ANOVA2.0-TPNC

Yuri Lavinas

July 14, 2016

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 métodos híbridos com e sem clusterização. Para cada um dos métodos temos algumas variações nas varíaveis utilizadas. Variamos os anos (2005-2010), as regiões (Kanto, EastJapan), a profundidade (<100km) e finalmente o catálogo utilizado (SC da Yen-san).

Statistical Analysis

ANOVA test and HSD Tukey

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

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.

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.

Filtering

Seleciono os modelos com terremotos com profundidade <= 100 km.

```
subTabela = finalData[finalData$depths==100,]
summary(subTabela)
```

```
##  loglikeValues                          model      depths      years
##  Min.   :-5221.9   GAModel                 : 240   100:3840   2005:640
##  1st Qu.:-2321.8   ReducedGAModel          : 240   25 :   0   2006:640
##  Median :-1990.4   EMP-GAModel             : 240   60 :   0   2007:640
##  Mean   :-1993.4   EMP-ReducedGAModel      : 240   RI :   0   2008:640
##  3rd Qu.:-1613.2   EMP-GAModelWindow       : 240              2009:640
##  Max.   : -865.4   EMP-ReducedGAModelWindow: 240              2010:640
##                    (Other)                 :2400
##       regions
##  Kanto    :960
```

```
##  Kansai   :960
##  Tohoku   :960
##  EastJapan:960
##
##
##
```

ANOVA - Specific analysis somente com Cluster.

Seleciono somente as áreas com dados do SC e os modelos apropriados.

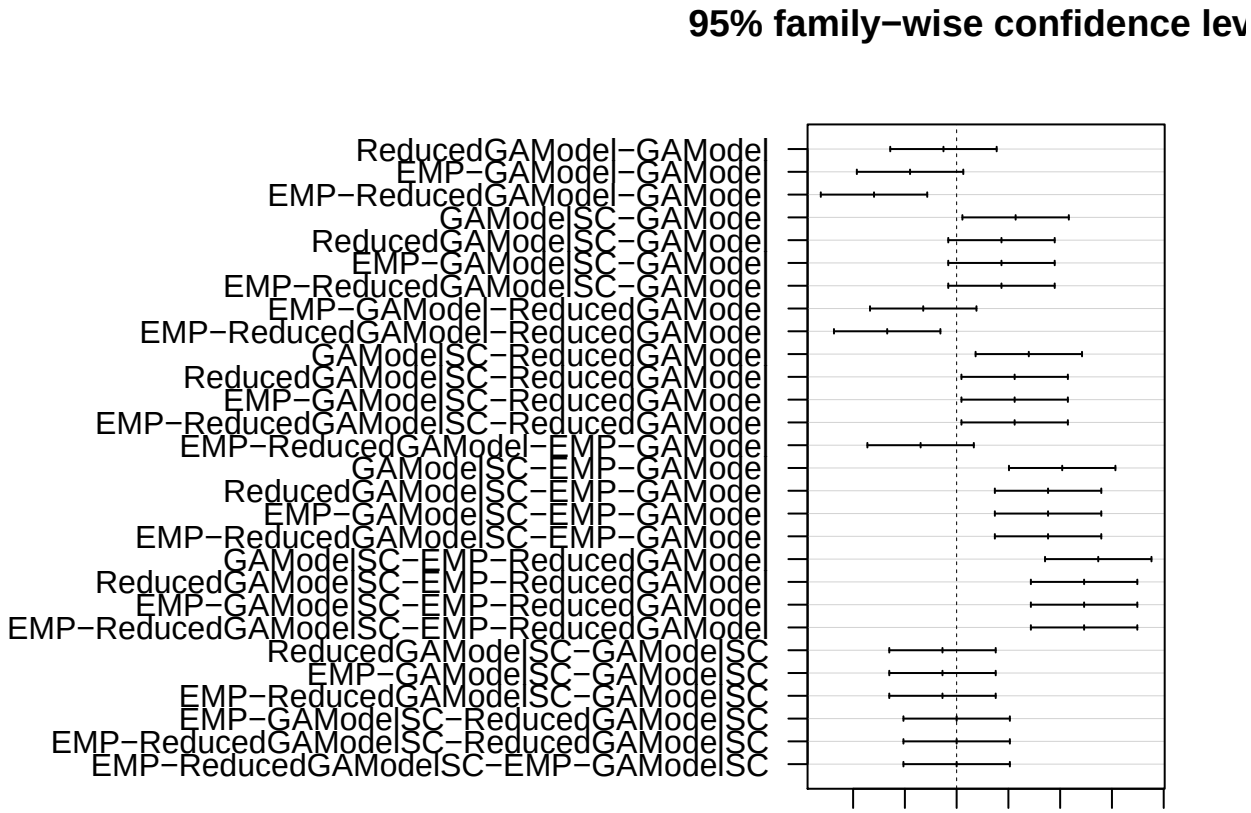
```
subTabela3 = subTabela[subTabela$model=='ReducedGAModelSC'|subTabela$model=='GAModelSC'|
                           subTabela$model=='EMP-ReducedGAModelSC'|subTabela$model=='EMP-GAModelSC'|
                           subTabela$model=='ReducedGAModel'|subTabela$model=='GAModel'|
                           subTabela$model=='EMP-ReducedGAModel'|subTabela$model=='EMP-GAModel',]
summary(subTabela3)
```

