

MÉTODOS DE SELEÇÃO E VALIDAÇÃO DE MODELOS DE REGRESSÃO VIA *BOOTSTRAP*

Adriano Soares Koshiyama
UFRRJ – Rodovia BR 465 Km 7, CEP 23890-000 – Seropédica/RJ
as.koshiyama@gmail.com

Maria Cristina Affonso Lorenzon
UFRRJ – Rodovia BR 465 Km 7, CEP 23890-000 – Seropédica/RJ
lorenzon@ufrj.br

Wagner de Souza Tassinari
UFRRJ – Rodovia BR 465 Km 7, CEP 23890-000 – Seropédica/RJ
tassinari@ufrj.br

Resumo: Em estudos de fenômenos quantitativos onde se pretende obter respostas para causas e efeitos, o uso de modelos de regressão é o mais indicado. Um dos critérios mais utilizados na seleção do modelo é o Critério de Informação de Akaike (*AIC*). Apesar de tal critério trazer bons resultados metodológicos, este carece de um teste de hipótese como forma de avaliar a significância da escolha do modelo mais parcimonioso. O uso de técnicas de *bootstrap* visam preencher essa lacuna, como forma de justificar a decisão da escolha por um modelo. O objetivo deste trabalho é ajustar e validar o modelo mais adequado, selecionado a partir do uso das técnicas de *bootstrap*. Modelos espaciais e lineares foram ajustados para avaliar os fatores que influenciam o preço do mel no Brasil. As estimativas do *AIC*, geradas através do método *bootstrap*, demonstraram que o modelo mais adequado foi o de aditivos generalizados (*GAM*), sendo o único a apresentar uma diferença significativa do *AIC* dentre os demais modelos.

Palavras-Chave: Modelos de Regressão, *Bootstrap*, *AIC*, Modelo de Aditivos Generalizados

Abstract: In the study of quantitative phenomena that the objective is to seek responses for causes and effects, the use of regression models are the best choice. One of the most applied criterion in model selection is the Akaike Information Criterion (*AIC*). Therefore the good methodological results, this has a lack in hypothesis test as a way to evaluate the significance of choose of the most parsimonious model. The bootstrap techniques can be used to measure the validity and justify the decision to select a model. The objective of this work is to fit and validity the most parsimonious model, selected through the use of bootstrapping techniques. Spatial and linear models were fitted to identify the influence factors of honey prices in Brazil. The *AIC* estimates, generated from bootstrap method, demonstrated that the most adequate model was the generalized additive model (*GAM*), where the only that show *AIC* significantly difference between the others models.

Keywords: Regression Models, *Bootstrap*, *AIC*, Generalized Additive Models

1. Introdução

Em estudos de fenômenos quantitativos onde se pretende respostas para causas e efeitos, o uso de modelos de regressão é o mais indicado. Mas, dentre uma enorme gama de modelos de regressão, escolher o mais parcimonioso requer bons critérios de qualidade do ajuste do modelo. Uma das medidas mais utilizadas é o Critério de Informação de Akaike (*AIC*) (AKAIKE, 1974).

O *AIC* é baseado no conceito de entropia da informação, tal critério oferece uma medida relativa da informação perdida quando um modelo é usado para descrever a realidade. Este descreve principalmente o *trade-off* entre precisão e complexidade do modelo (TAKEUCHI, 1976). As principais vantagens do *AIC* residem na possibilidade de comparar modelos com diferentes famílias de distribuições (modelos não aninhados) e por não precisar de mais inferências sobre o modelo para corroborar em seu resultado (BURNHAM e ANDERSON, 2004). Além disso, é de fácil avaliação, pois o melhor modelo é o que possui o menor valor do *AIC*. Este critério tem sido amplamente usado em diferentes áreas do conhecimento, como biometria, econometria, pesquisa operacional, séries temporais, regressão espacial, entre outras (BURNHAM e ANDERSON, 2002).

Apesar das vantagens na utilização do *AIC*, este possui algumas lacunas, como a impossibilidade de comparar modelos com variáveis respostas em transformações distintas (logaritmo, exponencial, linear, etc.) e com dados desbalanceados. Fora esses, o de maior relevância é a impossibilidade de realizar um teste de hipótese, pois *AIC* é um critério de informação e não um teste estatístico onde seria possível observar se o *AIC* de um modelo é significativamente menor do que outro modelo (CLAESKENS e HJORT, 2008). O método baseado em simulação e reamostragem *Bootstrap* (EFRON e TIBSHIRANI, 1993), pode ser usado para fazer tal avaliação dos modelos estatísticos.

