ANOVA Test

Yuri Lavinas

June 8, 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 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).

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.

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.

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

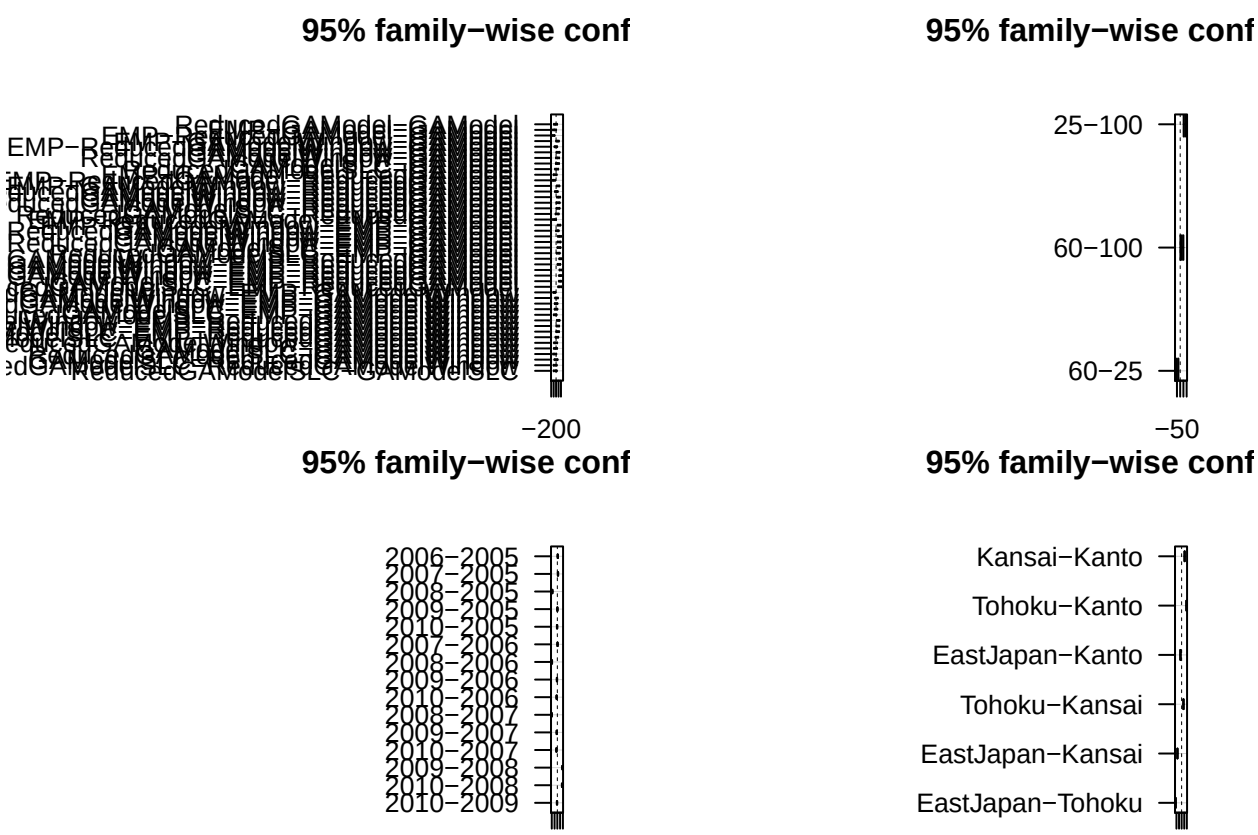
```
##  loglikeValues                          model      depths      years
##  Min.   :-4617.4   EMP-GAModelWindow       :1440   100:2880   2005:1440
##  1st Qu.:-2229.0   EMP-ReducedGAModelWindow:1440   25 :2880   2006:1440
##  Median :-1925.3   GAModel                 : 720   60 :2880   2007:1440
##  Mean   :-1903.1   ReducedGAModel          : 720              2008:1440
##  3rd Qu.:-1613.1   EMP-GAModel             : 720              2009:1440
##  Max.   : -863.6   EMP-ReducedGAModel      : 720              2010:1440
##                    (Other)                 :2880
##       regions
##  Kanto    :2160
##  Kansai   :2160
```

```
##  Tohoku   :2160
##  EastJapan:2160
##
##
##
```

```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          9 2.918e+07   3242678   23.24  < 2e-16 ***
## depths         2 8.760e+06   4379929   31.39 2.62e-14 ***
## years          5 1.067e+09 213368956 1528.98  < 2e-16 ***
## regions        3 1.834e+09 611182301 4379.68  < 2e-16 ***
## Residuals   8620 1.203e+09    139550
## ---
## 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(1,17,4,2) + 0.1)
#Função para gerar o gráfico
plot(tuk,las=1)
```



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

ANOVA - Specific analysis somente com Cluster.

Faço o ANOVA somente para os modelos "clusterizados" e abaixo de 25km.