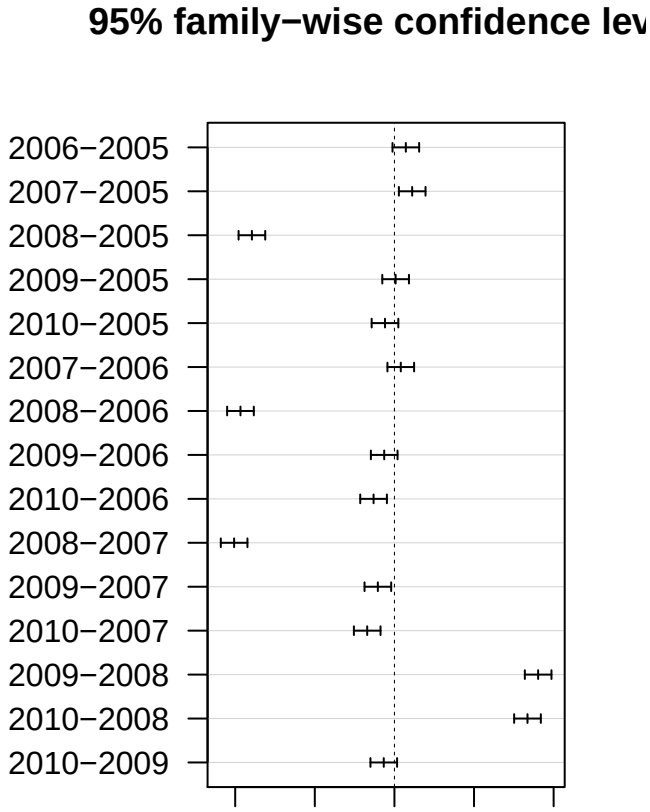
```
##  loglikeValues                    model     depths      years
##  Min.   :-4617.4   GAModel           :240   100:1920   2005:320
##  1st Qu.:-2292.9   ReducedGAModel    :240   25 :   0   2006:320
##  Median :-1960.9   EMP-GAModel       :240   60 :   0   2007:320
##  Mean   :-1947.4   EMP-ReducedGAModel:240   RI :   0   2008:320
##  3rd Qu.:-1612.4   GAModelSC         :240               2009:320
##  Max.   : -872.2   ReducedGAModelSC  :240               2010:320
##                    (Other)           :480
##       regions
##  Kanto    :480
##  Kansai   :480
##  Tohoku   :480
##  EastJapan:480
##
##
##
```