Esta técnica foi introduzida por Efron (1979) como abordagem alternativa para a estimação intervalar de parâmetros, em circunstâncias onde outras técnicas paramétricas tradicionais não são aplicáveis, na maioria das vezes quando o número de amostras é reduzido. A técnica de *bootstrap*, em geral, tenta realizar o que seria desejável realizar na prática: a repetição do experimento. O funcionamento dessa ferramenta (figura 1) é baseado nas *B* reamostragens das observações originais do banco de dados em estudo. Tais reamostragens são realizadas de forma aleatória, com reposição e as estimativas de cada calculadas. Baseado nas replicações é possível determinar intervalos de confiança e teste de hipóteses sobre os estimadores gerados.

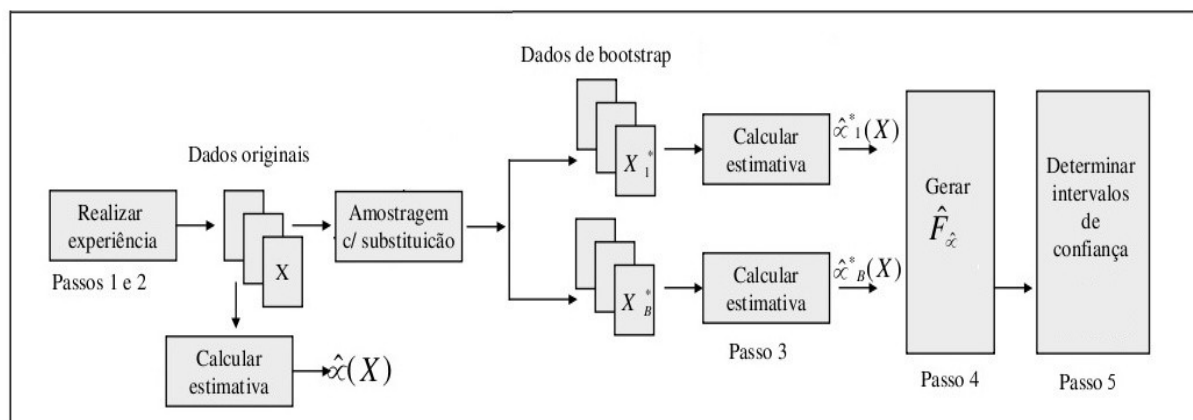


Figura 1. Procedimento de *bootstrap* para realização de reamostragens de uma determinada amostra, baseado em Efron (1979).

A utilização do *bootstrap* visa gerar um intervalo de confiança e, conseqüente teste de hipótese para o *AIC*, pois destas *B* réplicas de cada modelo podemos realizar inferências

através dos estimadores gerados (média, mediana, variância, etc...) (SHAO e TU, 1995). Portanto esta técnica preencherá a lacuna do *AIC*, com a possibilidade de demonstrar se um determinado *AIC* é significativamente menor, conseqüentemente possibilitando a escolha correta do modelo mais parcimonioso.

Utilizando modelos de regressão linear e espacial presentes em Koshiyama et al. (2011) para o estudo dos fatores que influenciam o preço do mel no Brasil, aplicaremos técnicas de *bootstrap* com o intuito de selecionar e validar o melhor modelo de regressão. O objetivo deste trabalho é ajustar e validar o modelo mais parcimonioso, selecionado a partir do uso das técnicas de *bootstrap*.

2. Metodologia

O estudo foi baseado nos modelos de regressão linear e espaciais, construídos para o estudo dos fatores associados na variabilidade do preço do mel no Brasil, contidos em Koshiyama et al. (2011). Esse estudo correspondeu uma análise prática, usando dados do Censo Agropecuário de 2006 (IBGE, 2006), distribuídos em 136 mesorregiões do Brasil.

Foram utilizadas técnicas de *bootstrap* não-paramétrico para realizar replicações das estimativas provenientes dos modelos de regressão. Após feitas as replicações, aplicamos testes não-paramétricos com intuito de verificar a existência de diferença significativa entre os *AIC* dos modelos.

