ANOVA test

Yuri Cossich Lavinas

May, 2016

Contents

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,a versão com listas, os sistemas híbridos (hybrid_gaModel e hybrid_lista). 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=>clustered).

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*, *versão com Listas*, *hybrid_gaModel* e *hybrid_lista* 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.

É importante resaltar que para todos os casos, aplico uma função de limite, que altera os valores do bins com mais que 12 ocorrências para 12. ## 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)
```

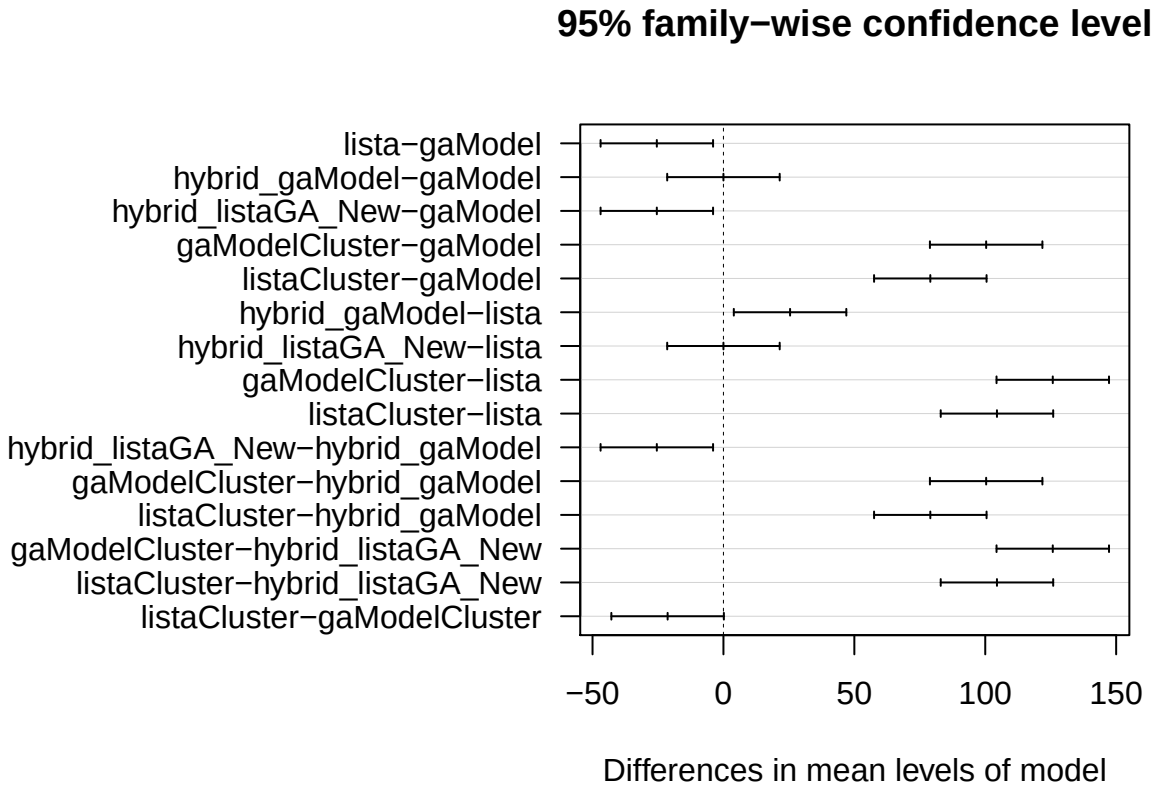
```
##  loglikeValues                  model     depths      years
##  Min.   :-2904   gaModel           :720   100:1440   2005:720
##  1st Qu.:-2055   lista             :720   25 :1440   2006:720
##  Median :-1658   hybrid_gaModel    :720   60 :1440   2007:720
```

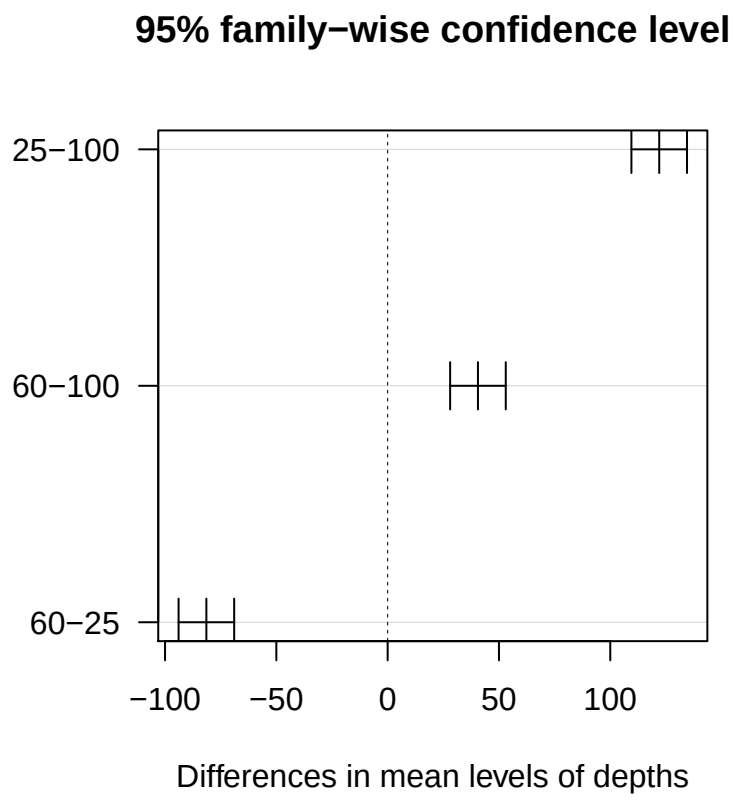
```
##  Mean   :-1669   hybrid_listaGA_New:720              2008:720
##  3rd Qu.:-1601   gaModelCluster    :720              2009:720
##  Max.   : -800   listaCluster      :720              2010:720
##       regions
##  Kanto    :1080
##  Kansai   :1080
##  Tohoku   :1080
##  EastJapan:1080
##
##
```

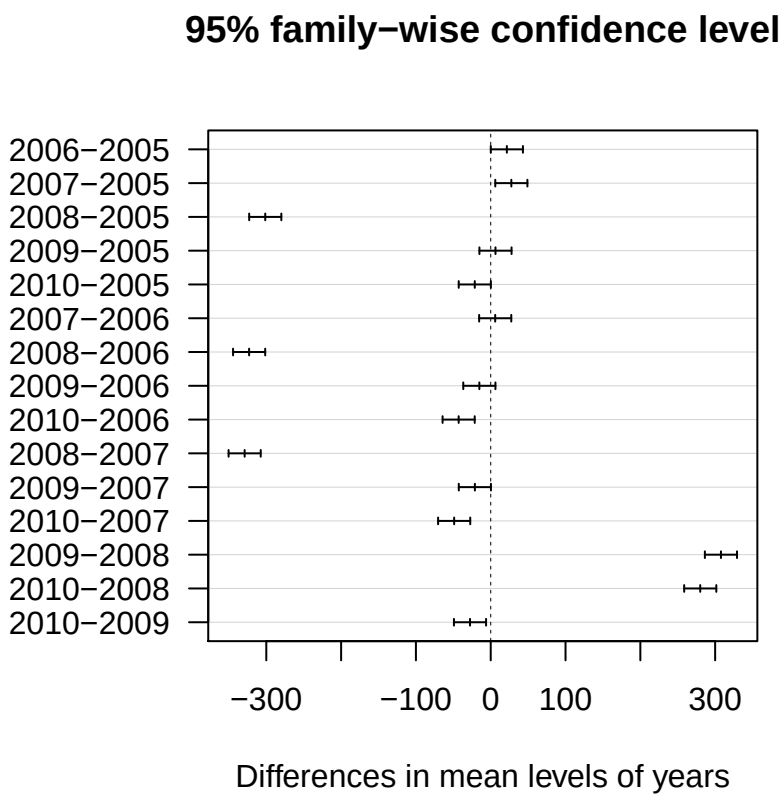
```r
#Primeira vez aplicando ANOVA
resultANOVA = aov(loglikeValues~model+depths+years+regions, data = finalData)
summary(resultANOVA)
```

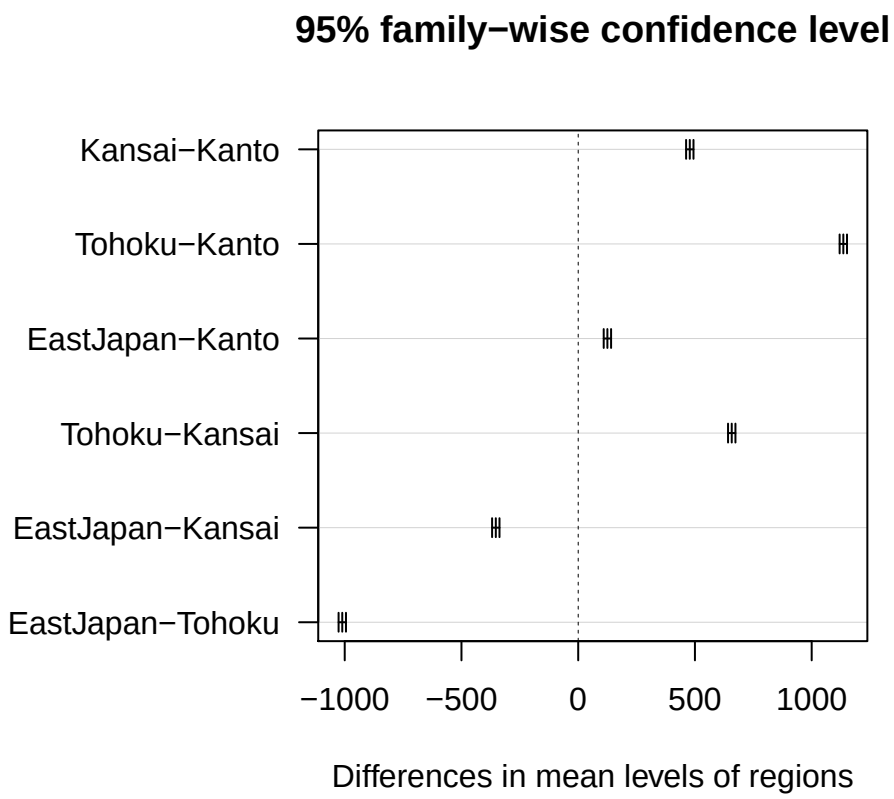
```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## model          5  10697549   2139510   104.7 <2e-16 ***
## depths         2  11116052   5558026   272.0 <2e-16 ***
## years          5  58073358  11614672   568.3 <2e-16 ***
## regions        3 840109826 280036609 13702.1 <2e-16 ***
## Residuals   4304  87962995     20437
## ---
## 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(5,16,4,2) + 0.1)
#Função para gerar o gráfico
plot(tuk,las=1)
```


95% family−wise confidence level
lista−gaModel
hybrid_gaModel−gaModel
hybrid_listaGA_New−gaModel
gaModelCluster−gaModel
listaCluster−gaModel
hybrid_gaModel−lista
hybrid_listaGA_New−lista
gaModelCluster−lista
listaCluster−lista
