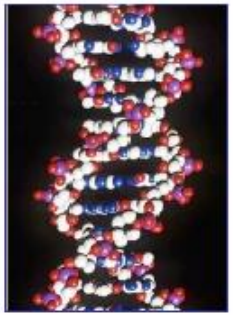
# Phénotypage haut débit des plantes : Prétraitement et analyses

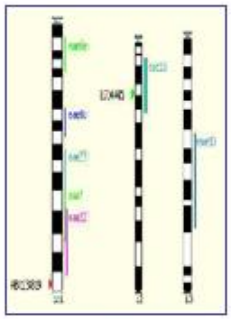
## Journées NETBIO – Montpellier décembre 2018

Nadine HILGERT (UMR MISTEA - Montpellier)

14 / 12 / 2015

# Qu’est ce que le phénotypage ?





Génotypes de plante

Interactions

Environnement

Phénotype

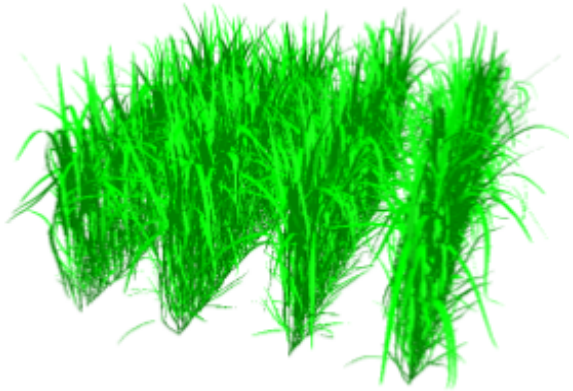
# Que mesurer ?

- **Climate**
- **Pathogen Pressure**
- **Soil**
  - Moisture/ Tension
  - Root Biomass and distribution

**Environnement**

- **Structure**
  - Leaf area (GAI per layer)
  - Clumping
  - Inclination/orientation of organs
  - Density of plants/stems/ears
  - Height (per layers)
  - FIPAR
  - Leaf rolling ….
- **Biochemical content**
  - Chlorophyll, water, dry matter, Nitrogen…
- **State**
  - Fluorescence
  - Skin temperature

**µ-plot**



# Phénotypage de plantes à haut débit

Enjeu : Comprendre la réponse des plantes aux facteurs de l'environnement pour permettre de répondre à la demande d'une production plus importante, de meilleure qualité et plus économe en intrants

PHENOME Réseau Français Phénomique végétale FPPN **Le réseau français de phénomique végétale**

- ➔ Infrastructure nationale https://www.phenome-emphasis.fr/
- ➔ Caractériser de grandes séries de génotypes nécessaires pour les études de variabilité génétique, dans des scénarios environnementaux divers

European Plant Phenotyping Network EPPN 2020

https://eppn2020.plant-phenotyping.eu/

# Un défi pour les maths - info

- De grandes masses de données observées dans le temps, hétérogènes, de nombreuses covariables

- Gérer / analyser ces données :
  - Organiser données et connaissance
  - Prédire / Expliquer en fonction des cofacteurs