2.1 Modelos de Regressão

A ferramenta estatística que estuda o relacionamento existente entre duas ou mais variáveis, de maneira que uma delas possa ser descrita ou predita seu valor a partir das demais, é denominada análise de regressão. Segundo Greene (2003), a análise de regressão está preocupada com o estudo da dependência de uma variável (Variável Exógena ou Dependente, Y_i), em relação a uma ou mais variáveis (Variáveis Endógenas ou Independentes, X_1, X_2, \dots, X_i), com uma visão de estimação, seja para associação ou previsão a partir dos valores fixos conhecidos destas variáveis explicativas. A equação de regressão linear clássica pode ser representada da seguinte forma:

$$Y_i = \beta_0 + \sum_{k=1}^K \beta_k X_{ki} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad (1)$$

$$i = 1, 2, \dots, n \text{ e } k = 1, 2, \dots, K$$

Onde Y_i é a variável exógena ou dependente relativa a i -ésima observação; o β_0 intercepto, X_{ki} a matriz de linha k e coluna i de variáveis endógenas, os β_k respectivos parâmetros de cada variável endógena e ε_i é a perturbação estocástica, distribuído normalmente, com o primeiro momento nulo e o segundo constante e finito.

O modelo de regressão linear leva em conta certos pressupostos, que especificam os procedimentos para estimação e inferência dos parâmetros. Esses pressupostos estão ligados em sua maioria com ε_i , tais como: ausência de correlação serial entre as perturbações estocásticas, homogeneidade das variâncias dos erros, o processo de geração dos ε_i não está relacionado com X_1, X_2, \dots, X_i e ε_i tende a seguir uma distribuição normal, com média zero e variância constante e finita.

Quando os dados são referenciados no espaço geográfico, podendo ocorrer a dependência espacial entre os eventos (autocorrelação espacial), portanto deve se levar em conta a estrutura espacial dessas observações (CRESSIE, 1993). O não uso dessa condição pode afetar a significância dos parâmetros, tornando-os inadequados e a existência da

dependência geográfica em larga escala pode induzir a presença de associações espúrias (DRUCK et al., 2004).

Outro aspecto é da caracterização da dependência espacial, que demonstra como as observações se correlacionam no espaço geográfico. Como forma de aferir a autocorrelação espacial, o Índice de Moran Global (ANSELIN, 1988) foi usado; o valor do índice varia entre -1 a 1, e indica o grau e a relação de semelhança de uma área com sua respectiva vizinhança.

Dado o fenômeno da dependência espacial é plausível o uso de modelos de regressão espacial. Para este estudo foram ajustados o *Simultaneous Autoregressive Model (SAR)*, o *Conditional Autoregressive Model (CAR)* e *Generalized Additive Model (GAM)*. Estes podem ser definidos como:

$$Y_i = \beta_0 + \sum_{k=1}^K \beta_k X_{ki} + \rho w_{sj} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad (2)$$

$$Y_i = \beta_0 + \sum_{k=1}^K \beta_k X_{ki} + \varepsilon_i, \quad \varepsilon_i = \rho w_{sj} \varepsilon_i + \xi_i, \quad \xi_i \sim N(0, \sigma^2) \quad (3)$$

$$Y_i = \beta_0 + \sum_{k=1}^K \beta_k X_{ki} + S(\text{lat}_s, \text{long}_j) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \quad (4)$$

$$i = 1, 2, \dots, n, \quad k = 1, 2, \dots, K, \quad j = 1, 2, \dots, J \quad e \quad s = 1, 2, \dots, S$$

O modelo SAR (2) considera a dependência espacial, através da adição de um novo termo na forma de uma relação entre a matriz de vizinhança (w_{sj}), combinada com um coeficiente espacial autorregressivo ρ (ANSELIN, 2002). No modelo CAR (3), os efeitos espaciais são um ruído ou perturbação, ou seja, fator que precisa ser removido. Neste caso, os efeitos da autocorrelação espacial são associados ao termo do erro. Onde a combinação de $\rho w_{sj} \varepsilon_i$ é considerado os erros com efeitos espaciais, e ξ_i passa a representar os erros aleatórios com média zero e variância desconhecida (BIVAND, 2002).

Nestes dois últimos modelos, 2 e 3, a ideia é incorporar a autocorrelação espacial como componente do modelo, onde a hipótese nula para a não existência de autocorrelação é $\rho = 0$, ou seja, o termo de erro não é espacialmente correlacionado.

No GAM (4) é incorporado uma função de suavização *spline* bivariada com as coordenadas geográficas (latitude e longitude) do centroide de cada área, função essa definida pelo termo $S(\text{lat}_s, \text{long}_j)$ (WOOD, 2006).

