

# Final ANOVA and Paired Design

*Yuri Lavinas*

*May 31, 2016*

## Contents

Summary . . . . .	1
Statistical Analysis . . . . .	1
ANOVA - Specific analysis somente com Cluster. . . . .	5
Conclusion . . . . .	15

## 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 variáveis 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 métodoJanela.JMA=>clustered).

## Statistical Analysis

### ANOVA test and HSD Tukey

Vou utilizar o ANOVA para nos dados obtidos para verificar qual composição de variáveis e métodos mais influenciam 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 ANOVA para 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.

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.

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.

```
#Taking a look at the data
summary(finalData)
```

```
## loglikeValues      model      depths      years
## Min.      :-3158   gaModel      :720   100:1440   2005:720
## 1st Qu.   :-2079   lista      :720   25 :1440   2006:720
## Median   :-1679   hybrid_gaModel :720   60 :1440   2007:720
## Mean     :-1702   hybrid_listaGA_New:720      2008:720
## 3rd Qu.  :-1602   gaModelCluster :720      2009:720
## Max.      : -800   listaCluster :720      2010:720
##      regions
## Kanto     :1080
## Kansai    :1080
## Tohoku    :1080
## EastJapan:1080
##
##
```

*#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   31424058    6284812   255.0 <2e-16 ***
## depths        2   16077491    8038746    326.2 <2e-16 ***
## years         5   57908014   11581603    470.0 <2e-16 ***
## regions       3   878253346  292751115  11879.4 <2e-16 ***
## Residuals    4304 106066400      24644
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*#Segunda vez aplicando ANOVA, como a variável years influencia menos os dados foram removidos do teste .*