  ➔ Méthodes génériques

  Efficaces en temps de calcul

⇒ Développer une nouvelle génération d'outils / méthodes transférables

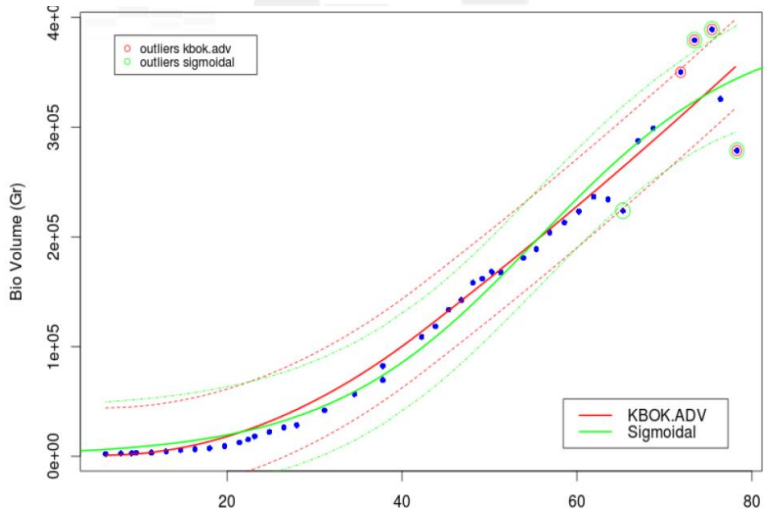
# Prétraitement et analyses

1/ Qualité – validation des données

2/ Extraction et inférence de connaissances à partir de données temporelles

INRA
SCIENCE & IMPACT

# 1. Outlier detection

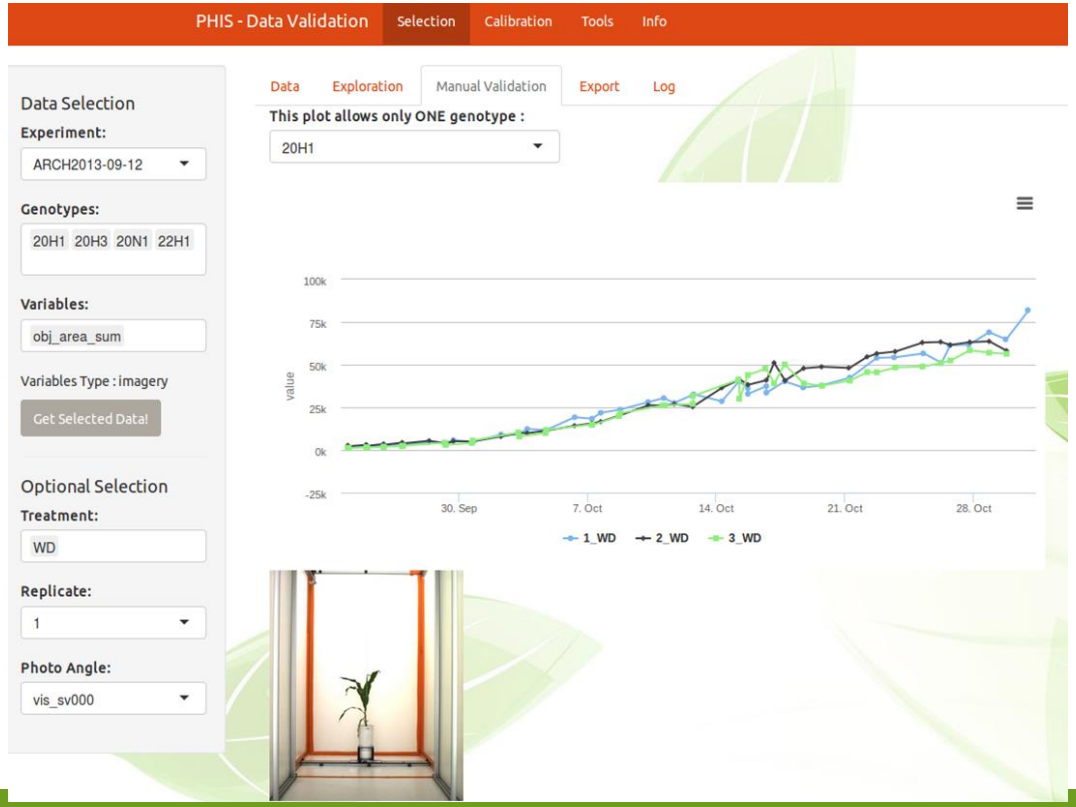



How to handle 1000s curves ?
How to trace it ? (reversible)

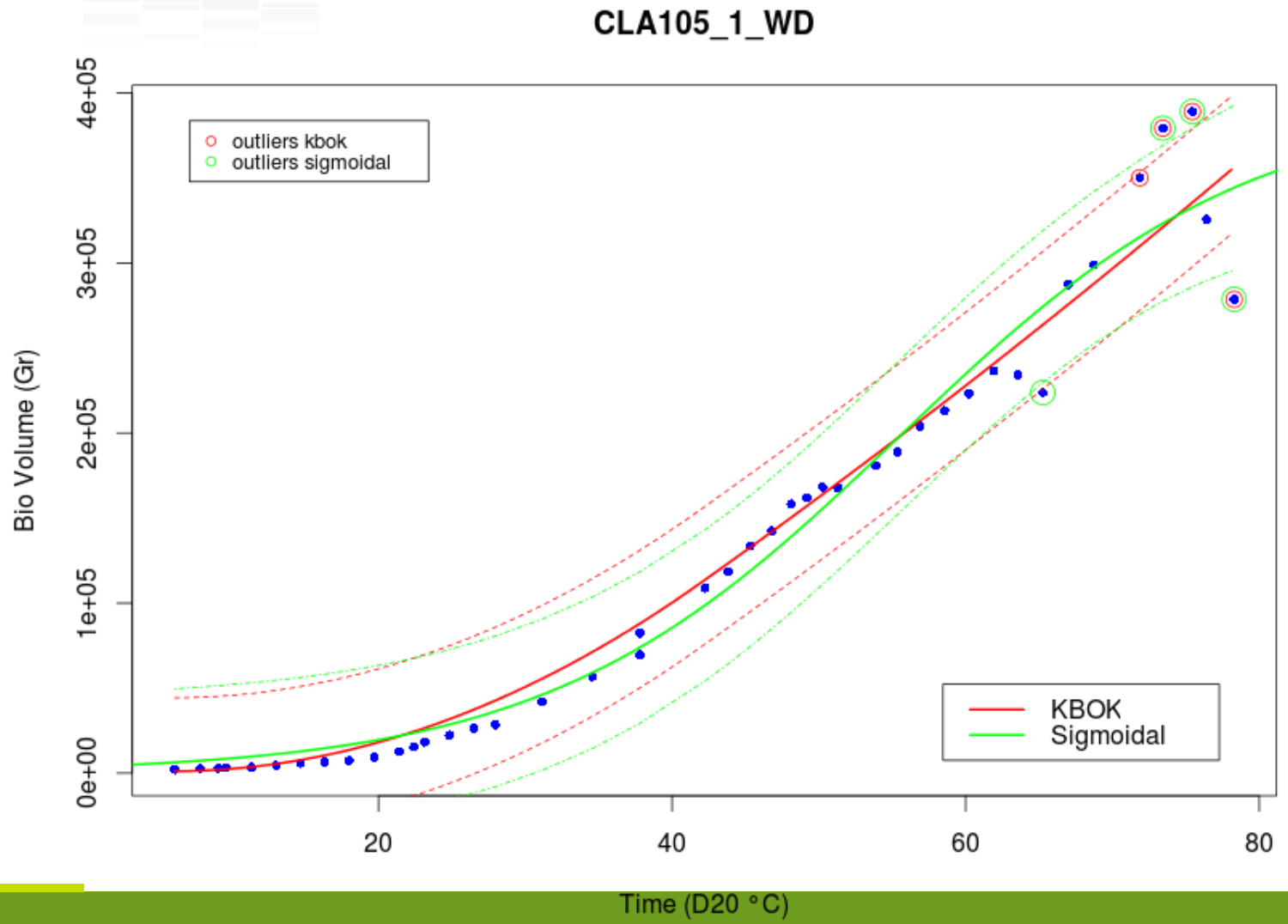
Automatic detection
(R, algorithms within DB)

Manual validation

Cleaning = confidence index
Associated with a person / a method
Reversible (all data kept)



