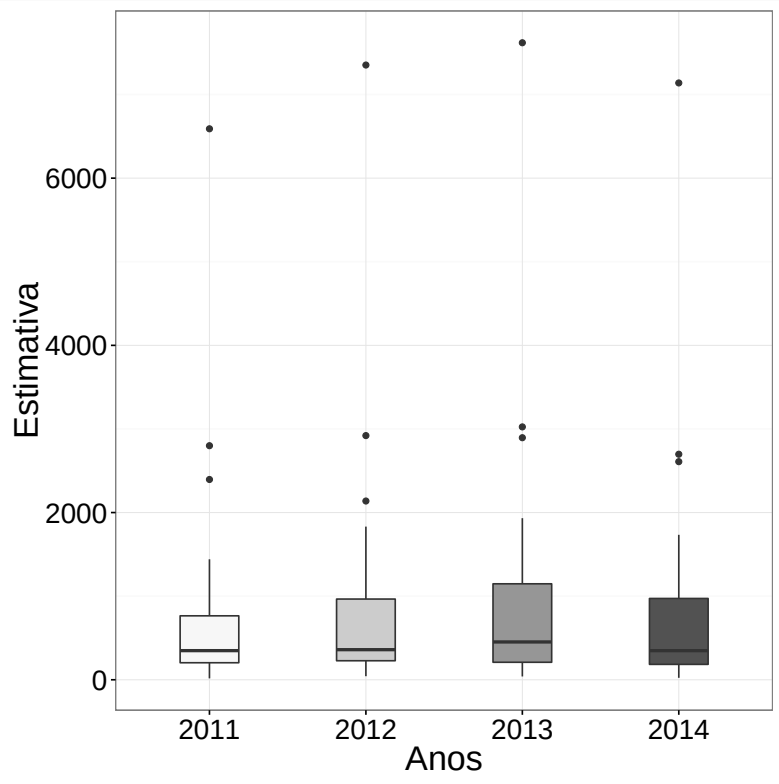


Figura 2: Acesso a internet.

Análise descritiva - Variáveis de interesse

Distribuição dos coeficientes de variação das estimativas diretas.

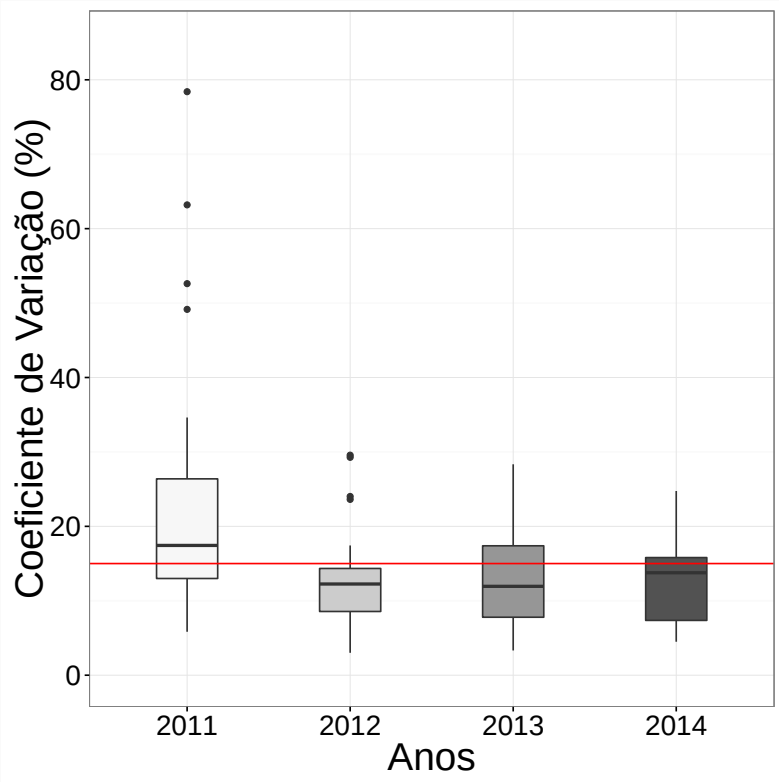


Figura 3: Acesso a computador.

