TD « Teneur en protéines »

Régression linéaire pour prédire la teneur en protéines du blé

Contexte

- Teneur en protéines des grains de blé :
  - Critère de qualité important pour les entreprises qui collectent et stockent les récoltes de blé;
  - Détermine le type d'utilisation industrielle d'une récolte (panification, fabrication de biscuits, alimentation animale, etc.)

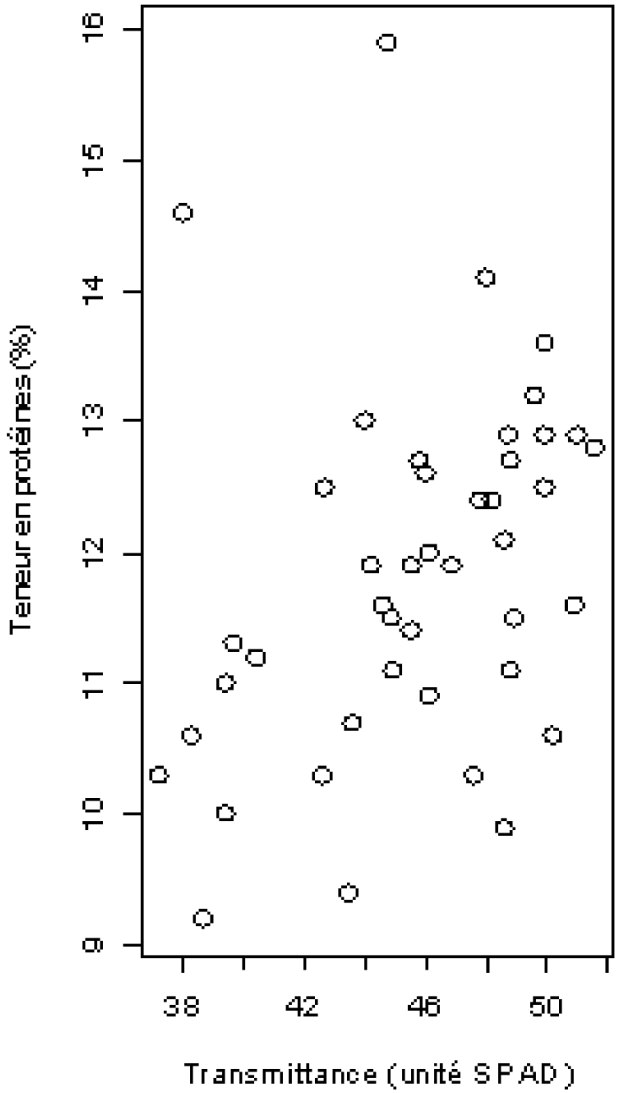
- Important de pouvoir prédire, avant la récolte, la qualité du blé afin d'organiser le stockage des grains en silos et de passer des contrats.