2.2 Método *Bootstrap*

O *bootstrap* foi introduzido como um método de computação intensiva para estimar o erro padrão (\hat{se}) de um estimador qualquer ($\hat{\theta}$) (EFRON, 1979). As vantagens do *bootstrap* residem na estimação do \hat{se} sem necessitar de extensos cálculos teóricos, está disponível não importando a complexidade matemática do $\hat{\theta}$ e dispensa a suposição de normalidade da variável em estudo.

O método *bootstrap* depende da noção de amostragem *bootstrap* e pode ser executado da seguinte forma: seja \hat{F} uma distribuição empírica qualquer e as observações $x_i, i = 1, 2, \dots, n$, com probabilidade $1/n$ para todo x_i observado. Uma amostra *bootstrap* é definido com uma amostra aleatória de tamanho n , retirada de \hat{F} , digamos $x^* = (x_1^*, x_2^*, \dots, x_n^*)$.

O conjunto de dados do *bootstrap* ($x_1^*, x_2^*, \dots, x_n^*$), são uma amostra aleatória de tamanho n , independente e retirada com reposição da população de n observações (x_1, x_2, \dots, x_n). O conjunto de dados do *bootstrap* consiste de membros do conjunto observado, alguns

aparecendo nenhuma, uma ou mais de uma vez.

Correspondente a base de dados do *bootstrap* (x^*) é a replicação do *bootstrap* para $\hat{\theta}^*$, definido para $\hat{\theta}^*(b) = s(x^{*b})$, $b = 1, 2, \dots, B$. Sendo $s(x^{*b})$ uma injeção de x^* em x . Portanto para todo estimador de x (média, variância, etc...), haverá um estimador correspondente x^* do conjunto de replicações do *bootstrap*. O $s\hat{e}_b$ das B replicações é definido por:

$$s\hat{e}_b = \left\{ \frac{\sum_{b=1}^B [\hat{\theta}(b) - \hat{\theta}(\cdot)]^2}{B-1} \right\}^{1/2}, \text{ onde } \hat{\theta}(\cdot) = \frac{\sum_{b=1}^B \hat{\theta}(b)}{B} \quad (5)$$

No limite, $s\hat{e}_b$ quando B tende ao infinito, $s\hat{e}_b$ se torna a estimativa ideal do *bootstrap* para $s\hat{e}_{\hat{F}}$. Mas na prática B termina antes do infinito e normalmente é superior a 500 replicações. $s\hat{e}_{\hat{F}}$ e sua aproximação $s\hat{e}_b$, são chamados de estimativas de *bootstrap* não-paramétrico, pois são baseados em \hat{F} , uma estimativa da população F , onde *a priori* não assumimos distribuição de probabilidade alguma.

A técnica de *bootstrap*, baseia-se na amostragem com reposição, acarretando que as observações devem ser independentes. Porém, se os dados apresentarem uma estrutura de dependência fraca, então as replicações do *bootstrap* serão válidas. Porém, em ambientes onde a estrutura de dependência exerce forte influência, o desconhecimento desse suposto com o procedimento de *bootstrap*, deve inevitavelmente eliminar a estrutura de correlação, gerar associações espúrias e estimadores enviesados (DAVISON e HINKLEY, 1997).

Em dados que possuam estrutura de autocorrelação, seja temporal ou espacial, é necessário levar em conta essa restrição no método *bootstrap* para não gerar estimativas enviesadas e portanto ter em conta a estrutura de autocorrelação é de crítica importância. Para esse propósito o *bootstrap* espacial (SOLOW, 1985; LOH e STEIN, 2004) deve ser aplicado como abordagem correta para gerar replicações que levam em conta a estrutura de correlação geográfica.

Baseado no *bootstrap* não-paramétrico, foram realizadas 5000 replicações do *AIC* de cada modelo ajustado, para gerar diversas estimativas dos *AIC*.

2.3 Seleção de Variáveis e do Modelo

A seleção dos modelos foi baseado no Critério de Informação de Akaike (*AIC*) (AKAIKE, 1974), expresso por:

$$AIC = -2 * \log\text{like} + 2k \quad (6)$$

Onde *loglike* o logaritmo da máxima verossimilhança e k o número de parâmetros de regressão. O funcionamento do *AIC* se baseia na inevitável perda de informação genérica, devido ao uso de um dos possíveis modelos para representar o "verdadeiro" modelo. Então devemos selecionar o modelo que minimize a quantidade esperada de informação perdida.

