ANOVA test

Yuri Cossich Lavinas

April, 2016

Contents

Algorithm 1 Obtain a Poisson deviate from a $[0, 1)$ value

```
Parameters 0 ≤ x < 1, μ ≥ 0
L ← exp(−μ), k ← 0, prob ← 1
repeat
    increment k
    prob ← prob * x
until prob > L
return k
```

Summary

O objeto é descobrir se existem variações ente os métodos e quais são as variáveis mais influentes.

Os métodos utilizados para comparação são o gaModel e a versão com listas. Para cada um dos métodos temos algumas variações nas varíaveis utilizadas. Variamos os anos (2005-2010), as regiões (Kanto, EastJapan, Touhoku e Kansai), a profundidade (<25km, <60km, <100km) e finalmente o catálogo utilizado (JMA X métodoJanelaJMA).

Experimental design

Vou utilizar o ANOVA para nos dados obtidos para verificar qual composição de variáveis e métodos mais influênciam no resultado final.

Para isso executei o *gaModel* e *versão com Listas* para cada conjunto de variáveis 10 vezes. Cada grupo para um método é composto por: região, ano, profundidade e catálogo. Um grupo para um cenário será chamado cenário de execução.

Após as execuções vou aplicar o ANOVA em uma data.frame composto pelos dados das **médias dos melhores indivíduos da última geração** para cada cenário de execução.

Caso uma variável esteja fora do intervalo de confiança ($P < 0.05$), vou aplicar novamente o ANOVA retirando essa variável do teste.

Aplico um teste post hoc nos resultados do ANOVO oara especificar quais são os grupos que diferem. O teste utilizado foi o Tukey teste.

Statistical Analysis

Começo a análise carregando o data.frame com os dados, seguindo para a aplicação do teste ANOVA e finalizando com o uso do Tukey teste.

```r
#Loading data
load("data.Rda")
#Taking a look at the data
summary(finalData)
```

```
##  loglikeValues                model     depths     years
##  Min.   :-3276.2   gaModel       :720   100:960   2005:480
##  1st Qu.:-1616.3   lista         :720   25 :960   2006:480
##  Median : -997.9   gaModelCluster:720   60 :960   2007:480
##  Mean   :-1199.1   listaCluster  :720             2008:480
##  3rd Qu.: -601.1                                  2009:480
##  Max.   : -237.7                                  2010:480
##       regions
##  Kanto    :720
##  Kansai   :720
##  Tohoku   :720
##  EastJapan:720
##
##
```

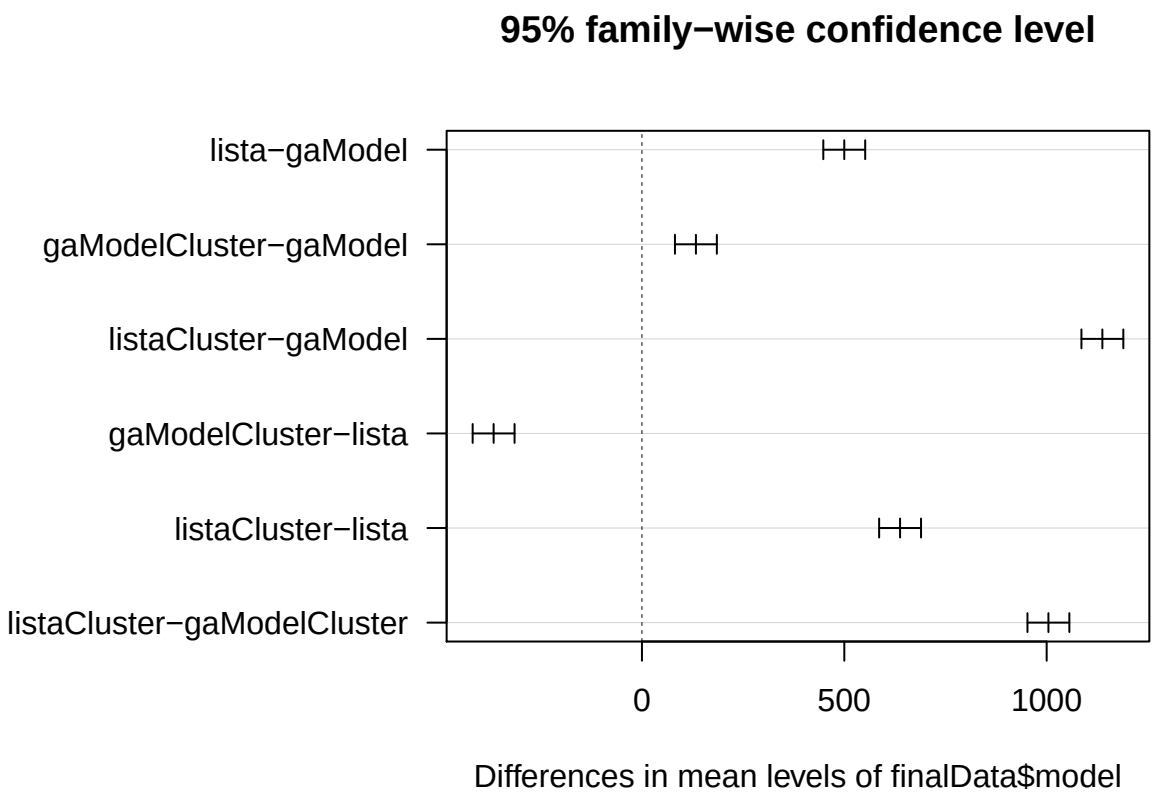
```r
#Primeira vez aplicando ANOVA
resultANOVA = aov(finalData$loglikeValues~finalData$model+finalData$depths+finalData$years+finalData$reg
summary(resultANOVA)
```