# 1A. Outlier point detection in time courses



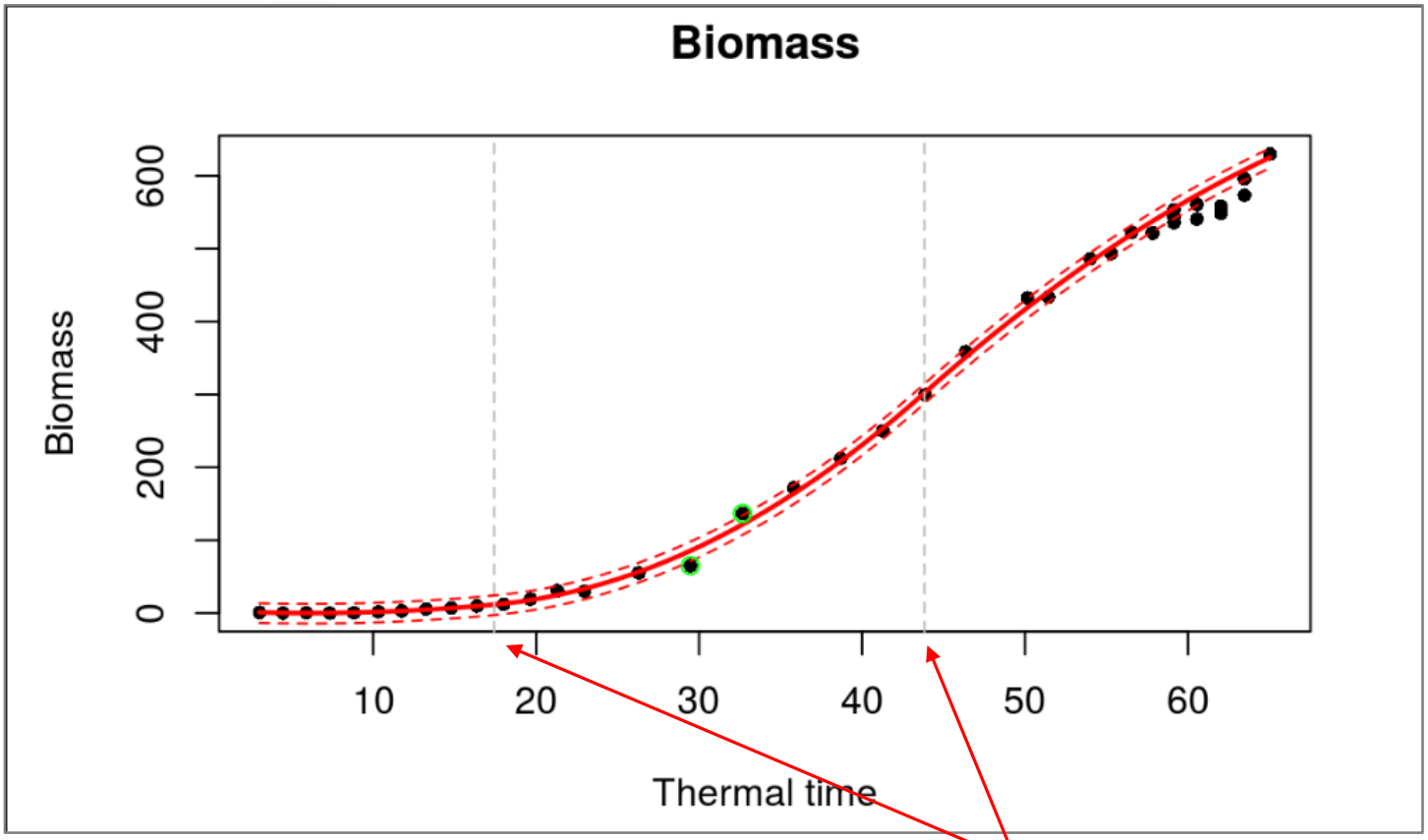
# 1A. Outlier point detection in time courses

**Detection done on a single curve at a time**

**STEP 1**: fitting of a model

- Parametric models : Gompertz, sigmoidal…
- Non parametric models: no a priori model is required
  e.g. local regression *(« locfit » function in R)*
  projection on a spline basis. *A spline function is a piecewise polynomial function. The places where the pieces meet are known as* ***knots.*** *KBOK is a method that optimizes knot positions.*

# Outlier point detection with KBOK




Projection on a spline basis with **Optimal knot positions**

# 1A. Outlier point detection in time courses

**Detection done on a single curve at a time**

**STEP 1**: fitting of a model

- Parametric models : Gompertz, sigmoidal…
- Non parametric models: no a priori model is required
  e.g. local regression *(« locfit » function in R)*
  projection on a spline basis. *A spline function is a piecewise polynomial function. The places where the pieces meet are known as* ***knots.*** *KBOK is a method that optimizes knot positions.*

**STEP 2**: A point is detected as outlier if the corresponded residual is greater than a given level

➔ **No spatial nor temporal effect** is taken into account for the detection

➔ **Genotype repetitions are not considered** as such.

# 1A. Outlier point detection in time courses

## In case of spatial heterogeneity

**Approach:** fitting of a mixed model with spatial effects

(many R packages available: **CARBayesST, SPATS,** …)

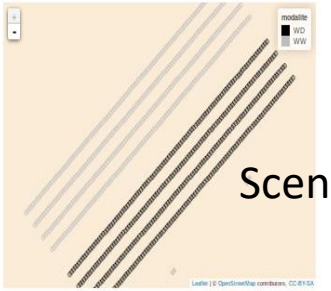
- That analyses all data at once (all genotypes, all scenarios, all dates)
- That takes non-homogeneous environmental variables into account: light or soil water reserve…
- Simple statistical models