Na prática, dado um conjunto de candidatos para modelar um determinado evento, o melhor modelo é o que possuir menor *AIC*; ainda, este somente recompensa a precisão do ajuste, mas também inclui uma penalidade que aumenta a cada parâmetro incluído no modelo. Essa penalidade visa desencorajar o excesso de ajuste (*overfitting*). O *AIC* portanto proporciona um meio eficaz para comparação de modelos (BURNHAM e ANDERSON, 2004).

Mas as desvantagens principais são: não possuir um teste de hipótese, objetivando escolher o melhor modelo significativamente; e não poder mensurar a qualidade do ajuste em um sentido absoluto. Mas apesar desses fatos, o benefício proporcionado pelo uso desse

método supera as desvantagens. Um outro meio para seleção dos modelos foi a soma dos quadrados dos resíduos (*SQR*). Segundo este critério, o melhor modelo é o que possui menor valor de *SQR*.

Dada a ausência *á priori* de distribuição de probabilidade nas replicações, foram usados os testes não-paramétricos de Kruskal-Wallis e o pós teste de Dunn (MYLES e DOUGLAS, 1973), para identificar a existência de diferença significativa entre as distribuições dos *AIC* de cada modelo.

Todas as análises estatísticas foram desenvolvidas no pacote estatístico R (R DEVELOPMENT CORE TEAM, 2011).

3. Resultados

A tabela 1 mostra os 4 modelos utilizados no estudo, seguidos de alguns critérios de ajuste. O modelo *GAM* se mostrou mais parcimonioso que os demais, seja por ter um *AIC* e *SQR* menor, devido à quantidade de explicação que a função de suavização *spline* bivariada acrescentou para o modelo, mesmo apesar desta não ter sido significativa (p-valor > 0,05). Cabe ressaltar que o modelo linear tem *AIC* menor que do modelo *CAR*, isto ocorre devido ao emprego do efeito espacial embutido no modelo, já o modelo *SAR* apresenta *SQR* e *AIC* menores que ambos os modelos. De toda maneira, o teste F aponta que todos os modelos são significativamente válidos na utilização para o estudo (p-valor < 0,001).

Tabela 1. Principais resultados dos modelos de regressão.

| Modelos | <i>SQR</i> | <i>AIC</i> | Teste F (p-valor) | <i>Spline</i> (p-valor) |
|------------|------------|------------|-------------------|-------------------------|
| Linear | 863,63 | 987,8527 | < 0,001 | - |
| <i>SAR</i> | 846,5131 | 986,7353 | < 0,001 | - |
| <i>CAR</i> | 855,0915 | 988,9955 | < 0,001 | - |
| <i>GAM</i> | 689,437 | 973,5894 | < 0,001 | 0,116 |

Para seleção e validação do modelo mais adequado, foram utilizadas técnicas de amostragens via *bootstrap* não-paramétrico. Foram feitas 5000 replicações de cada modelo e os resultados estão dispostos na figura 2, 3 e tabela 2.

Segundo os histogramas de cada modelo (figura 2), apesar de *á priori* não assumirmos distribuição de probabilidade específica alguma, os *AIC* tenderam a se distribuir em torno da média, principalmente no modelo linear. Isso se deve à Lei dos Grandes Números (JAMES, 2010). A Lei dos Grandes Números trata da convergência de uma sequência de variáveis aleatórias X_n integráveis, independentes e identicamente distribuídas onde
$$\frac{[\sum X_n - E(X_n)]}{n}$$
, converge para zero, quando se aumenta infinitamente o número de amostras. Nesse caso, existe uma tendência dos *AIC* (a variável aleatória) se concentrarem em torno da média amostral. Em adicional, se padronizado estes *AIC* e sob certas hipóteses gerais, sua distribuição tende a Gaussiana.

O *AIC* observado de cada um dos modelos mostra consistência por possuírem os respectivos *AIC* observado dentro do intervalo percentil (5%; 95%), construído pelo *bootstrap* não-paramétrico.

Os *boxplot* das replicações de cada modelo (figura 3) se mostraram bem similar no formato, indicando distribuição equitativa, seja na área interna e externa dos quartis. A presença de *outliers* nas replicações foi pequena, em média de 1,5% do total em cada modelo. Cabe ressaltar, que o modelo *GAM* tem o valor do *AIC* mediano bem abaixo dos demais modelos.

