

Module « Modélisation statistique »
Examen très redouté
2009

Tous les documents et la calculatrice sont autorisés. Merci de ne pas utiliser vos téléphones portables, même pour son application calculatrice. Merci de ne rien échanger avec vos voisins pendant la durée de l'épreuve.

Devant surveiller un autre examen, je ne suis pas présent pour vous observer, vous plongeant dans cet exercice que j'ai voulu teinté d'exotisme afin que vous puissiez travailler dans les conditions les plus agréables. Un seul mot d'ordre: pas de panique! Bonne réflexion.

1. Evaluation de la teneur en Carbone Organique du sol à partir de spectrométrie proche infra-rouge et d'images de télédétection rapprochée

L'évaluation de la teneur en Carbone Organique (CO) d'un sol est un enjeu important en matière de quantification des niveaux de pollution. La méthode la plus fiable pour mesurer cette teneur en Carbone Organique repose sur une analyse physico-chimique coûteuse, qui n'est pas envisageable pour une évaluation rapide du niveau de pollution d'une surface importante. D'autres méthodes sont donc envisagées dont la spectrométrie proche infra-rouge et l'imagerie par télédétection rapprochée à l'aide d'un capteur ASTER.

Afin d'évaluer les niveaux de précision de ces méthodes, on procède à une expérience sur une parcelle située dans le sud tunisien, sur un chott (un chott est un désert salé, propice aux mirages - c'est maintenant le moment exotique, profitez-en, il est court). La figure 1 donne une visualisation aérienne de la zone d'étude. Sur cette zone, on définit une grille régulière de 144 points de mesure (voir figure 2) sur lesquels on prélève un échantillon de sol pour une analyse déterminant sa teneur en CO (g/kg). Chaque échantillon de sol est également soumis à spectrométrie proche infra-rouge (on dispose d'une mesure spectrale par nm entre 400 et 2500 nm). Enfin, en chaque site de mesure, on extrait 9 valeurs spectrales d'une image par télédétection rapprochée avec un capteur ASTER, correspondant aux longueurs d'onde 556, 661, 807, 1656, 2167, 2209, 2262, 2336 et 2400 nm.

La figure 3 donne un histogramme de la teneur en CO sur la parcelle d'étude.

1. Proposez un modèle statistique permettant d'étudier la relation entre la teneur en CO et l'analyse par spectrométrie proche infra-rouge d'un échantillon de sol.
2. A partir de quelle taille d'échantillon peut-on, au moins numériquement, estimer les paramètres d'un tel modèle avec la méthode usuelle (celle implémentée dans la fonction `lm` de R)? Comment s'appelle cette méthode d'estimation?
3. Proposez une méthode permettant d'estimer les paramètres du modèle précédent. Comment expliqueriez-vous cette méthode à un non-spécialiste?

Une des méthodes d'estimation possibles, très utilisée par les chimiométriciens, s'appuie sur l'extraction d'un certain nombre q de variables latentes à partir du spectre.

4. Si L_1, L_2, \dots, L_q désignent les variables latentes, comment celles-ci sont-elles calculées à partir d'un spectre proche infra-rouge? Quel modèle permet de relier la teneur en CO aux variables latentes?
5. Donnez un critère permettant de comparer entre eux les modèles obtenus avec différents nombres de variables latentes? Comment choisir q à partir de ce critère.

Le critère que l'on choisit dans la suite a l'avantage de mesurer la qualité d'ajustement d'un modèle dans la même unité que la variable à prédire (g/kg). On estime ce critère par une procédure de validation croisée.