➔ **Outlier point detection is made from the residuals**

# Illustration on a field dataset

Soil Water Reserve

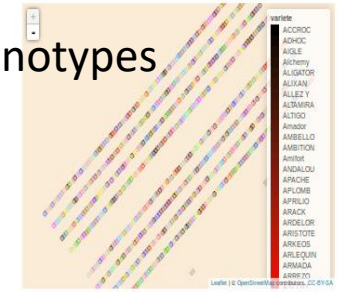
2.1 Experimental design: scenario



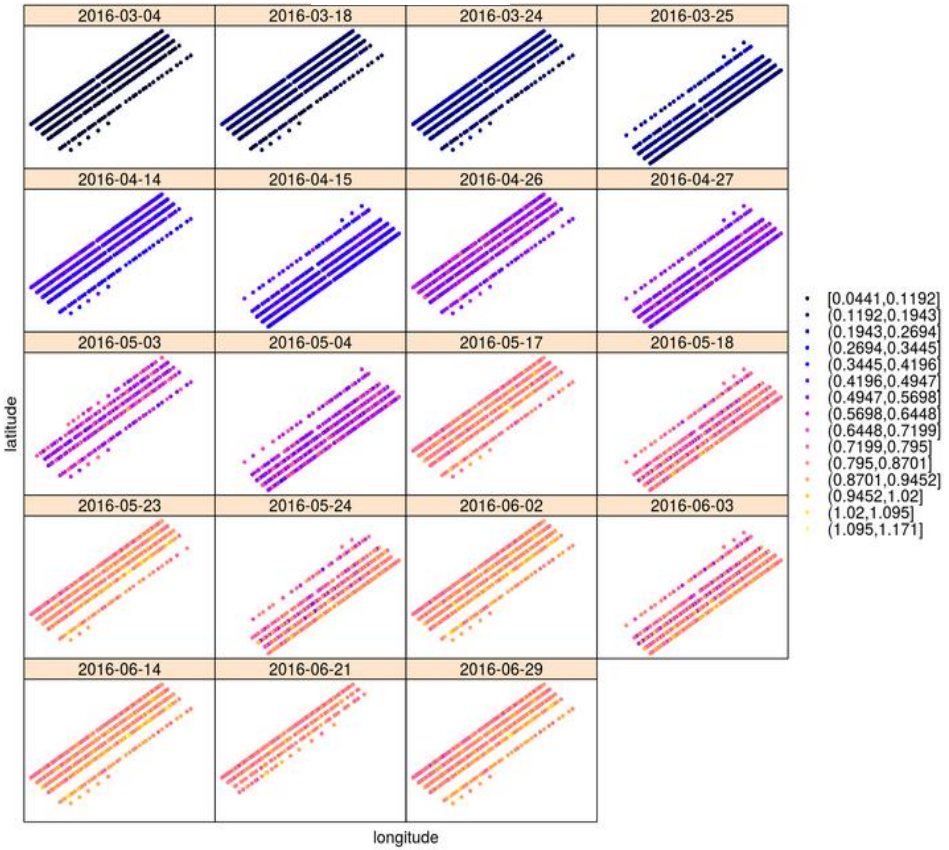

Scenario

2.2 Experimental design: variete

Genotypes

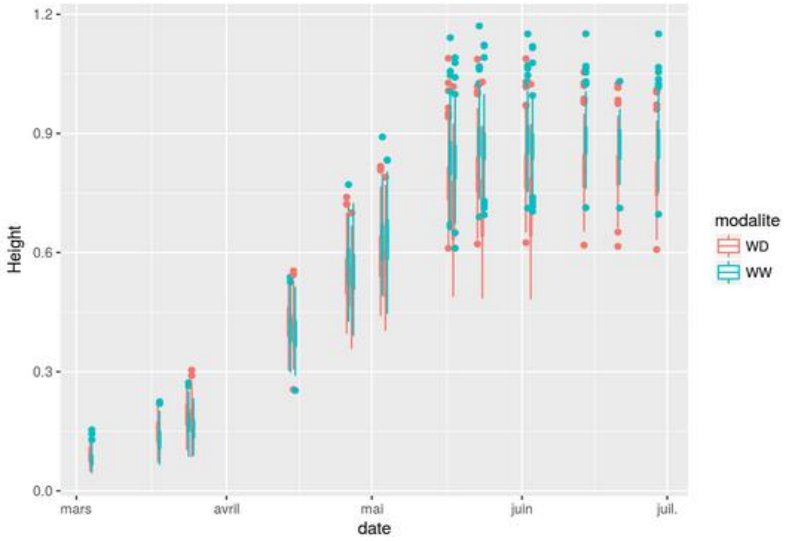


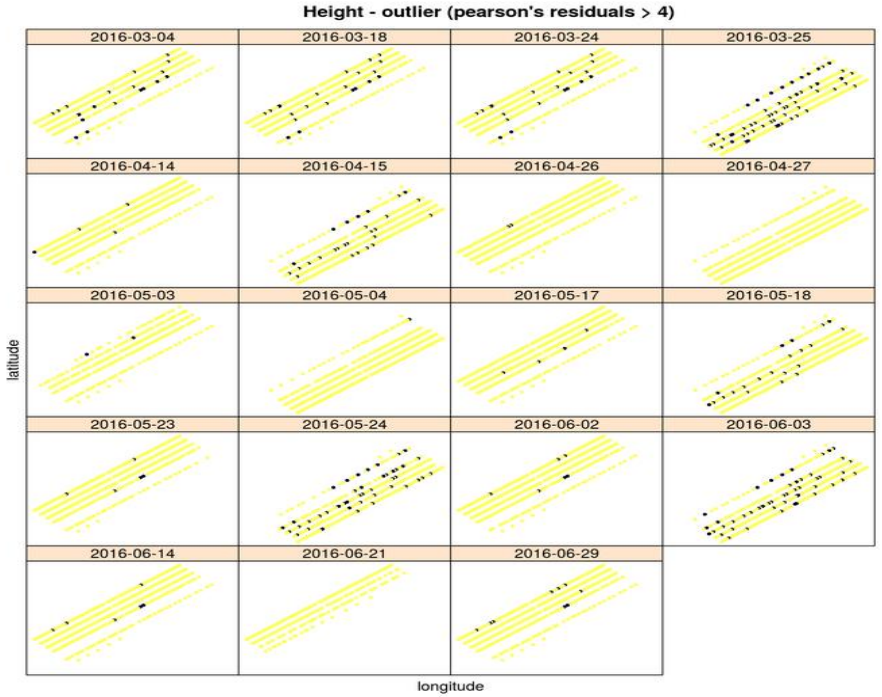

Height



