Université de **La Rochelle**

Qualification des données SOMLIT : cap sur l'objectif « boite à outil » R

Sauriau P.-G., Pédemay L.*, Pineau P.
et l'équipe SOMLIT de La Rochelle

* Master 2 SPE – GEEL
Univ. La Rochelle

Journée scientifique 28 septembre 2017

Le réseau SOMLIT : 20 ans d'âge

Station Antioche : 6 ans
- Créée en 2011 avec N. Savoye
- Opérationnelle depuis juin 2011

- SOMLIT depuis 2012

- Personnel : tâches attribuées
  Pineau, Lachaussée, Aubert, Bréret, Guillou, Beaugeard, Arnaud Agogué, Le Breton, Sauriau.

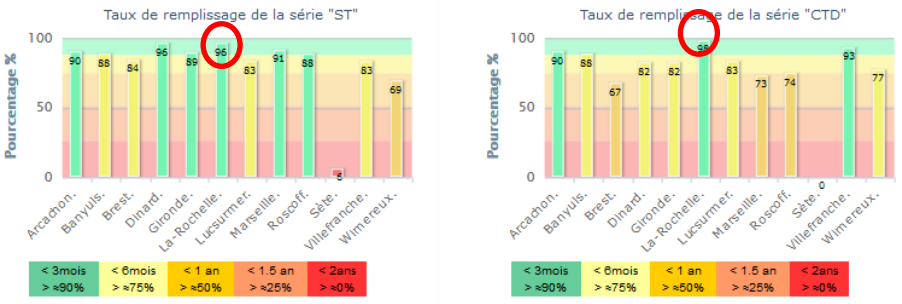
- Coûts élevés de gestion navire (> 10.000 € / an) : flotte station ?

- Mise en base régulière des résultats et
- Qualification au fil de l'eau

Question : la qualification est-elle juste / homogène / reproductible ?

Source : somlit.epoc.u-bordeaux1.fr

Source : F. Massard cargos-paquebots,net



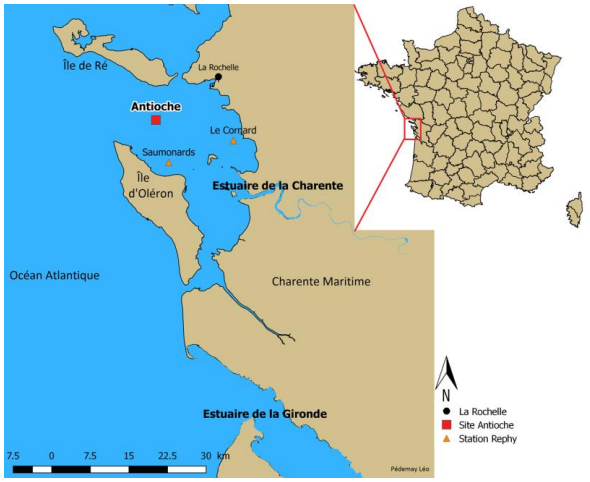
Source : somlit.epoc.u-bordeaux1.fr

Objectifs du stage M2

1- **Transcrire sous R les raisonnements menant à la qualification des données**

https://www.rstudio.com/

2- **Mettre au point les scripts (Antioche) et tester autres stations (Arcachon, Brest)**



Source : Pédemay (2017)

Source : somlit.epoc.u-bordeaux1.fr

~~Objectif du stage de~~ M2

3- Site web interactif : « boite à outils » R

Idée à suivre : MySOMLIT par David *et al.* (*cf*. l'atelier du 27/09/2017) programmation *via* SHINY de R Studio

MYSOMLIT

SOMLIT - Service d'Observation en Milieu LITtoral

DONNÉES BASSE FRÉQUENCE

développé par Valérie David (UMR EPOC 5805 / OASU)

Données de Pleine-Mer - Surface. Les données de code qualité 1, 4, 5 et 9 ne sont pas considérées.

Choisir sa série...

Station

Antioche

Paramètre

Température

Par défaut, toute la série est considérée: depuis le début du suivi (6/1997) jusqu'à aujourd'hui. Pour une période plus courte, les dates peuvent etre changée manuellement en respectant le format.

Choix de la période

01/06/1997 - 14/09/2017

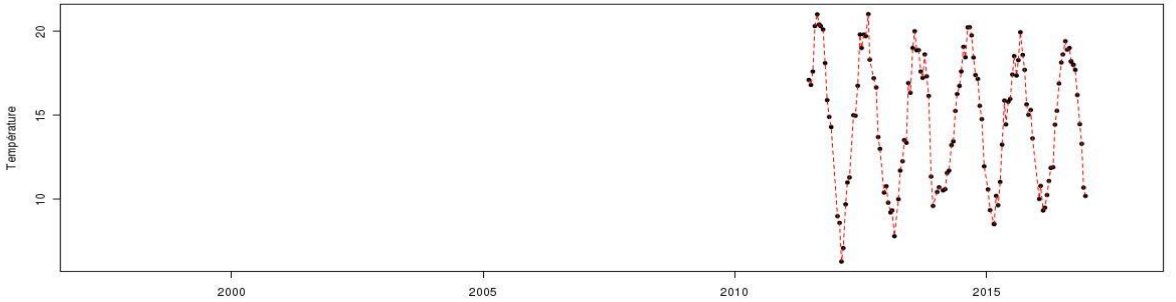
Logiciel et analyses statistiques utilisés

Toutes les analyses sont realisées avec le logiciel R: la régularisation avec la fonction regul(), la moyenne mobile avec tsd() de la librairie [pastecs], les tendances de Mann-Kendall avec la fonction mkTrend() corrigée de l'autocorrelation de la library [fume], les shifts avec Fstats(), efp() et breakpoints() de la librairie [strucchange], les régressions linéaires avec la fonction glm() type gaussien, la détection de cycle(s) avec la fonction spectrum(), la comparaison de la saisonnalité avec la fonction lm() pour un modèle linéaire de type ANOVA 2 facteurs fixes croisés avec interaction.

Choix de la série | Variabilité interannuelle | Saisonnalité

Antioche / Température - données brutes

-Sur l'ensemble des données de la base, 6 années disponibles (juin 2011 - déc. 2016) avec 21 données par an en moyenne (points noirs sur le graphique)



-Votre choix se porte sur 6 ans (juin 2011 - déc. 2016) avec 21 données par an en moyenne (pointillés rouges sur le graphique)

L'analyse de cette série (variabilité interannuelle, saisonnalité) est jugée non viable.

PRÉCAUTIONS À PRENDRE POUR COMPARER DES SÉRIES

Les analyses sont faites **série par série.** S'il s'agit d'analyser une série, la période temporelle totale disponible peut etre considérée mais **une comparaison entre série nécessite la considération d'une période commune** . Le choix de la fenetre temporelle commune peut etre fait à l'aide des tableaux ci-dessous (stations à comparer pour un paramètre donné ou paramètre à comparer pour une station donnée) **Préférer des années entières** du 1er janvier au 31 décembre.

Codes qualité SOMLIT

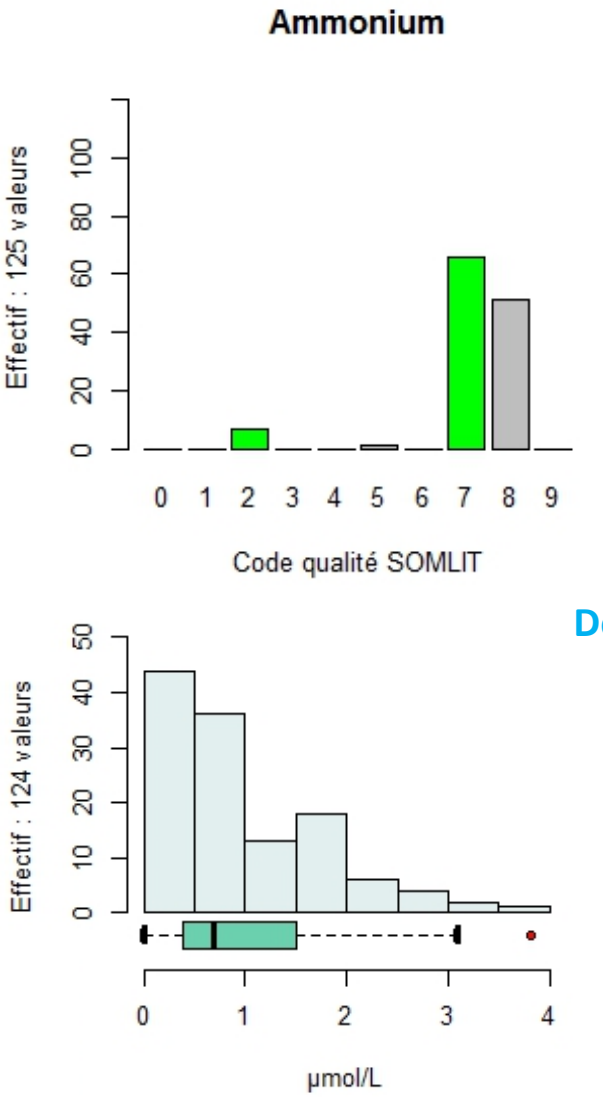
Code Qualité	Description	Code Couleur
0	Donnée en dessous de la limite de détection	
1	Prélèvement effectué, mais mesure non réalisée	NA
2	Mesure bonne, échantillon non répliqué	
3	Mesure douteuse	
4	Mesure mauvaise	
5	Prélèvement effectué, mais valeur pas encore reportée	NA
6	Mesure bonne (moyenne de plusieurs réplicats)	
7	Mesure bonne (valeur acquise hors protocoles SOMLIT)	
8	Donnée non qualifiée	
9	Échantillon non prélevé	NA

Source : Garcia et Oriol (2017) : protocoles SOMLIT

Codes qualité SOMLIT

Code Qualité	Description	Code Couleur
0	Donnée en dessous de la limite de détection	
1	Prélèvement effectué, mais mesure non réalisée	NA
2	Mesure bonne, échantillon non répliqué	
3	Mesure douteuse	
4	Mesure mauvaise	
5	Prélèvement effectué, mais valeur pas encore reportée	NA
6	Mesure bonne (moyenne de plusieurs réplicats)	
7	Mesure bonne (valeur acquise hors protocoles SOMLIT)	
8	Donnée non qualifiée	
9	Échantillon non prélevé	NA

