Validation de Données en Phénotypage Végétal

Antoine Schorgen
Ingénieur de Recherche (CDD), UMR MISTEA

Rapide Aperçu

1. Un peu de contexte
2. Procédé de nettoyage des données
3. Validation Statistique
4. Quelques Résultats
5. A faire!

La Serre de Phénotypage à Haut Débit

Plateforme automatisée montée par LemnaTec

- PhénoArch : croissance de la plante dans sa globalité
- PhénoDyn : feuilles ...

Les données

Une expérience-type:

- 2000 pots au maximum
- 200 Génotypes testés x 5 Répétitions
- x 2 Scénari : bien irrigué (WW) ou hydriquement stressé (WD)
- Plantes: Maïs, Pommiers ...
- Durée d'environ 3 mois (autour de 90j)

Objectifs

Que cherchent les biologistes?

- Quelle variété résiste le mieux au stress hydrique?
- Comment se développe la plante et ses feuilles?
- Et plein d'autres choses ...

Des données aux résultats

De l'acquisition des données à leur valorisation

```
Acquisition -> Nettoyage -> Validation -> Interpretation -> Valorisation
```

- Comment obtient-on les données?
- Que veut-on en faire? A quelles questions veut-on répondre?
- Quelle information doit-on tirer des données pour avancer?
- De combien de temps on dispose?
- Quels moyens humains a-t-on?

Cheminement des données

Partie Acquisition

- Les biologistes
- Les machines

Partie Nettoyage

- Les biologistes en font.
- Les statisticiens en font.

Partie Validation

- Les biologistes valident leurs données.
- Les statisticiens aussi.

Et les stats?

La partie "stat" concerne surtout la détection des valeurs aberrantes (VA) ainsi que le remplacement des valeurs manquantes (NA).

Mais également:

- organisation des données
- automatisation des traitements/procédures
- visualisation (package `ggplot2`)
- procédures statistiques (paramétriques, non paramétriques) pour détection des anomalies (valeurs aberrantes, NA, ...)

Code R : Données brutes

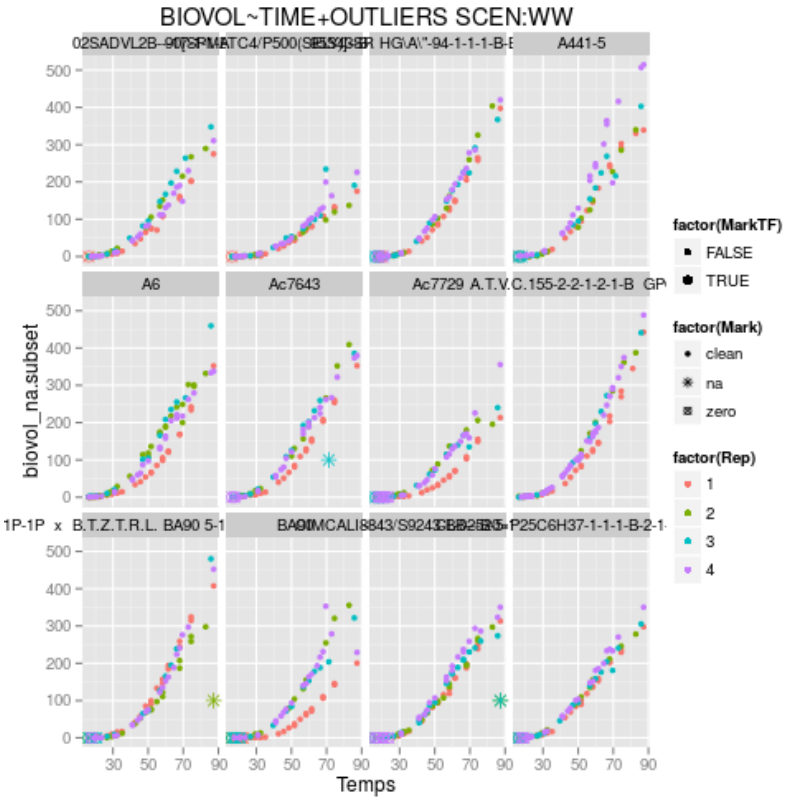
```
names(df.semiclean.names.order.flag )
```

```
##  [1] "Plant"               "Pot"                 "Geno"
##  [4] "SeedLot"             "Scenario"            "Rep"
##  [7] "Line"                "Position"            "Measurement.Label"
## [10] "Snapshot.Time.Stamp" "Day"                 "Hour"
## [13] "Temps"               "Bio1"                "Bio2"
## [16] "BioCor"              "BioHarvesting"       "BioFinal"
## [19] "Biomass"             "Flag0"
```