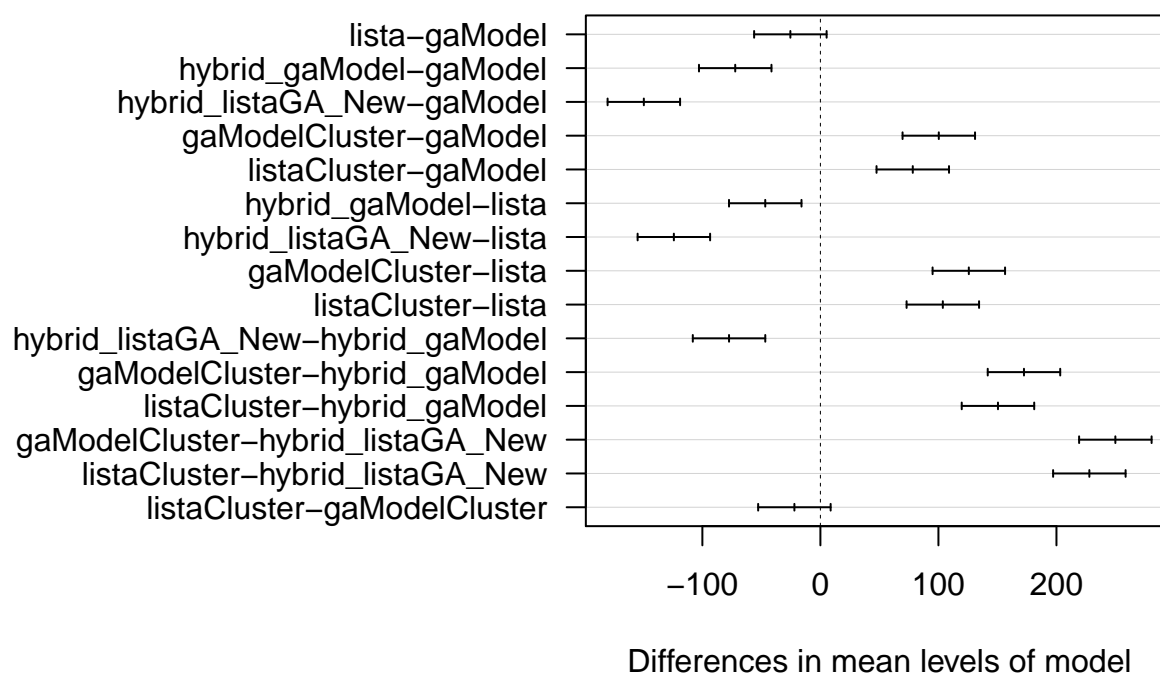
```
resultANOVA = aov(loglikeValues~model+regions , data = finalData)
summary(resultANOVA)
```

```
##              Df      Sum Sq   Mean Sq F value Pr(>F)
## model         5   31424058    6284812   150.5 <2e-16 ***
## regions       3   878253346  292751115   7009.4 <2e-16 ***
## Residuals    4311 180051906      41766
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

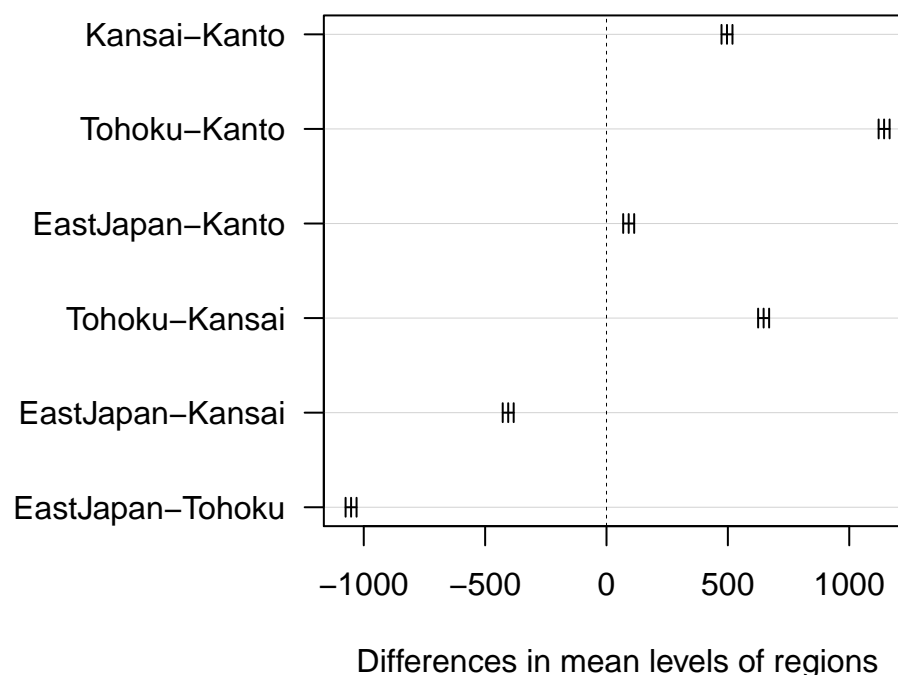
*#Especificando quais são os grupos que diferem*

```
tuk = TukeyHSD(resultANOVA)
#Variáveis para configuração do gráfico
# par(mfrow=c(2,2))
op <- par(mar = c(5,15,4,2) + 0.1)
#Função para gerar o gráfico
plot(tuk,las=1)
```

## 95% family-wise confidence level



## 95% family-wise confidence level



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

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = loglikeValues ~ model + regions, data = finalData)
##
## $model
##
```