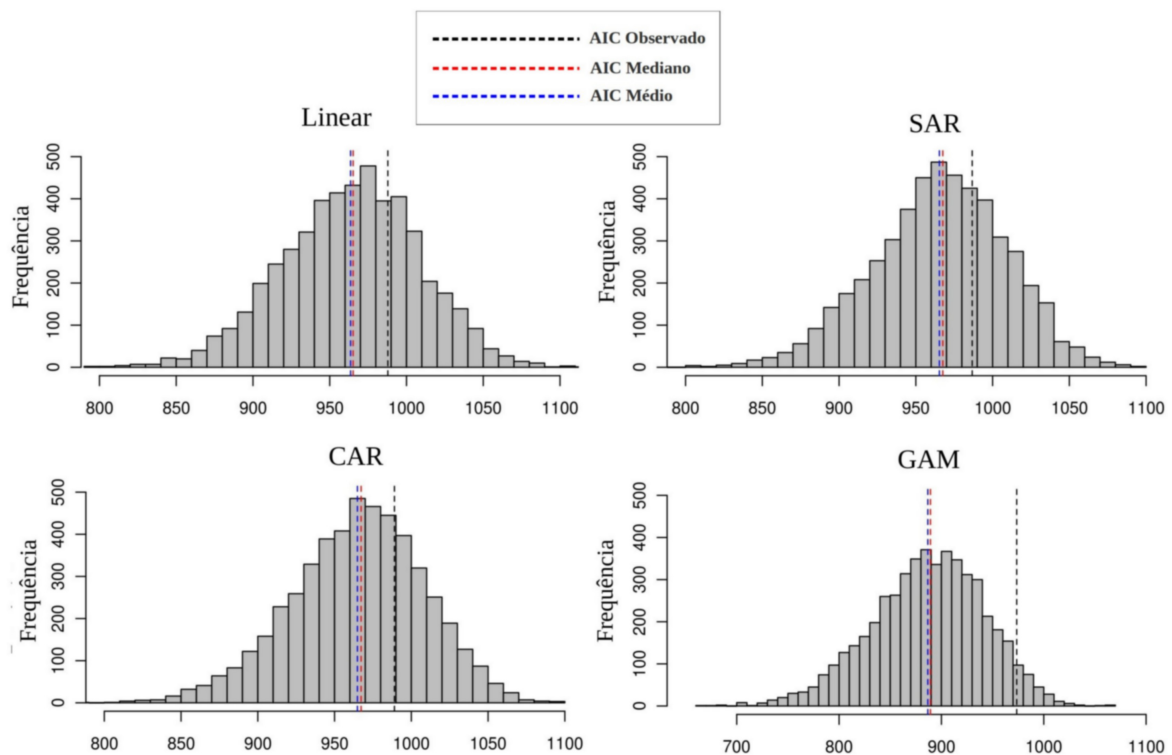


Figura 2. Histograma das distribuições das 5000 replicações das estimativas do *AIC* via *bootstrap* não-paramétrico dos modelos de regressão.

Na tabela 2, observa-se que o valor médio e mediano não estão muito afastados, corroborando com a tese da aproximação com a distribuição Gaussiana, pois esta possui o valor médio, mediano e modal idênticos. O modelo *GAM* teve o maior coeficiente de variação em relação aos outros modelo, como maior distância entre o percentil 5% até o 95%. Mas apesar da dispersão maior em torno da média, o valor mediano do *AIC* está próximo da média, similarmente como ocorre nos demais modelos.

Tabela 2. Medidas de posição e dispersão das estimativas do *AIC* via *bootstrap* não-paramétrico dos modelos de regressão.

| Modelos | Perc. 5% | 1º Quartil | Mediana | Média | 3º Quartil | Perc. 95% | C.V* (%) |
|---------|----------|------------|---------|-------|------------|-----------|----------|
| Linear | 873,9 | 934,5 | 965,2 | 963,6 | 994,6 | 1046 | 4,6251 |
| SAR | 890,24 | 937,4 | 967,6 | 965,3 | 995,3 | 1032,83 | 4,5189 |
| CAR | 888,57 | 937,1 | 967,3 | 964,9 | 994,8 | 1033,24 | 4,5015 |
| GAM | 773,2 | 850,9 | 889,3 | 886,7 | 925,2 | 985,3 | 6,1656 |