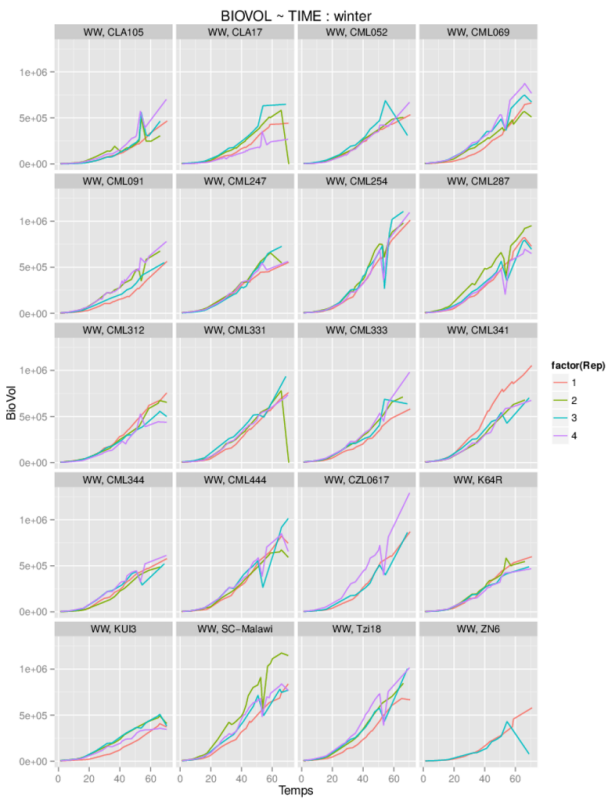
Code R : utilisation de {ggplot2}

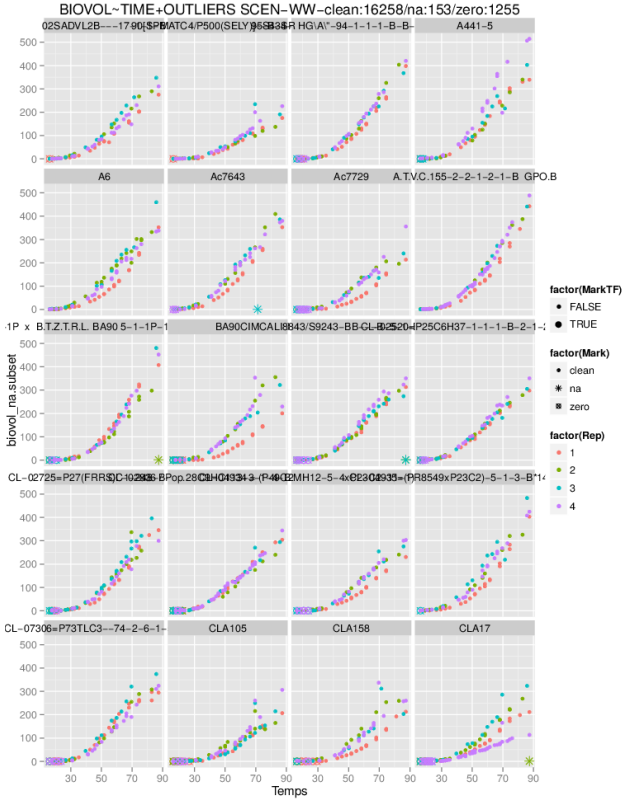
```
library(ggplot2)
# DEFINITION DES VARIABLES A DESSINER
p <- ggplot(dframe.subset, aes(Temps, biovol_na.subset,
                                color = factor(Rep),
                                shape = factor(Mark),
                                size  = factor(MarkTF)))
# DESSIN
p + geom_point() +
scale_size_manual(values = c(2, 3)) +
scale_shape_manual(values = c("clean" = 20, "na" = 8, "zero" = 13)) +
facet_wrap( ~ Geno, nrow = 3, ncol = 4) +  # layout=5x4=20 (grid.nb)
ggtitle(paste("BIOVOL~TIME+OUTLIERS SCEN", scen, sep=":"))
```