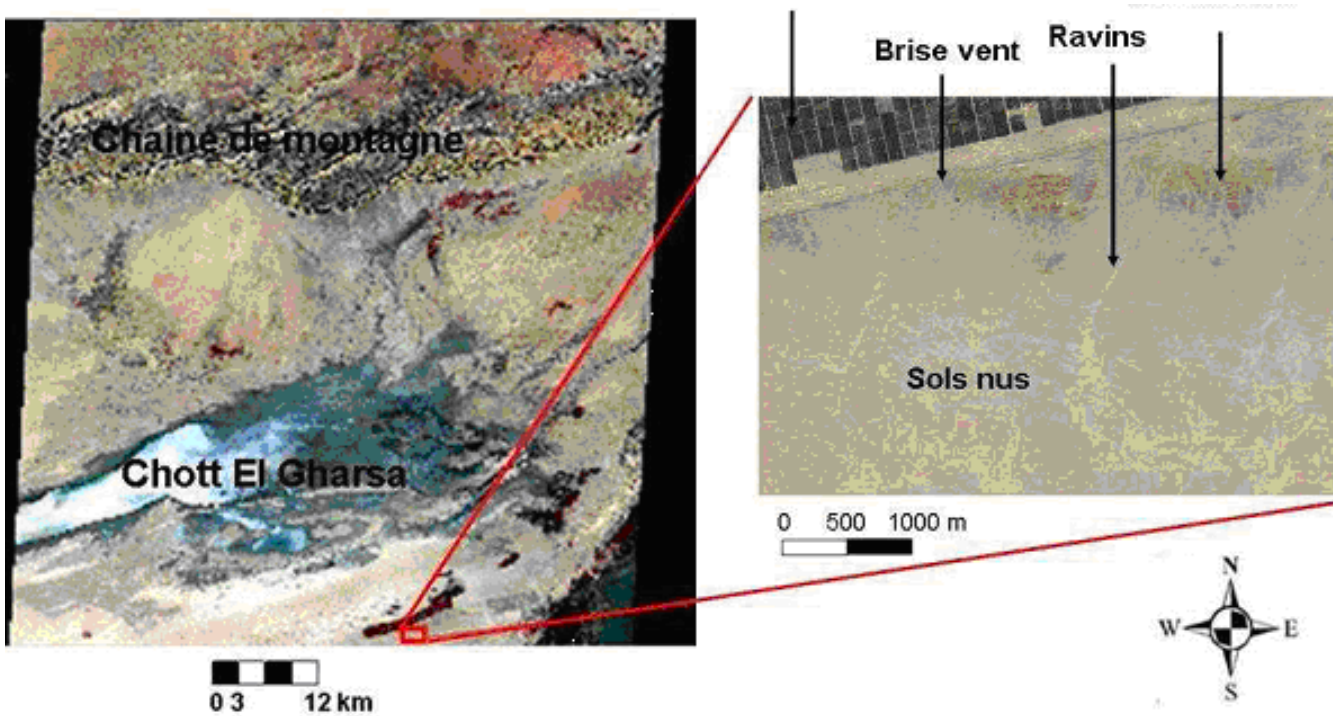


FIG. 1: Scène de l'image ASTER couvrant la partie septentrionale de la région du Djérid, visualisée en fausses couleurs RVB 321 avec une résolution de 60X60 km. Image prise le 27-10-2006. Le rectangle rouge délimite le site d'étude expérimental, son aperçu est affiché à l'aide de l'outil Google map de Google Earth.

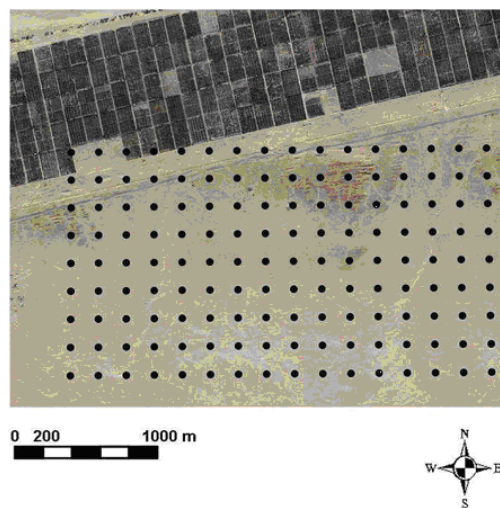


FIG. 2: Schéma d'échantillonnage systématique aux noeuds d'une grille carrée régulière de 200 m de côté. Points échantillonnés, visualisés sur un fond d'image satellite extraite à l'aide de l'outil Google maps de Google Earth.