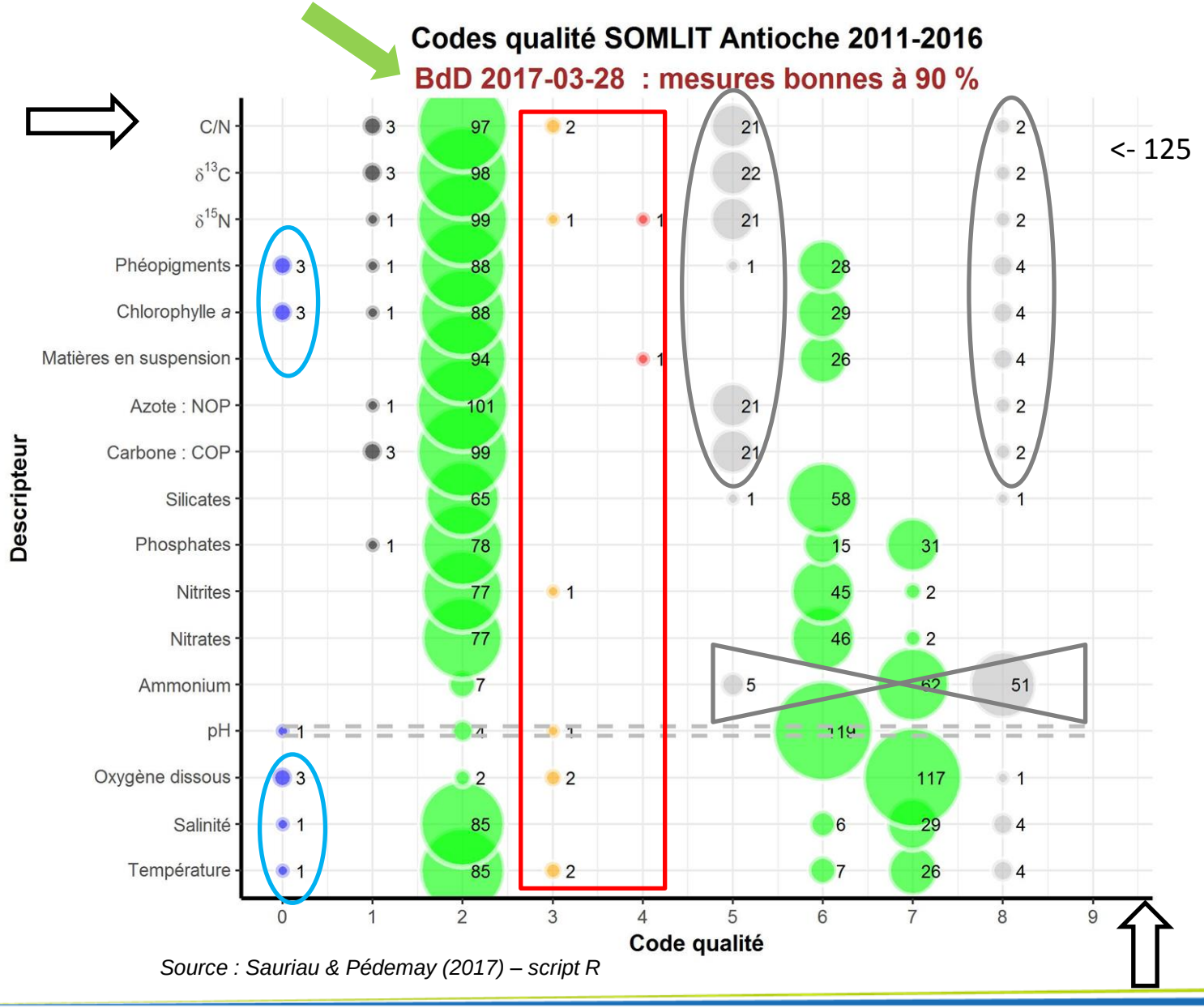
Source : Garcia et Oriol (2017) : protocoles SOMLIT



Donnée initiales 28/03/2017

Source : Sauriau & Pédemay (2017) – script R

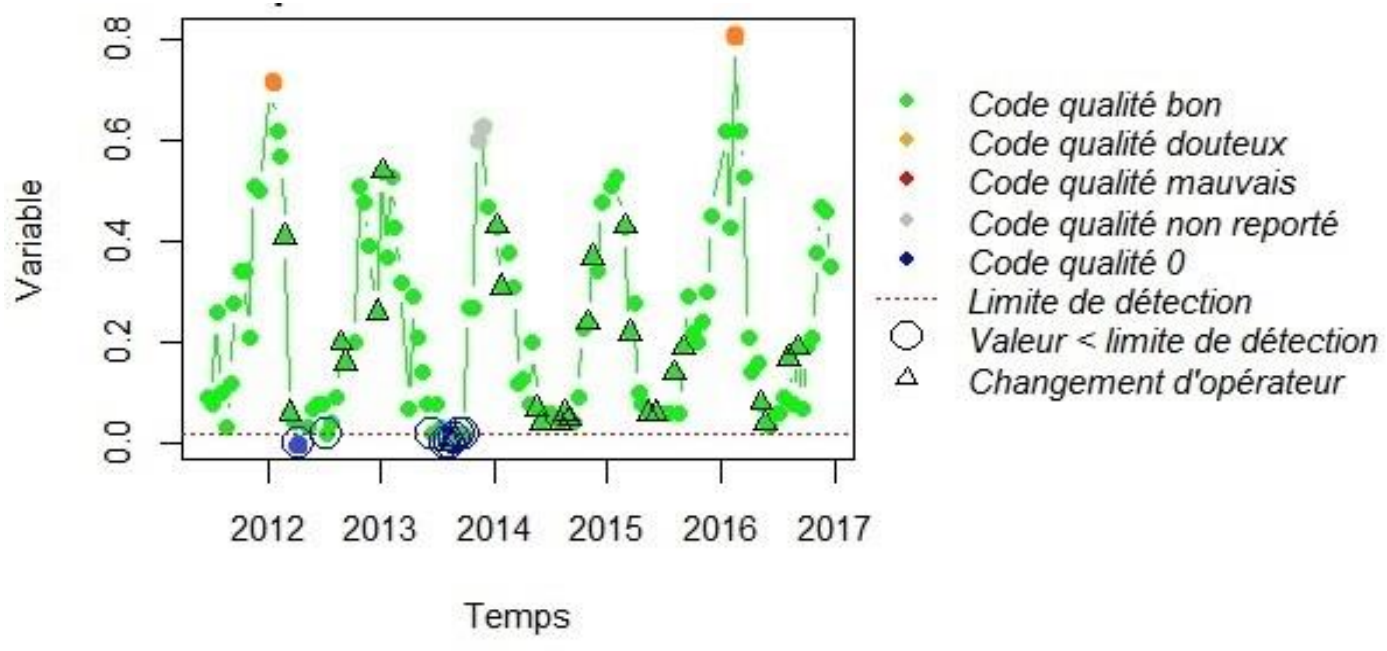
Bilan Antioche
état initial 28-03-2017



Source : Sauriau & Pédemay (2017) – script R

1- Analyses qualitatives : ☺ aux valeurs extrêmes

Données initiales



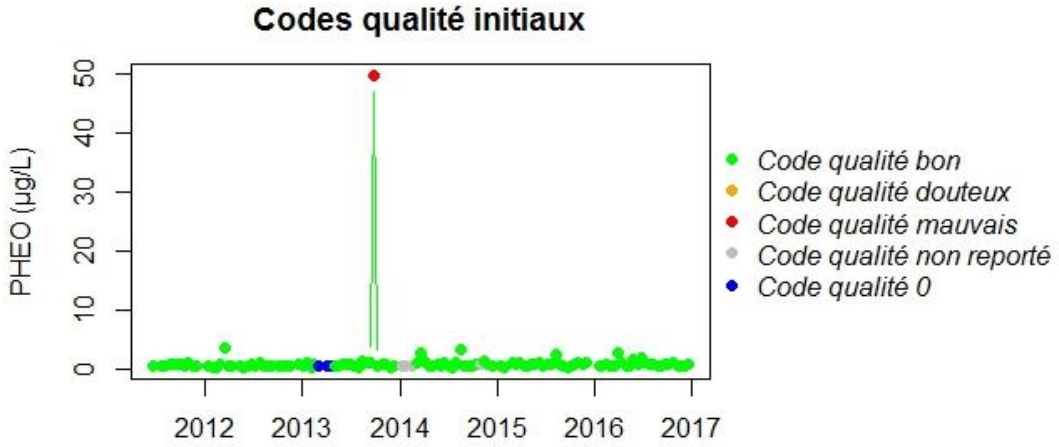
- Vérification du code qualité 0 : « < limite de détection » (Aminot & Kérouel, 2007)

- Effet du changement d'opérateur → test de Kolmogorov-Smirnov (Marsaglia *et al.*, 2003)

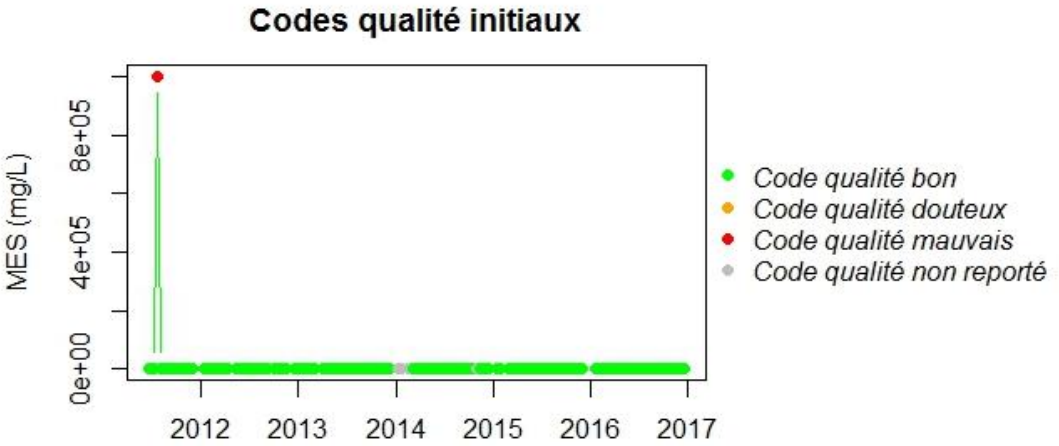
- QUESTION : les valeurs code 3 sont elles réellement douteuses ?