hybrid_listaGA_New−hybrid_gaModel
gaModelCluster−hybrid_gaModel
listaCluster−hybrid_gaModel
gaModelCluster−hybrid_listaGA_New
listaCluster−hybrid_listaGA_New
listaCluster−gaModelCluster
−50
0
50
100
150
Differences in mean levels of model


95% family−wise confidence level
25−100
60−100
60−25
−100
−50
0
50
100
Differences in mean levels of depths


95% family−wise confidence level
2006−2005
2007−2005
2008−2005
2009−2005
2010−2005
2007−2006
2008−2006
2009−2006
2010−2006
2008−2007
2009−2007
2010−2007
2009−2008
2010−2008
2010−2009
−300
−100
0
100
300
Differences in mean levels of years




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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = loglikeValues ~ model + depths + years + regions, data = finalData)
##
## $model
##                                          diff        lwr        upr
## lista-gaModel                     -2.544743e+01 -46.928636  -3.966217
## hybrid_gaModel-gaModel            -9.094947e-13 -21.481209  21.481209
## hybrid_listaGA_New-gaModel        -2.544743e+01 -46.928636  -3.966217
## gaModelCluster-gaModel             1.003448e+02  78.863606 121.826024
## listaCluster-gaModel               7.902067e+01  57.539461 100.501879
## hybrid_gaModel-lista               2.544743e+01   3.966217  46.928636
## hybrid_listaGA_New-lista          -6.821210e-13 -21.481209  21.481209
## gaModelCluster-lista               1.257922e+02 104.311032 147.273451
## listaCluster-lista                 1.044681e+02  82.986887 125.949306
## hybrid_listaGA_New-hybrid_gaModel -2.544743e+01 -46.928636  -3.966217
## gaModelCluster-hybrid_gaModel      1.003448e+02  78.863606 121.826024
## listaCluster-hybrid_gaModel        7.902067e+01  57.539461 100.501879
## gaModelCluster-hybrid_listaGA_New  1.257922e+02 104.311032 147.273451
## listaCluster-hybrid_listaGA_New    1.044681e+02  82.986887 125.949306
## listaCluster-gaModelCluster       -2.132415e+01 -42.805355   0.157064
##                                       p adj
```

```