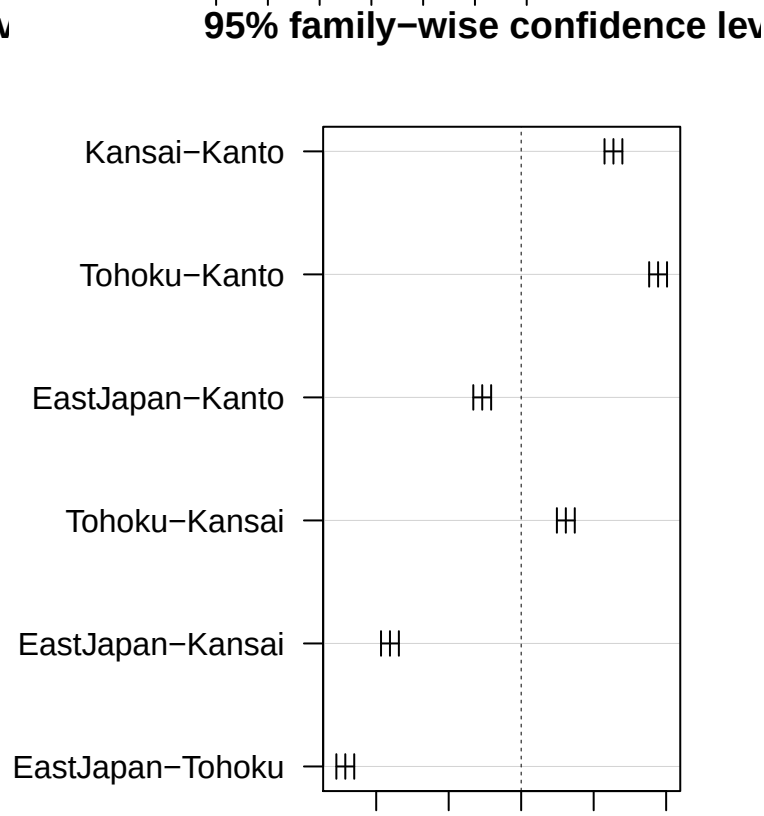
```
resultANOVA = aov(loglikeValues~model+years+regions , data = subTabela3)
summary(resultANOVA)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## model          7  16438151   2348307   17.05 <2e-16 ***
## years          5 232103284  46420657  336.99 <2e-16 ***
## regions        3 450628337 150209446 1090.45 <2e-16 ***
## Residuals   1904 262275213    137750
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tuk = TukeyHSD(resultANOVA)
# par(mfrow=c(2,2))
op <- par(mar = c(1,21,4,2) + 0.1)
plot(tuk,las=1)
```


95% family−wise confidence lev
ReducedGAModel−GAModel
EMP−GAModel−GAModel
EMP−ReducedGAModel−GAModel
GAModelSC−GAModel
ReducedGAModelSC−GAModel
EMP−GAModelSC−GAModel
EMP−ReducedGAModelSC−GAModel
EMP−GAModel−ReducedGAModel
EMP−ReducedGAModel−ReducedGAModel
GAModelSC−ReducedGAModel
ReducedGAModelSC−ReducedGAModel
EMP−GAModelSC−ReducedGAModel
EMP−ReducedGAModelSC−ReducedGAModel
EMP−ReducedGAModel−EMP−GAModel
GAModelSC−EMP−GAModel
ReducedGAModelSC−EMP−GAModel
EMP−GAModelSC−EMP−GAModel
EMP−ReducedGAModelSC−EMP−GAModel
GAModelSC−EMP−ReducedGAModel
ReducedGAModelSC−EMP−ReducedGAModel
EMP−GAModelSC−EMP−ReducedGAModel
EMP−ReducedGAModelSC−EMP−ReducedGAModel
ReducedGAModelSC−GAModelSC
EMP−GAModelSC−GAModelSC
EMP−ReducedGAModelSC−GAModelSC
EMP−GAModelSC−ReducedGAModelSC
EMP−ReducedGAModelSC−ReducedGAModelSC
EMP−ReducedGAModelSC−EMP−GAModelSC


95% family−wise confidence lev
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


95% family−wise confidence lev
Kansai−Kanto
Tohoku−Kanto
EastJapan−Kanto
Tohoku−Kansai
EastJapan−Kansai
EastJapan−Tohoku

```r
# print(tuk)
```

Como sempre é mais interessante utilizar o SC, refaço as análises só com eles para garantir que são estatisticamente iguais.