1- Analyses qualitatives : ☺ aux valeurs extrêmes

- Code 4 confirmé en l'absence d'une bonne raison de penser à un processus explicatif connu



- Attention aux possibles phénomènes météo-océaniques rares

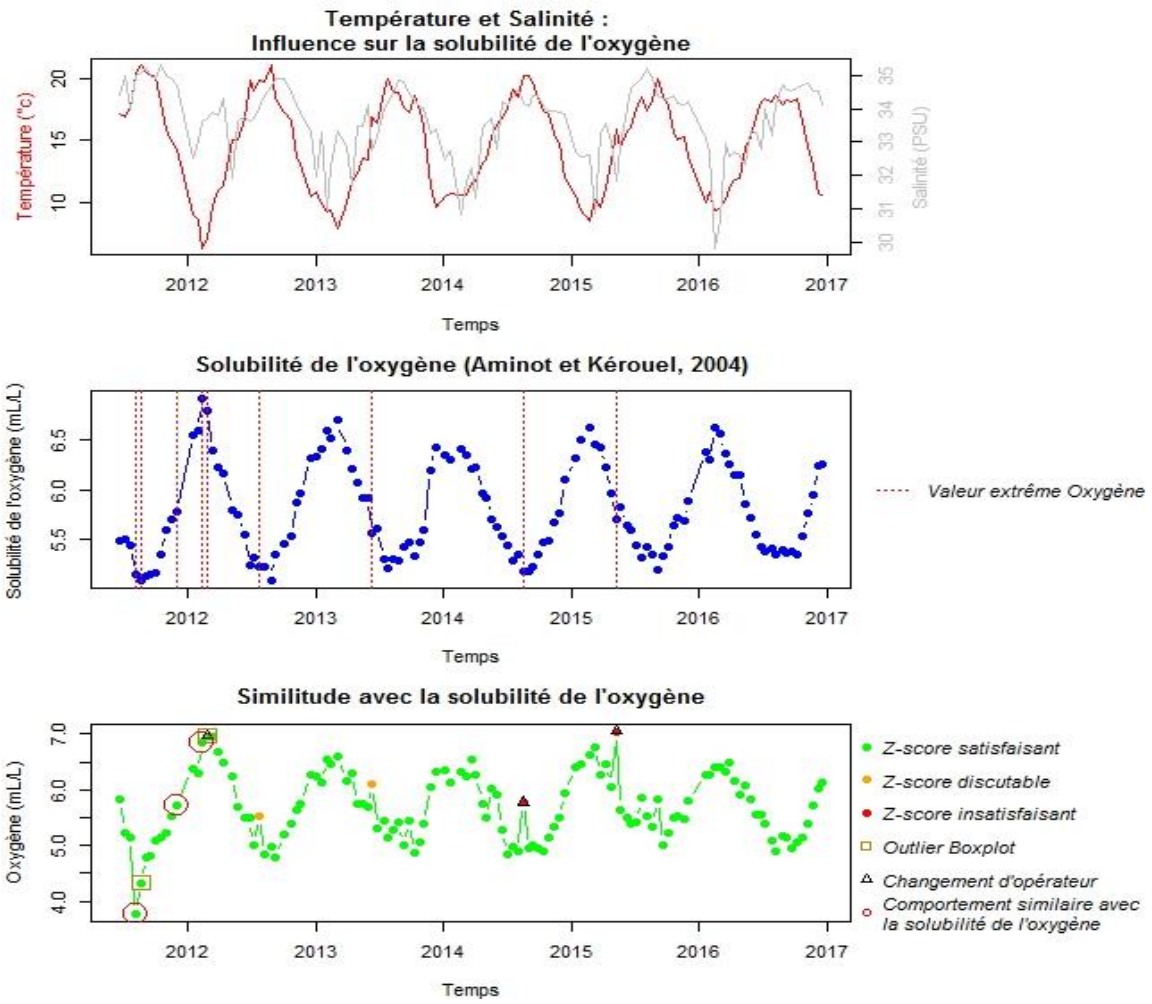


Idée à suivre

Besoin de données externes…
cf. météo, blooms du phytoplancton

2- Analyses quantitatives : ☺ aux chroniques

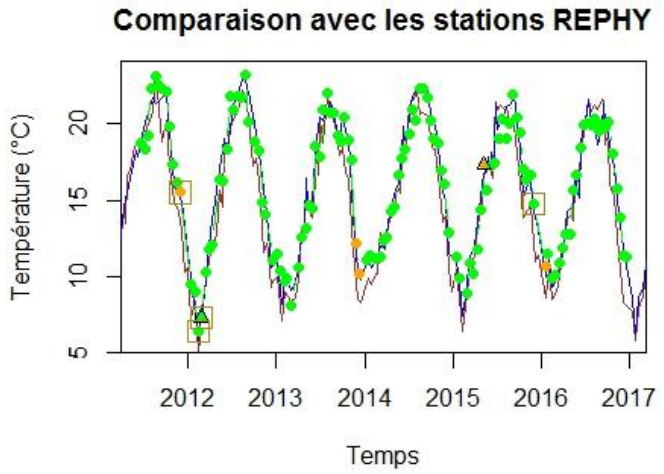
Solubilité O_2 dissous
(Aminot & Kérouel, 2004)

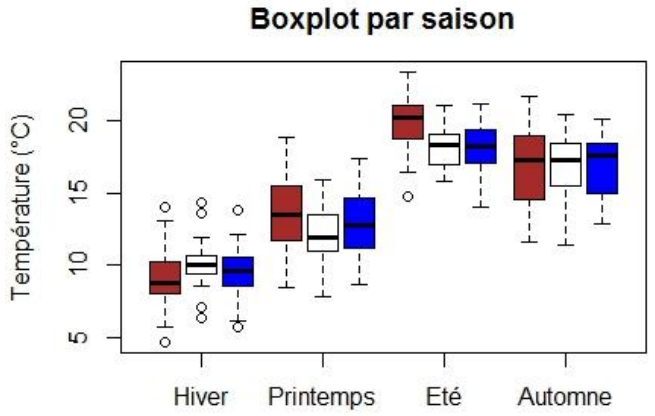


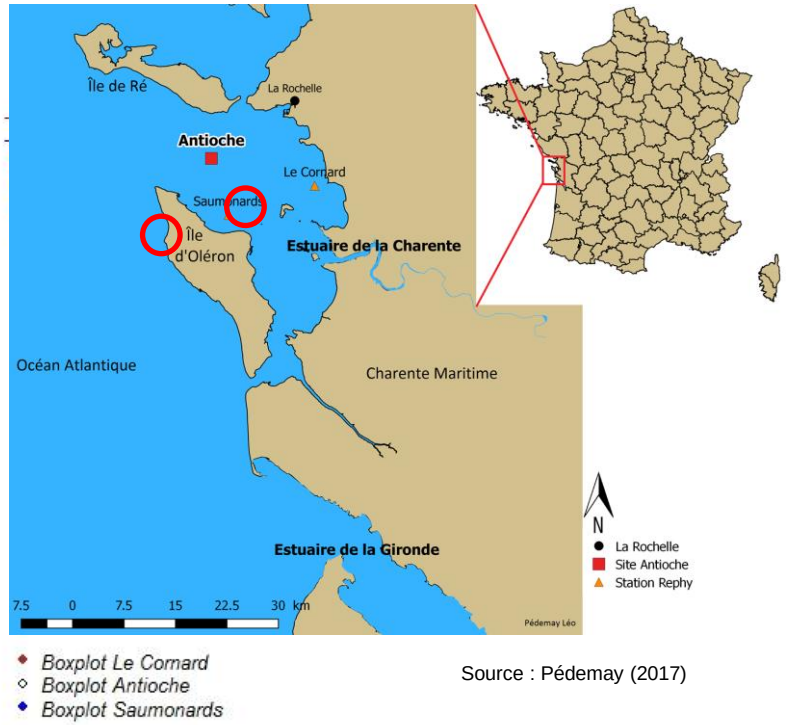
Idée à suivre

Régression entre variables SOMLIT +
+ fichier opérateur SOMLIT à créer…
+ fichier blooms du phytoplancton à créer…

2- Analyses quantitatives : ☺ aux chroniques







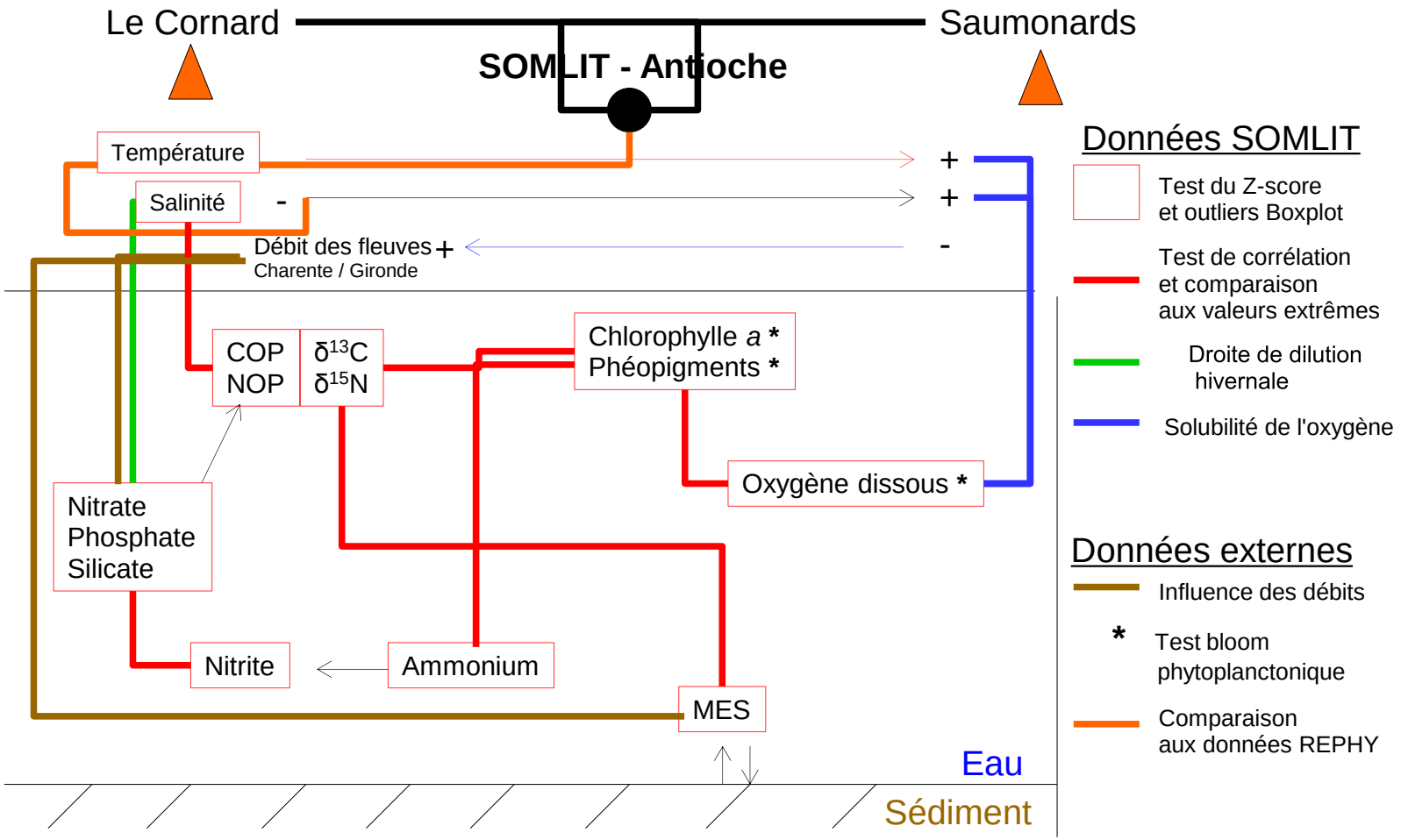
◆ *Boxplot Le Cornard*
◇ *Boxplot Antioche*
◆ *Boxplot Saumonards*