# Illustration on a field dataset

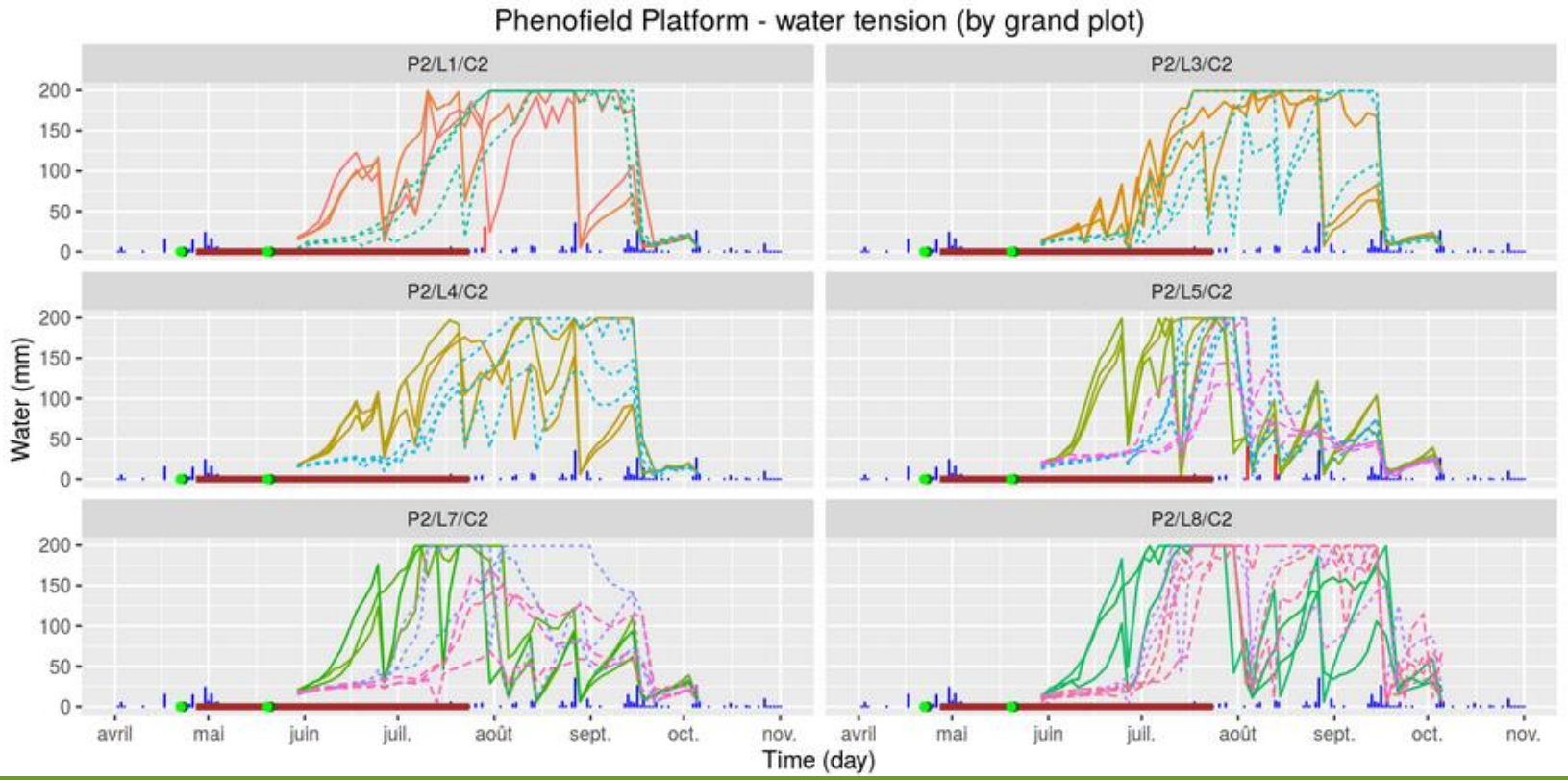
Point outlier detection made with a Bayesian approach which takes the spatio-temporal structure into account: CARBayesST library



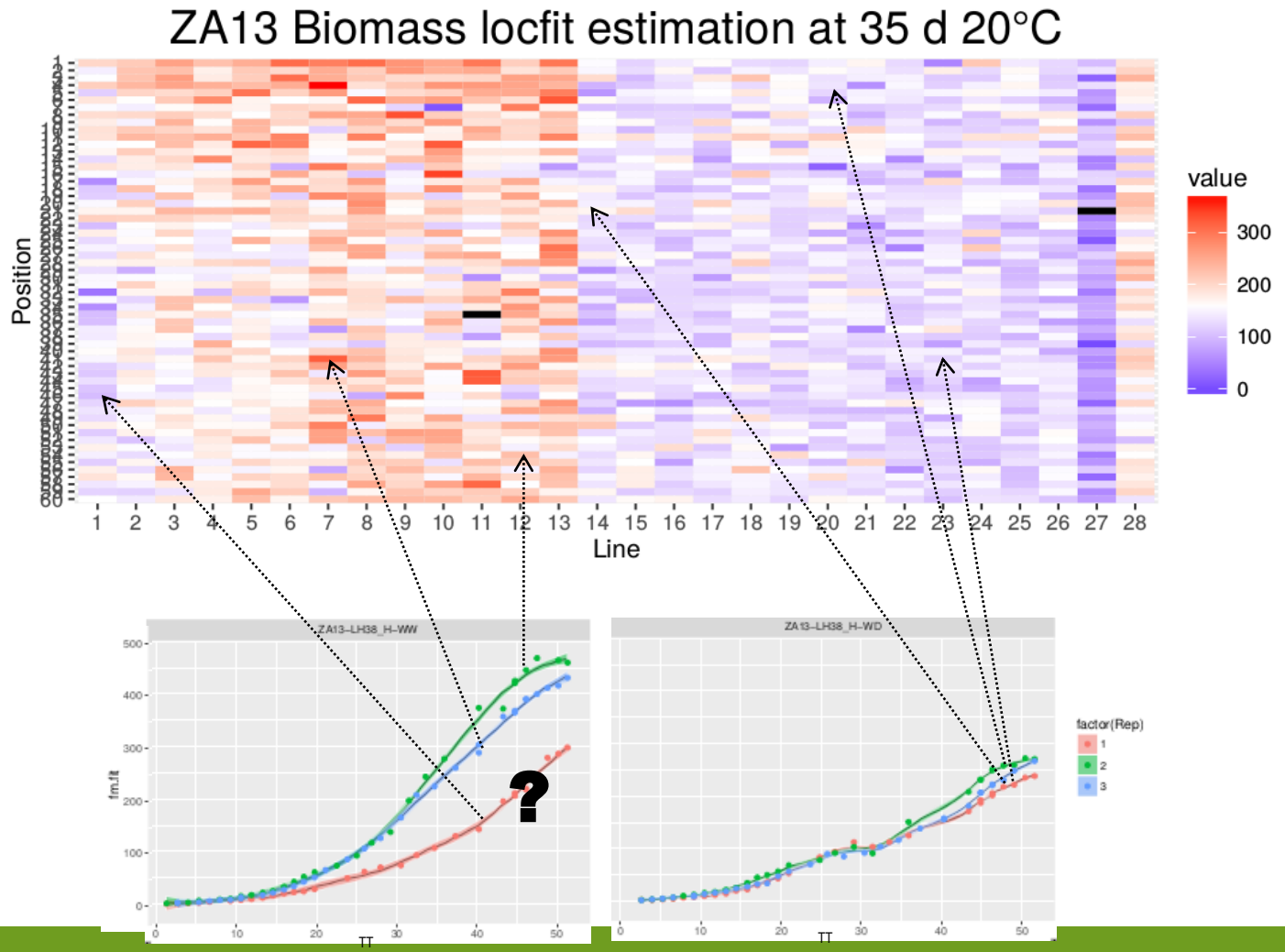



# Monitoring of environmental data

: rain (mm)
: irrigations (mm)
: green house management
: treatments (herbicides, engrais…)



