Estudo comparativo de métodos de crítica estatística para índices de preços

IBGE – COINP – SINAPI

Rio de Janeiro, 08 de novembro de 2018.

SINAPI – Sistema nacional de pesquisa de custos e índices da construção Civil

- Dois módulos:
  - Índice de Custos (projetos do setor de habitação)
  - Orçamentação (Engloba habitação, saneamento e infraestrutura - Caixa).
- Cooperação técnica:
  - IBGE (Coleta e Apuração)
  - Caixa (Base técnica de engenharia, projetos, insumos, etc.)
- Pesquisa realizada nas 26 UF's e DF

- Coletar os mais variados insumos da construção civil
- Interesse em divulgar níveis de preços (mediana)

SINAPI – Sistema nacional de pesquisa de custos e índices da construção Civil

- Descrição básica + complementação (marca e modelo) + Unidade
- Exemplo:
- 1185 - TUBO DE PVC RÍGIDO (BRANCO), ROSCÁVEL, PARA SISTEMAS PREDIAIS DE ÁGUA FRIA DE 3/4", NORMA PECP34 - VARA DE 6 M (INDICAR FABRICANTE)
  - Preços diferentes por painel:
    - Marca 1 – R$42,24
    - Marca 2 – R$32,71
    - Marca 3 – R$25,02

Crítica estatística automatizada

- Um tema de fundamental importância para pesquisas estatísticas é a detecção e o tratamento de observações que sejam discrepantes, outliers.
- A escolha de um algoritmo deve levar em conta as particularidades da pesquisa e os efeitos que uma contaminação de outiliers gera na estatística de interesse
- Devido à latência do processo de geração de preços de um produto, não é possível aplicar um método determinístico para avaliar a validade de um valor coletado.
- Por isso são aplicados métodos estatísticos de crítica para determinar se um preço é destoante.

Crítica estatística automatizada

- Pesquisas de índices de preços possuem diversas peculiaridades que precisam ser observadas para a construção de um método de crítica adequado. Algumas questões:
  - dificuldade de encontrar uma distribuição adequada para representar os preços de um produto;
  - a variedade de produtos que contribuem para a cesta de bens e serviços pesquisada;
  - número de observações com preços repetidos em períodos consecutivos de pesquisa.
- Motivação:
  - Assegurar a qualidade dos valores coletados
  - Balizar um resultado para facilitar decisão do analista sobre a validade de um preço
  - Evitar:
    - Viés nos estimadores
    - Orçamentação Inflada (Sinapi)

Variáveis base para a crítica

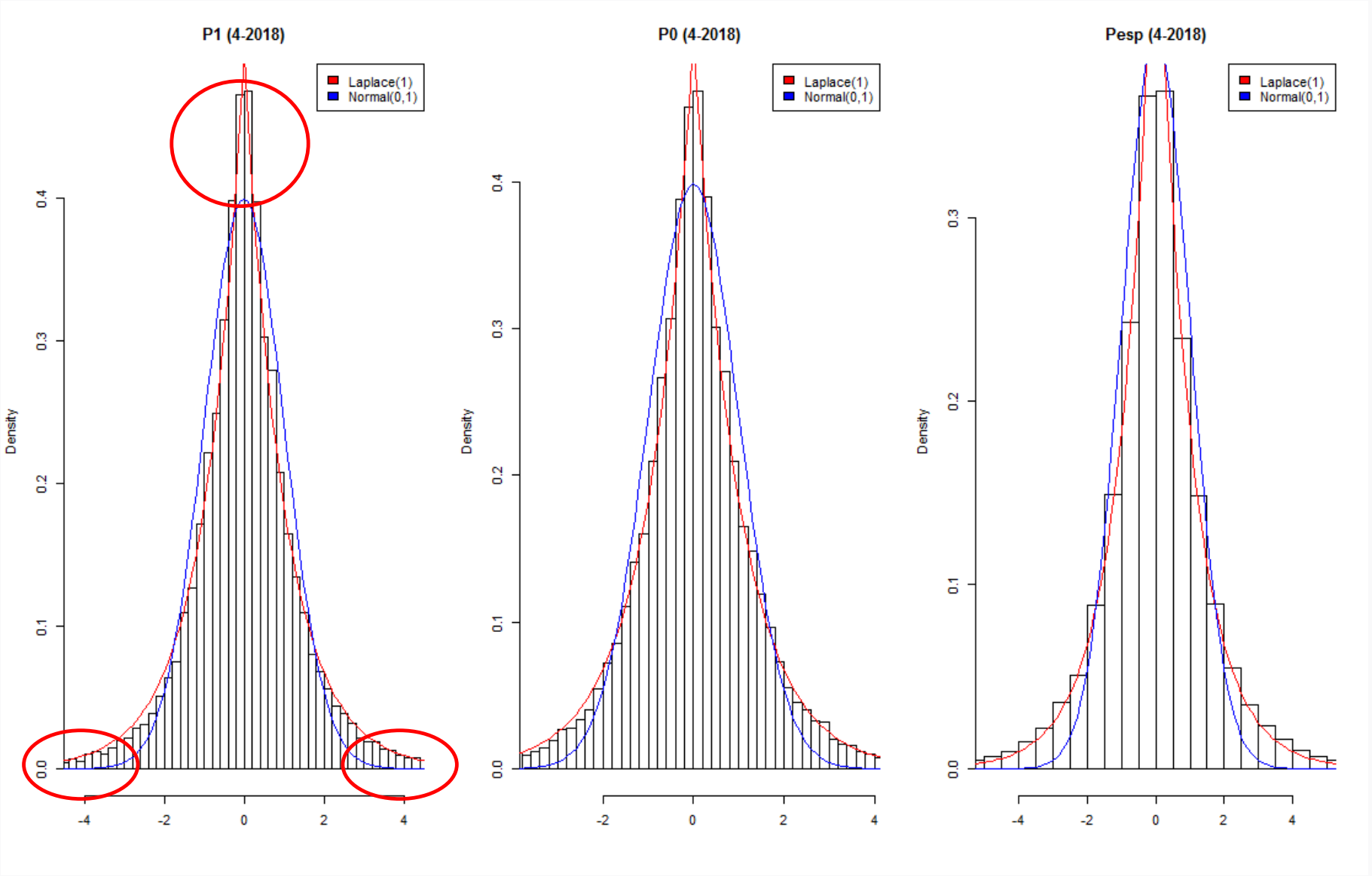
- Uso do log de desvios medianos:
- Vantagens:
  - Os desvios medianos são invariantes ao nível
  - Possibilidade de juntar dados coletados em estados diferentes para ter uma amostra maior para a critica
  - Simetria em torno de zero
  - Formato de densidade conhecido
- Definição:

$$P_t = \log\left(\frac{Preço_t}{Mediana(Preços\ de\ um\ agregado\ geográfico\ no\ mês\ t)}\right)$$

$$P_{t-1} = \log\left(\frac{Preço_{t-1}}{Mediana(Preços\ de\ um\ agregado\ geográfico\ no\ mês\ t-1)}\right)$$

$$Pesp_{t-1} = \log\left(\frac{Preço_{t-1}}{Mediana(Preços\ de\ uma\ qualidadede\ um\ agregado\ geográfico\ no\ mês\ t-1)}\right)$$

Diagrama de dispersão agregado da pesquisa



Transformação para tratar efeito de caudas pesadas

- Lambert Way (The Scientific World Journal, 2015): Transformação para caudas pesadas.

$$Y = Xexp\left(\frac{\delta}{2}\,U^2\right), \delta \in \mathbb{R}$$

- Para se determinar a transformação basta calcular a inversa da função acima, utilizando a função W de Lambert:

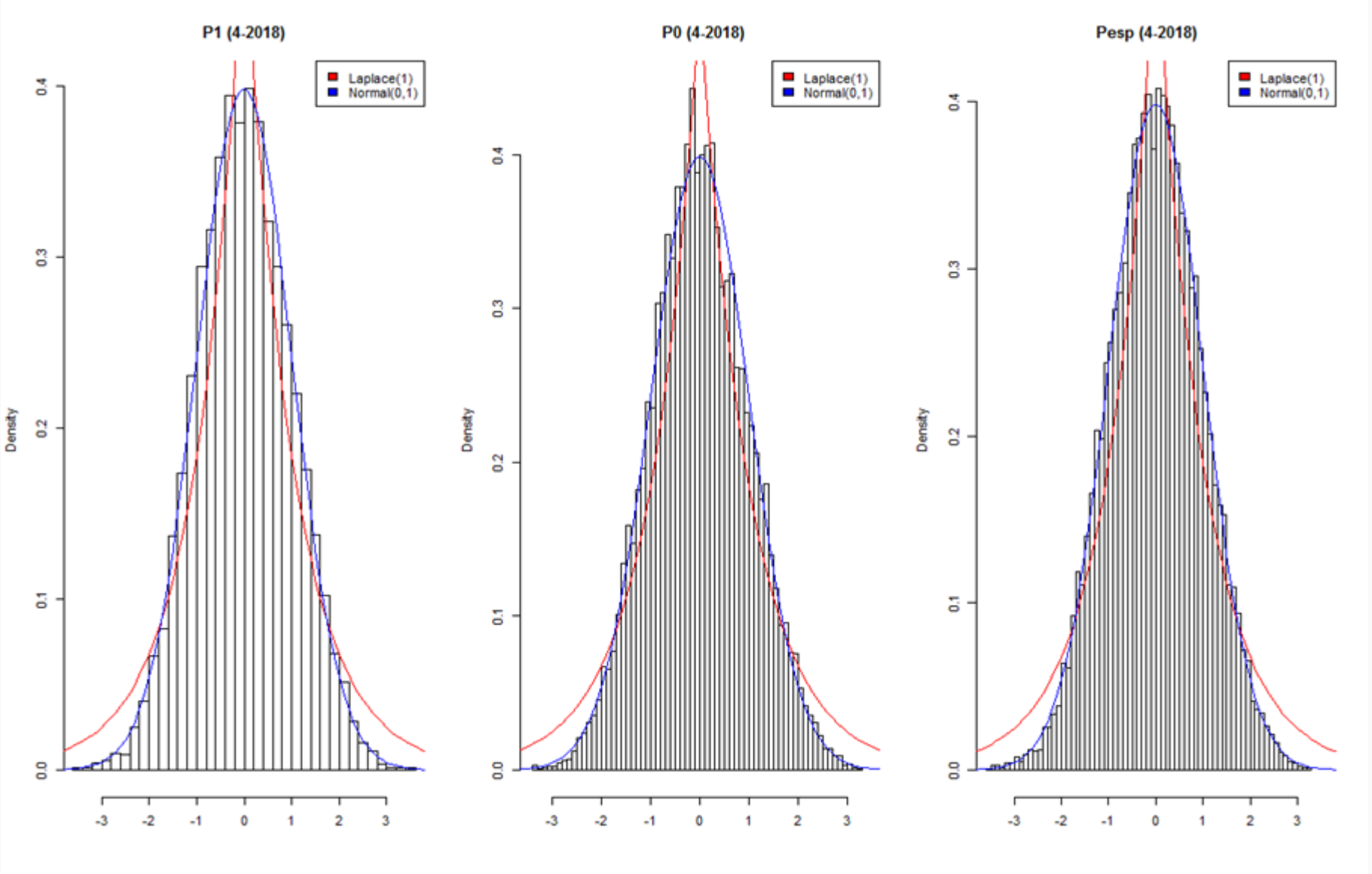
$$W_\tau(Y) \coloneqq \sigma_X W_\delta(z) + \mu_X \qquad W_\delta(z) = sinal(z)\left(\frac{W(\delta z^2)}{\delta}\right)^{\frac{1}{2}}$$

- A distribuição dos dados fica na forma:

$$g_Y(y \mid \beta, \delta) = f_X(W_\delta(z)\sigma_x + \mu_x \mid \beta)\frac{W_\delta(z)}{z[1 + W(\delta(z)^2)]}$$

- Implementação no pacote LambertW no software R.