6. Pourquoi ce recours à la validation croisée ? Pour un nombre fixé de variables latentes, décrivez point par point la procédure de calcul du critère selon une méthode de validation croisée reposant sur un découpage de l'ensemble des données en 3 segments.

La figure 4 montre l'évolution du critère en fonction du nombre de variables latentes. A partir de ce graphique, on choisit de retenir 7 variables latentes dans le modèle.

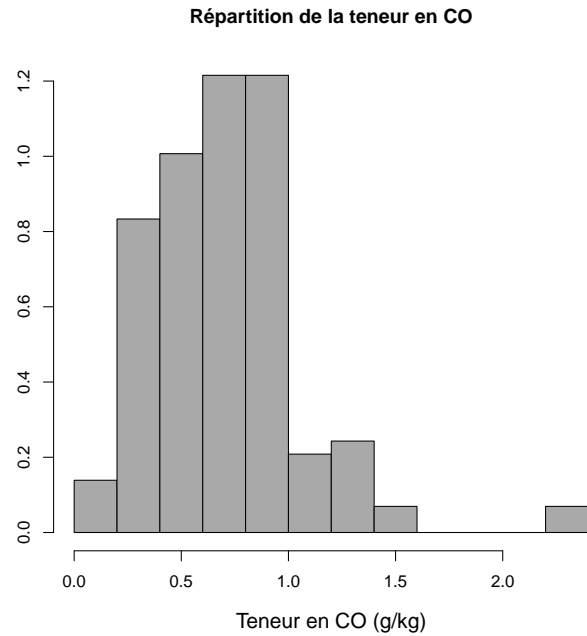


FIG. 3: Répartition de la teneur en CO (g/kg).

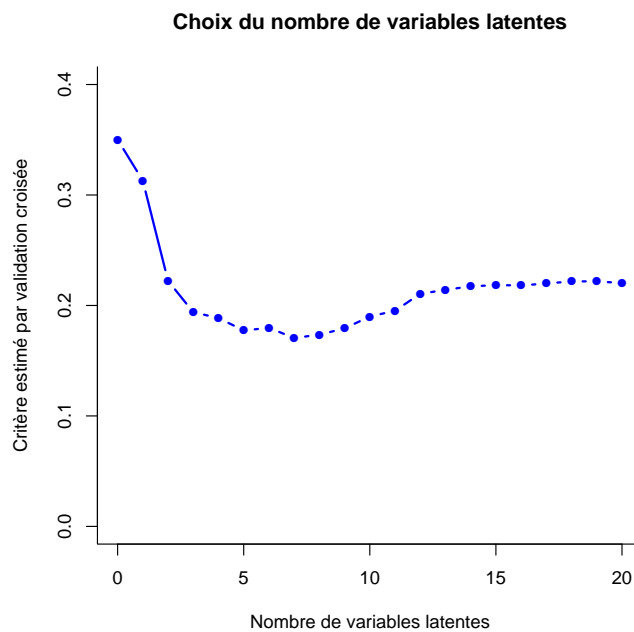


FIG. 4: Choix du nombre de variables latentes pour le modèle de prédiction de la teneur en CO à partir de la spectrométrie proche infra-rouge.

Pour donner un sens aux deux premières variables latentes construites par le modèle, on s'appuie sur un graphique donnant les coordonnées de ces variables (*loadings*) en fonction de la longueur d'onde. La figure 5 donne ce graphique, ainsi qu'un découpage des spectres en 5 zones, suggéré par les ruptures de courbes.