Source : Pédemay (2017)

Idée à suivre

Besoin de données externes T, S en contrôle…

Analyses quantitatives : synopsis



Idée à suivre

Boxplot, régression, Z score, droite de dilution, corrélation… processus commun

Fonctionnement d'un script R avec R Studio

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

validation_SIOH4.R

```
38  # Date au bon format :
39  data$date<-as.Date(data$DATE)
40  data$annee<-format(data$date, format = "%Y")
41  data$mois<-format(data$date, format = "%m")
42  data$jour<-format(data$date, format = "%d") # Format de la date (Ymd)
43
44  # Choix du paramètre : T S O PH NH4 NO3 NO2 PO4 SIOH4 COP NOP MES CHLA PHEO DN15 DC13
45  data$par<-data$SIOH4
46  data$qual<-data$qSIOH4
47
48  # Supprimer les data 999.999 ou 999999 ou NA avec codes qualité 1, 5 et 9, supprimer l
49  data$qual<-ifelse(is.na(data$qual),5,data$qual)
50  data$par<-ifelse(data$qual==1,NA,
51                   ifelse(data$qual==5,NA,
52                          ifelse(data$qual==4,NA,
53                                 ifelse(data$qual==9,NA,data$par))))
54
55  # Changement d'opérateur :
56  data$diffOperateur<-abs(c(0, diff(data$OPERATEUR)))
57  data$coloperateur<-ifelse(data$diffOperateur>0,"black",NA) # couleur diff-opérateur
58  data$pchOperateur<-ifelse(data$diffOperateur>0,17,16) # forme diff-opérateur
59
```

Console C:/Users/Utilisateur/Desktop/SOMLIT/donnee_somlit/

```
> data<-merge(donnees,operateur,by= DATE )
> data$date<-as.Date(data$DATE)
> data$annee<-format(data$date, format = "%Y")
> data$mois<-format(data$date, format = "%m")
> data$jour<-format(data$date, format = "%d") # Format de la date (Ymd)
> data$par<-data$SIOH4
> data$qual<-data$qSIOH4
> data$qual<-ifelse(is.na(data$qual),5,data$qual)
> data$par<-ifelse(data$qual==1,NA,
+                  ifelse(data$qual==5,NA,
+                         ifelse(data$qual==4,NA,
+                                ifelse(data$qual==9,NA,data$par))))
>
```

Environment History

Import Dataset List

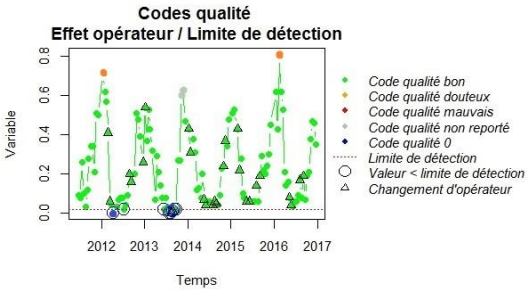
Global Environment

Data

data	125 obs. of 102 variables
donnees	125 obs. of 89 variables
operateur	125 obs. of 8 variables