	diff	lwr	upr
lista-gaModel	-25.44743	-56.15563	5.260775
hybrid_gaModel-gaModel	-72.18184	-102.89005	-41.473643
hybrid_listaGA_New-gaModel	-149.60620	-180.31441	-118.898003
gaModelCluster-gaModel	100.31468	69.60648	131.022879
listaCluster-gaModel	78.27546	47.56726	108.983663
hybrid_gaModel-lista	-46.73442	-77.44262	-16.026217
hybrid_listaGA_New-lista	-124.15878	-154.86698	-93.450577
gaModelCluster-lista	125.76210	95.05390	156.470306
listaCluster-lista	103.72289	73.01469	134.431089
hybrid_listaGA_New-hybrid_gaModel	-77.42436	-108.13256	-46.716158
gaModelCluster-hybrid_gaModel	172.49652	141.78832	203.204724
listaCluster-hybrid_gaModel	150.45731	119.74910	181.165508
gaModelCluster-hybrid_listaGA_New	249.92088	219.21268	280.629084
listaCluster-hybrid_listaGA_New	227.88167	197.17346	258.589868
listaCluster-gaModelCluster	-22.03922	-52.74742	8.668985

```
##
## p adj
```

```
## lista-gaModel 0.1697991
## hybrid_gaModel-gaModel 0.0000000
## hybrid_listaGA_New-gaModel 0.0000000
## gaModelCluster-gaModel 0.0000000
## listaCluster-gaModel 0.0000000
## hybrid_gaModel-lista 0.0002122
## hybrid_listaGA_New-lista 0.0000000
## gaModelCluster-lista 0.0000000
## listaCluster-lista 0.0000000
## hybrid_listaGA_New-hybrid_gaModel 0.0000000
## gaModelCluster-hybrid_gaModel 0.0000000
## listaCluster-hybrid_gaModel 0.0000000
## gaModelCluster-hybrid_listaGA_New 0.0000000
## listaCluster-hybrid_listaGA_New 0.0000000
## listaCluster-gaModelCluster 0.3163227
##
## $regions
## diff lwr upr p adj
## Kansai-Kanto 496.37177 473.7695 518.9740 0
## Tohoku-Kanto 1143.63742 1121.0352 1166.2397 0
## EastJapan-Kanto 91.40506 68.8028 114.0073 0
## Tohoku-Kansai 647.26565 624.6634 669.8679 0
## EastJapan-Kansai -404.96671 -427.5690 -382.3645 0
## EastJapan-Tohoku -1052.23236 -1074.8346 -1029.6301 0
```

## ANOVA - Specific analysis somente com Cluster.

Faço o ANOVA somente para os modelos “clusterizados”

Primeiro crio o data frame somente com os modelos citados

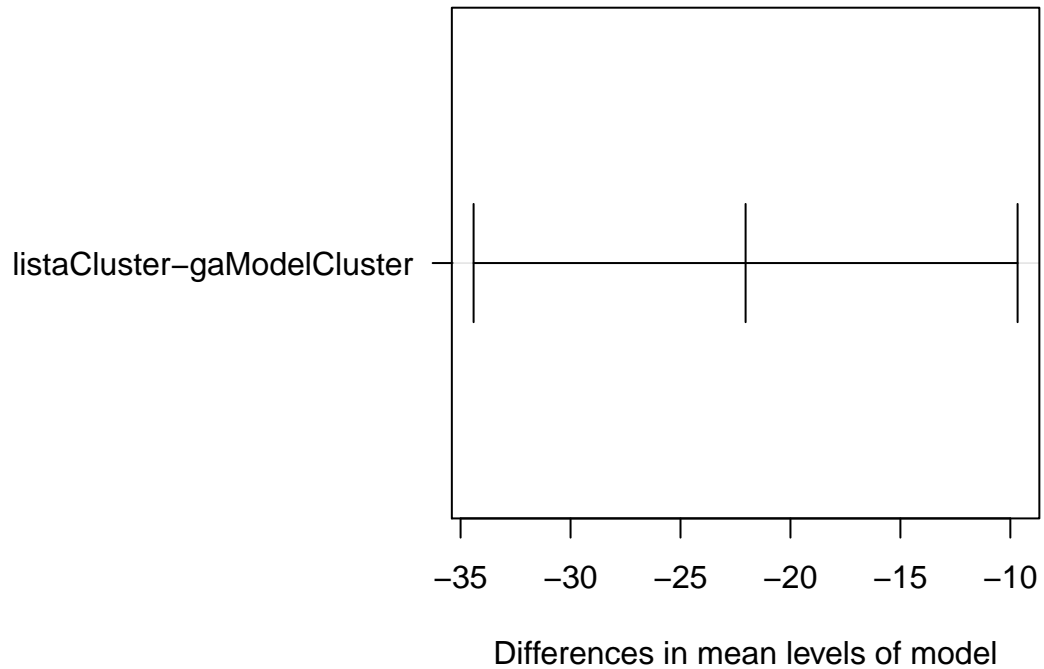
```
subTabela = finalData[finalData$model=='gaModelCluster'|finalData$model=='listaCluster',]
summary(subTabela)
```

```
## loglikeValues model depths years
## Min. :-2420 gaModel : 0 100:480 2005:240
## 1st Qu.:-2032 lista : 0 25 :480 2006:240
## Median :-1634 hybrid_gaModel : 0 60 :480 2007:240
## Mean :-1601 hybrid_listaGA_New: 0 2008:240
## 3rd Qu.:-1574 gaModelCluster :720 2009:240
## Max. : -800 listaCluster :720 2010:240
## regions
## Kanto :360
## Kansai :360
## Tohoku :360
## EastJapan:360
##
##
```