Graph obtenu : BioVol ~ Temps

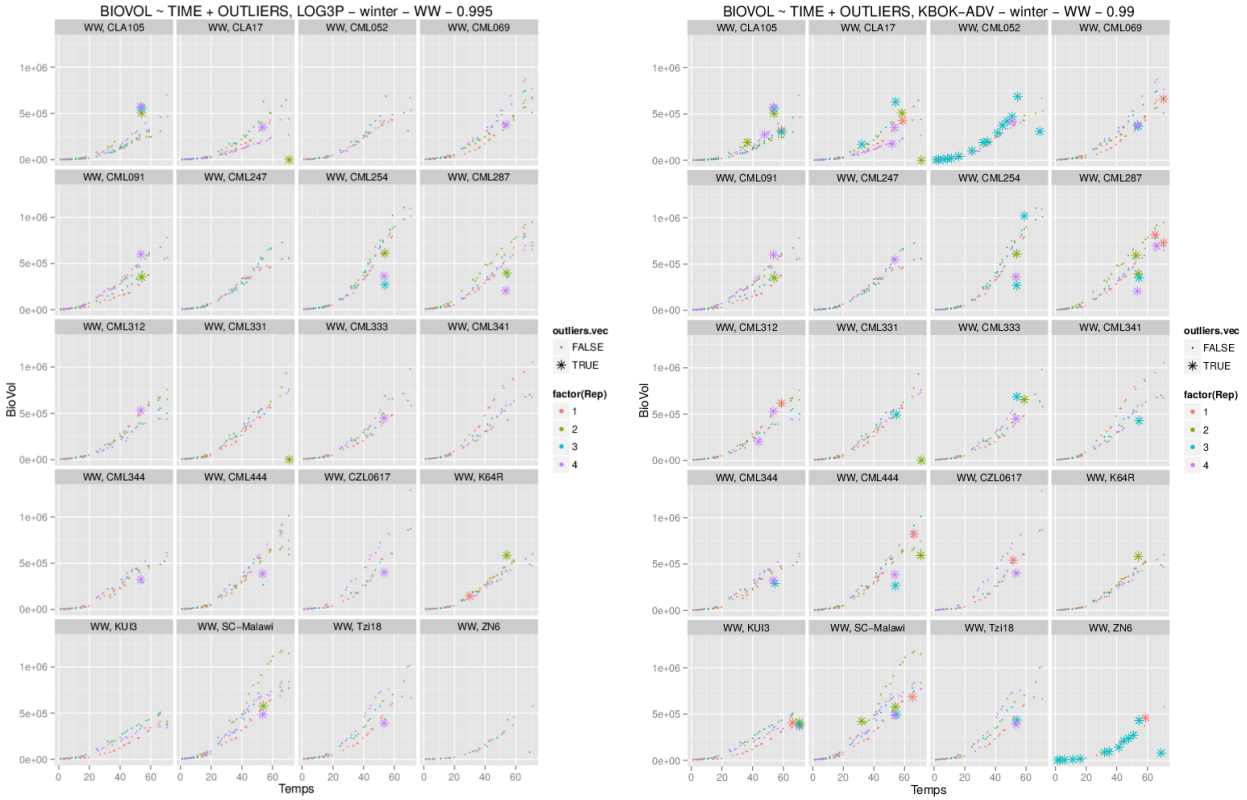


Visualisation : données "brutes"