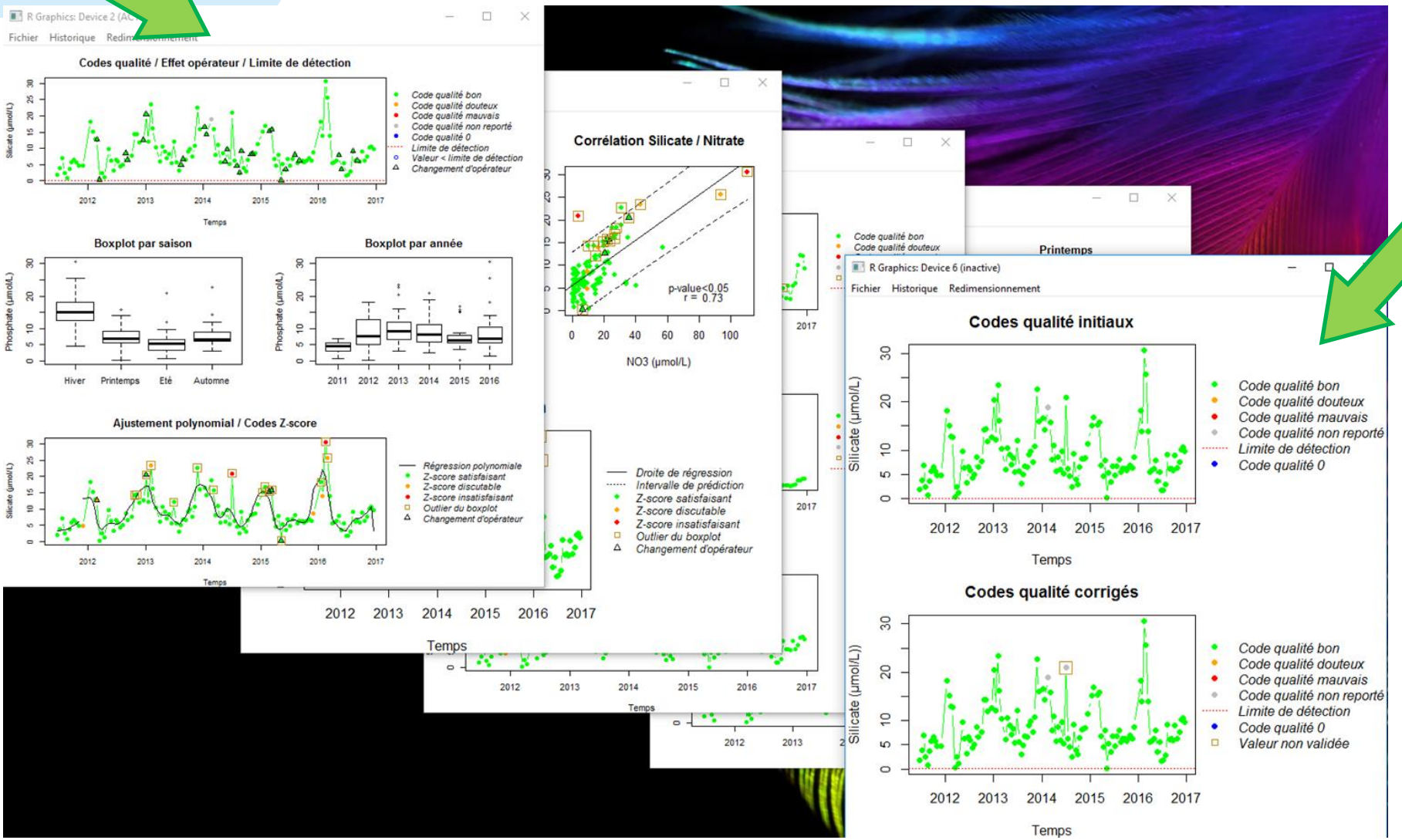
Files Plots Packages Help Viewer

Zoom Export



Source : https://www.rstudio.com/
Pedemay (2017) - Scripts

Résultats obtenus



Source : https://www.rstudio.com/
Pedemay (2017) - Scripts

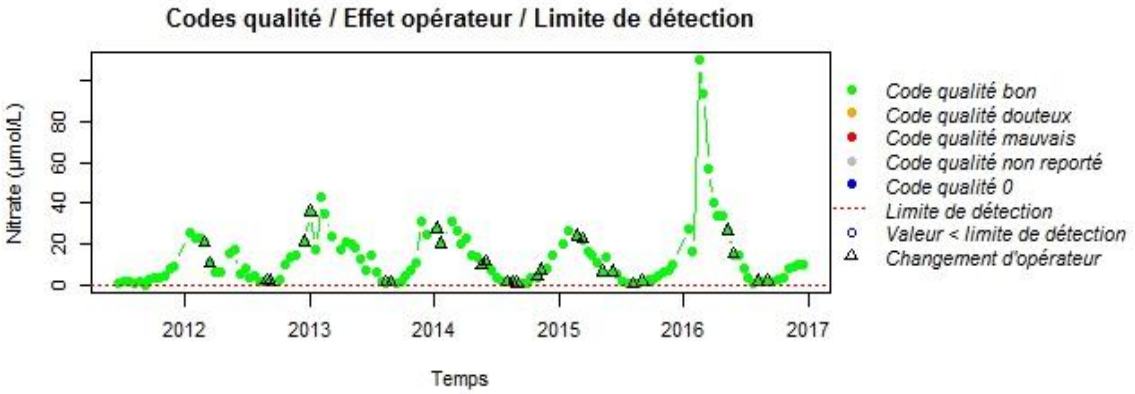
Nitrate

Analyses qualitatives

Données initiales

→ **Limite de détection : 0**

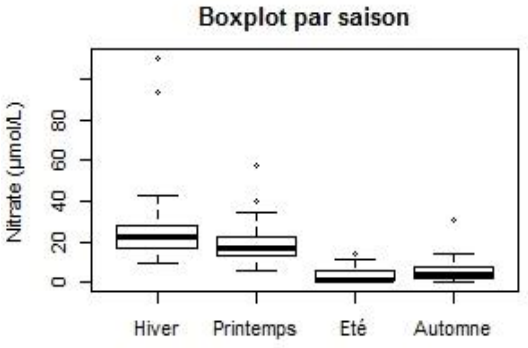
→ **Opérateur : NS**

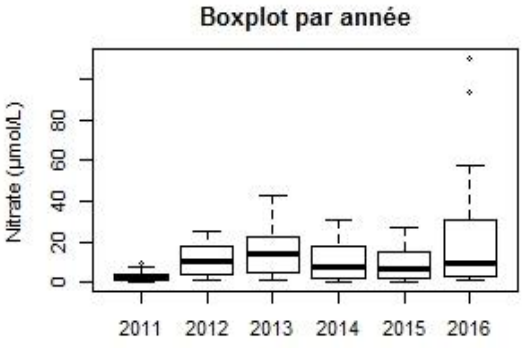


Analyses quantitatives

Valeurs extrêmes

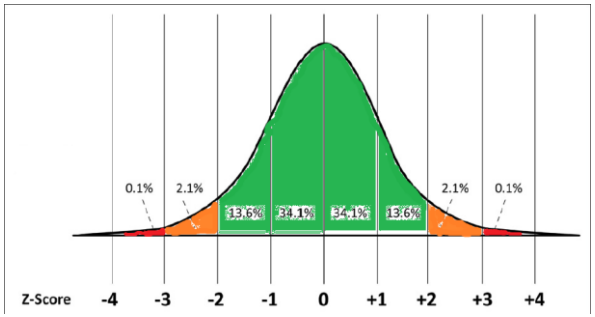
→ **Boxplots : 8**

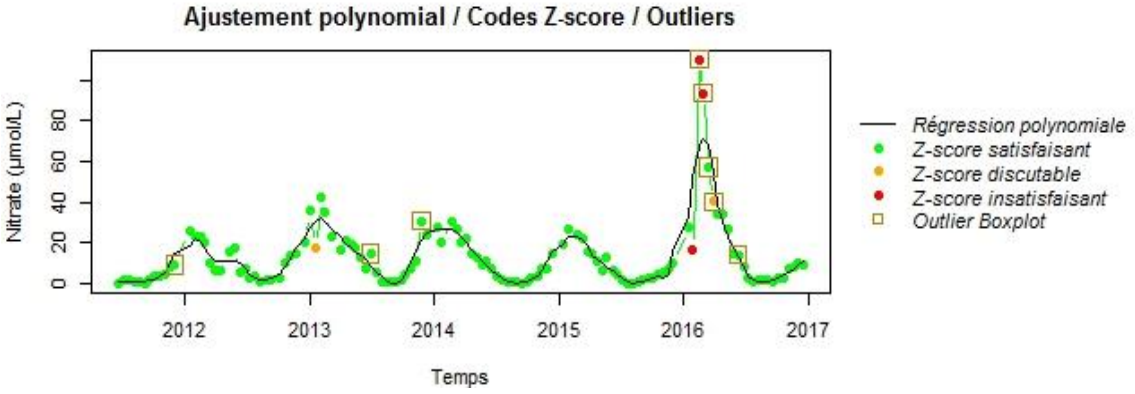




→ **Polynomiale LOESS**

→ **Z-score : 2 + 3 = 5**





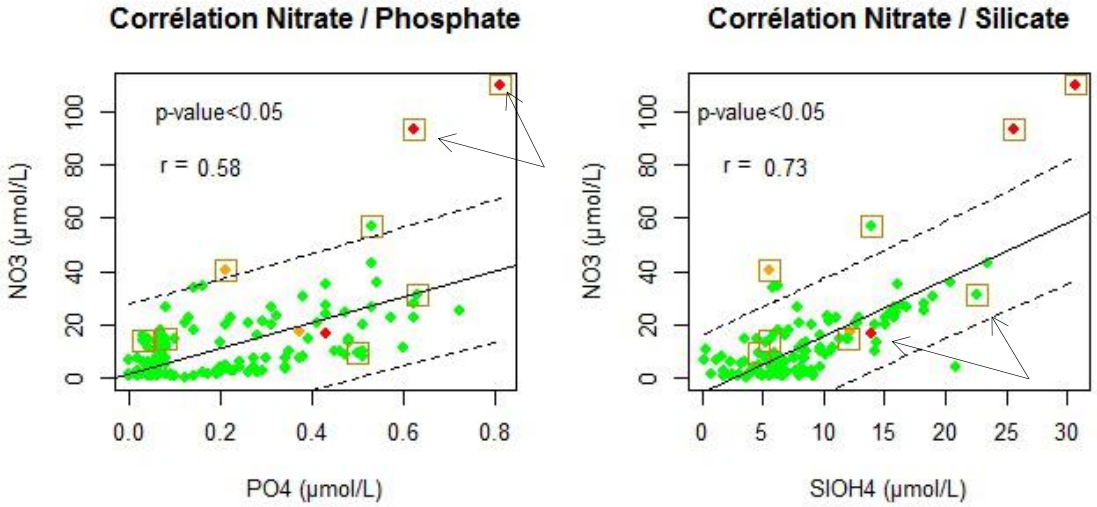
Nitrate

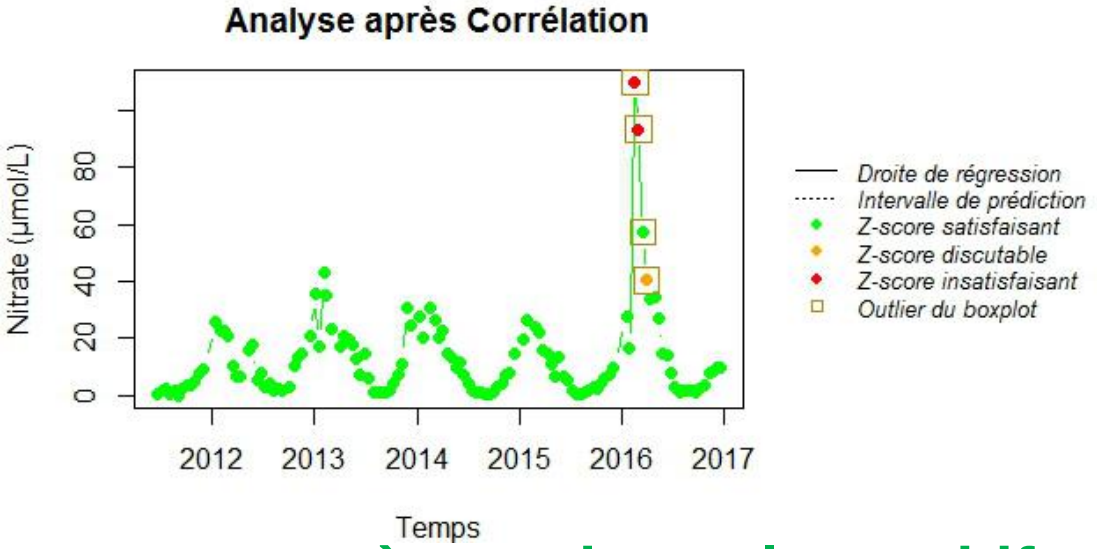
Analyses quantitatives

Méthode des corrélations
(Best & Roberts, 1975)
R de Spearman

→ Intervalle de prédiction 95%
de la droite de régression

Valeurs extrêmes : **4**





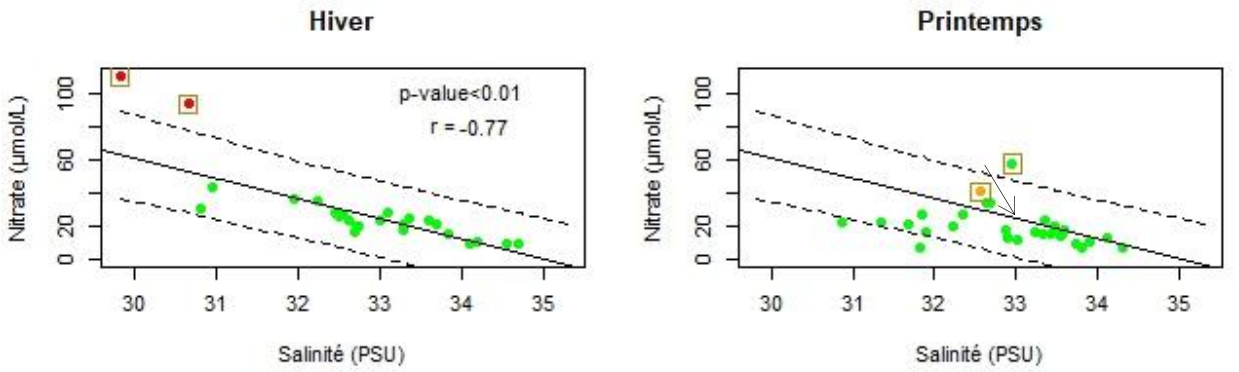
Idée à suivre

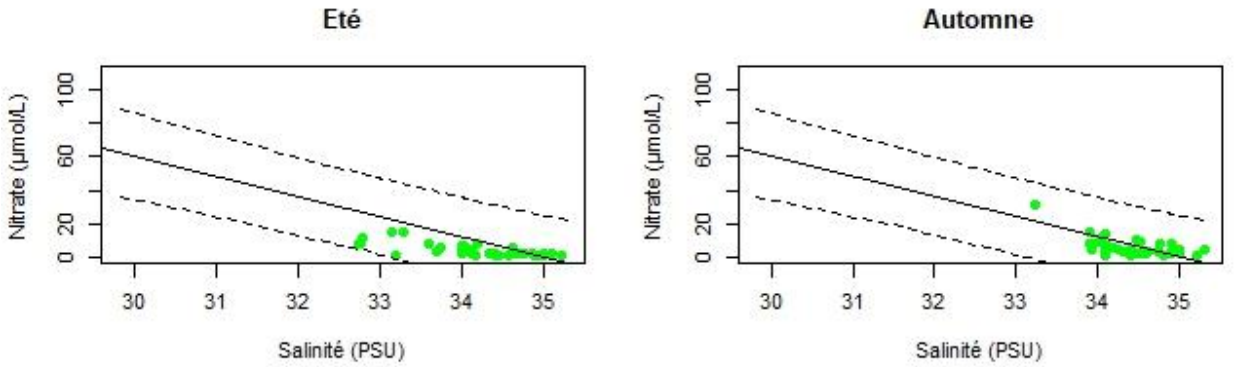
Processus commun à certains sels nutritifs