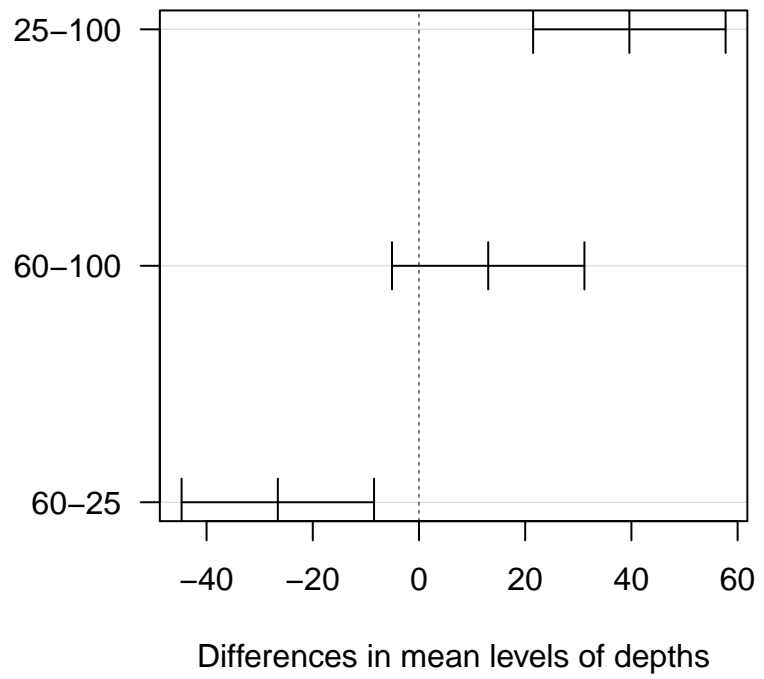
Aplico o anova, com a regressão para modelos, profundidades, anos e regiões.

```
resultANOVA = aov(loglikeValues~model+depths+years+regions , data = subTabela)
tuk = TukeyHSD(resultANOVA)
op <- par(mar = c(5,15,4,2) + 0.1)
plot(tuk,las=1)
```