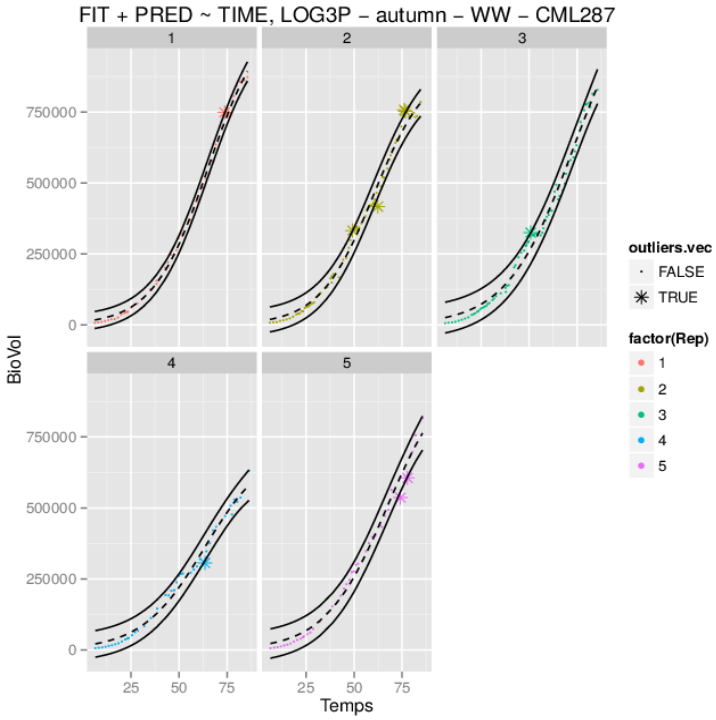




Détection VA, vue globale



Détection VA, vue individuelle



Objectifs

- Fluidifier les différentes étapes du cheminement des données
- Automatiser les procédures
- Mettre à l'épreuve "à grande échelle" notre procédure de validation des données (détection des VA + remplacement des NA)
- Fournir du code R "prêt à l'emploi" et facilement utilisable
- Interfaces graphiques web

Bonus

Utilisation de markdown avec R

```
library(knitr)
library(markdown)
knit("test.Rmd")
markdownToHTML("test.md","test.html")
```

Ou bien directement

```
library(knitr)
knit2html("test.Rmd")
```

Installation de Rmarkdown :

```
install.package("rmarkdown")
```

MAIS nécessite l'installation de la dernière version de pandoc:

```
sudo apt-get install haskell-platform
cabal update
cabal install pandoc
```

et le rajout du chemin suivant dans /.bashrc :

```
export PATH=$PATH:~/cabal/bin
```

Utilisation de Rmarkdown :

```
library(rmarkdown)
render("input.Rmd") # si format précisé dans l'entête
render("input.Rmd", html_document())
render('input.Rmd', pdf_document())
```

Format possibles en sortie: html_document, pdf_document, word_document, md_document, beamer_presentation, ioslides_presentation, slidy_presentation

Entête (metadata section) d'un fichier .Rmd:

```
---
title: "Sample Document"
output: pdf_document
---
```

Installation de Slidify:

```
library(devtools)
install_github("ramnathv/slidify")
install_github("ramnathv/slidifyLibraries")
```

ATTENTION : il peut y avoir un problème à l'installation de {devtools}

```
sudo apt-get -y build-dep libcurl4-gnutls-dev
sudo apt-get -y install libcurl4-gnutls-dev
```

Utilisation de Slidify

```
library(slidify)
slidify("slides.Rmd")
```

Avec en entête du fichier .Rmd :

```
---
title  : Validation de Données en Phénotypage Végétal
author : Antoine Schorgen
framework : io2012  # {html5slides, shower, dzslides, ...}
widgets: mathjax
mode : selfcontained  # {standalone, draft}
---
```

Passage en beamer

Et si l'on veut obtenir un beamer (pdf) à partir du fichier markdown .md, on utilise pandoc dans le terminal (Rmarkdown v1) :

```
pandoc -t beamer my_source.md -o my_beamer.pdf
```

ou mieux avec la commande suivante directement dans R (Rmarkdown v2) :

```
rmarkdown::render("slides.Rmd", beamer_presentation())
```

Code R

```
print(sessionInfo(),locale=FALSE)
```

```
## R version 3.0.2 (2013-09-25)
## Platform: x86_64-pc-linux-gnu (64-bit)
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_1.0.0   slidify_0.4.5   rmarkdown_0.3.3 vimcom_1.0-0
## [5] setwidth_1.0-3  colorout_1.0-3
##
## loaded via a namespace (and not attached):
##  [1] codetools_0.2-8  colorspace_1.2-2 digest_0.6.4     evaluate_0.5.5
##  [5] formatR_1.0      grid_3.0.2       gtable_0.1.2     htmltools_0.2.6
##  [9] knitr_1.8        labeling_0.3     markdown_0.7.4   MASS_7.3-29
## [13] munsell_0.4.2    plyr_1.8.1       proto_0.3-10     Rcpp_0.11.0
## [17] reshape2_1.4     scales_0.2.4     stringr_0.6.2    tools_3.0.2
## [21] whisker_0.3-2    yaml_2.1.13
```

Un peu de maths?

...pour faire beau...

$$\hat{\theta}_n = \sum_{i=0}^{n} X_i e^{2\pi}$$

25/25