# 1B. Outlier Plant: How to identify atypical plant growth curves?



# 1B. Outlier Plant: How to identify atypical plant growth curves?

After removal of outlier points

## A purely statistical-based cleaning method (SC)

- **Growth curves** of each plant are **fitted with smoothing splines**
- Curves are compared through an ANOVA smoothing spline
  - repetition effect taken into account
  - but **no spatial effect** to model heterogeneity in the platform

➔ Detection is made one genotype after the other

# 1B. Outlier Plant: How to identify atypical plant growth curves?

After removal of outlier points

## A combined cleaning method (CC)

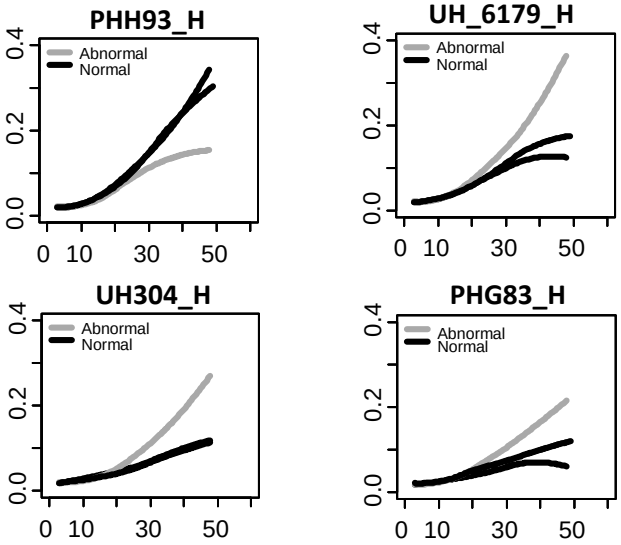
The **decision criteria combined various phenotypes** (PH, biomass… depending on the plant species)

- Measured **at a given time**
- Modelled with **a mixed model** with random genotypic, repetition and spatial effects

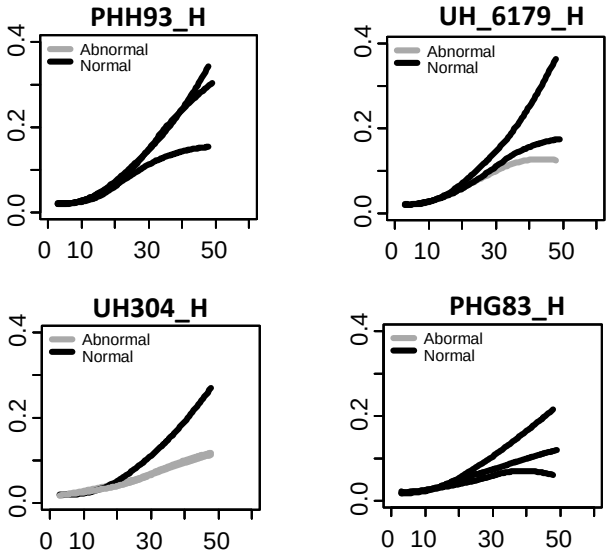
➔ Model residuals for each phenotype are combined for the detection of outlier plants.

# 1B. Outlier Plant detection: How to choose between methods ?

Illustration on a maize experiment (PhenoArch LEPSE)



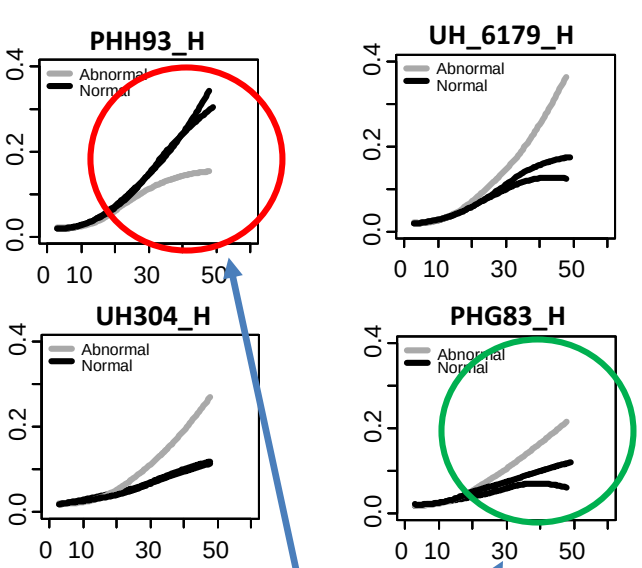

**Combined Cleaning**




**Statistical Cleaning**

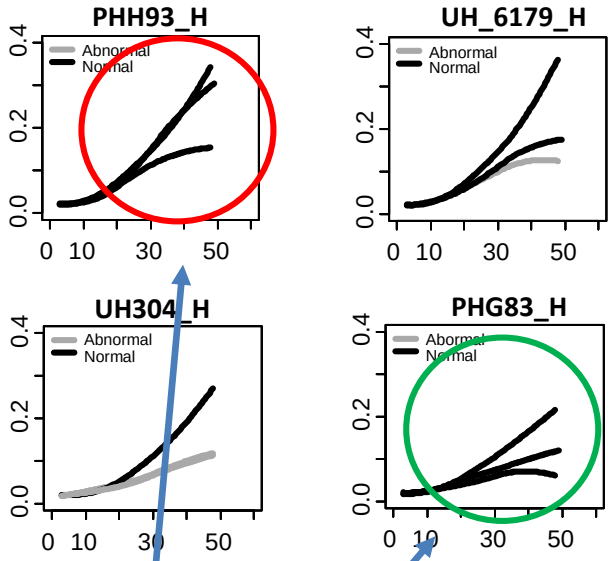
# 1B. Outlier Plant detection: How to choose between methods ?