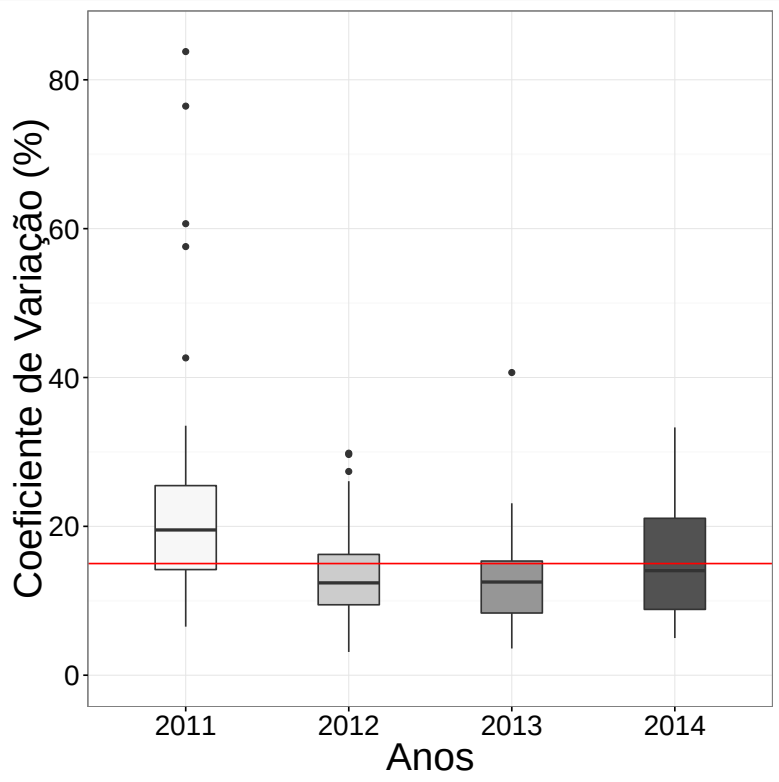


Figura 4: Acesso a internet.

Análise descritiva - Variáveis auxiliares

- As variáveis auxiliares devem ser escolhidas de forma que sejam relevantes para explicar o fenômeno a ser estudado e tenham altas correlações com as variáveis de interesse.
- As variáveis foram escolhidas com base em informações relacionadas à situação econômica das UFs, escolaridade dos indivíduos, distribuição de internet nos municípios das UFs, além de variáveis sociodemográficas.
- Para a escolha das variáveis auxiliares que seriam utilizadas no modelo de pequenas áreas foram ajustados modelos lineares simples e múltiplos.

Conforme os resultados foram utilizadas as seguintes variáveis auxiliares:

- Número de instituições de ensino superior,
- Taxa de distorção série-idade no ensino fundamental, e
- Receita bruta em serviços de TIC *per capita*.