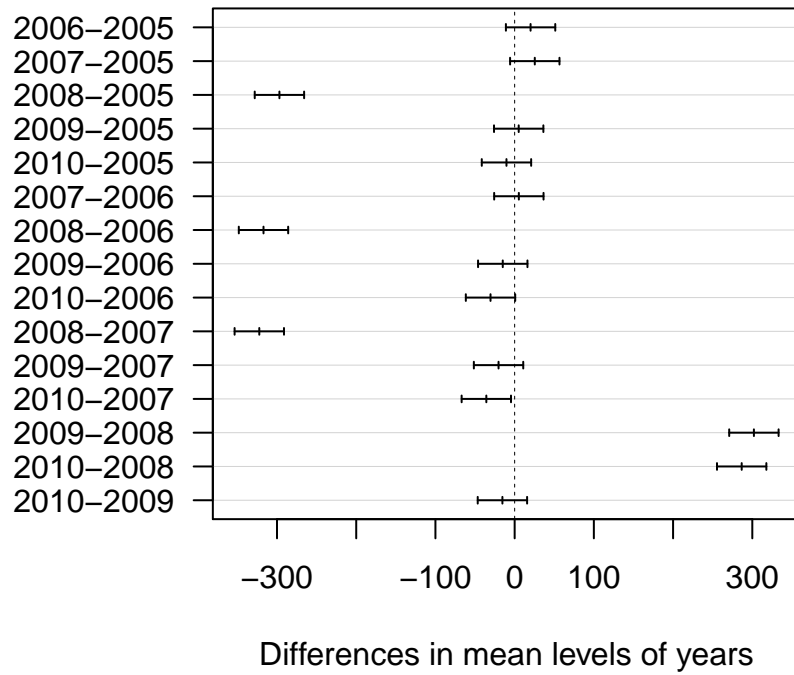
### 95% family-wise confidence level



### 95% family-wise confidence level

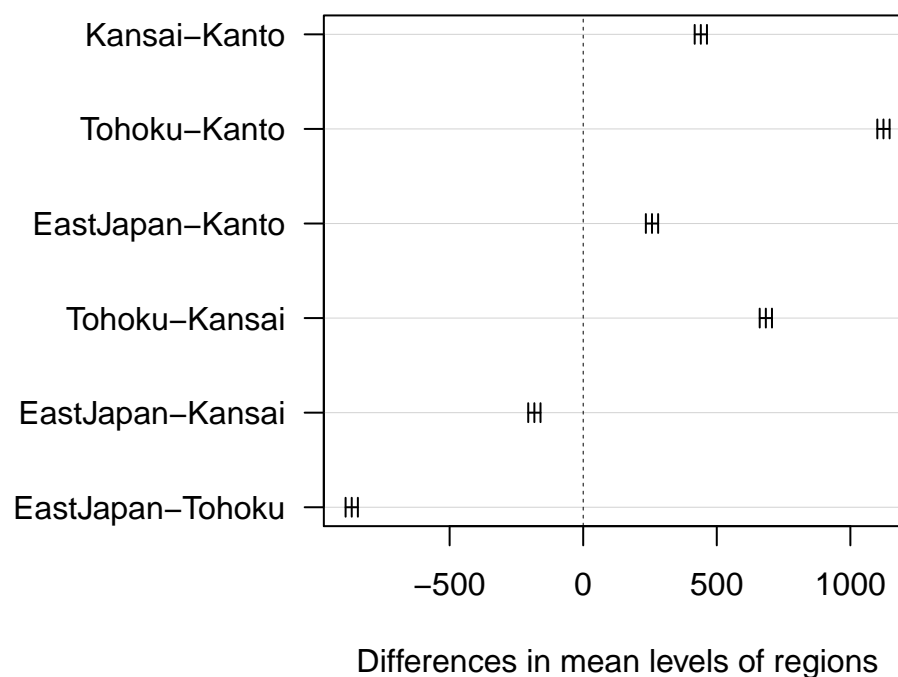


### 95% family-wise confidence level





## 95% family-wise confidence level



```
print(tuk)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = loglikeValues ~ model + depths + years + regions, data = subTabela)
##
## $model
##               diff      lwr      upr      p adj
## listaCluster-gaModelCluster -22.03922 -34.4088 -9.669629 0.0004884
##
## $depths
##           diff      lwr      upr      p adj
## 25-100  39.61761  21.498345  57.736877 0.0000010
## 60-100  13.03637  -5.082893  31.155638 0.2101135
## 60-25  -26.58124 -44.700504  -8.461972 0.0017206
##
## $years
##           diff      lwr      upr      p adj
## 2006-2005  20.120125 -11.046557  51.2868058 0.4388396
## 2007-2005  25.428493  -5.738188  56.5951741 0.1833601
## 2008-2005 -296.931458 -328.098139 -265.7647768 0.0000000
## 2009-2005   5.110504 -26.056177  36.2771851 0.9972103
## 2010-2005 -10.329327 -41.496008  20.8373539 0.9344671
## 2007-2006   5.308368 -25.858313  36.4750495 0.9966586
```

```
## 2008-2006 -317.051583 -348.218264 -285.8849014 0.0000000
## 2009-2006 -15.009621 -46.176302 16.1570605 0.7425591
## 2010-2006 -30.449452 -61.616133 0.7172293 0.0599571
## 2008-2007 -322.359951 -353.526632 -291.1932697 0.0000000
## 2009-2007 -20.317989 -51.484670 10.8486923 0.4273269
## 2010-2007 -35.757820 -66.924501 -4.5911390 0.0138287
## 2009-2008 302.041962 270.875281 333.2086431 0.0000000
## 2010-2008 286.602131 255.435450 317.7688119 0.0000000
## 2010-2009 -15.439831 -46.606512 15.7268500 0.7187745
##
## $regions
##          diff          lwr          upr p adj
## Kansai-Kanto    440.2862   417.3493   463.2231    0
## Tohoku-Kanto    1123.9321  1100.9952  1146.8690    0
## EastJapan-Kanto  257.4748   234.5379   280.4117    0
## Tohoku-Kansai    683.6459   660.7090   706.5828    0
## EastJapan-Kansai -182.8114  -205.7483  -159.8745    0
## EastJapan-Tohoku -866.4573  -889.3942  -843.5204    0
```

## Paired Design - Student t-test

Agora faço o Paired Design t.test aplicando para todas as combinações possíveis de modelos, em todas as regiões e profundidades, para todos os anos.