7. Que représentent les coordonnées des variables latentes? Que déduit-on du graphique de la figure 5?

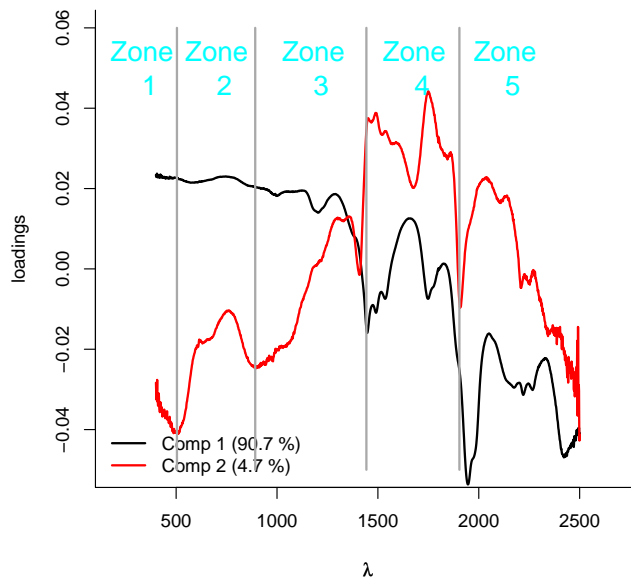


FIG. 5: *Coordonnées des deux premières variables latentes.*

On complète l'interprétation des variables latentes par le cercle des corrélations donné dans la figure 6.

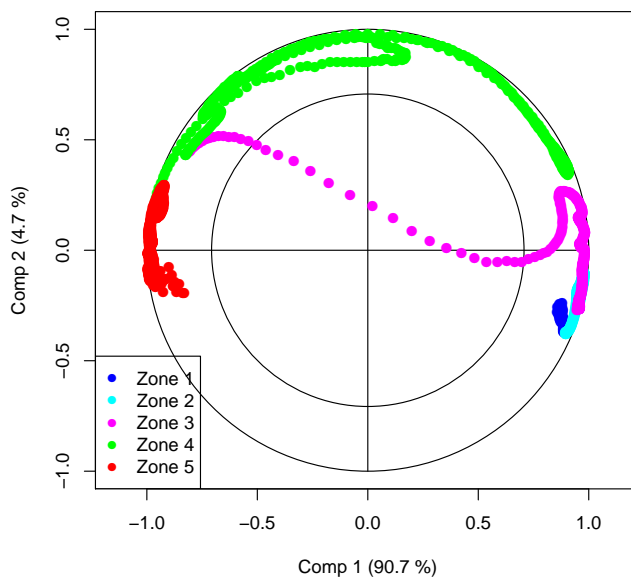


FIG. 6: *Cercle des corrélations des deux premières variables latentes.*

8. *Que représentent les coordonnées des points dans cette représentation? Que déduit-on du graphique de la figure 6, concernant l'interprétation des variables latentes?*

Le graphique de la figure 7 permet de visualiser le lien entre la teneur en Carbone Organique et la 1ère variable

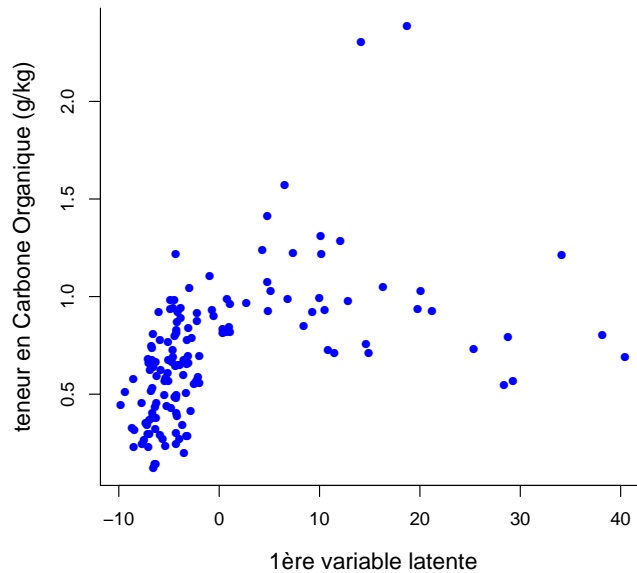


FIG. 7: Teneur en Carbone Organique selon la valeur de la 1ère variable latente.

latente.

