Phenoarch platform - Cleaning procedure - Curve level - gss package

I.Sanchez

mai 28, 2020

Objective of cleaning procedure using smoothing splines anova

Smoothing spline analysis of variance on each genotype-scenario of an experiment. Detection of outlier repetition if significant TT*Rep (thermal time by repetition) interaction using a Kullblack-Leibler projection (KL). I consider a genotype-scenario as outlier:

- biovolume: if KL > 0.05
- plantHeight: if KL > 0.05
- leafArea: if KL > 0.05

The input dataset must contain the following columns:

- experimentAlias
- genotypeAlias
- scenario
- repetition
- thermalTime (for thermal time)
- parameter of interest (biovolume, plantHeight etc. . .)

The five first column names are standard names extracted from the web service.

Import of data

```
library(ggplot2)
library(lubridate)
library(tidyr)
library(dplyr)
library(gss)
library(openSilexStatR)

myreport<-substr(now(),1,10)
```

```
data(plant3)
cat("-------------- plant3 dataset ---------------\n")
```

```
## -------------- plant3 dataset ---------------
```

```
printExperiment(datain=plant3)
```

```
## Experiment: manip3
## Genotypes: 10
##  [1] "A3_H"     "A310_H"   "11430_H"  "A554_H"   "A374_H"   "A347_H"
##  [7] "B100_H"   "A375_H"   "AS5707_H" "A347"
## Scenario: 2
## [1] "WW" "WD"
## Repetition-scenario: 6
## [1] "1-WW" "2-WW" "3-WW" "1-WD" "2-WD" "3-WD"
## Pots (number of plants): 60
## Line: 25
## Position: 42
```

```
# Import data, here is a dataset in the phisStatR package, You have to import your own dataset
# using a read.table() statement or a request to the web service
# You can add some datamanagement statements...
#------------------------------------------------------------------------
# Please, add the 'Ref' and 'Genosce' columns if don't exist.
# 'Ref' is the concatenation of experimentAlias-Line-Position-scenario
# 'Genosce' is the concatenation of experimentAlias-genotypeAlias-scenario
#------------------------------------------------------------------------

mydata<-unite(plant3,Genosce,experimentAlias,genotypeAlias,scenario,sep="-",remove=FALSE)
mydata<-arrange(mydata,Genosce)
```

```
# For one parameter, for example biovolume
resbio<-fitGSS(datain=mydata,trait="biovolume",loopId="Genosce")
```

Curves by genotype-scenario

Biovolume

```
outlierbio<-printGSS(object=resbio[[2]],threshold = 0.05)
klbio<-printGSS(object=resbio[[2]],threshold = NULL)

cat("Detection of outlier curve with KL projection:\n")
```

```
## Detection of outlier curve with KL projection:
```

```
print(outlierbio)
```