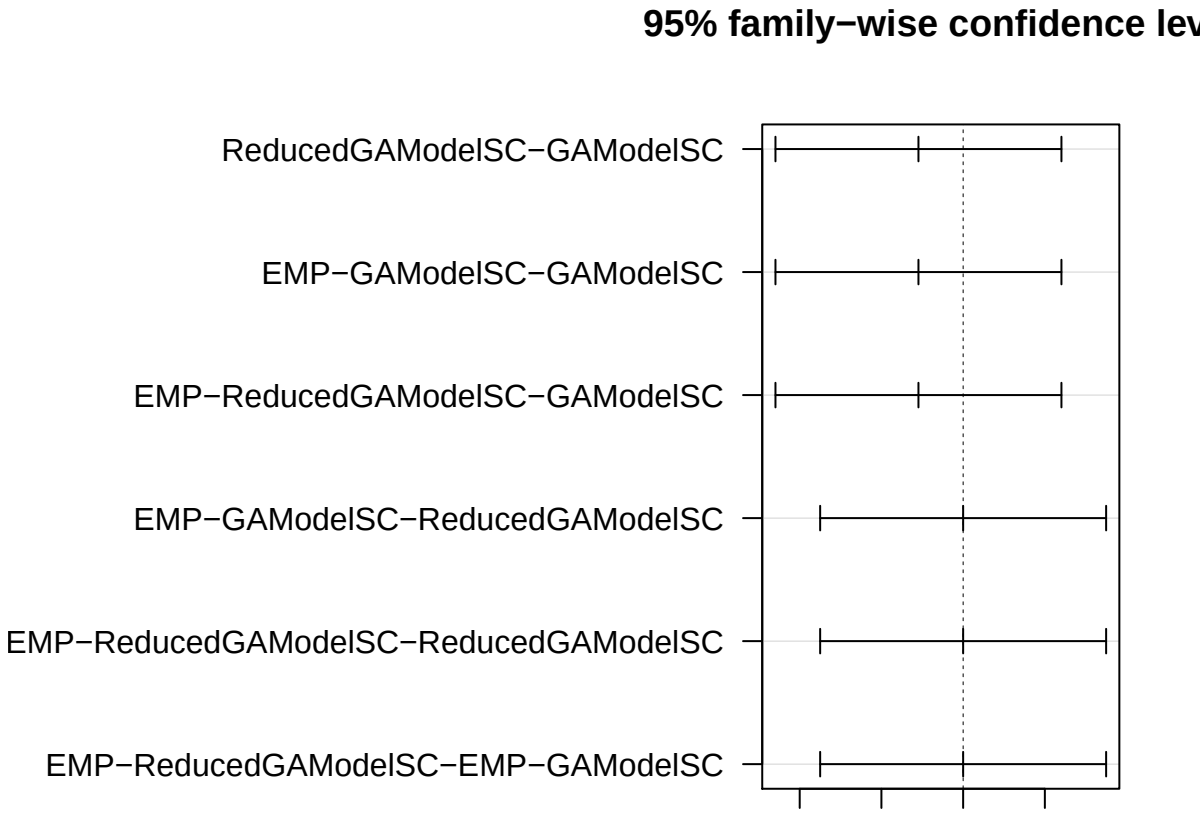
```r
subTabela3 = subTabela3[subTabela3$model=='ReducedGAModelSC'|subTabela3$model=='GAModelSC'|
                            subTabela3$model=='EMP-ReducedGAModelSC'|subTabela3$model=='EMP-GAModelSC',]
summary(subTabela3)
```