Análise descritiva - Variáveis auxiliares

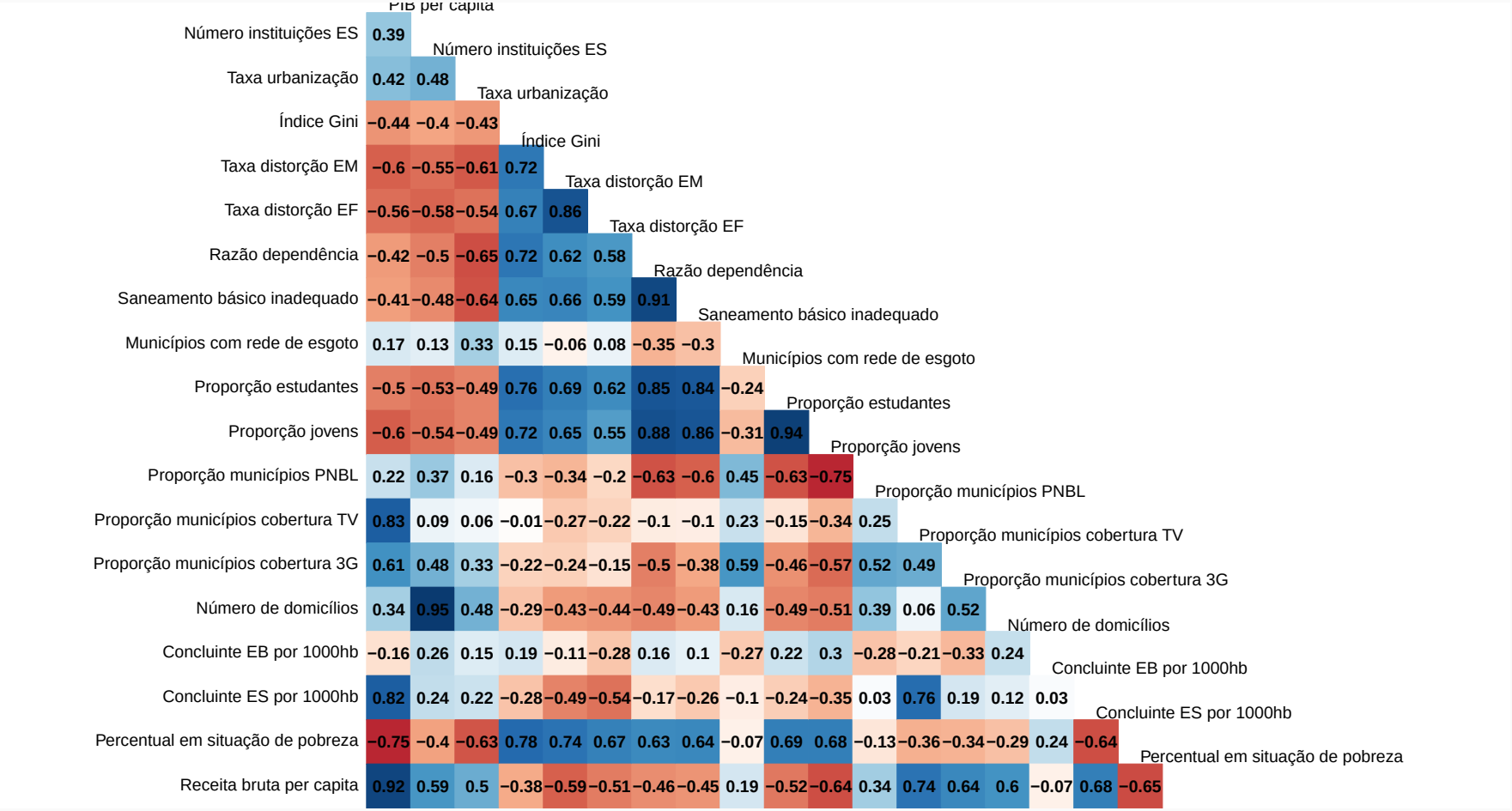


Figura 5: Ilustração da matriz de correlação entre as variáveis auxiliares - 2014.

Modelo para estimação em pequenas áreas

Modelo para estimação em pequenas áreas

- Apesar das variáveis utilizadas serem significativas também para o modelo de pequenas áreas, os resultados obtidos a partir do ajuste do modelo básico de Fay-Herriot não mostraram melhoria na precisão.

- Para aprimorar as estimativas em pequenas áreas, decidiu-se utilizar o modelo espaço-temporal, que utiliza, além das estimativas diretas, a distribuição do fenômeno no espaço e as diferentes observações ao longo do tempo.

- Para a avaliação da dependência espacial foi calculado o I de Moran que indicou a existência de tal dependência.

- Para o ajuste dos modelos foi utilizado o pacote **sae** existente no R.

Modelo para estimação em pequenas áreas

Distribuição espacial das estimativas diretas para a variáveis de interesse em cada UF.