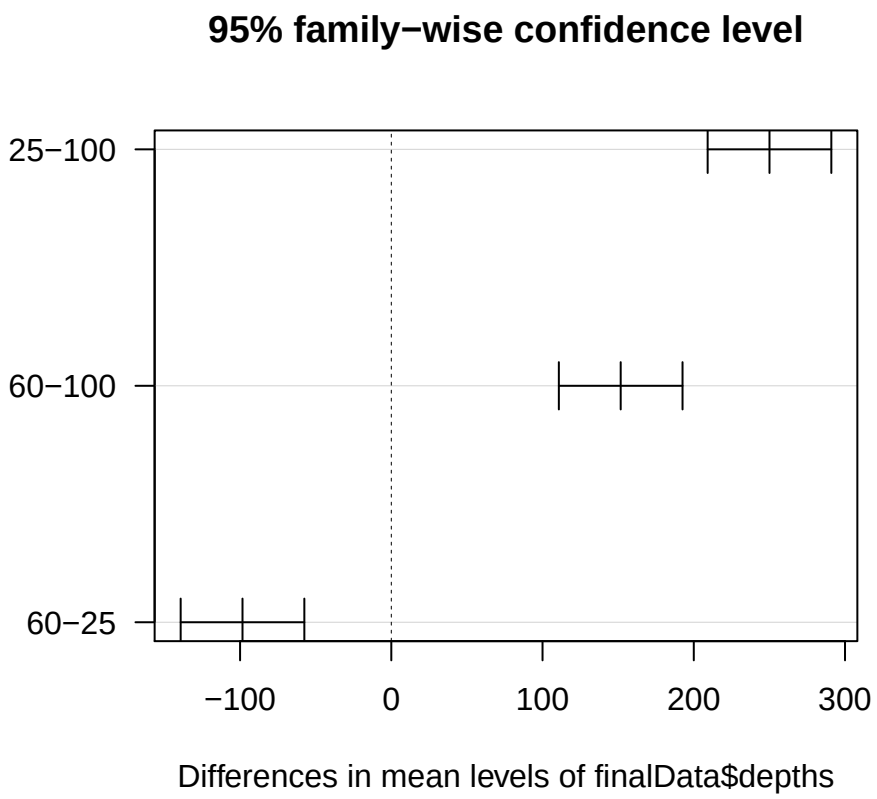
```
##                      Df    Sum Sq   Mean Sq  F value Pr(>F)
## finalData$model       3 560157617 186719206 1282.742 <2e-16 ***
## finalData$depths      2  30470265  15235133  104.664 <2e-16 ***
## finalData$years       5   1860545    372109    2.556 0.0257 *
## finalData$regions     3 306892155 102297385  702.772 <2e-16 ***
## Residuals          2866 417182410    145563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

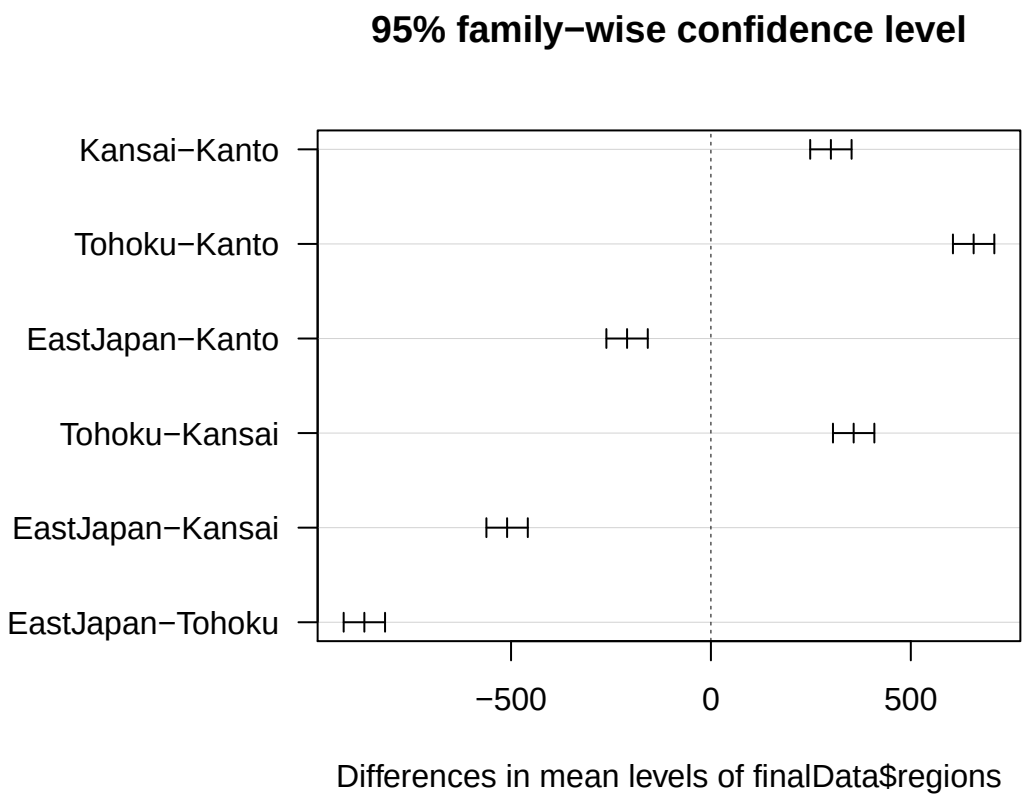
```r
#Segunda vez aplicando ANOVA, como a variável years influencia menos os dados foram removidos do teste A
resultANOVA = aov(finalData$loglikeValues~finalData$model+finalData$depths+finalData$regions)
summary(resultANOVA)
```