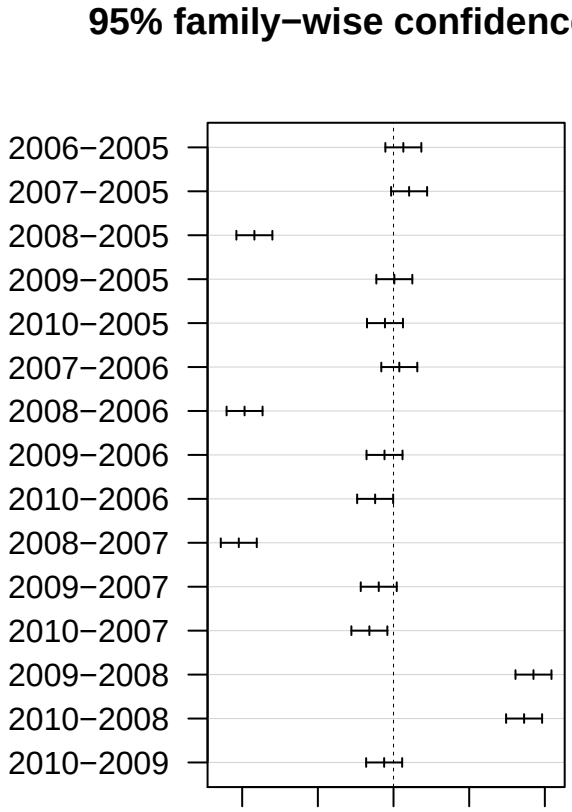
```
##  loglikeValues                      model     depths     years
##  Min.   :-4015.8   GAModelSC           :240   100:960   2005:160
##  1st Qu.:-2175.8   ReducedGAModelSC    :240   25 :  0   2006:160
##  Median :-1907.3   EMP-GAModelSC       :240   60 :  0   2007:160
##  Mean   :-1866.3   EMP-ReducedGAModelSC:240   RI :  0   2008:160
##  3rd Qu.:-1611.1   GAModel             :  0             2009:160
##  Max.   : -872.2   ReducedGAModel      :  0             2010:160
##                    (Other)             :  0
##       regions
##  Kanto    :240
##  Kansai   :240
##  Tohoku   :240
##  EastJapan:240
##
##
##
```

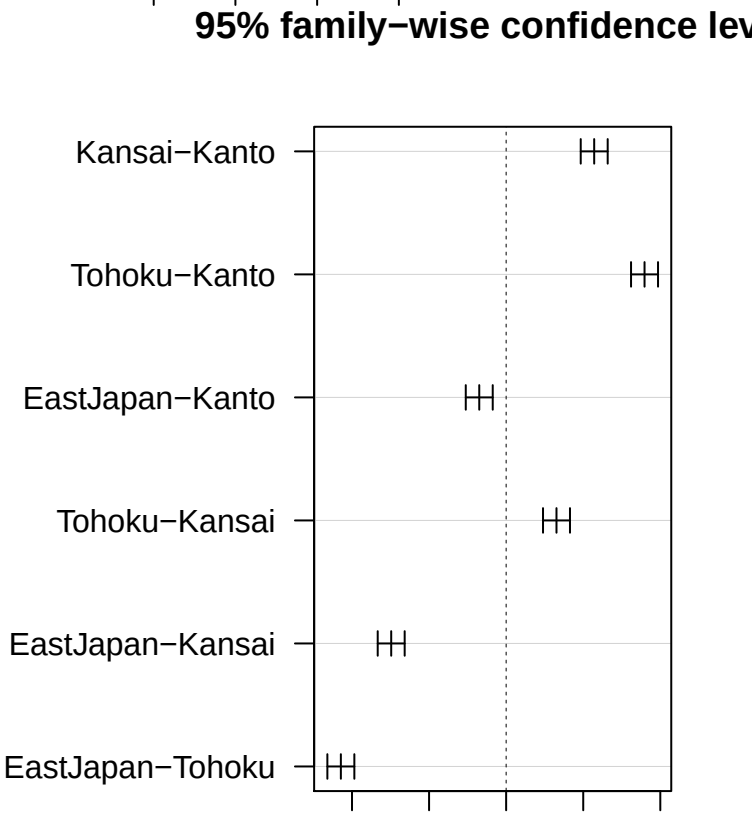
```r
resultANOVA = aov(loglikeValues~model+years+regions , data = subTabela3)
summary(resultANOVA)
```

```
##              Df    Sum Sq  Mean Sq F value Pr(>F)
## model         3    134524    44841   0.323  0.809
## years         5 120969002 24193800 174.452 <2e-16 ***
## regions       3 178516884 59505628 429.071 <2e-16 ***
## Residuals   948 131473255   138685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
tuk = TukeyHSD(resultANOVA)
# par(mfrow=c(2,2))
op <- par(mar = c(1,21,4,2) + 0.1)
plot(tuk,las=1)
```