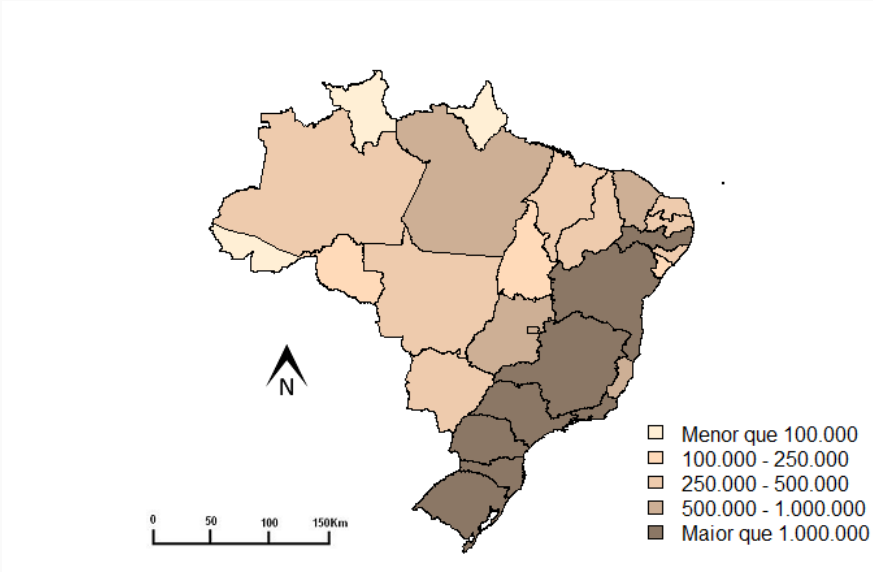


Figura 6: Acesso a computador.

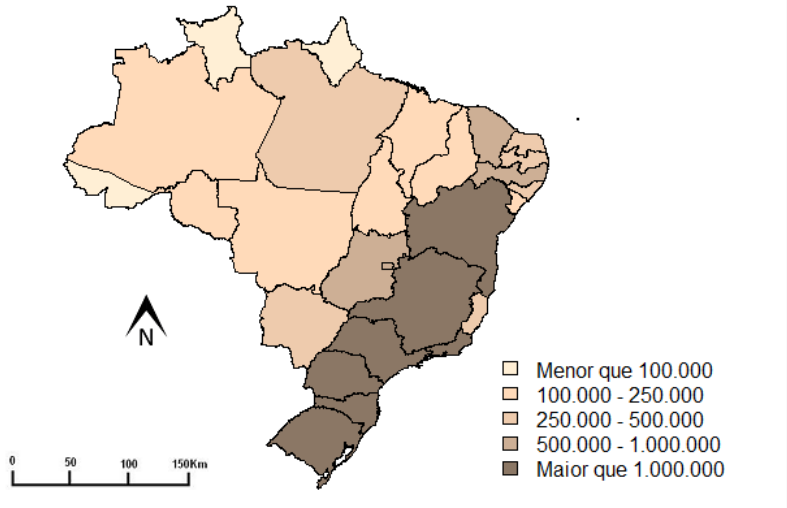
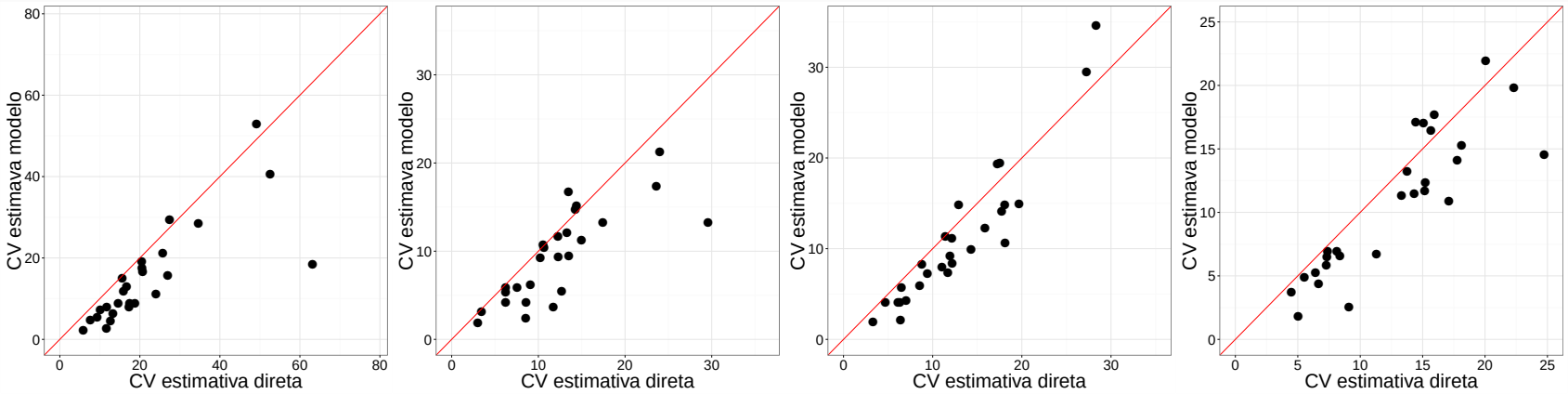