Dados após a transformação Lambert Way



Resultados: Teste de normalidade por insumo

- Proporção de insumos com p-valor do teste de Shapiro-Wilk maior ou igual ao nível de significância:

Nível de Significância	Variável	Sem transformação	Box-Cox	Lambert Way
10%	P0	38,54%	56,83%	81,19%
	P1	29,24%	46,70%	78,88%
	Pesp	25,25%	39,21%	86,80%
5%	P0	46,84%	64,76%	90,43%
	P1	34,22%	53,30%	90,10%
	Pesp	29,24%	43,17%	96,04%
1%	P0	57,81%	74,45%	99,67%
	P1	42,52%	59,03%	99,67%
	Pesp	40,20%	54,63%	99,67%

- Lambert Way mais eficiente que Box-Cox para transformação dos dados da pesquisa.

Objetivos

- Há na literatura muitos métodos de crítica
- O objetivo do presente estudo é avaliar um método que tenha melhor performance para a nossa pesquisa
- O que determina o melhor método?

Inferência baseada nos parâmetros populacionais	Resultado da crítica	Diagnóstico	Resultado
Outlier	Marca	Situação ideal	Não há erro de marcação
Não é outlier	Não marca		
Não é outlier	Marca	Falso positivo	Erro de marcação
Outlier	Não marca	Falso negativo	

Método Estudados

- Métodos não paramétricos:
  - Algoritmo de Tukey (TA)
  - Algoritmo de Epidemia (EA)
  - Método de Crítica Estatística Automatizada COINP (CEA)

- Métodos paramétricos baseados na distância de Mahalanobis com estimação robusta de parâmetros dados por:
  - Transformed Rank Correlation (TRC)
  - Passo R (PR)

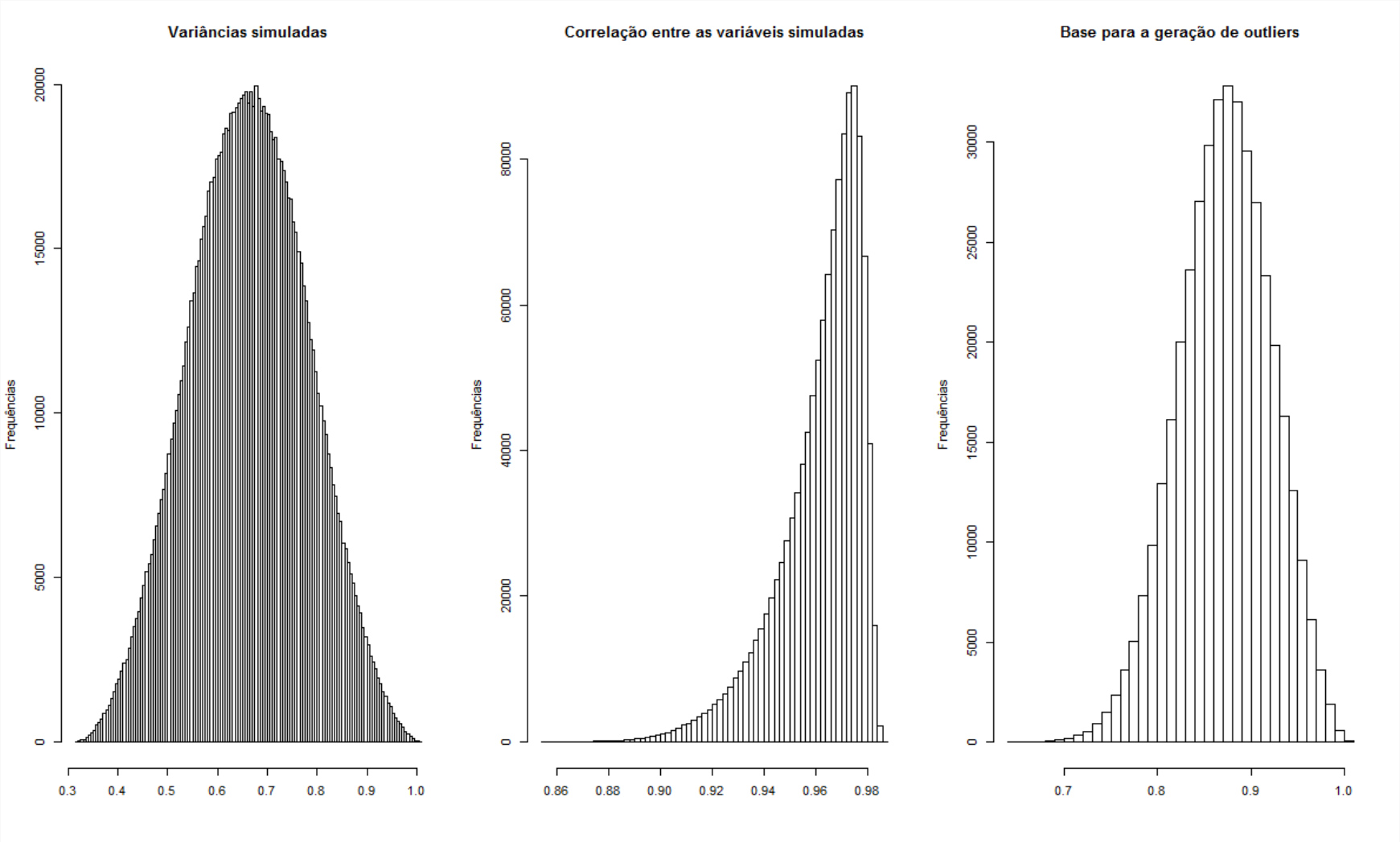
Descrição das simulações

- São geradas amostras da normal multivariada de tamanhos de 10, 15, 20, 25, 30, 50, 100, 200 e 300 observações. A análise dos resultados para amostras pequenas é importante para comparação com resultados de uma amostra real da pesquisa.
- As distribuições normais multivariadas são construídas da seguinte maneira:
  - Vetor de médias gerados de uma Uniforme (-0,05; 0,05);
  - Para a matriz de variâncias e covariâncias, as variâncias são obtidas de uma distribuição simétrica, com valores entre 0,3 e 1;
  - Como na pesquisa as variáveis são altamente correlacionadas, as correlações geradas pertencem ao intervalo [85%, 99%].