Nitrate

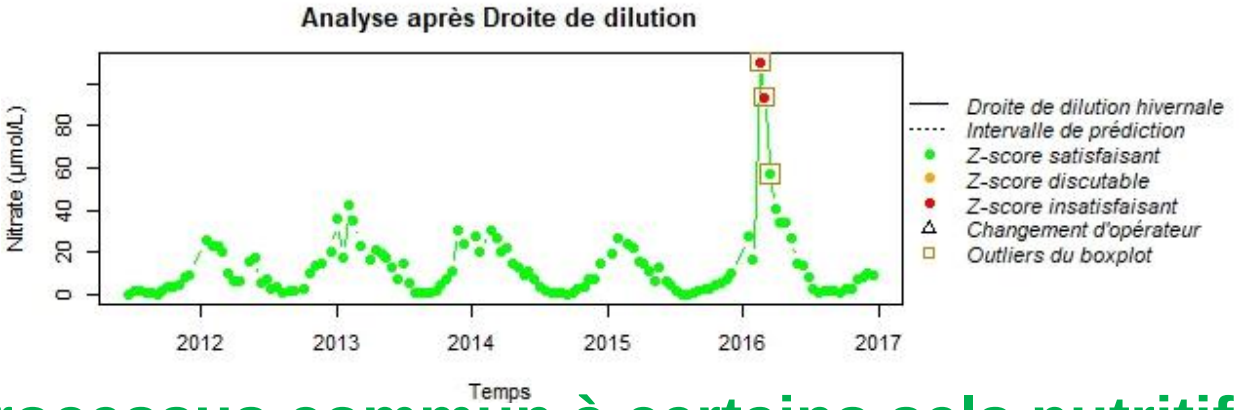
Analyses quantitatives

Méthode de la droite de dilution hivernale





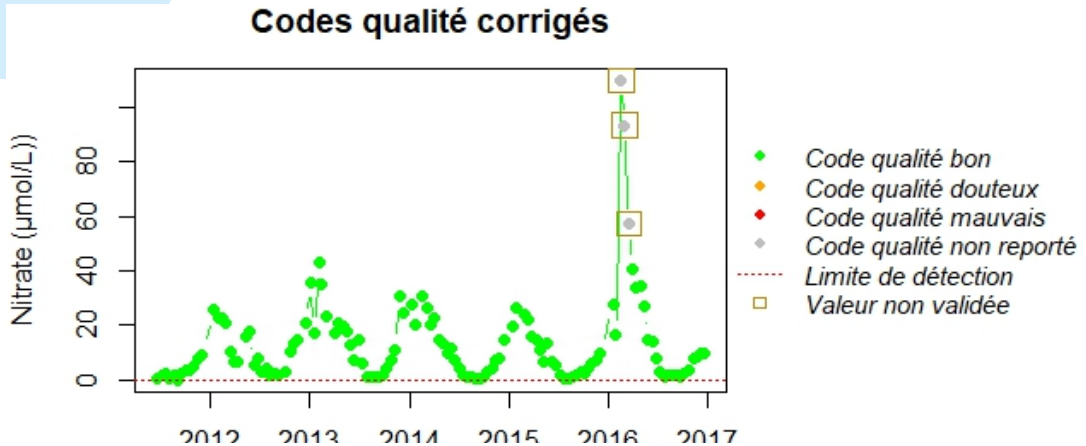
Valeurs extrêmes : **3**

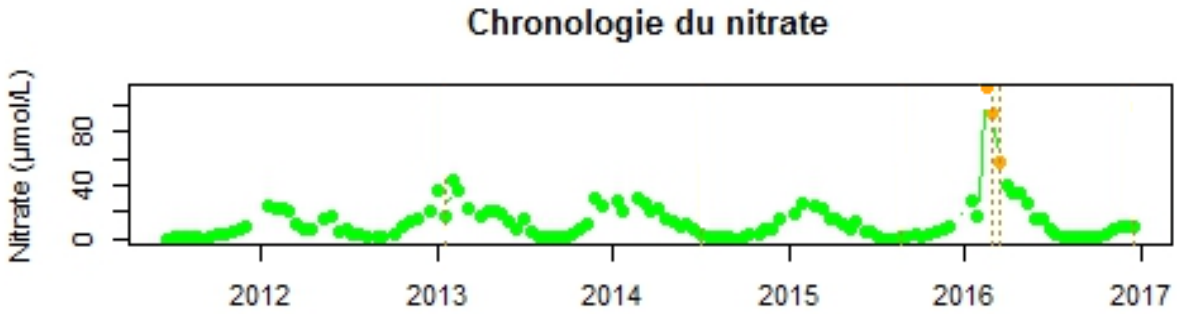


Idée à suivre

Processus commun à certains sels nutritifs

Nitrate



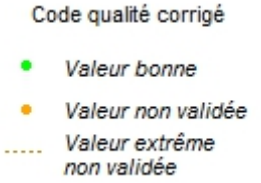


Données SOMLIT

Restent : 3 valeurs extrêmes

Données externes

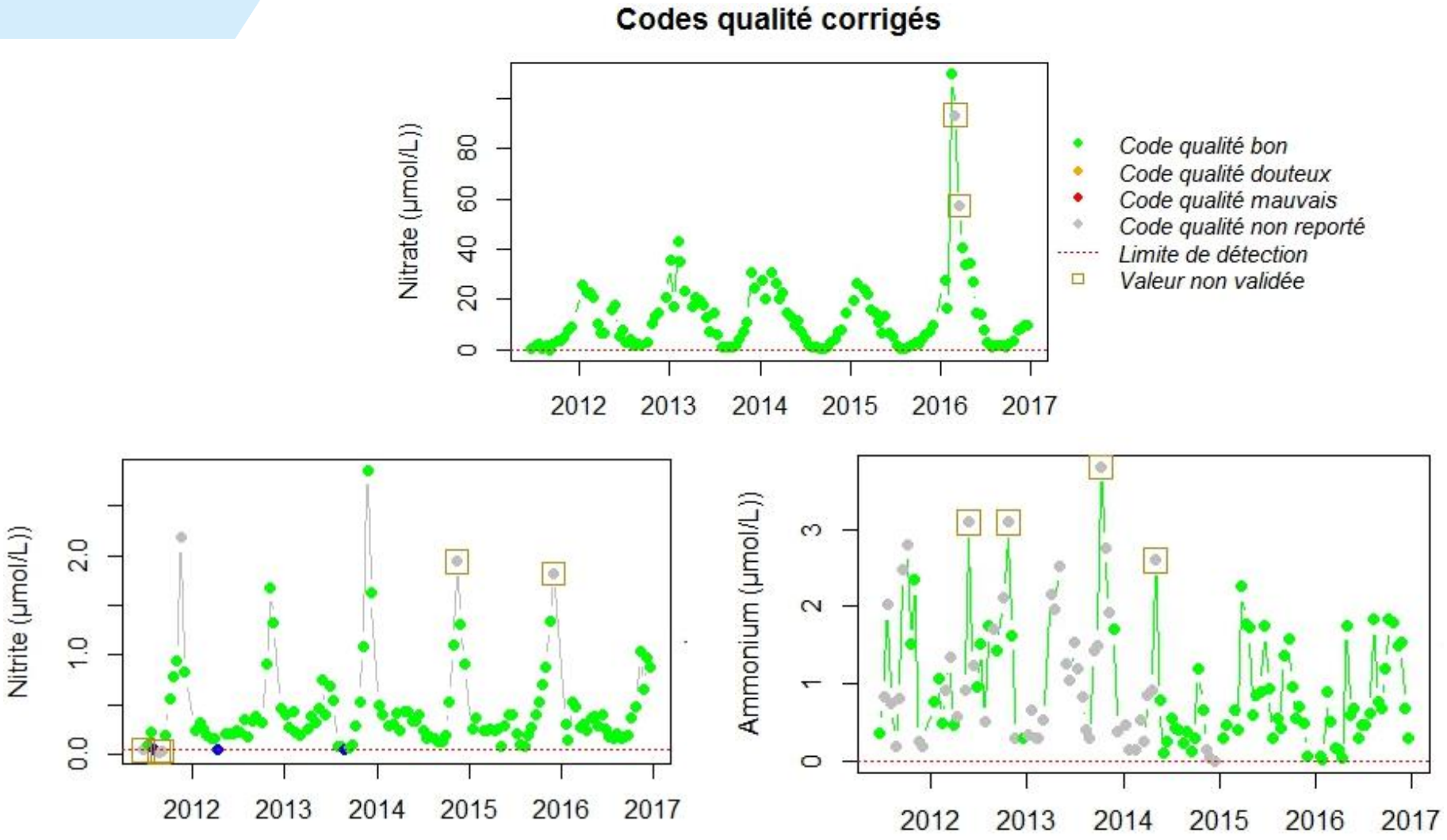
Au final : valeurs validées



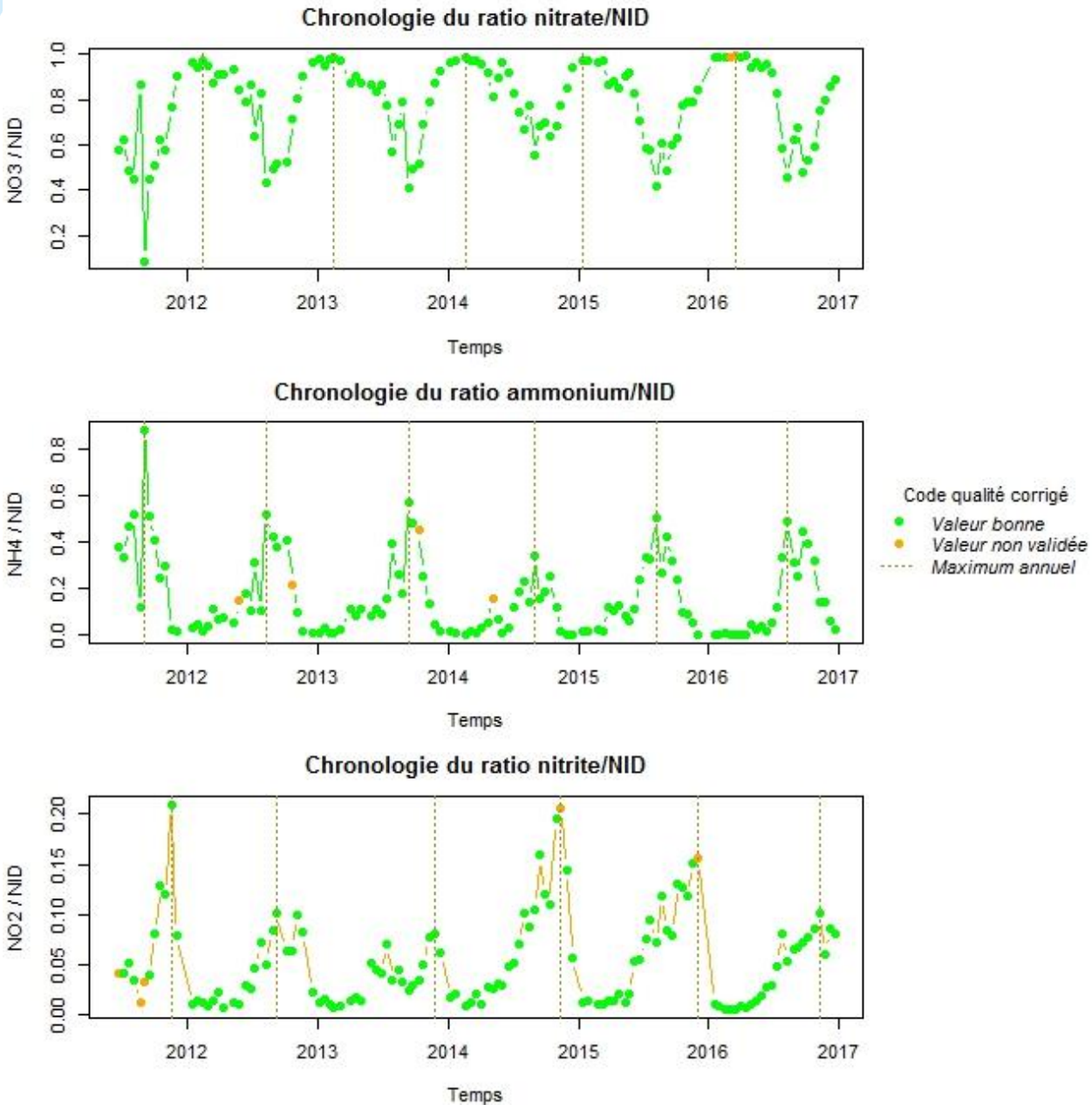
Idée à suivre

Besoin de données externes débits fluviaux

Cycle de l'azote

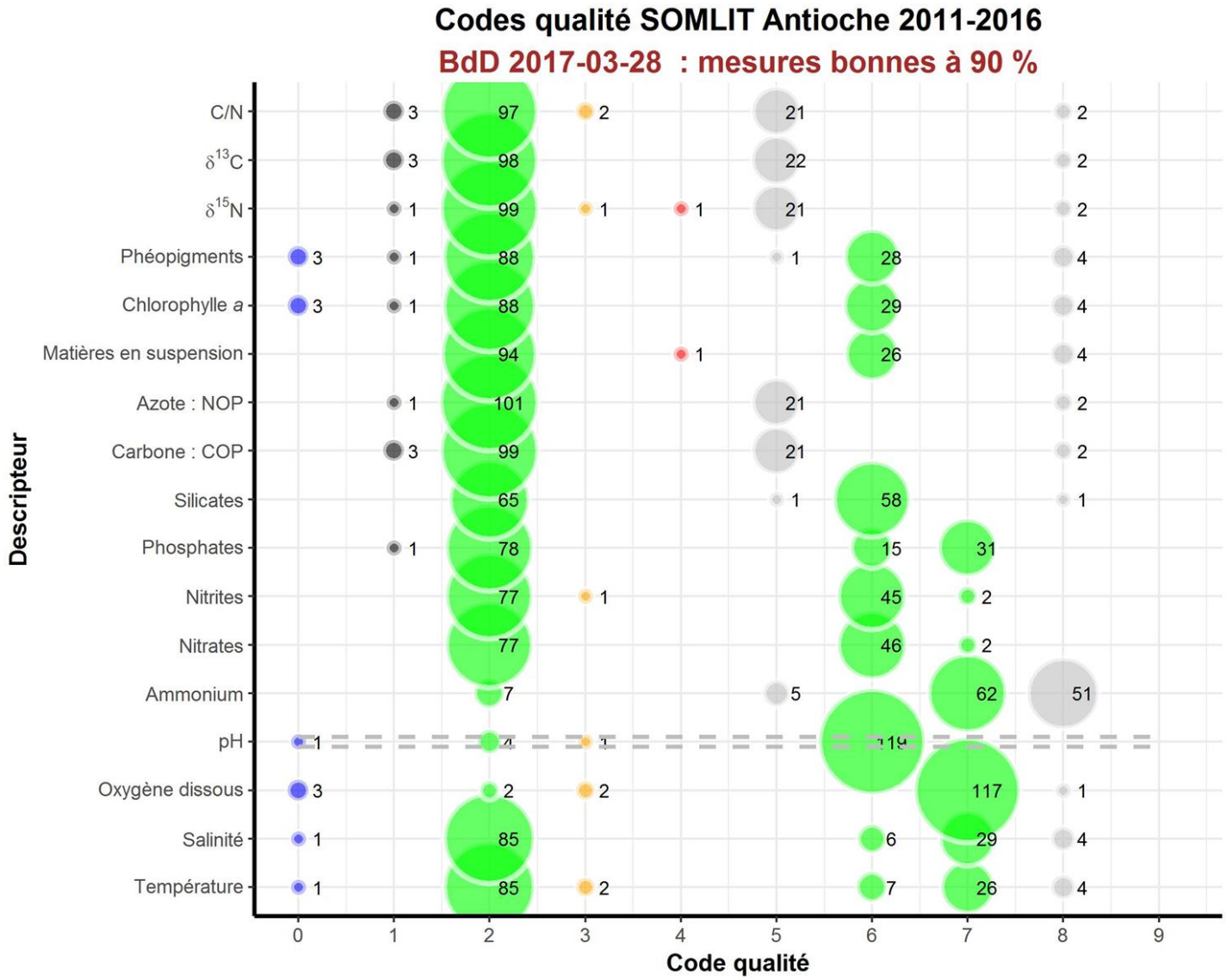


Cycle de l'azote



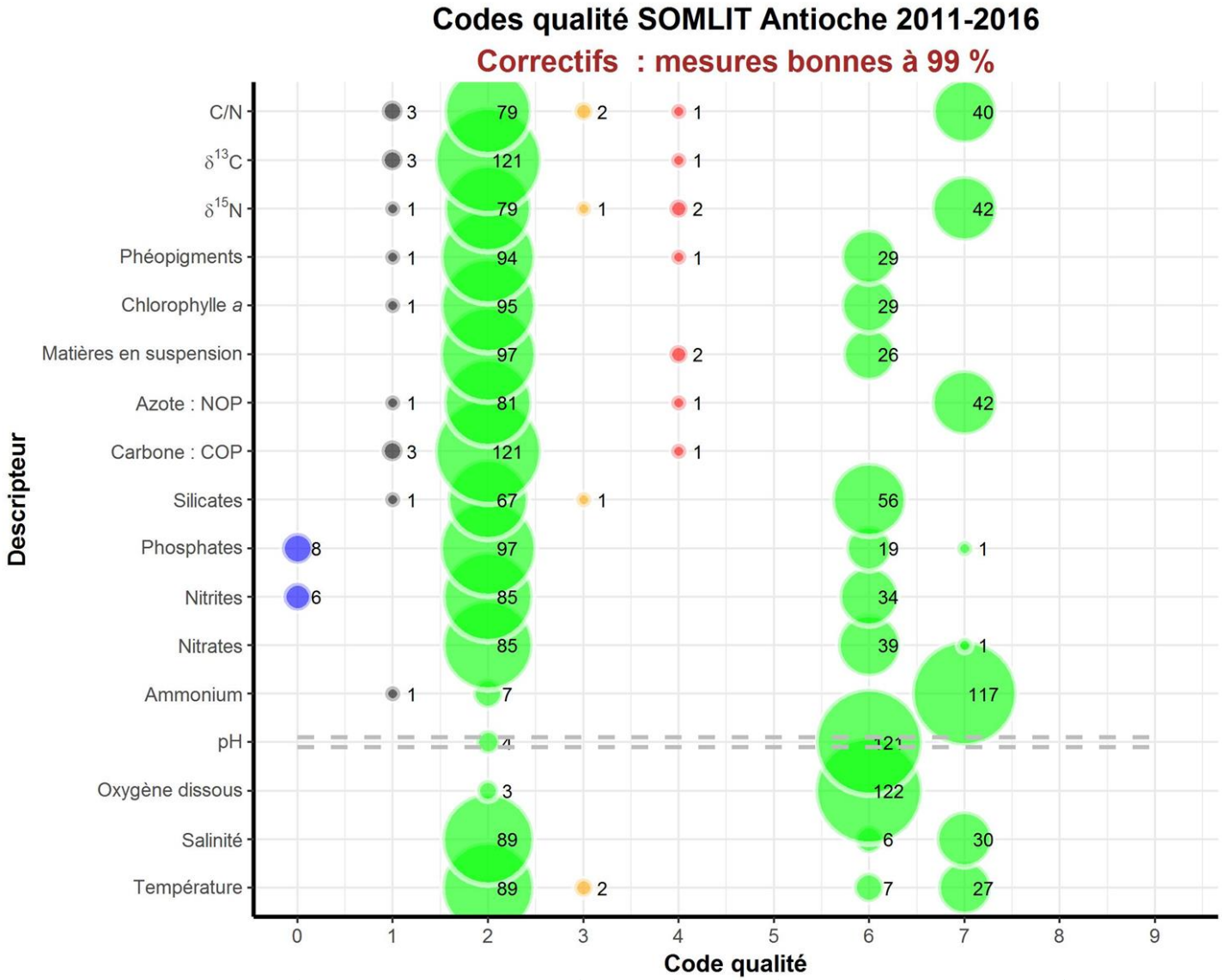
Objectif 1

Bilan sur Antioche : état initial



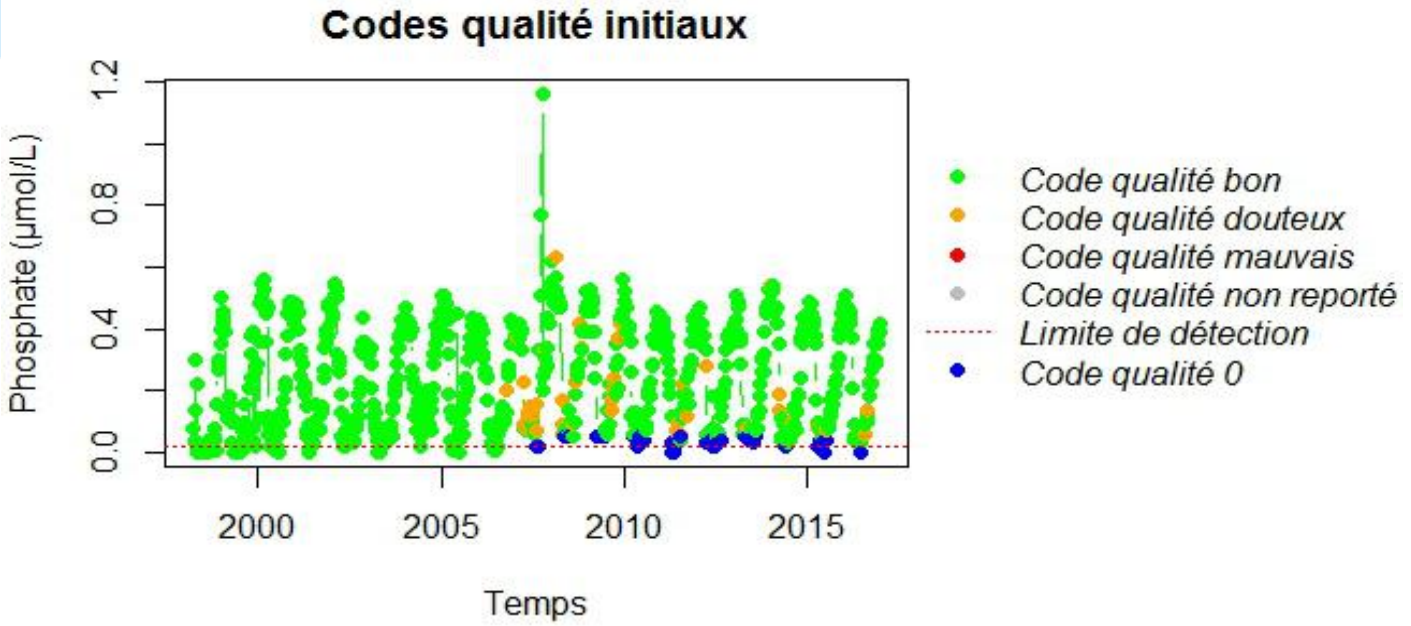
Objectif 1

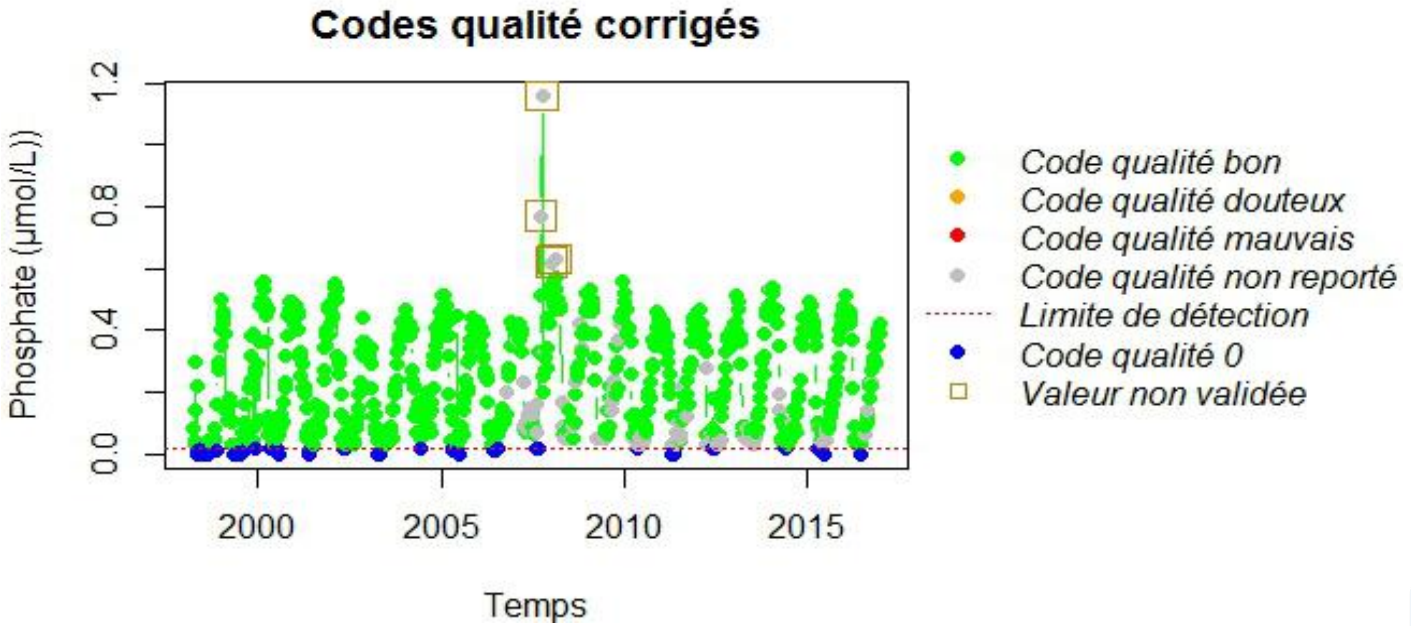
Bilan sur Antioche : état après analyse



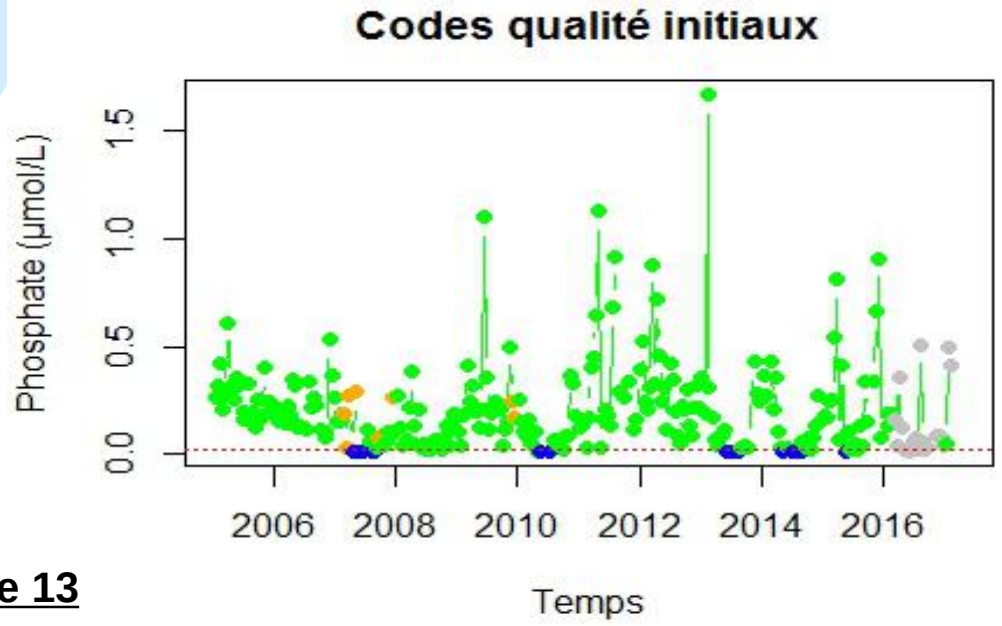
Objectif 2 : test des scripts sur autres sites