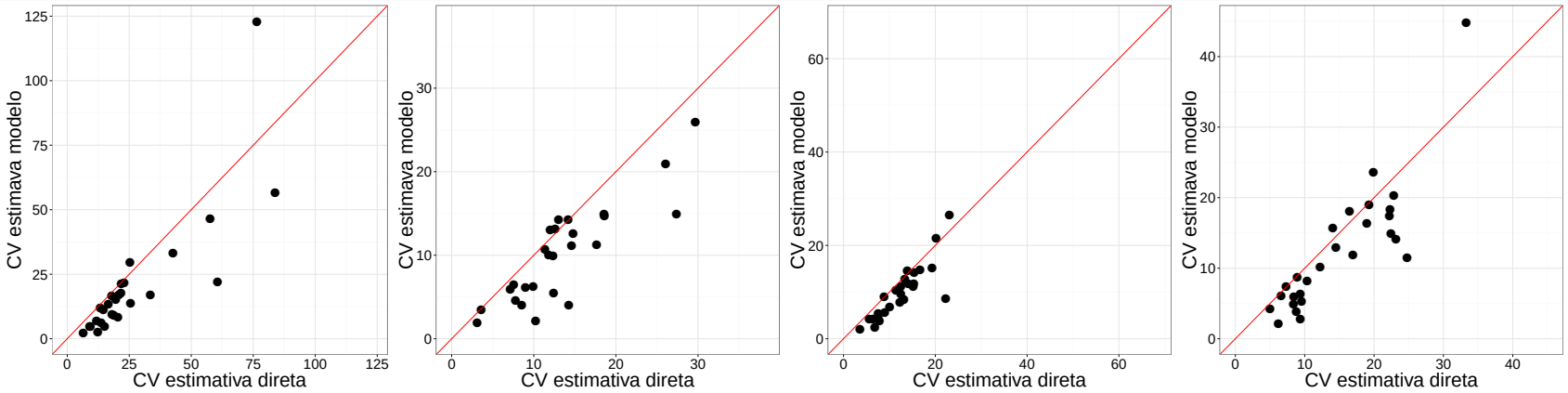
Figura 7: Acesso a internet.

Modelo para estimação em pequenas áreas

Gráfico de dispersão entre os CVs das estimativas diretas e do modelo.



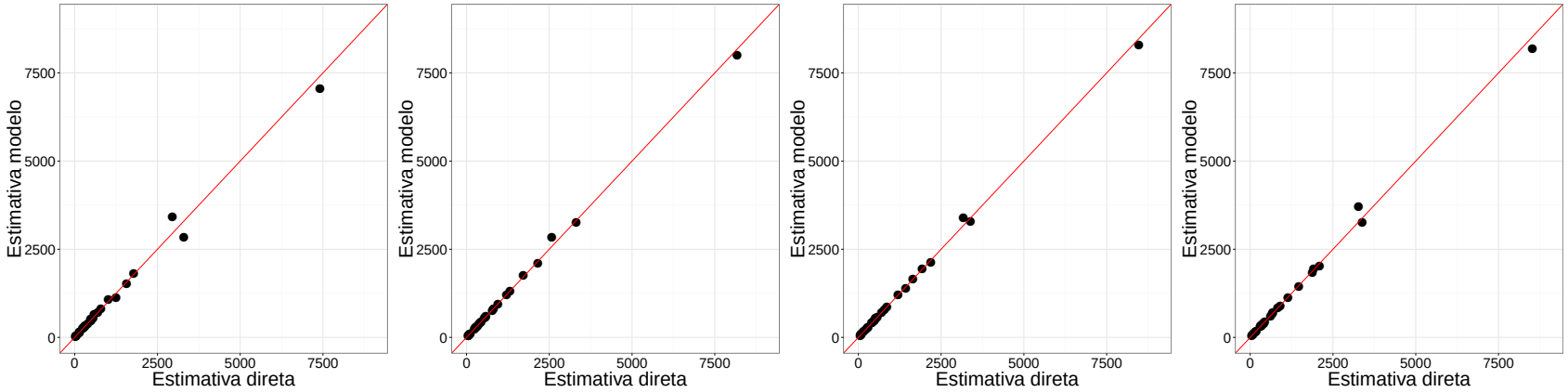
Acesso a computador - 2011 a 2014.