## lista-gaModel                      0.0096156
## hybrid_gaModel-gaModel             1.0000000
## hybrid_listaGA_New-gaModel         0.0096156
## gaModelCluster-gaModel             0.0000000
## listaCluster-gaModel               0.0000000
## hybrid_gaModel-lista               0.0096156
## hybrid_listaGA_New-lista           1.0000000
## gaModelCluster-lista               0.0000000
## listaCluster-lista                 0.0000000
## hybrid_listaGA_New-hybrid_gaModel 0.0096156
## gaModelCluster-hybrid_gaModel      0.0000000
## listaCluster-hybrid_gaModel        0.0000000
## gaModelCluster-hybrid_listaGA_New 0.0000000
## listaCluster-hybrid_listaGA_New   0.0000000
## listaCluster-gaModelCluster        0.0530072
##
## $depths
##             diff       lwr       upr p adj
## 25-100 121.99302 109.50193 134.48411     0
## 60-100  40.56372  28.07263  53.05481     0
## 60-25  -81.42930 -93.92039 -68.93821     0
##
## $years
##                  diff           lwr          upr     p adj
## 2006-2005   21.562653    0.08144347   43.0438620 0.0484981
## 2007-2005   27.555041    6.07383211   49.0362506 0.0035149
## 2008-2005 -301.395222 -322.87643141 -279.9140129 0.0000000
## 2009-2005    6.371453  -15.10975629   27.8526622 0.9588818
## 2010-2005  -21.306557  -42.78776612    0.1746524 0.0533533
## 2007-2006    5.992389  -15.48882062   27.4735979 0.9683884
## 2008-2006 -322.957875 -344.43908415 -301.4766656 0.0000000
## 2009-2006  -15.191200  -36.67240903    6.2900095 0.3331095
## 2010-2006  -42.869210  -64.35041886  -21.3880003 0.0000002
## 2008-2007 -328.950264 -350.43147278 -307.4690543 0.0000000
## 2009-2007  -21.183588  -42.66479766    0.2976209 0.0558258
## 2010-2007  -48.861598  -70.34280750  -27.3803890 0.0000000
## 2009-2008  307.766675  286.28546586  329.2478844 0.0000000
## 2010-2008  280.088665  258.60745603  301.5698746 0.0000000
## 2010-2009  -27.678010  -49.15921909   -6.1968006 0.0033056
##
## $regions
##                        diff        lwr       upr p adj
## Kansai-Kanto       477.9282   462.1174  493.7391     0
## Tohoku-Kanto      1135.5248  1119.7139 1151.3357     0
## EastJapan-Kanto    125.0368   109.2259  140.8477     0
## Tohoku-Kansai      657.5966   641.7857  673.4075     0
## EastJapan-Kansai  -352.8914  -368.7023 -337.0806     0
## EastJapan-Tohoku -1010.4880 -1026.2989 -994.6771     0
```