Brest - Portzic

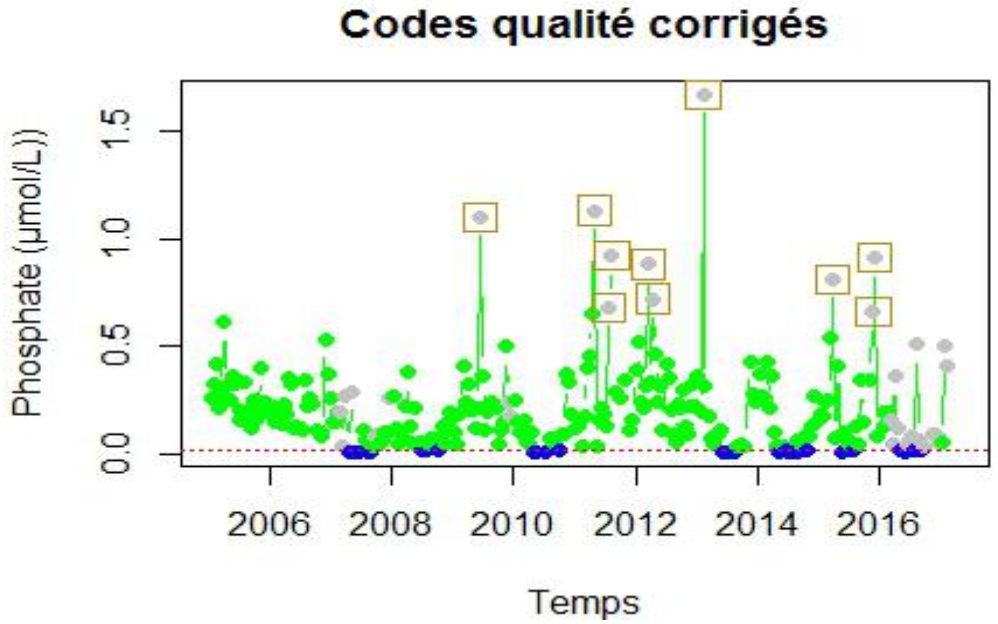




Objectif 2 : test des scripts sur autres sites



Arcachon – Bouée 13



Bilan

1/ Approche méthodologique

- **Concept des valeurs extrêmes** — Valides si bonne raison de penser qu'un processus physique vs biologique sous-jacent explique les valeurs
- **Concept analyse des chroniques** — Dépendance temporelle des données
- **Limites des approches méthodologiques** — Intervalle de prédiction à 95%
Corrélations linéaires
Potentiel décalage temporel
Forçages météo ignorés

2/ Utilisation des scripts

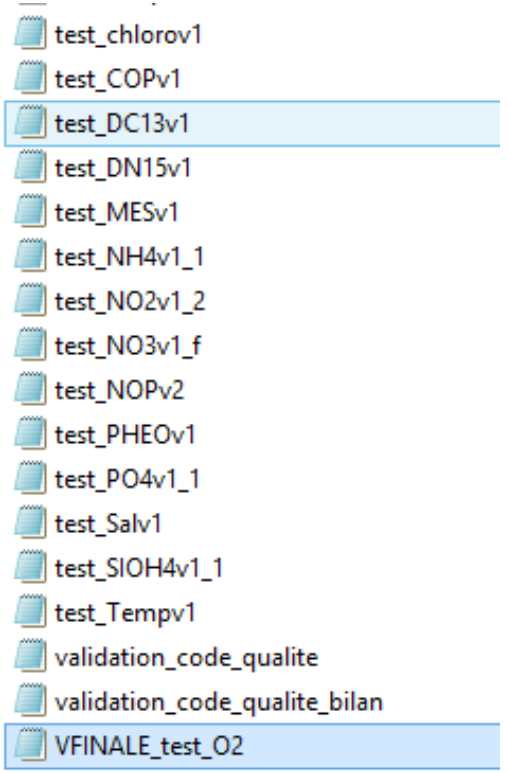
- **Validation future des données**
- **Expertise approfondie des valeurs extrêmes**
- **Algorithme des scripts** — Création 2 fichiers : opérateur / station
Identification des valeurs extrêmes
Stations océaniques *vs* estuariennes
- **Données externes** — Besoin de comparaison
Météo-France, Eau-France, CANDHIS, Quadrige², Pelagos-Phytobs, modèles physiques (trajets masse d'eau)

Bilan

3/ Liste des 17 scripts : 430 (184-735) lignes / script

- **Code qualité et histogramme des valeurs / variable**
- **Bilan des codes qualité / variable**
- **Température**
- **Salinité**
- **O_2**
- **Sels nutritifs : NO_3, NO_2, NH_4, PO_4, $SI(OH)_4$**
- **MES**
- **Chloro- et phéopigments**
- **COP, NOP**
- **$\delta^{13}C$, $\delta^{15}N$**
- **C/N déduit des 4 précédents**

- **Pas de script**
- **pH**
- **Pico-nanoplancton**



4/ Reprendre l'idée SHINY pour un site web intéractif

Idée à suivre

MySOMLIT par David *et al.* (*cf*. l'atelier du 27/09)
programmation *via* SHINY de R Studio

Conclusion sous forme d'un questionnaire

- **Utilité de la démarche :** OUI NON
- **Pertinence analyses statistiques :** Lesquelles
- **Autres analyses statistiques :** Lesquelles
- **Test des scripts :** Des volontaires ?
- **Perspective d'utilisation interface web (Shiny + R Studio) :** POUR CONTRE