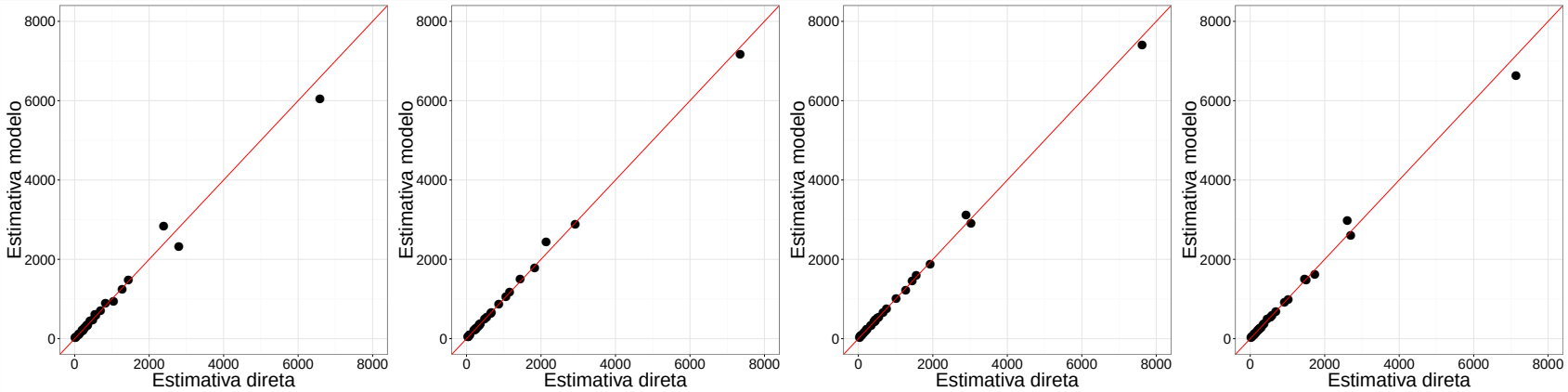
Acesso a internet - 2011 a 2014.

Modelo para estimação em pequenas áreas

Gráfico de dispersão entre as estimativas diretas e do modelo.



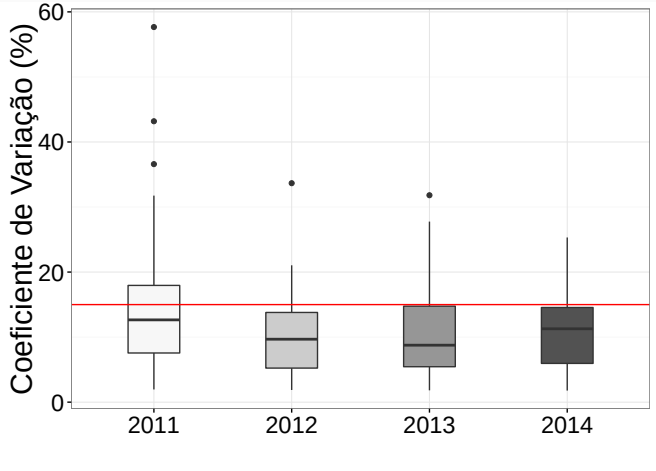
Acesso a computador - 2011 a 2014.

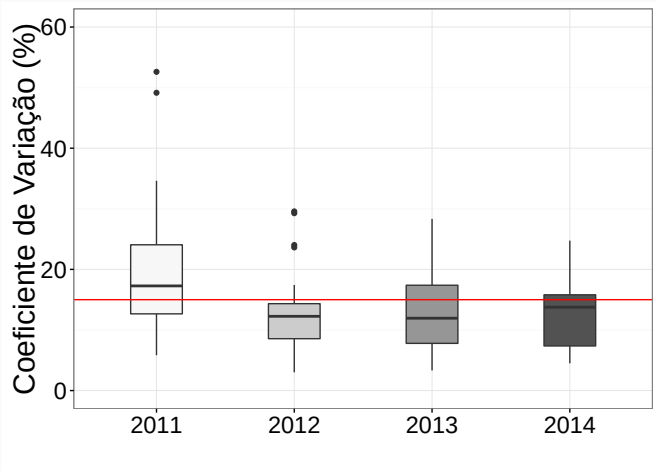


Acesso a internet - 2011 a 2014.