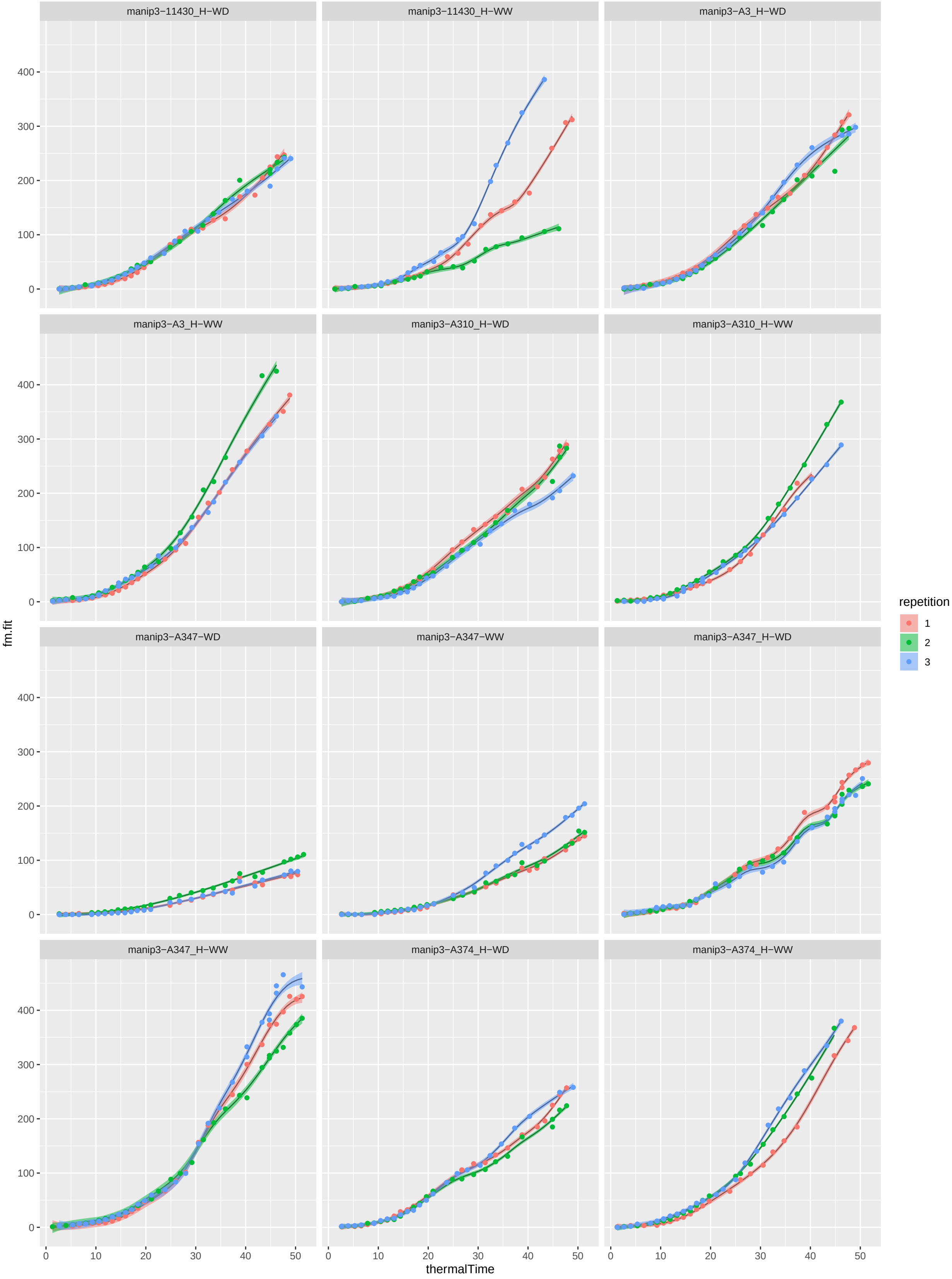
```
##              Genosce      ratio        kl     check
## 1  manip3-11430_H-WW 0.15015910 1175.7782 0.9999887
## 2 manip3-AS5707_H-WD 0.07205993  633.0729 0.9999874
```

```
#------------------------------------------------
# You can export these two datasets
# suppress the comments
#------------------------------------------------
#write.table(outlierbio,paste0(myreport,"outlier_gss_biovolume.csv"),row.names = FALSE,sep="\t")
#write.table(klbio,paste0(myreport,"KLprojection_gss_biovolume.csv"),row.names = FALSE,sep="\t")
```