Illustration on a maize experiment (PhenoArch LEPSE)




**Combined Cleaning**

One repetition detected

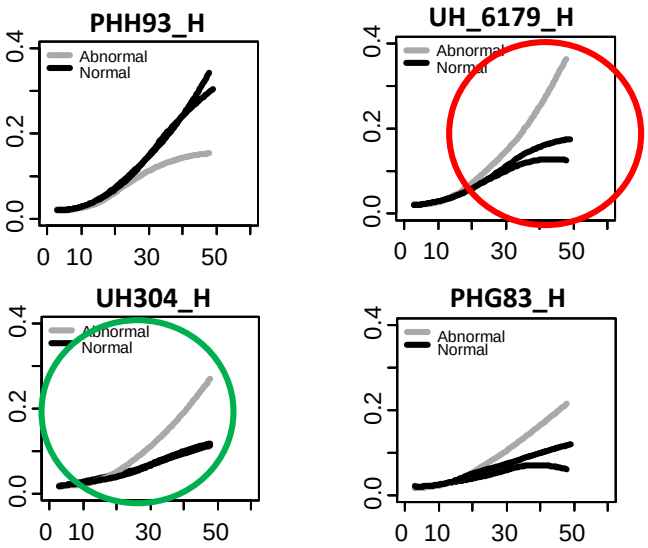



**Statistical Cleaning**

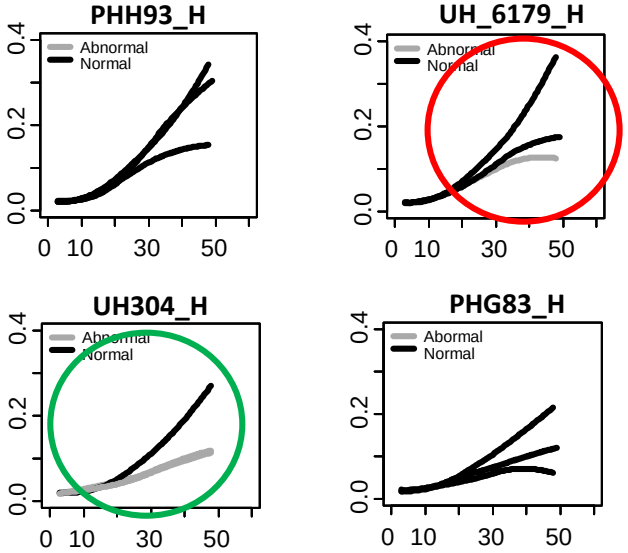
No detection

# 1B. Outlier Plant detection: How to choose between methods ?

Illustration on a maize experiment (PhenoArch LEPSE)




**Combined Cleaning**




**Statistical Cleaning**

Two distinct repetitions detected

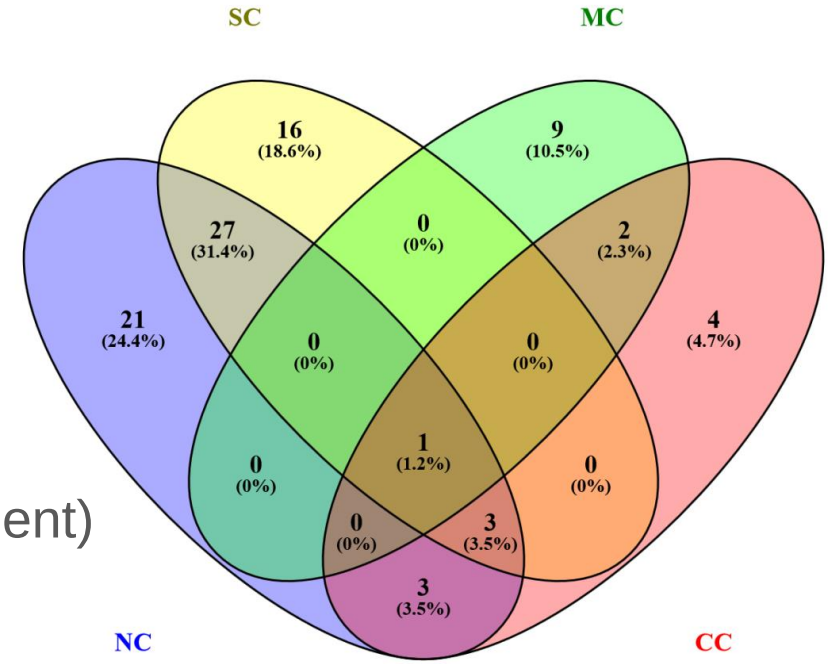
# 1B. Outlier Plant detection: How to choose between methods ?

Joint work with I. Sanchez (MiSTEA),
S. Alvarez Prado, Ll. Cabrera-Bosquet, C. Welker and F. Tardieu (LEPSE)

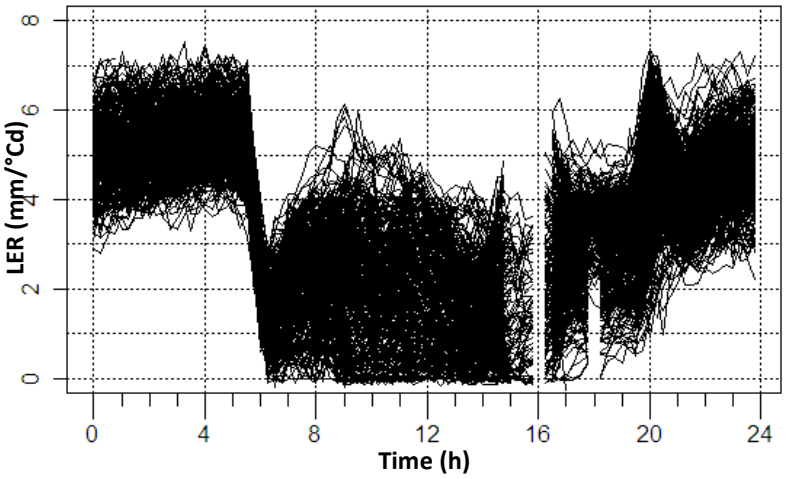
QTL detection on 4 distinct datasets:

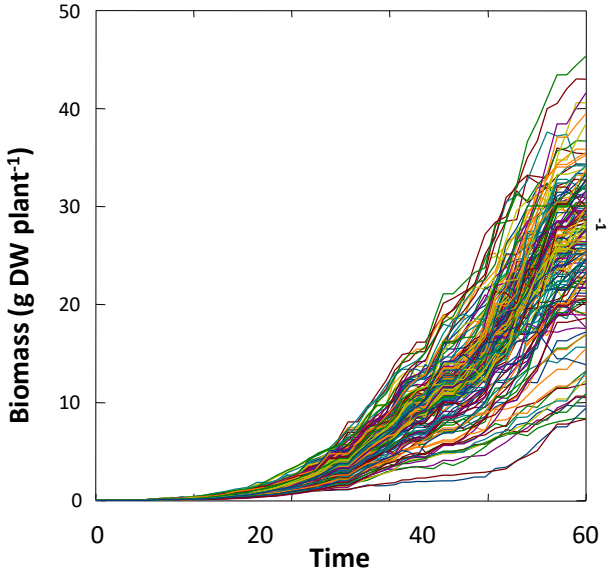
- No Cleaning
- Manual Cleaning
- Statistical Cleaning
- Combined Cleaning