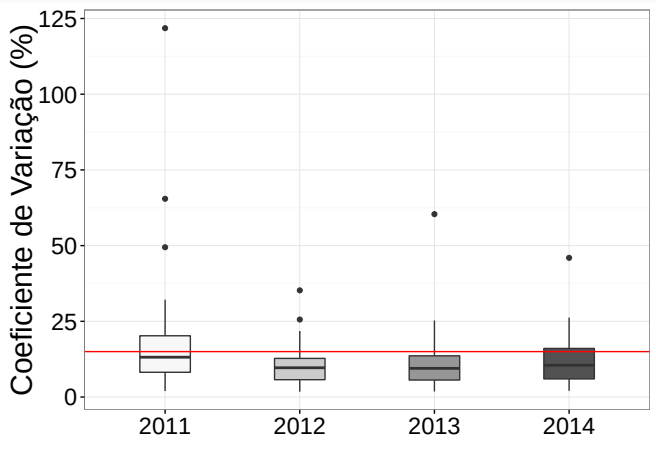
Modelo para estimação em pequenas áreas

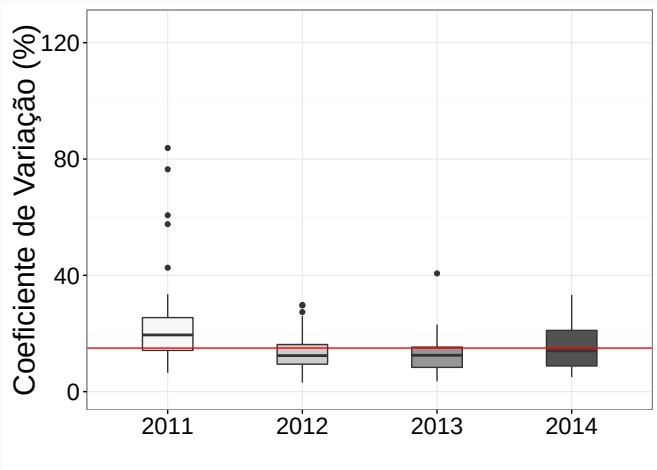
Distribuição do coeficiente de variação da estimativa do modelo e direta.





Acesso a computador - 2011 a 2014.





Acesso a internet - 2011 a 2014.

Conclusão

Conclusões e trabalhos futuros

- O modelo que incorpora a distribuição espaço-temporal do fenômeno de interesse na estimação apresentou melhor precisão para nova estimativa, e a precisão encontrada é boa o suficiente para possível publicação.
- Grande dificuldade para encontrar variáveis auxiliares que não apresentassem erros amostrais e fossem relacionadas ao tema.
- Outras variáveis de interesse foram testadas (número de domicílios com banda larga, proporção de domicílios com computador e internet).
- Utilização de modelos que sejam mais flexíveis em relação a escolha das variáveis auxiliares, isto é, que permitam o uso de variáveis com erro amostral.
- Utilização de modelos de caudas pesadas e modelo multivariado.

[7, 5, 3, 1, 2, 6, 4]

Referências I


CGI.
Pesquisa sobre o uso de tecnologias da informação e da comuniação no brasil 2014: Tic domicílios e tic empresas.
Technical report, Comitê Gestor da Internet da Internet no Brasil, São Paulo, 2015.


Y. Marhuenda, I. Molina, and D. Morales.
Small area estimation with spatio-temporal fay–herriot models.

Computational Statistics & Data Analysis, 58:308–325, 2013.


I. Molina and Y. Marhuenda.
sae: An r package for small area estimation.
R Journal, Under revision, 2015.

Referências II


I. Molina, N. Salvati, and M. Pratesi.
Bootstrap for estimating the mse of the spatial eblup.
Computational Statistics, 24(3):441–458, 2009.


D. Pfeffermann et al.
New important developments in small area estimation.
Statistical Science, 28(1):40–68, 2013.


N. Prasad and J. Rao.
The estimation of the mean squared error of small-area estimators.
Journal of the American statistical association, 85(409):163–171, 1990.


J. N. Rao.
Small area estimation.
Wiley Online Library, New York, 2003.