Baseado nos arquivos que explicam o Paired Desing, escrevi o código a seguir. Porém não entendi porque ao fazer desta forma pode ser considerado um teste pareado. Os slides comparam duas formas de realizar este tipo de teste. Uma delas tem *seta* um parametro da função com **True**, explicitando que é um teste pareado. Já para o outra forma, esse parametro fica com **False**.

```
summary(finalData)
```

```
## loglikeValues          model      depths      years
## Min.      :-3158   gaModel          :720   100:1440   2005:720
## 1st Qu.   :-2079   lista            :720   25 :1440   2006:720
## Median    :-1679   hybrid_gaModel    :720   60 :1440   2007:720
## Mean      :-1702   hybrid_listaGA_New:720           2008:720
## 3rd Qu.   :-1602   gaModelCluster    :720           2009:720
## Max.      : -800   listaCluster      :720           2010:720
##          regions
## Kanto      :1080
## Kansai     :1080
## Tohoku     :1080
## EastJapan:1080
##
##
```

```
# Summarize the n=30 repeated measures on each Problem:Algorithm combination by their mean value
ttestPaired= function(region){
  subTabela = finalData[finalData$depths==25&finalData$regions==region,]
  aggfinaldata<-aggregate(loglikeValues~years:model, data=subTabela,FUN=mean)
  # Perform paired t-test
  cat('in', region, 'the t.test between the models gaModel and lista is: ')
  difTimes<-with(aggfinaldata,loglikeValues[1:6]-loglikeValues[7:12])
```

```

print(t.test(difTimes))
cat('in', region, 'the t.test between the models gaModel and hybrid_gaModel is: ')
difTimes<-with(aggfinaldata,loglikeValues[1:6]-loglikeValues[13:18])
print(t.test(difTimes))
cat('in', region, 'the t.test between the models gaModel and hybrid_listaGA_New is: ')
difTimes<-with(aggfinaldata,loglikeValues[1:6]-loglikeValues[19:24])
print(t.test(difTimes))
cat('in', region, 'the t.test between the models gaModel and gaModelCluster is: ')
difTimes<-with(aggfinaldata,loglikeValues[1:6]-loglikeValues[25:30])
print(t.test(difTimes))
cat('in', region, 'the t.test between the models gaModel and listaCluster is: ')
difTimes<-with(aggfinaldata,loglikeValues[1:6]-loglikeValues[31:36])
print(t.test(difTimes))
}

ttestPaired('Kansai')