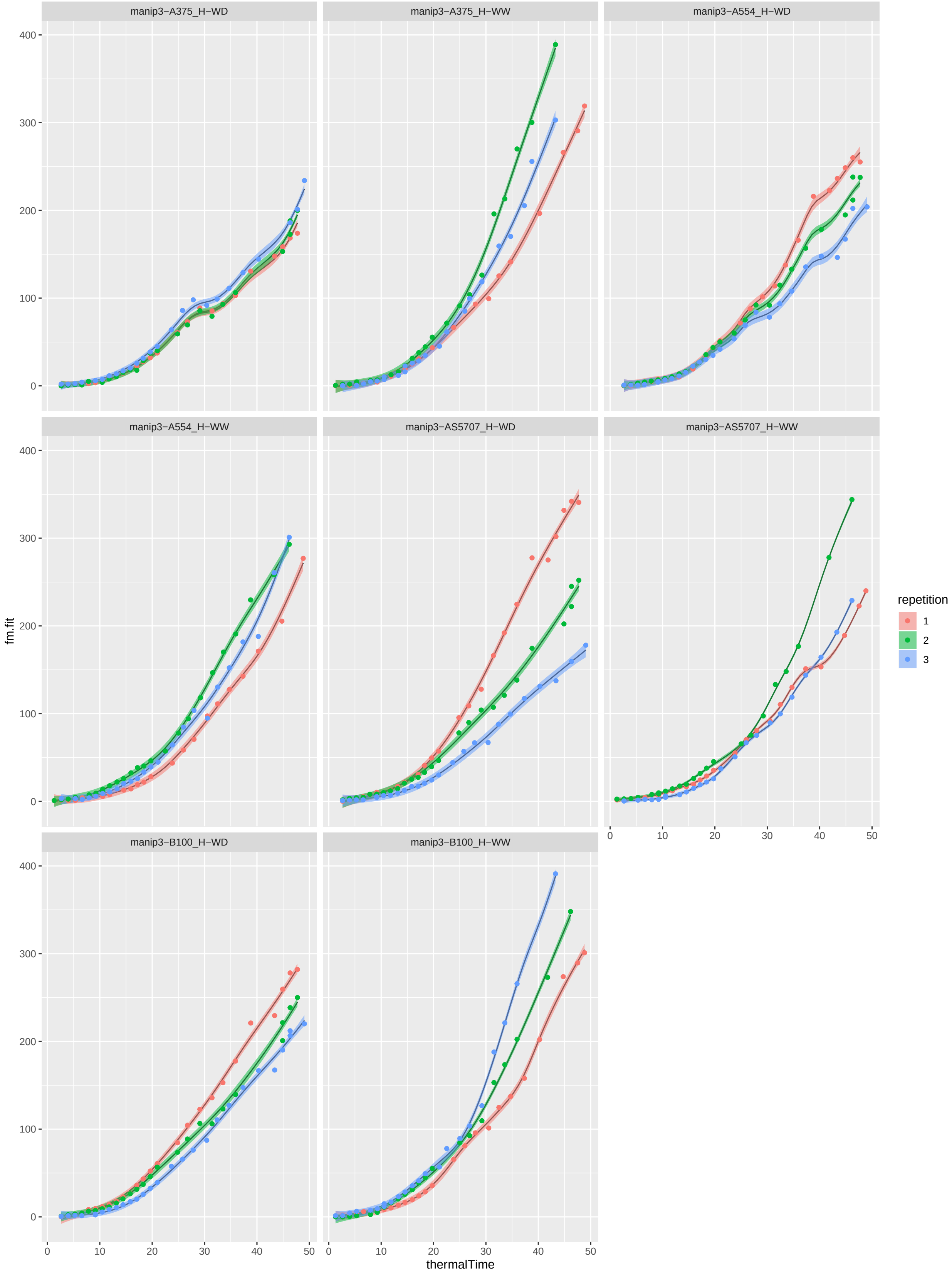
I take a threshold of 0.05 for this example. We can take a more conservative threshold like 0.01 or 0.02 to detect more outlier curves…

```
# plot of the smoothing splines by genotype-scenario
for(i in seq(1,length(unique(mydata[,"Genosce"])),by=12)){
  myvec<-seq(i,i+11,1)
  myvec<-myvec[myvec<=length(unique(mydata[,"Genosce"]))]
  print(plotGSS(datain=mydata,modelin=resbio[[1]],trait="biovolume",myvec=myvec,lgrid=50))
```

```
    cat("\n\n")
  }
```



manip3−11430_H−WD
manip3−11430_H−WW
manip3−A3_H−WD
manip3−A3_H−WW
manip3−A310_H−WD
manip3−A310_H−WW
manip3−A347−WD
manip3−A347−WW
manip3−A347_H−WD
manip3−A347_H−WW
manip3−A374_H−WD
manip3−A374_H−WW
fm.fit
0
100
200
300
400
thermalTime
0
10
20
30
40
50
repetition
● 1
● 2
● 3


manip3−A375_H−WD
manip3−A375_H−WW
manip3−A554_H−WD
manip3−A554_H−WW
manip3−AS5707_H−WD
manip3−AS5707_H−WW
manip3−B100_H−WD
manip3−B100_H−WW
fm.fit
0
100
200
300
400
thermalTime
0
10
20
30
40
50
repetition
1
2
3