9. Comment obtient-on les abscisses du graphique de la figure 7 lorsque l'on dispose des coordonnées des variables latentes ?

10. En quoi le graphique de la figure 7 remet-il en question le modèle proposé initialement ?

11. Quelle autre méthode d'estimation peut permettre de corriger le problème révélé par le graphique de la figure 7 ? Décrire cette méthode en insistant sur les différences et les points communs avec l'approche mise en œuvre ci-dessus.

Dans cette nouvelle méthode d'estimation, un spectre est vu comme une fonction régulière de la longueur d'onde. Les graphiques de la figure 8 donnent une représentation de chacun des spectres ainsi que du spectre moyen.

Dans un premier temps, on cherche à approcher chaque spectre par une fonction spline de degré 3.

12. Quand dit-on qu'une fonction est une spline de degré 3 ? Donnez les principes d'ajustement d'une telle fonction à un spectre observé.

Le graphique de la figure 9 montre l'évolution de la qualité d'ajustement d'une spline de degré 3 au spectre moyen en fonction du nombre de nœuds dans la partition de la plage des longueurs d'onde. On choisit de retenir un ajustement à partir d'une partition en 30 nœuds.

13. Comment est calculé le critère de qualité d'ajustement, en ordonnée du graphique de la figure 9 ?

Le choix d'une partition en trente nœuds conduit à réduire chaque spectre à 32 coefficients.

14. En quoi ces coefficients portent-ils (presque) la même information qu'un spectre complet ?

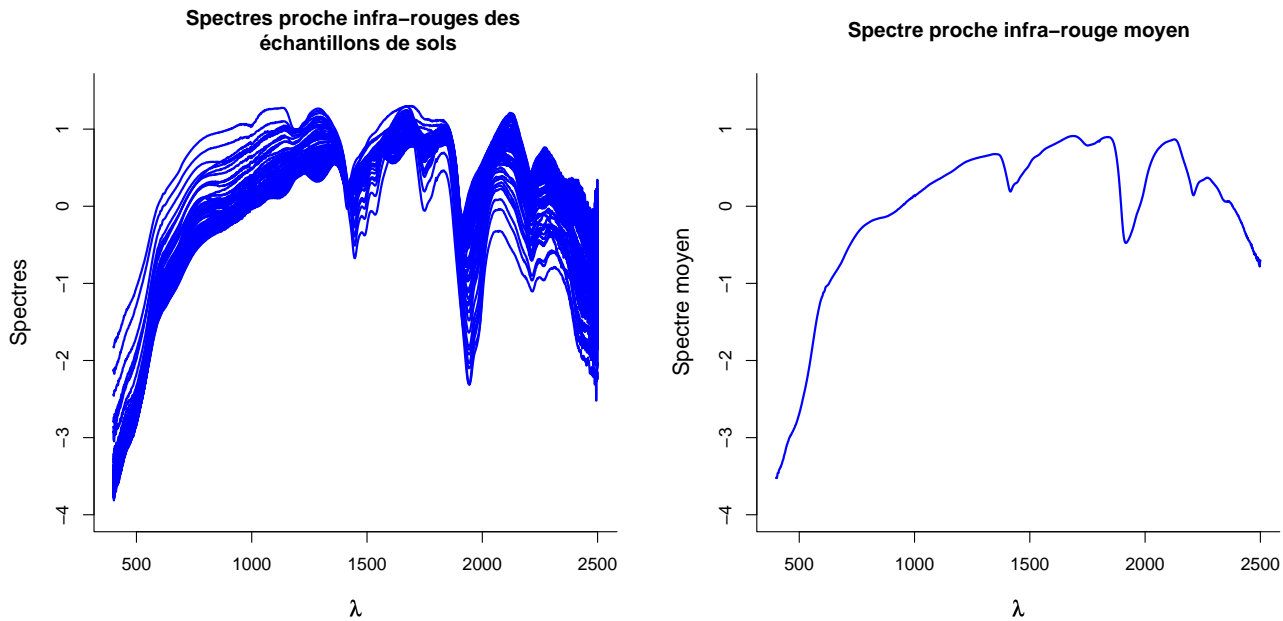


FIG. 8: Spectres proche infra rouges (graphique de gauche) et spectre moyen (graphique de droite).