* Coeficiente de Variação Robusto - $CV = \frac{S}{Md}$

Ao se utilizar o teste não-paramétrico de Kruskal-Wallis, observa-se a existência de diferença significativa (p-valor < 0,001) entre algum dos modelos ajustados. Usando o pós teste de Dunn verificou-se diferença significativa entre o modelo linear e o *SAR* (p-valor < 0,05), onde o modelo linear por ter um *AIC* mediano menor do que o *SAR*, seria o melhor a ser ajustado. E ainda, diferença significativa entre o modelo *GAM* frente todos os outros modelos (p-valor < 0,001). Isso indica que o modelo de aditivos generalizados (*GAM*) é significativamente (p-valor < 0,001) mais parcimonioso que os demais, devendo este ser utilizado para compreender o fenômeno estudado.

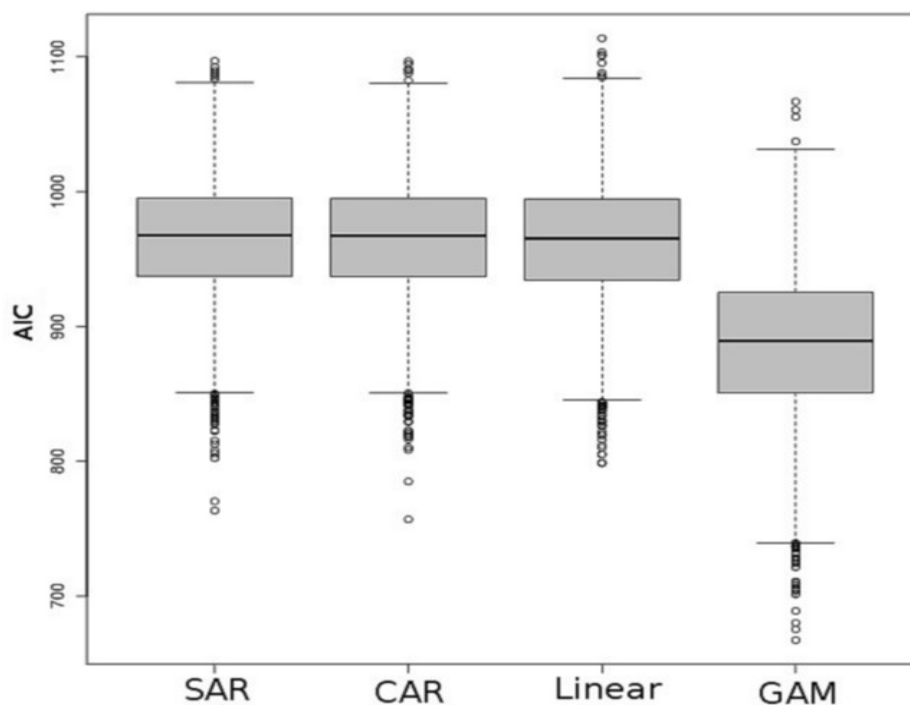


Figura 3. *Boxplot* das distribuições das 5000 replicações das estimativas do *AIC* via *bootstrap* não-paramétrico dos modelos de regressão.

4. Conclusão

Este estudo explorou o uso de técnicas de *bootstrap* não-paramétrico com a finalidade da criação de um intervalo de confiança e, conseqüente, teste de hipótese para comparação de modelos através do *AIC*. O *AIC* é um critério largamente utilizado (HOSMER e LEMESHOW, 1989; BIVAND, 2002; KOOP, 2003; WILKS, 2006; BROCKWELL e DAVIS, 2009) na eleição do melhor modelo estatístico, sendo este utilizado nos mais diversos estudos empíricos. Ao preenchermos essa lacuna, tal metodologia proporciona a escolha do modelo com menor *AIC* significativamente, e não somente por haver um valor menor que do outro.

Através de métodos de *bootstrap* aplicados nos modelos de regressão, foi possível validar a consistência dos modelos quanto ao *AIC* e possibilitou eger o modelo de aditivos generalizados (*GAM*) como o que significativamente demonstrou maior parcimônia. Outro ponto importante, resultou na comparação do modelo linear com o *SAR* e *CAR*, onde o modelo linear não foi significativamente diferente do *CAR*, mas significativamente melhor que o modelo *SAR*, mostrando que com o uso dos efeitos espaciais, nas matrizes de correlação, não foi observado melhora significativa. Isto corrobora com a ideia de Akaike (1974) de que nem sempre o melhor modelo é o mais complexo.