Descrição das simulações

- Para verificar a qualidade dos métodos, essas amostras são contaminadas por pontos influentes.
- Devido à experiencia dos analistas da pesquisa, foi constatada a proporção de um outlier para cada dez preços coletados.
- O valor máximo da diagonal principal da matriz de variâncias e covariâncias (denotado como σ_{MAX}) servirá de base para os outliers:
  - $Outilerj \sim Unif(2\sigma_{MAX}, 3\sigma_{MAX})$ j=1,2,3. Com isto garante-se que os valores selecionados pertencem à cauda direita da distribuição normal, em acordo com o fato de os valores $2\sigma_{MAX}$ e $3\sigma_{MAX}$ serem as posições dos quantis 97,73% e 99,87%, respectivamente, da distribuição normal.
  - É gerado um vetor de outliers com 3 posições (uma para cada variável): Out=[$Outlier1$, $Outlier2$, $Outlier3$];
  - Em seguida, cada entrada dos vetores é multiplicada por -1 ou 1, 50% de chance cada, para atribuir valores negativos ou positivos aos outliers e evitar viés.

Descrição das simulações



Descrição das simulações: detalhes para algoritmos de tolerância

- Para os algoritmos não paramétricos é necessário trabalhar com os dados na escala original, isto é, antes da transformação Lambert Way.
- Isto requer que para estes casos os dados gerados das distribuições normais sejam “destransformados”.
- A transformação Lambert Way é bijetora, ou seja, admite inversa, assim, basta escolher, conjuntamente, parâmetros de média, variância e curtose de um dos insumos aleatoriamente para levar os dados simulados à escala original.
- Os métodos da CEA e do Algoritmo de Tukey criticam o logaritmo dos relativos dos preços além dos desvios medianos.
- Porém, neste caso os relativos não são conhecidos e precisam ser estimados.

Descrição das simulações: detalhes para algoritmos de tolerância

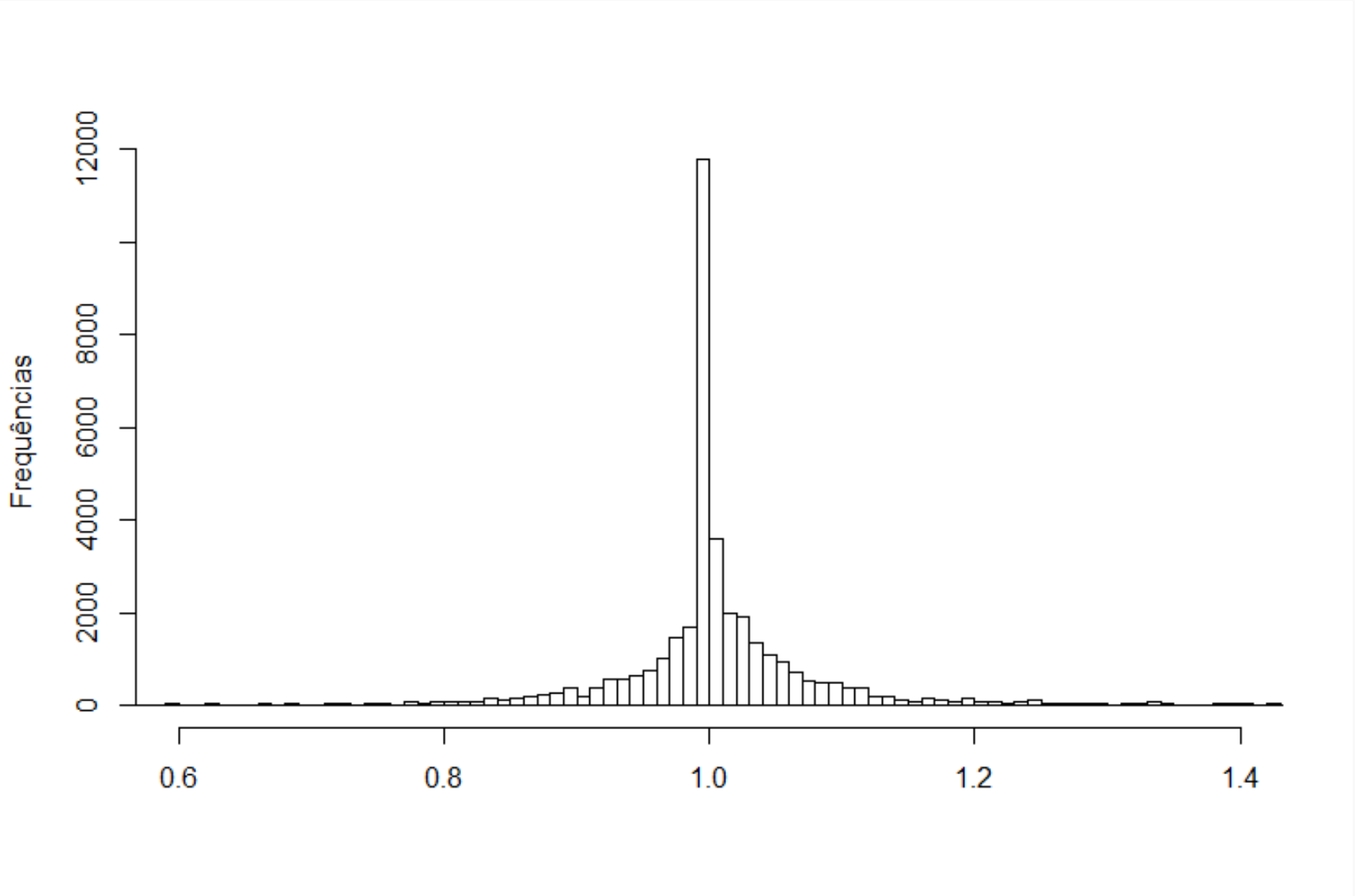
$$Desvio\ Mediano \rightarrow Logarítmo \rightarrow Lambert\ Way \rightarrow Simulação\ da\ dist.Normal$$