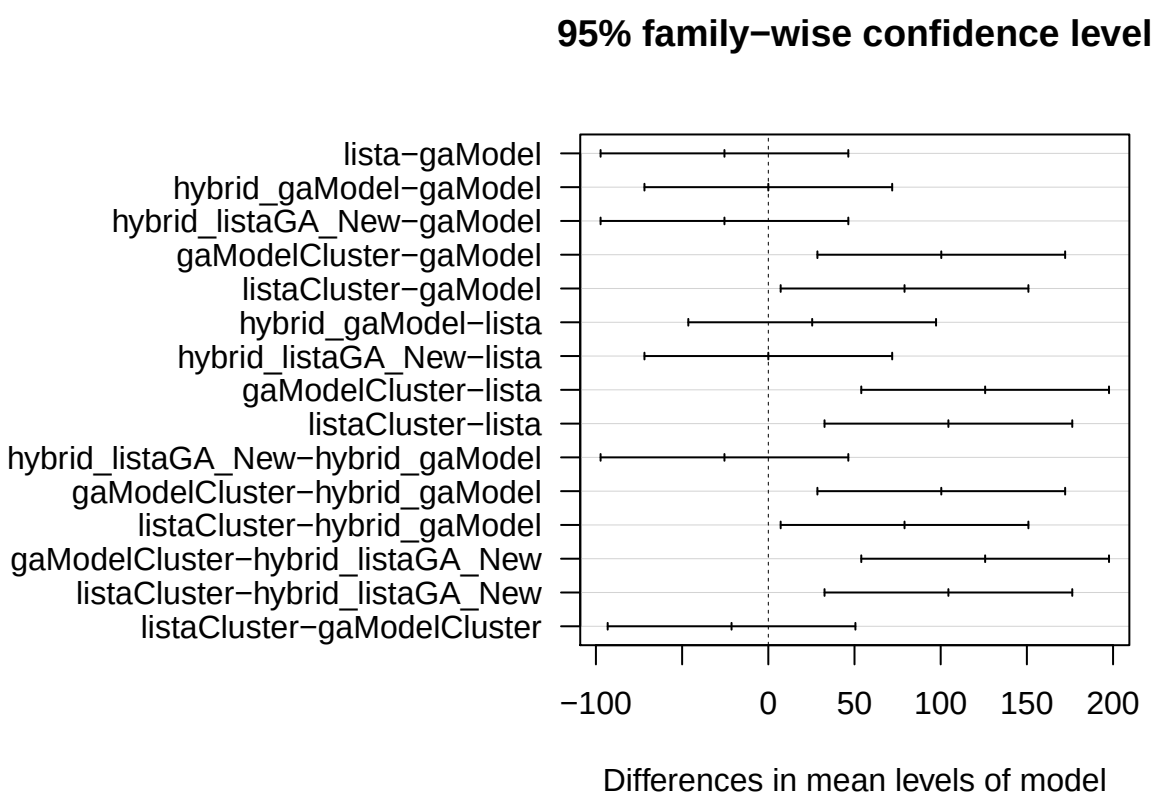
Dado que para os primeiros resultados, temos que todas as variáveis estão dentro do intervalo de confiança. Porém, entendo que tanto para os anos quanto para as regiões, essas variações já era previstas e poucom acrescentam ao estudo.