```
##                      Df    Sum Sq   Mean Sq F value Pr(>F)
## finalData$model       3 560157617 186719206  1279.3 <2e-16 ***
## finalData$depths      2  30470265  15235133   104.4 <2e-16 ***
## finalData$regions     3 306892155 102297385   700.9 <2e-16 ***
## Residuals          2871 419042955    145957
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Especificando quais são os grupos que diferem
tuk = TukeyHSD(resultANOVA)
#Variáveis para configuração do gráfico
# par(mfrow=c(2,2))
op <- par(mar = c(5,12,4,2) + 0.1)
#Função para gerar o gráfico
plot(tuk,las=1)
```




95% family−wise confidence level
25−100
60−100
60−25
−100
0
100
200
300
Differences in mean levels of finalData$depths




```r
#Mostrando os resultados também em texto
print(tuk)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = finalData$loglikeValues ~ finalData$model + finalData$depths + finalData$regions)
##
## $`finalData$model`
##                                   diff        lwr       upr p adj
## lista-gaModel                499.8954  448.13652  551.6544     0
## gaModelCluster-gaModel       133.3367   81.57777  185.0956     0
## listaCluster-gaModel        1137.7082 1085.94930 1189.4671     0
## gaModelCluster-lista        -366.5588 -418.31768 -314.7998     0
## listaCluster-lista           637.8128  586.05385  689.5717     0
## listaCluster-gaModelCluster 1004.3715  952.61261 1056.1305     0
##
## $`finalData$depths`
##             diff       lwr       upr p adj
## 25-100 250.06673  209.1765 290.95700 0e+00
## 60-100 151.67471  110.7844 192.56498 0e+00
## 60-25  -98.39202 -139.2823 -57.50176 1e-07
##
## $`finalData$regions`
##                       diff       lwr       upr p adj
## Kansai-Kanto      300.1009  248.3419  351.8598     0
```

```
## Tohoku-Kanto      657.1879  605.4289  708.9468     0
## EastJapan-Kanto  -209.7388 -261.4978 -157.9799     0
## Tohoku-Kansai     357.0870  305.3281  408.8459     0
## EastJapan-Kansai -509.8397 -561.5986 -458.0808     0
## EastJapan-Tohoku -866.9267 -918.6856 -815.1678     0
```

Conclusions