- Para chegar nos desvios medianos bastam realizar as transformações inversas do log e Lambert Way
- Para o relativo (razão de preço do mês corrente com o anterior):

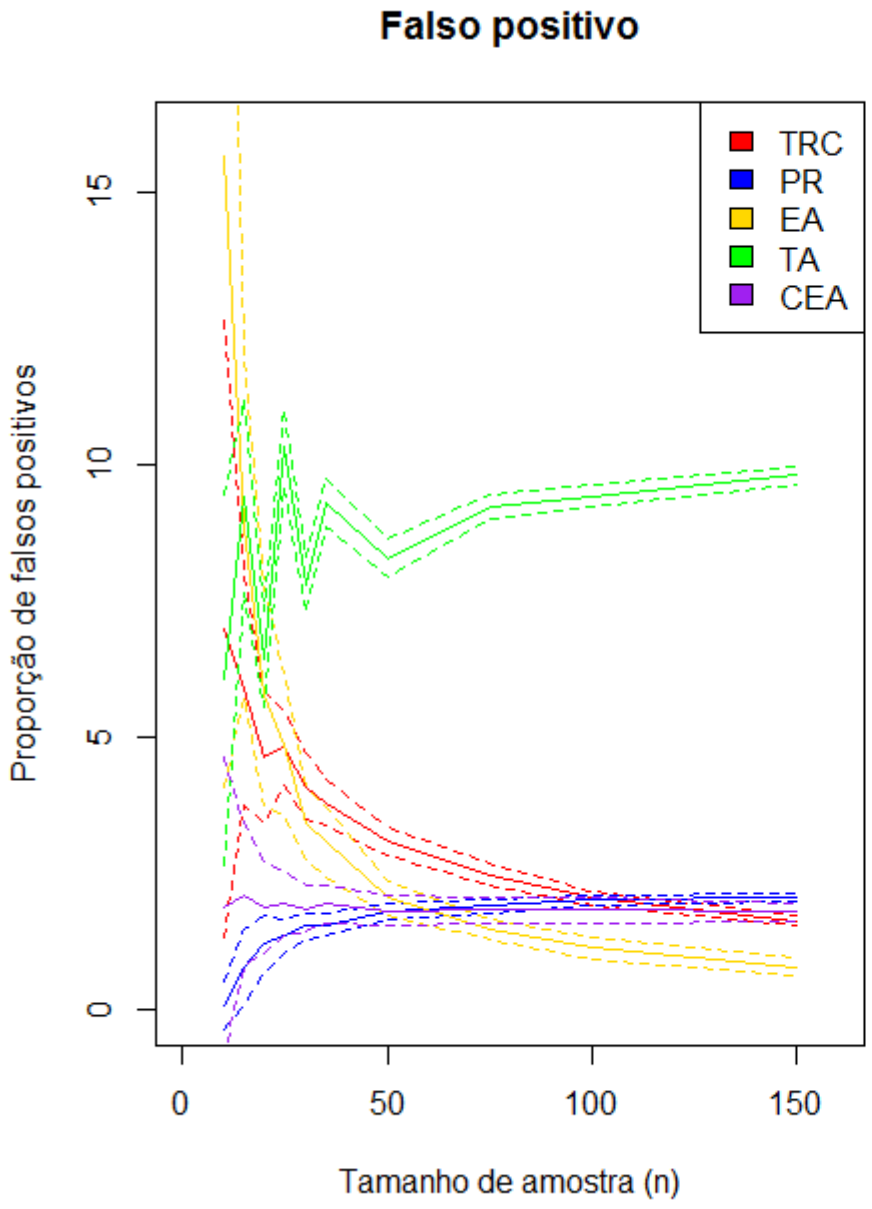
$$\frac{Preço_t / Mediana_t}{Preço_{t-1} / Mediana_{t-1}} = \frac{Preço_t / Preço_{t-1}}{Mediana_t / Mediana_{t-1}} = \frac{relativo}{variação\ mediana}$$