```

```

## in Kansai the t.test between the models gaModel and lista is:
## One Sample t-test
##
## data: difTimes
## t = 10.637, df = 5, p-value = 0.000127
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 24.06675 39.40531
## sample estimates:
## mean of x
## 31.73603
##
## in Kansai the t.test between the models gaModel and hybrid_gaModel is:
## One Sample t-test
##
## data: difTimes
## t = 1.1955, df = 5, p-value = 0.2855
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -4.693085 12.852947
## sample estimates:
## mean of x
## 4.079931
##
## in Kansai the t.test between the models gaModel and hybrid_listaGA_New is:
## One Sample t-test
##
## data: difTimes
## t = 17.138, df = 5, p-value = 1.238e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 79.70227 107.83100
## sample estimates:
## mean of x
## 93.76664
##

```

```

## in Kansai the t.test between the models gaModel and gaModelCluster is:
## One Sample t-test
##
## data: difTimes
## t = -3.2157, df = 5, p-value = 0.02358
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -12.753743 -1.422024
## sample estimates:
## mean of x
## -7.087883
##
## in Kansai the t.test between the models gaModel and listaCluster is:
## One Sample t-test
##
## data: difTimes
## t = 4.7105, df = 5, p-value = 0.005287
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 6.07558 20.67227
## sample estimates:
## mean of x
## 13.37392

```

```
ttestPaired('Tohoku')
```

```

## in Tohoku the t.test between the models gaModel and lista is:
## One Sample t-test
##
## data: difTimes
## t = -1.622, df = 5, p-value = 0.1657
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -13.769220 3.115127
## sample estimates:
## mean of x
## -5.327047
##
## in Tohoku the t.test between the models gaModel and hybrid_gaModel is:
## One Sample t-test
##
## data: difTimes
## t = 6.624, df = 5, p-value = 0.001181
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 25.53039 57.91218
## sample estimates:
## mean of x
## 41.72128
##
## in Tohoku the t.test between the models gaModel and hybrid_listaGA_New is:
## One Sample t-test
##
## data: difTimes

```

```
## t = 3.3329, df = 5, p-value = 0.02071
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 8.308453 64.338805
## sample estimates:
## mean of x
## 36.32363
##
## in Tohoku the t.test between the models gaModel and gaModelCluster is:
## One Sample t-test
##
## data: difTimes
## t = -9.4035, df = 5, p-value = 0.0002294
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -36.89030 -21.05111
## sample estimates:
## mean of x
## -28.97071
##
## in Tohoku the t.test between the models gaModel and listaCluster is:
## One Sample t-test
##
## data: difTimes
## t = -6.257, df = 5, p-value = 0.001529
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -33.37542 -13.93769
## sample estimates:
## mean of x
## -23.65656
```

```
ttestPaired('EastJapan')
```

```
## in EastJapan the t.test between the models gaModel and lista is:
## One Sample t-test
##
## data: difTimes
## t = 1.9129, df = 5, p-value = 0.114
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -10.07453 68.68307
## sample estimates:
## mean of x
## 29.30427
##
## in EastJapan the t.test between the models gaModel and hybrid_gaModel is:
## One Sample t-test
##
## data: difTimes
## t = 6.5282, df = 5, p-value = 0.001262
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 61.94934 142.42401
```

```

## sample estimates:
## mean of x
## 102.1867
##
## in EastJapan the t.test between the models gaModel and hybrid_listaGA_New is:
## One Sample t-test
##
## data: difTimes
## t = 11.564, df = 5, p-value = 8.482e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 156.1337 245.3855
## sample estimates:
## mean of x
## 200.7596
##
## in EastJapan the t.test between the models gaModel and gaModelCluster is:
## One Sample t-test
##
## data: difTimes
## t = -8.8802, df = 5, p-value = 0.0003012
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -109.10009 -60.11634
## sample estimates:
## mean of x
## -84.60822
##
## in EastJapan the t.test between the models gaModel and listaCluster is:
## One Sample t-test
##
## data: difTimes
## t = -5.4451, df = 5, p-value = 0.002837
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -93.31209 -33.46295
## sample estimates:
## mean of x
## -63.38752