Primeiro crio o data frame somente com os modelos citados

```
subTabela = finalData[finalData$depths==25,]
subTabela = subTabela[subTabela$model=='EMP-GAModelWindow'|subTabela$model=='GAModelWindow'|
                          subTabela$model=='EMP-ReducedGAModelWindow'|subTabela$model=='ReducedGAModelW
                          subTabela$model=='EMP-GAModelSLC'|subTabela$model=='GAModelSLC'|
                          subTabela$model=='EMP-ReducedGAModelSLC'|subTabela$model=='ReducedGAModelSLC'
summary(subTabela)
```

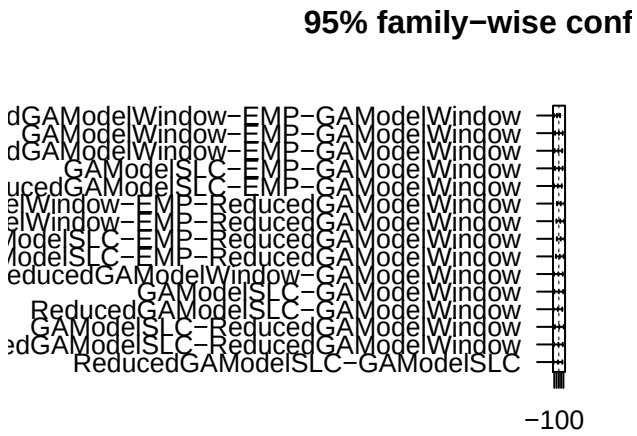
```
##  loglikeValues                          model     depths      years
##  Min.   :-4084.2   EMP-GAModelWindow       :480   100:   0   2005:320
##  1st Qu.:-2169.9   EMP-ReducedGAModelWindow:480   25 :1920   2006:320
##  Median :-1869.1   GAModelWindow           :240   60 :   0   2007:320
##  Mean   :-1847.4   ReducedGAModelWindow    :240               2008:320
##  3rd Qu.:-1611.3   GAModelSLC              :240               2009:320
##  Max.   : -868.1   ReducedGAModelSLC       :240               2010:320
##                    (Other)                 :  0
##       regions
##  Kanto    :480
##  Kansai   :480
##  Tohoku   :480
##  EastJapan:480
##
##
##
```

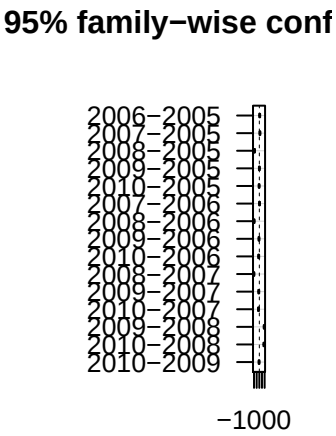
Aplico o anova, com a regressão para modelos, anos e regiões. mesma profundidade e só cluster.

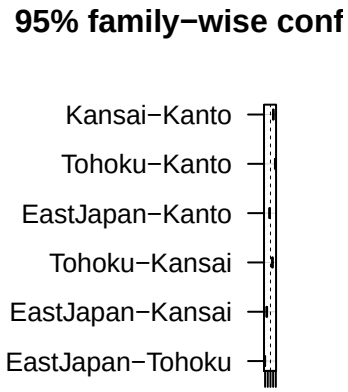
```
resultANOVA = aov(loglikeValues~model+years+regions , data = subTabela)
summary(resultANOVA)
```

```
##               Df    Sum Sq   Mean Sq F value Pr(>F)
## model          5    585681    117136   0.845  0.518
## years          5 241256980  48251396 348.049 <2e-16 ***
## regions        3 341161103 113720368 820.293 <2e-16 ***
## Residuals   1906 264236215    138634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
tuk = TukeyHSD(resultANOVA)
par(mfrow=c(2,2))
op <- par(mar = c(1,17,4,2) + 0.1)
plot(tuk,las=1)
# print(tuk)
```


95% family−wise conf
dGAModelWindow−EMP−GAModelWindow
GAModelWindow−EMP−GAModelWindow
dGAModelWindow−EMP−GAModelWindow
GAModelSLC−EMP−GAModelWindow
ucedGAModelSLC−EMP−GAModelWindow
elWindow−EMP−ReducedGAModelWindow
elWindow−EMP−ReducedGAModelWindow
ModelSLC−EMP−ReducedGAModelWindow
ModelSLC−EMP−ReducedGAModelWindow
educedGAModelWindow−GAModelWindow
GAModelSLC−GAModelWindow
ReducedGAModelSLC−GAModelWindow
GAModelSLC−ReducedGAModelWindow
edGAModelSLC−ReducedGAModelWindow
ReducedGAModelSLC−GAModelSLC
−100


95% family−wise conf
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
−1000


95% family−wise conf
Kansai−Kanto
Tohoku−Kanto
EastJapan−Kanto
Tohoku−Kansai
EastJapan−Kansai
EastJapan−Tohoku