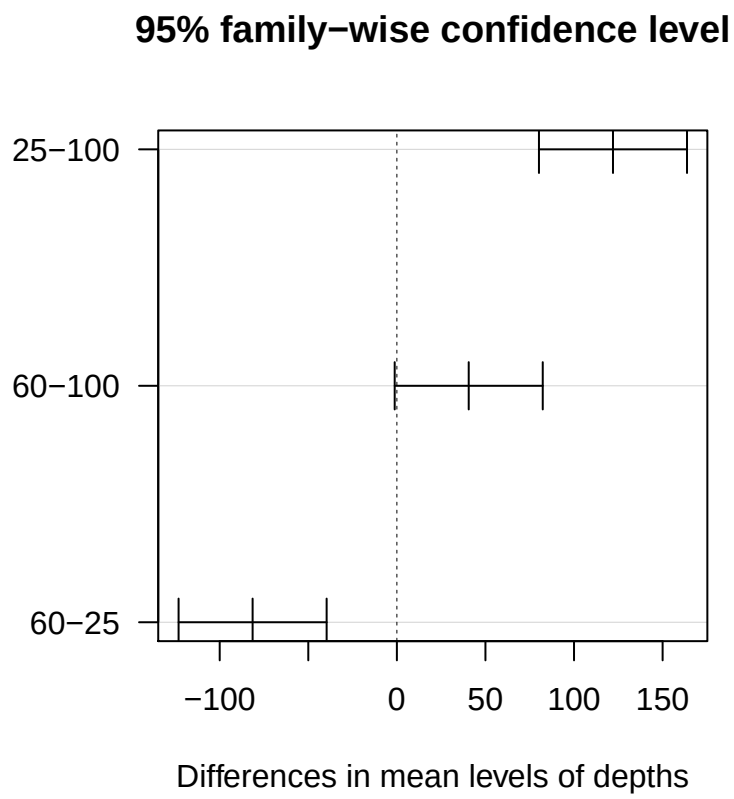
Baseado nisso, refiz os cálculos, seguindo o mesmo processo feito anteriormente, a fim de simplificar a análise dos resultados.

```r
#Segunda vez aplicando ANOVA, como a variável years e region tem incluência esperada, foram retiradas
resultANOVA = aov(loglikeValues~model+depths, data = finalData)
summary(resultANOVA)
```

```
##               Df    Sum Sq Mean Sq F value   Pr(>F)
## model          5  10697549 2139510   9.355 7.05e-09 ***
## depths         2  11116052 5558026  24.303 3.19e-11 ***
## Residuals   4312 986146178  228698
## ---
## 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(5,16,4,2) + 0.1)
#Função para gerar o gráfico
plot(tuk,las=1)
```