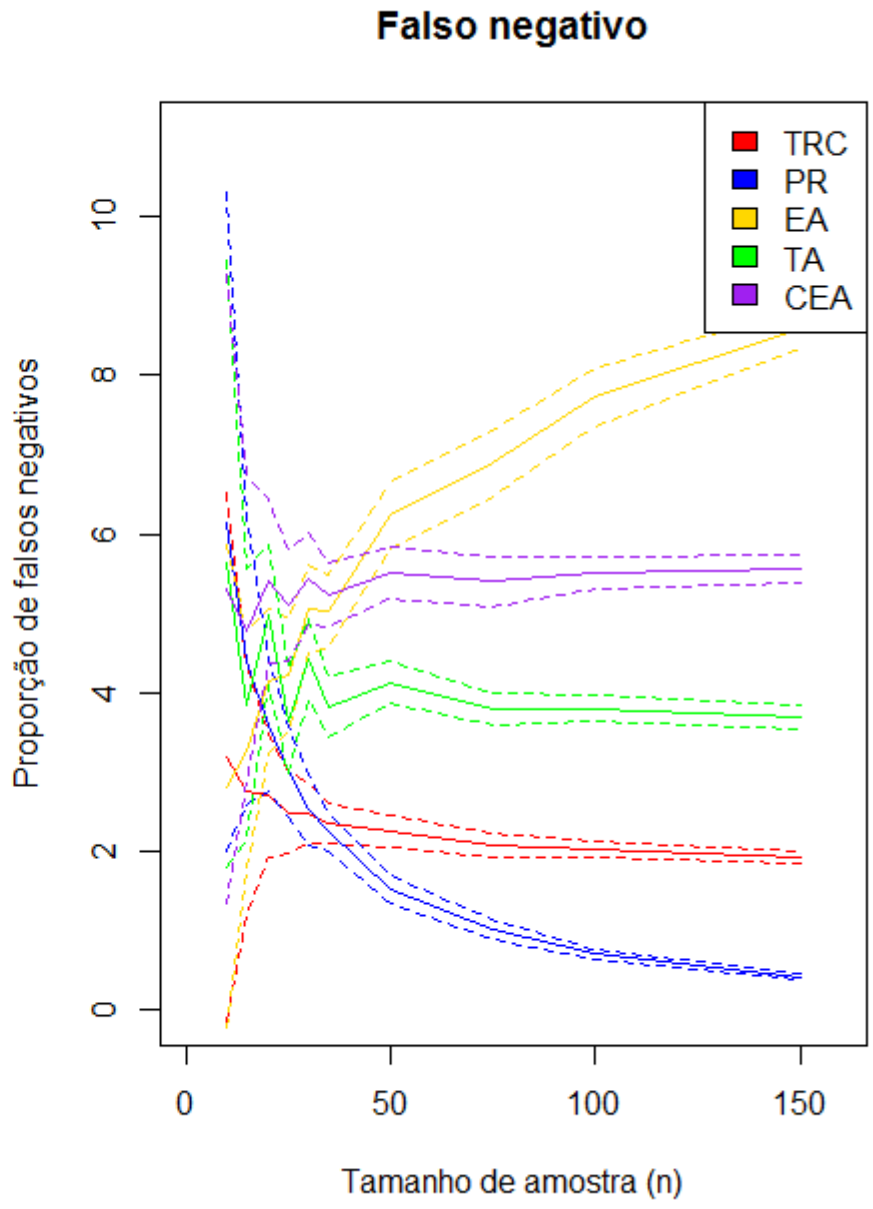
- A variação mediana é gerada escolhendo aleatoriamente 6 variações medianas nos preços coletados da pesquisa SINAPI
- O número seis representa o número médio de estados nos agregados geográficos da pesquisa.
- Como o número de observações geradas aleatoriamente da distribuição normal são superiores da seis estes valores são repetidos igualitariamente a fim de que se atinja o tamanho de amostra e assim se chegar nos relativos.

