Etude de l'impact des covariables sur la qualité de prédiction des modèles d'interactions génotype x environnement

V.Choisi and **C. Pontet**

Terres Inovia
Centre de recherche INRA de Toulouse, 24 Chemin de Borde Rouge, 31326 Castanet-Tolosan
v.choisi@terresinovia.fr

Mots clefs : interactions génotype x environnement, covariables environnementales, modèles de prédiction, tournesol, PLS, Random Forest, Lasso

L'évolution des conditions climatiques et la diversification des pratiques culturales génèrent des conditions de culture de plus en plus hétérogènes, avec une plus grande expression des stress environnementaux. Dans ce contexte, il devient de plus en plus important de comprendre la variabilité du comportement des variétés face à ces stress afin de pouvoir mieux adapter le choix variétal au milieu de culture. Pour ce faire, il est nécessaire d'être capable de comprendre et de prédire les interactions entre génotype et milieu (IGE) dans les réseaux d'essais variétaux.
Des travaux sur l'analyse des IGE [1] ont été réalisés en lien avec l'INRA et les autres instituts techniques(2). Différentes méthodes statistiques telles que la régression PLS [2] ou les forêts aléatoires [3] ont permis jusqu'à maintenant d'expliquer une part significative des IGE. Cependant, une évaluation des performances de prédiction de ces modèles montre au travers d'indicateurs divers qu'elles ne sont pas satisfaisantes [4]. Or, pour que ces modèles soient utiles au préconisateur, il est nécessaire de pouvoir prédire le comportement des variétés face aux stress dans des environnements non expérimentés. On se propose donc dans cette étude d'évaluer la qualité prédictive des modèles statistiques d'IGE en s'intéressant particulièrement aux covariables utilisées pour mesurer les niveaux des stress dans les environnements.

Les données exploitées dans cette étude sont issues du réseau d'essais variété tournesol 2015. Il est composé d'une dizaine de lieux, sur lesquels sont cultivées une dizaine de variétés, réparties de manière aléatoire en 3 répétitions. Pour chaque variété sur un lieu et une répétition, on mesure le rendement, qui correspond au poids de graines de tournesol par surface récoltée.
Dans un premier temps, le modèle mixte suivant est réalisé:

$$(1)\ Y_{ijk} = \mu + Var_i + Env_j + Var{:}Env_{ij} + Rep_k(Env_j) + \varepsilon_{ijk}$$

Où Y_{ijk} est le rendement de la variété i sur l'environnement j et la répétition k, Var_i est l'effet de la variété i, Env_j est l'effet de l'environnement j, $Var{:}Env_{ij}$ est l'effet de l'interaction entre la variété i et l'environnement j, $Rep_k(Env_j)$ est l'effet de la répétition k sur l'environnement j, et ε_{ijk} l'erreur du modèle.

Les termes d'interactions sont ensuite extraits de ce modèle pour être modélisés. Afin d'expliquer ces IGE, chaque environnement est caractérisé à l'aide de covariables quantifiant les stress qui se sont appliqués sur chacun d'eux (déficit ou excès d'eau, de températures, de rayonnement, d'azote …).

On compare dans cette étude trois types de calcul de covariables :

1) Calculées sur les différents stades sensibles du cycle du tournesol : levée, végétation, floraison, remplissage
    a. Les dates clé du cycle sont estimées à partir des sommes de températures nécessaires pour atteindre chaque stade. Les covariables calculées rendent compte surtout des stress climatiques
    b. Le modèle de culture SUNFLO [5] est utilisé afin de simuler l'évolution de la culture au cours de son cycle. Les dates clés sont estimées par le modèle de culture, qui donne également accès à des stress importants comme le stress azoté
2) Calculées de manière automatique, sur des périodes indépendantes des stades, par décades ou par mois.

Le terme « environnement » du modèle (1) est alors décomposé en somme des covariables, aboutissant alors au modèle suivant :

$$IGE = \mu + variété + \sum covariables + variété{:} \sum covariables + \varepsilon$$

La qualité prédictive de ce modèle sera évaluée grâce à l'analyse des erreurs de prédictions (procédure de validation croisée), résumée à l'aide de différents critères tels que le coefficient de Spearman ou la RMSEP (Root Mean Square Error of Prediction). Nous présenterons alors les résultats de cette évaluation en fonction des trois listes de covariables étudiées. Plusieurs méthodes statistiques seront également utilisées : régression PLS, régression Lasso [6], régression Ridge [7] et forêts aléatoires.

(2)Dans la cadre d'une action soutenue par le GIS GC HP2E (Groupement d'intérêt scientifique Grandes Cultures à Hautes Performances Economiques et Environnementales)

Références

[1]Debaeke, P., Casadebaig, P., Mestries, E., Palleau, J. P., Salvi, F., Bertoux, V., & Uyttewaal, V. (2011). Evaluer et valoriser les interactions variété-milieu-conduite en tournesol. Innovations Agronomiques, 14, pp. 77-90.

[2]Lazraq, A. & Cléroux, R. (2001). The PLS multivariate regression model: testing the significance of successive PLS components. Journal of chemometrics, 15(6), pp. 523-536.

[3]Genuer, R. (2010). Forêts aléatoires : aspects théoriques, sélection de variables et applications. (Doctoral dissertation, Université Paris Sud-Paris XI).

[4]D'orchymont, M. (2015). Rapport de stage de fin d'études de 3ème année, ENSAI.

[5]Casadebaig, P., Guilioni, L., Lecoeur, J., Christophe, A., Champolivier, L., & Debaeke, P. (2011). SUNFLO, a model to simulate genotype-specific performance of the sunflower crop in contrasting environments. Agricultural and forest meteorology, 151(2), pp. 163-178.

[6]Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pp. 267-288.

[7]Marquardt, D., Snee, R. (1975). Ridge regression in practice. The American Statistician. 29(1), pp. 3-20.