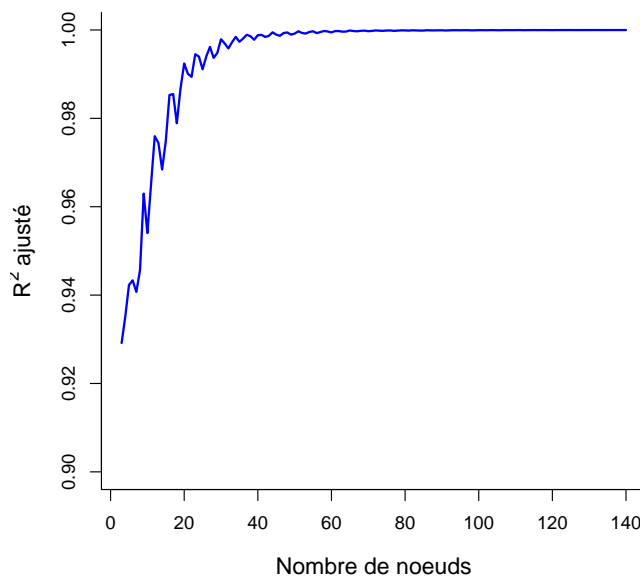


FIG. 9: Détermination du nombre de nœuds dans la partition de la plage des longueurs d'onde pour l'ajustement d'une spline de degré 3 au spectre moyen.

15. Quel modèle permet de relier la teneur en carbone à ces coefficients ?

Comme le nombre de coefficients est élevé, on suggère d'ajuster le modèle proposé à la question précédente par une méthode LASSO, consistant à estimer les coefficients de régression sous contrôle de la somme des valeurs absolues de ces coefficients.

16. En quoi cette méthode diffère-t-elle de la méthode d'ajustement classique d'un modèle de régression ?

Le graphique de la figure 10 donne la valeur des coefficients estimés du modèle en fonction du niveau de rétrécissement de ces coefficients.