Teste pareado para esses modelos: GAModelSC, GAModelWindow, ReducedGAModelSC, Emp-GAModelSC.

```r
ttestPaired= function(region){
    subTabela6 = subTabela[subTabela$regions==region,]
    aggfinaldata<-aggregate(loglikeValues~years:model, data=subTabela6,FUN=mean)
    # Perform paired t-test
    cat('in', region, 'the t.test between the models GAModelSC and ReducedGAModelSC is: ')
    difTimes<-with(aggfinaldata,loglikeValues[1:6]-loglikeValues[7:12])
    print(t.test(difTimes))
    cat('in', region, 'the t.test between the models GAModelSC and Emp-GAModelSC is: ')
    difTimes<-with(aggfinaldata,loglikeValues[1:6]-loglikeValues[13:18])
    print(t.test(difTimes))
    cat('in', region, 'the t.test between the models ReducedGAModelSC and Emp-GAModelSC is: ')
    difTimes<-with(aggfinaldata,loglikeValues[7:12]-loglikeValues[13:18])
    print(t.test(difTimes))
}

    ttestPaired('Kansai')
```

```
## in Kansai the t.test between the models GAModelSC and ReducedGAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 14.034, df = 5, p-value = 3.304e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  31.59156 45.75950
## sample estimates:
## mean of x
##  38.67553
##
## in Kansai the t.test between the models GAModelSC and Emp-GAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 1.2678, df = 5, p-value = 0.2607
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -4.389914 12.934284
## sample estimates:
## mean of x
##  4.272185
##
## in Kansai the t.test between the models ReducedGAModelSC and Emp-GAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = -6.9255, df = 5, p-value = 0.000963
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -47.17306 -21.63363
## sample estimates:
## mean of x
## -34.40335
```

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

```
## in Tohoku the t.test between the models GAModelSC and ReducedGAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 0.6473, df = 5, p-value = 0.546
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -9.940479 16.631600
## sample estimates:
## mean of x
##   3.34556
##
## in Tohoku the t.test between the models GAModelSC and Emp-GAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 32.446, df = 5, p-value = 5.225e-07
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  75.14398 88.07532
## sample estimates:
## mean of x
##  81.60965
##
## in Tohoku the t.test between the models ReducedGAModelSC and Emp-GAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 14.375, df = 5, p-value = 2.938e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  64.26880 92.25938
## sample estimates:
## mean of x
##  78.26409
```

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

```
## in EastJapan the t.test between the models GAModelSC and ReducedGAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 0.060707, df = 5, p-value = 0.9539
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -77.42738  81.17290
## sample estimates:
## mean of x
##  1.872764
##
## in EastJapan the t.test between the models GAModelSC and Emp-GAModelSC is:
```

```
##  One Sample t-test
##
## data:  difTimes
## t = 25.208, df = 5, p-value = 1.834e-06
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  174.6612 214.3277
## sample estimates:
## mean of x
##  194.4944
##
## in EastJapan the t.test between the models ReducedGAModelSC and Emp-GAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 5.1102, df = 5, p-value = 0.003738
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   95.7269 289.5165
## sample estimates:
## mean of x
##  192.6217
```

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

```
## in Kanto the t.test between the models GAModelSC and ReducedGAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 6.4009, df = 5, p-value = 0.00138
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  34.68124 81.23101
## sample estimates:
## mean of x
##  57.95612
##
## in Kanto the t.test between the models GAModelSC and Emp-GAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 13.918, df = 5, p-value = 3.441e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  64.90515 94.31048
## sample estimates:
## mean of x
##  79.60781
##
## in Kanto the t.test between the models ReducedGAModelSC and Emp-GAModelSC is:
##  One Sample t-test
##
## data:  difTimes
## t = 1.9367, df = 5, p-value = 0.1105
```

```
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -7.086516 50.389896
## sample estimates:
## mean of x
##  21.65169
```