Deve se estimular a utilização dessa metodologia em trabalhos que abordem outros tipos de modelos, como da família de modelos lineares generalizados, de séries temporais, sobrevida, entre outros que utilizam o Critério de Informação de Akaike na escolha do modelo melhor ajustado. Essa metodologia não deve somente se restringir ao *AIC*, mas ser estendido a outros outros critérios de seleção de modelos, como o *BIC* (*Bayesian Information Criterion*) (AKAIKE, 1980), critério de Hannan–Quinn (HQ), a soma do quadrado dos resíduos (*SQR*) e o coeficiente de determinação (R^2).

A utilização das técnicas de *bootstrap* para validar o modelo mais adequado na abordagem do estudo, devem ser feitas com bastante cautela pois apesar da facilidade de

implementação em relação a outros métodos de maior complexidade.

5. Bibliografia

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974. 19 (6): 716–723.

AKAIKE, H. *Likelihood and the Bayes procedure*. In: *Bayesian Statistics*, Ed. J.M Bernardo et al., Valencia: University Press, 1980. 143-166

ANSELIN, L. *Spatial econometrics: Methods and models*. Boston: Kluwer Academic Publishers, 1988.

ANSELIN, L. Under the hood: issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 2002.

BIVAND, R. S. Spatial econometrics functions in R: classes and methods. *Journal of Geographical Systems*, 2002. 4: 405–421.

BROCKWELL, P. J.; DAVIS, R. A. *Time Series: Theory and Methods*. 2 ed. New York: Springer, 2009.

BURNHAM, K. P.; ANDERSON, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2 ed. New York: Springer-Verlag, 2002.

BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding AIC and BIC in Model Selection. *Sociological Methods and Research* (33), 2004. 261-304.

CLAESKENS, G; HJORT, N. L. *Model Selection and Model Averaging*, Cambridge: Cambridge University Press, 2008.

CRESSIE, N. *Statistics for Spatial Data*. New York: John Wiley & Sons, 1993.

DRUCK, S.; CARVALHO, M.S.; CÂMARA, G.; MONTEIRO, A.V.M. *Análise Espacial de Dados Geográficos*. Brasília: EMBRAPA, 2004.

EFRON, B. Bootstrap methods: Another look at the Jackknife. *Ann. Statist*, 1979. (7):1-26.

EFRON, B.; TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. New York: Chapman & Hall, 1993.

GREENE, W. H. *Econometrics Analysis*. 4 ed. New Jersey: Pearson Education LTD, 2003.

HOSMER, D. W. Jr. ; LEMESHOW, S. *Applied logistic regression*. New York: Wiley, 1989.

JAMES, B. R. *Probabilidade: um curso em nível intermediário*. 3 ed. Rio de Janeiro: IMPA, 2010.

KOOP, G. *Bayesian Econometrics*. New York: Wiley, 2003.

KOSHIYAMA, A. S. ; CARVALHO, K. F. S. ; SOARES-NETO, J. ; LORENZON, M. C. A.;

TASSINARI, W. S. Utilização de modelos econométricos espaciais aplicados ao estudo do preço do mel no Brasil. *BJO&PM* (Submetido, 2011).

MYLES, H.; DOUGLAS, A. W. *Nonparametric Statistical Methods*. New York: John Wiley & Sons, 1973. 115–120

LOH, J. M.; STEIN, M. L. Bootstrapping a spatial point process. *Statist. Sinica*, 2004. 14, 69–101.

R DEVELOPMENT CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. URL <http://www.R-project.org>.

SHAO, J.; TU, D. *The Jackknife and Bootstrap*. New York: Springer-Verlag, 1995.

SOLOW, A. Bootstrapping Correlated Data. *Journal of the International Association for Mathematical Geology* (17), 1985. 769-775.

TAKEUCHI, K. Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, 1976. 12–18. (Em japônes).

WILKS, D. S. *Statistical Methods in the Atmospheric Sciences*. 2 ed. Elsevier, 2006.

WOOD, S. N. *Generalized Additive Models – An Introduction with R*. New York: Chapman & Hall/CRC, 2006.

6. Agradecimentos

Os autores agradecem ao CNPq pelo apoio financeiro concedido e a UFRRJ por toda a ajuda institucional.