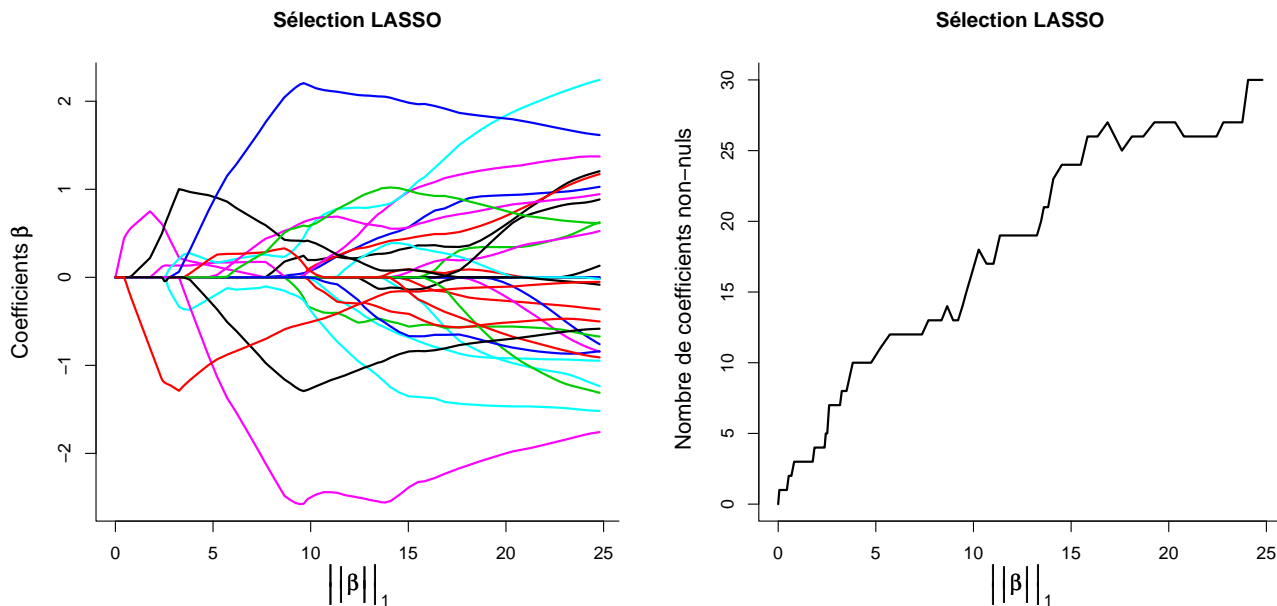


FIG. 10: Coefficients du modèle en fonction du niveau de rétrécissement (graphique de gauche) et nombre de coefficients non-nuls (graphique de droite).

17. Quelle méthode suggérez-vous pour choisir le bon niveau de rétrécissement ?

Le graphique de la figure 11 montre l'évolution d'un critère de la qualité d'ajustement du modèle en fonction du niveau de rétrécissement. Ce critère est calculé de manière similaire à celui utilisé dans le graphique de la figure 4.

18. Finalement, combien de coefficients garde-t-on dans le modèle ? La nouvelle méthode d'ajustement a-t-elle apporté des améliorations ?

Dans la dernière partie de l'exercice, on cherche à voir s'il est possible d'atteindre ce niveau de précision dans la modélisation de la teneur en CO (g/kg) à partir des 9 valeurs spectrales issues du capteur ASTER. A titre d'illustration, le graphique de la figure 12 décrit la relation entre la teneur en CO et la valeur spectrale à 2209 nm.

19. Si les relations marginales entre la teneur en CO et chacune des valeurs spectrales ASTER sont à l'image du graphique de la figure 12, le modèle de régression linéaire vous semble-t-il adapté ? Si non, pourquoi ? Quel modèle suggérez-vous ?

Les commandes R figurant dans le tableau 1 permettent d'ajuster un tel modèle.

20. Que signifie $s(\text{ASTER556}, 4)$ dans les commandes du tableau 1 ? Que réalisent ces commandes ?

Le graphique de la figure 13 représentent les effets marginaux présents dans le modèle retenu.

21. Déduire de la figure 13 une expression du modèle retenu.

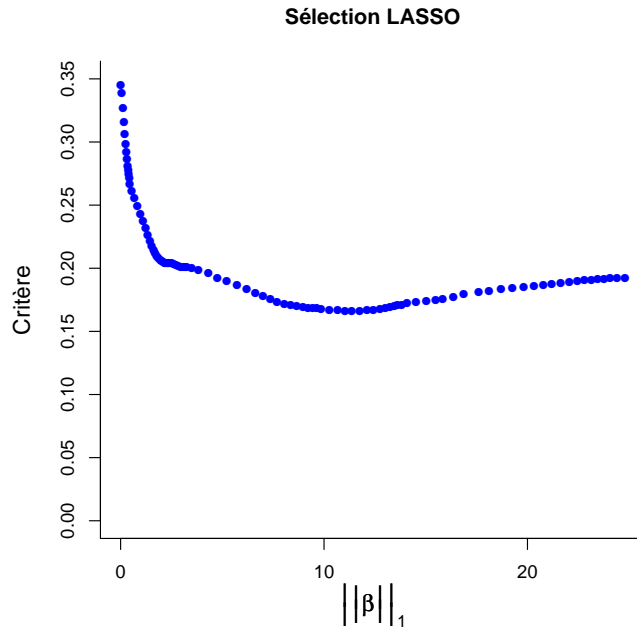


FIG. 11: Détermination du niveau de rétrécissement optimal pour la méthode LASSO.