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

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = loglikeValues ~ model + depths, data = finalData)
##
## $model
##                                          diff        lwr       upr
## lista-gaModel                     -2.544743e+01 -97.305532  46.41068
## hybrid_gaModel-gaModel            -9.094947e-13 -71.858105  71.85811
## hybrid_listaGA_New-gaModel        -2.544743e+01 -97.305532  46.41068
## gaModelCluster-gaModel             1.003448e+02  28.486710 172.20292
## listaCluster-gaModel               7.902067e+01   7.162564 150.87878
## hybrid_gaModel-lista               2.544743e+01 -46.410679  97.30553
## hybrid_listaGA_New-lista          -6.821210e-13 -71.858105  71.85811
## gaModelCluster-lista               1.257922e+02  53.934136 197.65035
## listaCluster-lista                 1.044681e+02  32.609991 176.32620
## hybrid_listaGA_New-hybrid_gaModel -2.544743e+01 -97.305532  46.41068
## gaModelCluster-hybrid_gaModel      1.003448e+02  28.486710 172.20292
## listaCluster-hybrid_gaModel        7.902067e+01   7.162564 150.87878
## gaModelCluster-hybrid_listaGA_New  1.257922e+02  53.934136 197.65035
## listaCluster-hybrid_listaGA_New    1.044681e+02  32.609991 176.32620
## listaCluster-gaModelCluster       -2.132415e+01 -93.182251  50.53396
##                                       p adj
```

```
## lista-gaModel                      0.9149041
## hybrid_gaModel-gaModel             1.0000000
## hybrid_listaGA_New-gaModel         0.9149041
## gaModelCluster-gaModel             0.0009838
## listaCluster-gaModel               0.0213881
## hybrid_gaModel-lista               0.9149041
## hybrid_listaGA_New-lista           1.0000000
## gaModelCluster-lista               0.0000093
## listaCluster-lista                 0.0004959
## hybrid_listaGA_New-hybrid_gaModel 0.9149041
## gaModelCluster-hybrid_gaModel      0.0009838
## listaCluster-hybrid_gaModel        0.0213881
## gaModelCluster-hybrid_listaGA_New 0.0000093
## listaCluster-hybrid_listaGA_New   0.0004959
## listaCluster-gaModelCluster        0.9587951
##
## $depths
##             diff         lwr       upr     p adj
## 25-100 121.99302   80.208296 163.77774 0.0000000
## 60-100  40.56372   -1.221001  82.34844 0.0593107
## 60-25  -81.42930 -123.214018 -39.64458 0.0000151
```

Conclusions