Descrição das simulações: base de parâmetros para as variações medianas

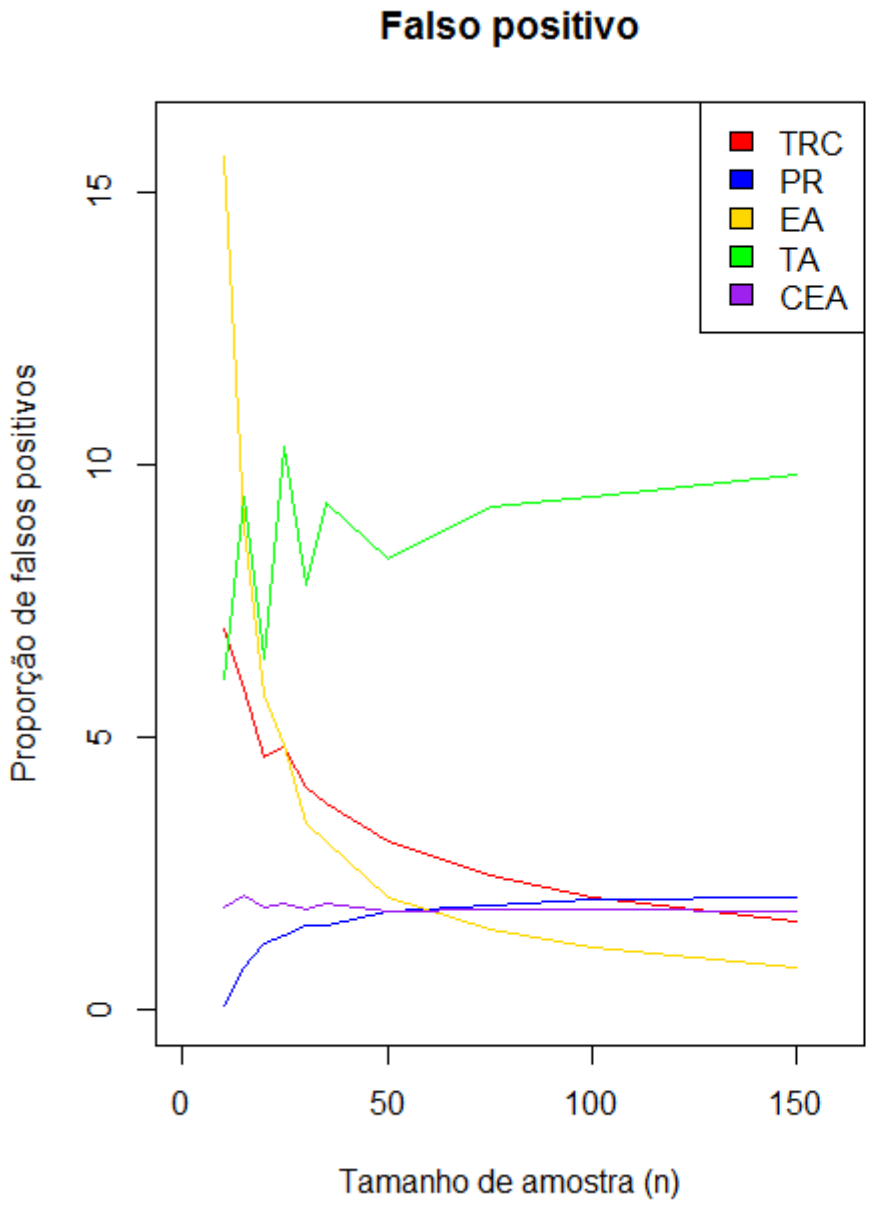


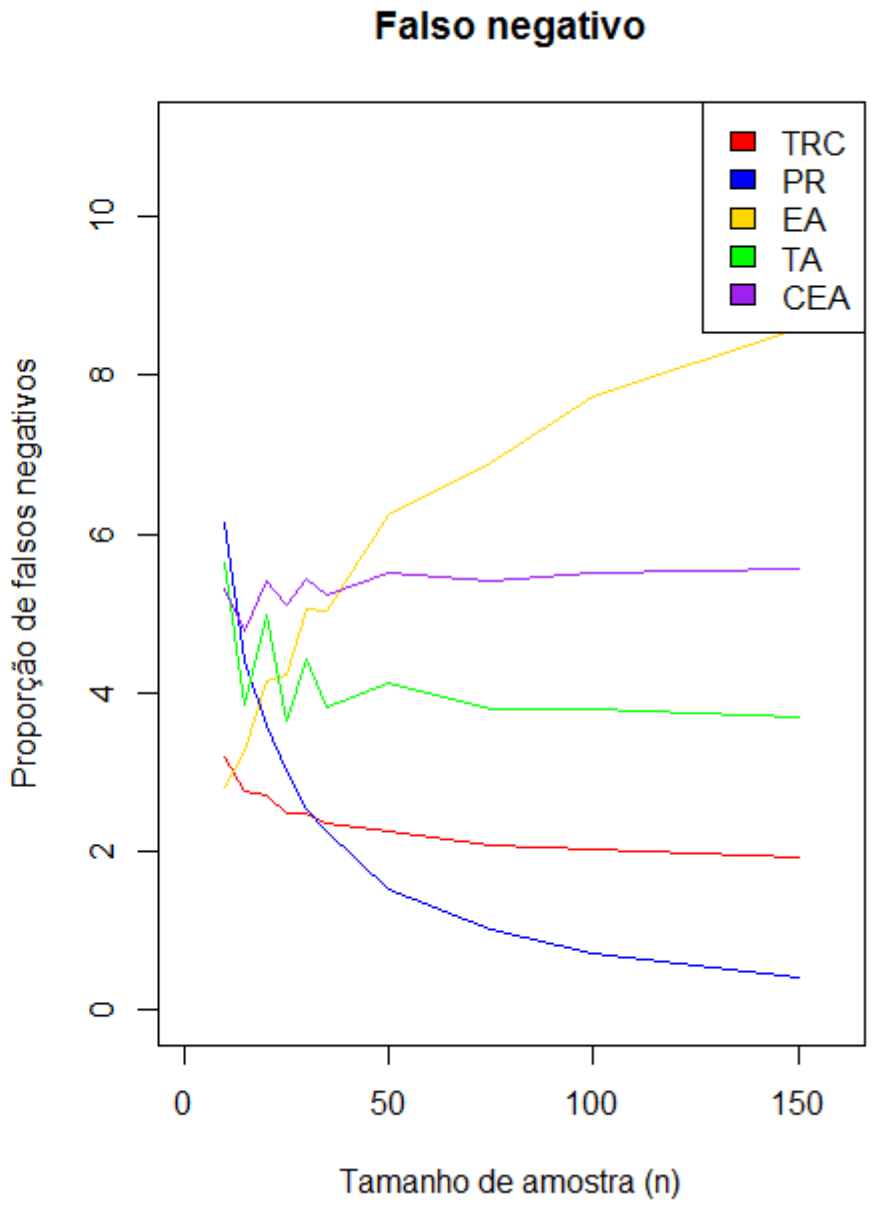
Resultados


Falso positivo
TRC
PR
EA
TA
CEA
Proporção de falsos positivos
0
5
10
15
0
50
100
150
Tamanho de amostra (n)