```

```
ttestPaired('Kanto')
```

```

## in Kanto the t.test between the models gaModel and lista is:
## One Sample t-test
##
## data: difTimes
## t = 4.1215, df = 5, p-value = 0.00916
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 4.985032 21.509870
## sample estimates:
## mean of x
## 13.24745
##

```

```

## in Kanto the t.test between the models gaModel and hybrid_gaModel is:
## One Sample t-test
##
## data: difTimes
## t = 1.3808, df = 5, p-value = 0.2259
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -1.542245 5.122177
## sample estimates:
## mean of x
## 1.789966
##
## in Kanto the t.test between the models gaModel and hybrid_listaGA_New is:
## One Sample t-test
##
## data: difTimes
## t = 5.8073, df = 5, p-value = 0.002136
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## 23.43230 60.65227
## sample estimates:
## mean of x
## 42.04228
##
## in Kanto the t.test between the models gaModel and gaModelCluster is:
## One Sample t-test
##
## data: difTimes
## t = -3.3043, df = 5, p-value = 0.02137
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -9.408156 -1.175006
## sample estimates:
## mean of x
## -5.291581
##
## in Kanto the t.test between the models gaModel and listaCluster is:
## One Sample t-test
##
## data: difTimes
## t = 1.1659, df = 5, p-value = 0.2963
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
## -3.474332 9.241409
## sample estimates:
## mean of x
## 2.883539

```

## Conclusion

A one-way between subjects ANOVA was conducted to compare the effects of the models, the depths, the years and regions on the log-likelihood value. In this study there are 6 options for model: lista, gaModel, hybrid\_gaModel, hybrid\_list, gaModelCluster and listaCluster. Based on the results of the test, there was a

not a significant effect of the depths or years variables. For both cases at the we obtained  $p > 0.05$  level for the depths condition [ $F(2) = 2.072$ ,  $p = 0.126$ ] and we also obtained  $p > 0.05$  for the years condition [ $F(5) = 0.050$ ,  $p = 0.999$ ]. There was a significant effect of the models condition ( $p > 0.05$  [ $F(5) = 9699.690$ ,  $p < 2e-16$ ]) and regions condition ( $p > 0.05$  [ $F(3) = 764.220$ ,  $p < 2e-16$ ]). Therefore, we conduct a new anova test, with only the last two variables to verify the influence of those conditions more accurately. The results only changed a little, maintaining the significant effect of both conditions,  $p > 0.05$  [ $F(5) = 9705.6$ ,  $p < 2e-16$ ] and  $p > 0.05$  [ $F(3) = 764.7$ ,  $p < 2e-16$ ], respectively.

Because we found statistically significant result, we applied a Post hoc comparisons using the Tukey HSD test. It compared each condition with all others. For example, it compares the values from the gaModel with the gaModelClustered. It indicated that the gaModelCluster and the listaCluster, when compared with all other models, achieve greater log-likelihood values. Furthermore, we noticed that the depths conditions show a greater influence when the depth is smaller or equal to 25 km.

When comparing the models from the lista method and from the gaModel against themselves, with or without using clustering techniques, we found that there is no statistically significant result between the methods. That implies that it can be considered that the methods are obtain statistically equal results.

Therefore, based on the result of the HSD test, we performed a new AVOVA test, considering only the gaModelClustered and the listaClustered. That was meant not only to verify the previous results but also to certify if the depth influence is preserved.

Taken together, these results suggest that the using cluster and depth smaller or equal to 25km showed the best results.