(Data: one experience and one treatment)



# 2/ Extraire et inférer des connaissances à partir de données temporelles




**Deux thématiques complémentaires** :

- **Fouille de données – visualisation** : Identifier automatiquement des régularités dans les données qui soient exploitables pour la prédiction
- **Modèles statistiques pour données fonctionnelles**

# Modèles pour données fonctionnelles

Difficultés : courbes = objets de dimension infinie
(données « fonctionnelles », pas classique)

- Transposer les méthodes statistiques classiques au cadre des données fonctionnelles
  - Problèmes méthodologiques de réduction de dimension
  - Nécessité de pouvoir « automatiser » les méthodes, d'avoir des temps de calcul raisonnables

- Interprétabilité des quantités estimées ?

  Les modèles permettent-ils de comprendre le phénomène ?

# Un modèle « explicable » à sortie scalaire

Thèse de P. Grollemund (P. Pudlo, C. Abraham, M. Baragatti)

**Objectif :** Estimer le paramètre de façon explicable (fonction constante par morceaux)

$$y = \mu + \langle X, \theta \rangle_H + \epsilon = \mu + \int_0^1 X(t)\theta(t)dt + \epsilon$$

# Un modèle « explicable » à sortie scalaire

Thèse de P. Grollemund (2017)

**Objectif :** Estimer le paramètre de façon explicable (fonction constante par morceaux)

➔Régression linéaire **Bayésienne** basée sur la projection du paramètre dans une **base d'histogrammes parcimonieuse et adaptative**

**Intérêts du Bayésien :**

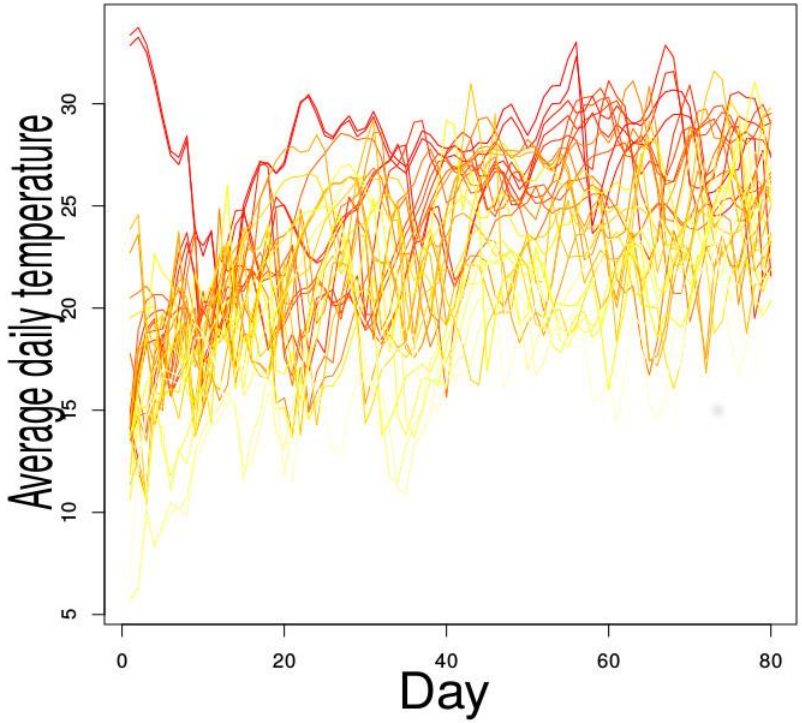
- Inférence plus complète (intervalle de confiance, erreur de prédiction,…)
- Inclut de la connaissance a priori

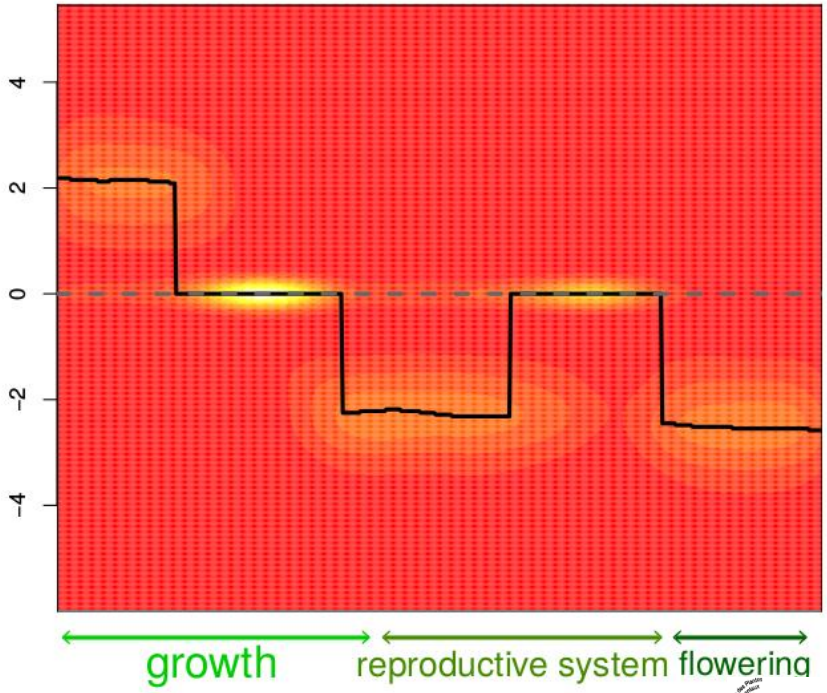
**Principe :** Méthode MCMC pour échantillonner la distribution a posteriori jointe de l'estimateur (échantillonneur de Gibbs)

# Illustration

$y$ : nombre de grains de maïs par plante

$X(t)$ : température pendant la saison




➔ Interprétation facilitée par la structure de l'estimation

Merci de votre attention !