Session info

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 7 x64 (build 7601) Service Pack 1
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=French_France.1252  LC_CTYPE=French_France.1252
## [3] LC_MONETARY=French_France.1252 LC_NUMERIC=C
## [5] LC_TIME=French_France.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] gss_2.1-12           locfit_1.5-9.1       ggplot2_3.2.1
## [4] tidyr_1.0.0          openSilexStatR_1.0.0 dplyr_0.8.3
## [7] lubridate_1.7.4
##
## loaded via a namespace (and not attached):
##   [1] colorspace_1.4-1   deldir_0.1-23      ellipsis_0.3.0
##   [4] class_7.3-15       leaflet_2.0.2      rgdal_1.4-7
##   [7] evd_2.3-3          rprojroot_1.3-2    fs_1.3.1
##  [10] rstudioapi_0.11    remotes_2.1.1      splines_3.6.1
##  [13] knitr_1.25         pkgload_1.0.2      spam_2.4-0
##  [16] shiny_1.4.0        compiler_3.6.1     backports_1.1.5
##  [19] assertthat_0.2.1   Matrix_1.2-17      fastmap_1.0.1
##  [22] lazyeval_0.2.2     cli_1.1.0          later_1.0.0
##  [25] htmltools_0.4.0    prettyunits_1.0.2  tools_3.6.1
##  [28] dotCall64_1.0-0    coda_0.19-3        gtable_0.3.0
##  [31] glue_1.4.0         CARBayesdata_2.1   maps_3.3.0
##  [34] gmodels_2.18.1     Rcpp_1.0.4         vctrs_0.2.4
##  [37] spdep_1.1-3        gdata_2.18.0       nlme_3.1-141
##  [40] crosstalk_1.0.0    xfun_0.10          stringr_1.4.0
##  [43] ps_1.3.2           testthat_2.2.1     mime_0.7
##  [46] lifecycle_0.1.0    gtools_3.8.1       devtools_2.2.2
##  [49] LearnBayes_2.15.1  MASS_7.3-51.4      scales_1.0.0
##  [52] promises_1.1.0     expm_0.999-4       RColorBrewer_1.1-2
##  [55] fields_9.9         yaml_2.2.1         memoise_1.1.0
##  [58] gridExtra_2.3      truncdist_1.0-2    reshape_0.8.8
##  [61] stringi_1.4.3      SpATS_1.0-11       desc_1.2.0
##  [64] e1071_1.7-2        boot_1.3-23        pkgbuild_1.0.6
##  [67] truncnorm_1.0-8    spData_0.3.2       rlang_0.4.5
##  [70] pkgconfig_2.0.3    matrixStats_0.55.0 evaluate_0.14
##  [73] lattice_0.20-38    purrr_0.3.3        sf_0.8-0
##  [76] htmlwidgets_1.5.1  labeling_0.3       processx_3.4.2
##  [79] tidyselect_1.0.0   GGally_1.5.0       plyr_1.8.4
##  [82] magrittr_1.5       R6_2.4.1           DBI_1.0.0
##  [85] pillar_1.4.2       foreign_0.8-72     withr_2.1.2
##  [88] units_0.6-5        shapefiles_0.7     sp_1.3-1
##  [91] tibble_2.1.3       crayon_1.3.4       CARBayesST_3.1
##  [94] KernSmooth_2.23-16 rmarkdown_1.16     usethis_1.5.1
##  [97] grid_3.6.1         data.table_1.12.6  callr_3.3.2
## [100] matrixcalc_1.0-3   digest_0.6.25      classInt_0.4-2
## [103] xtable_1.8-4       httpuv_1.5.2       stats4_3.6.1
## [106] munsell_0.5.0      sessioninfo_1.1.1
```

References

1. R Development Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
2. Chong Gu (2014). Smoothing Spline ANOVA Models: R Package gss. Journal of Statistical Software, 58(5), 1-25. URL http://www.jstatsoft.org/v58/i05/.