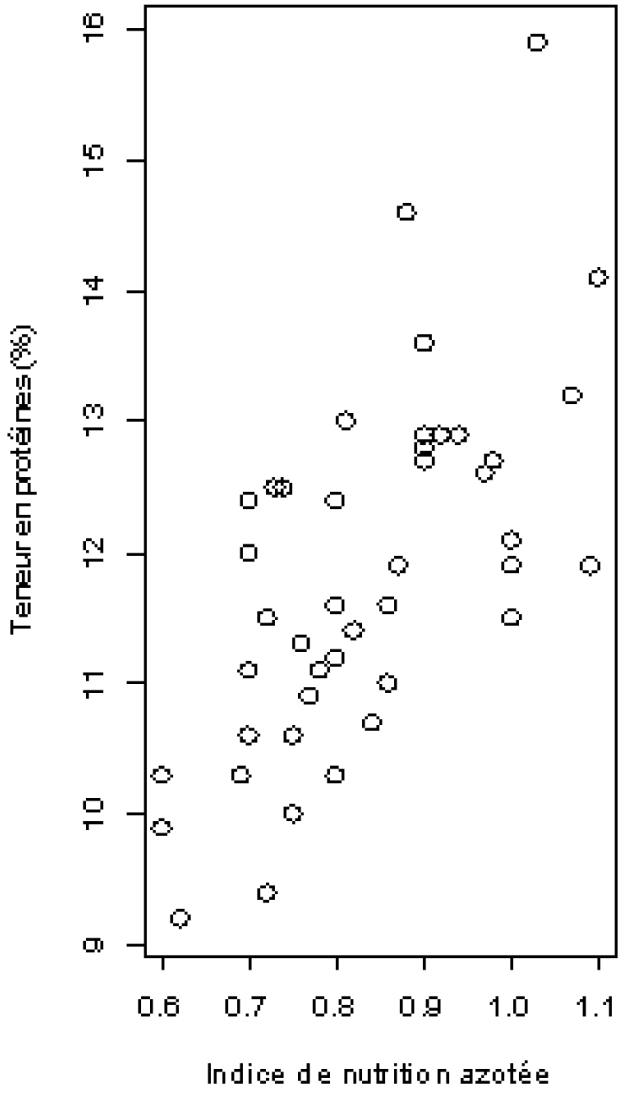
Objectif

- Proposer un modèle linéaire pertinent pour prédire la teneur en protéines du blé
- Utiliser des variables explicatives mesurables quelques semaines avant la récolte dans les parcelles cultivées par les agriculteurs :
  - une mesure de transmittance (SPAD) réalisée sur un échantillon de feuilles de blé avec le chlorophyl meter Minolta ;
  - une mesure de l'indice de nutrition azotée du blé (INN) à floraison.

Données

- Quarante-trois expérimentations ont été réalisées en exploitations agricoles pendant trois ans (2004, 2005, 2006) sur plusieurs sites en France.
- Chaque expérimentation est constituée d'une parcelle d'agriculteur sur laquelle l'INN, le SPAD et la teneur en protéines (%) ont été mesurés.


Teneur en protéines (%)
9
10
11
12
13
14
15
16
38
42
46
50
Transmittance (unité SPAD)


Teneur en protéines (%)
9
10
11
12
13
14
15
16
0.6
0.7
0.8
0.9
1.0
1.1
Indice de nutrition azotée

Questions

1- Ecrire les équations de plusieurs modèles linéaires permettant de prédire la teneur en protéines à partir:

- ✓ d'une variable explicative,
- ✓ de deux variables explicatives,
- ✓ d'aucune variable explicative.

2- Représenter graphiquement ces modèles

3- Quels sont les paramètres à estimer ?

Questions

4- Définir des distributions a priori pour les paramètres

5- Implémenter les modèles sous OpenBUGS ou rjags ou rstan afin d'estimer les paramètres à partir des données disponibles. Vérifier tout d'abord que les programmes fonctionnent en réalisant 5000 itérations MCMC

Questions

6- Pour chaque modèle : lancer trois chaînes de Markov. Utiliser le diagnostic de convergence de Gelman et Rubin et des représentations graphiques des chaînes pour déterminer le nombre d'itérations nécessaire pour espérer avoir atteint la convergence.

7- Analyser les auto-corrélations des chaînes. Faites de nouvelles itérations MCMC afin d'obtenir un échantillon de valeurs suffisamment représentatif de la loi a posteriori.

Questions

8- Représenter les densités a posteriori des paramètres et résumer ces distributions par leurs moyennes, médianes, et quelques percentiles.

9- Calculer le DIC des différents modèles proposés et identifier le modèle qui a le DIC le plus faible

10- Calculer le WAIC des différents modèles proposés et identifier le modèle qui a le WAIC le plus faible

Questions

11 - Utiliser le modèle sélectionné pour déterminer la distribution a posteriori de la teneur en protéines d'une nouvelle parcelle caractérisée par SPAD=50 et/ou INN=0.95