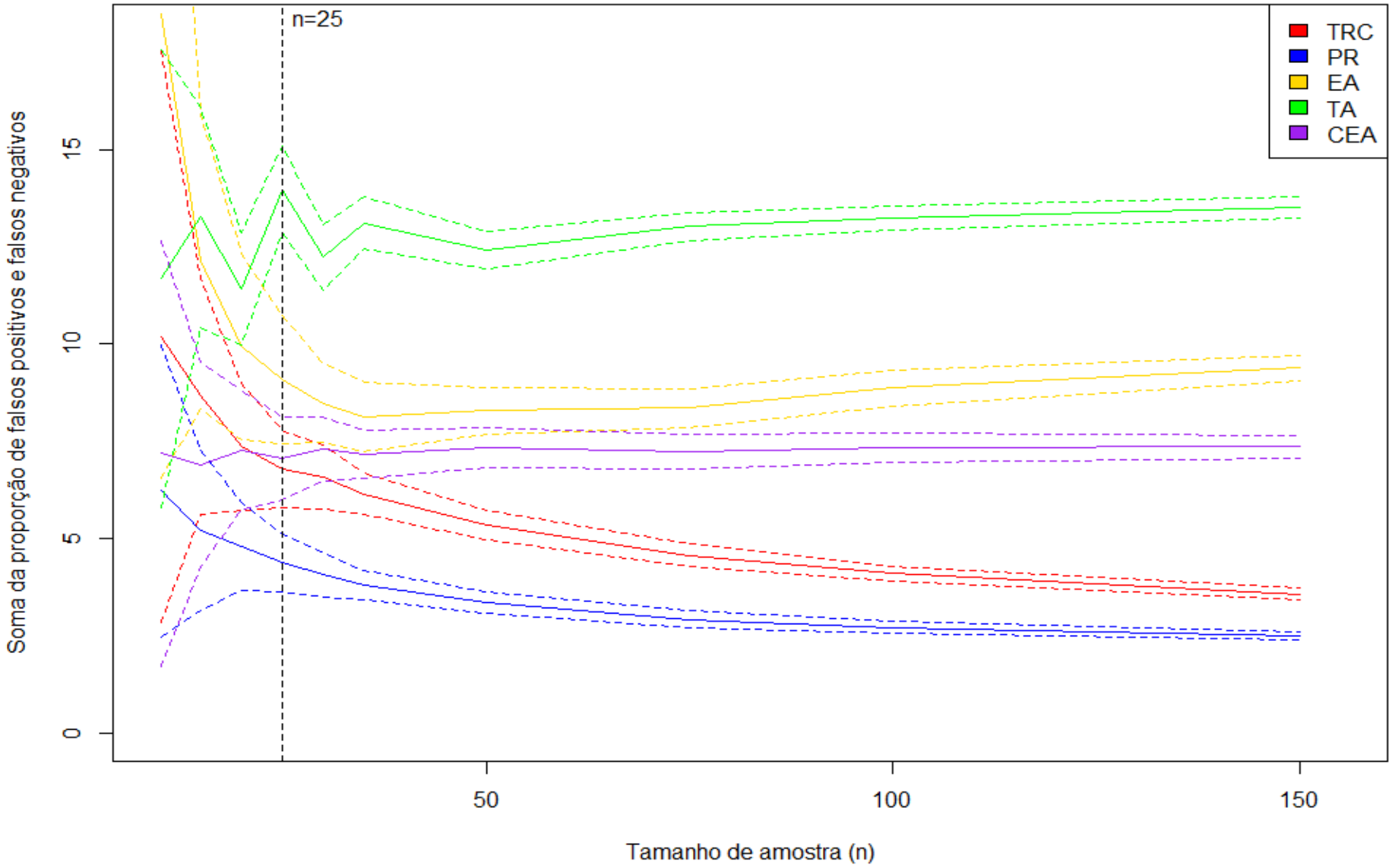

Falso negativo
TRC
PR
EA
TA
CEA
Proporção de falsos negativos
0
2
4
6
8
10
0
50
100
150
Tamanho de amostra (n)

Resultados


Falso positivo
TRC
PR
EA
TA
CEA
Proporção de falsos positivos
0
5
10
15
0
50
100
150
Tamanho de amostra (n)


Falso negativo
TRC
PR
EA
TA
CEA
Proporção de falsos negativos
0
2
4
6
8
10
0
50
100
150
Tamanho de amostra (n)

Resultados


n=25
TRC
PR
EA
TA
CEA
Soma da proporção de falsos positivos e falsos negativos
0
5
10
15
50
100
150
Tamanho de amostra (n)

Conclusões

- Presente estudo comparou diferentes métodos de detecção de outliers baseados na pesquisa do SINAPI.
- Para os tamanhos de amostra inferiores a 25 é difícil apontar o melhor método.
- Para tamanhos maiores que 25, o método do passo R apresentou os melhores resultados.
- O resultado encontrado mostra que o método do passo R pode ser de interesse para aplicação em dados oriundos de fontes de Big Data, onde as amostras possuem tamanhos grandes e métodos adequados de crítica ainda estão sendo desenvolvidos.
- Uma última observação é que para o algoritmo de Tukey apresentou os piores resultados para amostras grandes, mostrando-se inadequado para lidar com dados de Big Data e com grande contaminação de outliers.

Agradecimentos

Obrigado!

Bibliografia

Béguin, C.; Hulliger, B. (2003). Robust multivariate outlier detection and imputation with incomplete survey data. Deliverable D4/5.2.1/2 Part C, EUREDIT. Disponível em: https://www.cs.york.ac.uk/euredit/_temp/The%20Euredit%20Software/NAG%20Prototype%20platform/D45-2-12-C.pdf. Acesso em 30/08/2018

Casella, G., Berger, R. Statistical Inference. 2ª ed. California: Duxbury. 2002.

Goerg, G. 2015. The Lambert Way to Gaussianize heavy tailed data with the inverse of Tukey's h transformation as a special case. The Scientific World Journal. Disponível em: https://www.hindawi.com/journals/tswj/2015/909231/. Acesso em: 30/08/2018.

Hall P. e Wang Q. (2004). Exact convergence rate and leading term in central limit theorem for Student's t statistic. The Annals of Probability, Vol. 32, No. 2, páginas 1419–1437.

Hulliger B, Beguin C (2001). Detection of multivariate outliers by a simulated epidemic. In Proceedings of the ETK/NTTS Conference, pp667-676 Eurostat.

Bibliografia

IBGE, SINAPI: sistema nacional de pesquisa de custos e índices da construção civil, Manual metodológico, 2016.

Rais, S. (2008). Outlier Detection for the Consumer Price Index in Proceedings of the Survey Methods Section: Statistical Society of Canada Annual Meeting, May, 2008. 1–10 Ottawa, Ontario, Canadá. Disponível em: https://ssc.ca/sites/default/files/survey/documents/SSC2008_S_Rais.pdf. Acesso em 30/08/2018.

Silva, P. (1989). Crítica e imputação de dados quantitativos utilizando o SAS. Instituto de Matemática Pura e Aplicada.

Ververidis, C. e Kotropoulos, C. Gaussian mixture modeling by exploiting the Mahalanobis distance. IEEE Transactions on Signal Processing, Volume: 56 , Issue: 7 , July 2008.