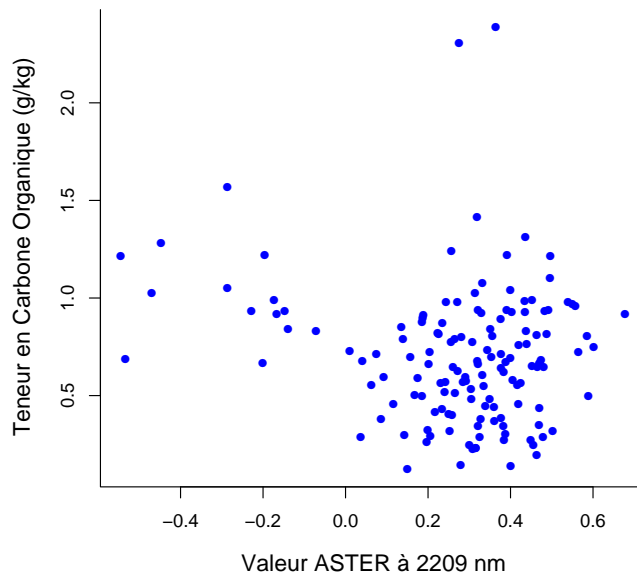


FIG. 12: Lien entre teneur en CO et valeur spectrale ASTER à 2209 nm.

22. Selon vous, le capteur ASTER permet-il une modélisation intéressante de la teneur en CO? Justifiez votre réponse.

```

> carbone.gam = gam(y~1, data=carbone)
> carbone.select = step.gam(carbone.gam,scope=list("ASTER556"=~1+ASTER556+s(ASTER556,4),
"ASTER661"=~1+ASTER661+s(ASTER661,4),"ASTER807"=~1+ASTER807+s(ASTER807,4),
"ASTER1656"=~1+ASTER1656+s(ASTER1656,4),"ASTER2167"=~1+ASTER2167+s(ASTER2167,4),
"ASTER2209"=~1+ASTER2209+s(ASTER2209,4),"ASTER2262"=~1+ASTER2262+s(ASTER2262,4),
"ASTER2336"=~1+ASTER2336+s(ASTER2336,4),"ASTER2400"=~1+ASTER2400+s(ASTER2400,4)))
> sqrt(cv.glm(glmfit=carbone.select,data=carbone,K=3)$delta[2])
0.2853012

```

TAB. 1: Ajustement d'un modèle de la teneur en CO.

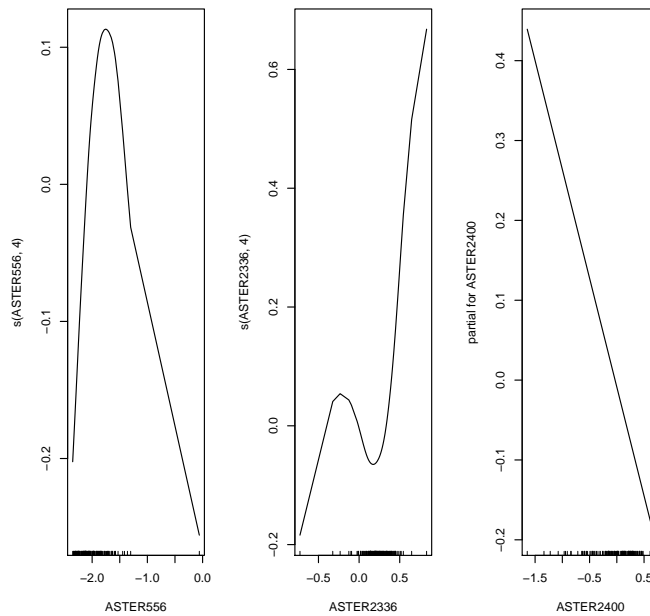


FIG. 13: Effets marginaux des valeurs ASTER sur la teneur en CO.