Règlement du « Crop Data Challenge 2018 - Prédiction des rendements du maïs en France »

Objectif

Le rendement agricole annuel d'une culture représente la quantité de produits récoltée par unité de surface une année donnée. Dans le cas du maïs, le rendement correspond à la quantité récoltée de grains, souvent exprimée en tonnes par hectare. Le rendement dépend des caractéristiques de la région où le maïs est cultivé et des conditions climatiques de l'année dans cette région (températures, rayonnements, précipitations etc.). La valeur du rendement est susceptible de varier fortement entre régions et entre années. Par exemple, le rendement pourra être anormalement faible une année présentant un déficit hydrique important à un stade clé du développement de la culture ou, au contraire, très élevé une année présentant des conditions climatiques optimales tout au long de la saison.

Il est important de prédire précisément le rendement avant la récolte. Celle-ci a généralement lieu en à l'automne pour le maïs. Des prédictions fiables avant la récolte offrent la possibilité aux opérateurs économiques régionaux de planifier leur récolte, de gérer leurs stocks et d'optimiser leurs contrats (achats et ventes de grains). Les prédictions de rendement constituent également une information stratégique utilisée par les acteurs opérants sur les marchés internationaux. Des prédictions de récoltes abondantes ou, au contraire, des prédictions de pertes importantes, peuvent fortement impacter les cours des marchés agricoles mondiaux.

L'objectif de ce challenge est de développer des outils permettant de prédire aussi précisément que possible les rendements du maïs en France.

Qui peut participer ?

Ce défi ouvert à la fois aux professionnels et aux étudiant.e.s. Tout.e étudiant.e inscrit.e dans une formation universitaire ou dans une école d'ingénieur française peut participer. Les stagiaires inscrits en master et les doctorants sont bienvenus. Tout professionnel peut participer. Un classement différent sera établi pour les professionnels et les étudiants.

Données

Le(a) participant.e dispose de plusieurs fichiers accessibles depuis le site du data challenge :

- **Un fichier de données** « **entrainement** » pour le maïs (TrainingDataSet_Maize) incluant (i) les anomalies rendements annuels de maïs en tonnes par ha pour différents départements français et pour 38 années tirées au hasard et (ii) les valeurs correspondantes de variables climatiques pour les mêmes départements et les mêmes années qui pourront être utilisées pour prédire les rendements. Une variable liée au pourcentage de surface agricole irriguée dans chaque département qui pourra également être prise en compte pour prédire le rendement. Dans ce fichier, les anomalies de rendement correspondent aux écarts entre les rendements observés et les tendances temporelles ajustées département par département. Une anomalie positive représente un gain de rendement par rapport au rendement attendu, et une anomalie négative correspond à une perte de rendement.

- **Un fichier « test »** maïs (TestDataSet_Maize_blind) incluant les valeurs des variables climatiques pour les mêmes départements mais pour 19 années non incluses dans les fichiers « entrainement ». Ce fichier n'inclut pas de valeurs de rendements. Les rendements obtenus au cours de ces années ont été retirés et seront utilisés par l'organisateur pour évaluer les performances des prédictions des participants.

Règles

Pour le challenge maïs, le(a) participant.e développe une méthode de prédiction des anomalies annuelles de rendement de maïs (ex : régression , forêt aléatoire, réseau de neurone) à l'échelle des départements français à partir du fichier « entrainement » en utilisant **le logiciel R (https://cran.r-project.org/)** ou **python**. Le résultat de la méthode doit correspondre à une valeur prédite de rendement par département et par année.

Une fois sa méthode au point, le(a) participant.e l'utilise pour **prédire les anomalies de rendement pour toutes les situations (départements*années) du fichier « test ».** Il/elle dépose un fichier incluant **les valeurs prédites des anomalies dans le même ordre que les situations du fichier test**. La précision de ces prédictions sera évaluée par les organisateurs à partir des anomalies de rendement observées, en calculant **le Root Mean Square Error (RMSE).**

Format des réponses au challenge

Le(a) candidat.e devra envoyer les éléments suivants à l'adresse cland_fcy@lsce.ipsl.fr :

- Un document au format .pdf décrivant de manière détaillée la méthode utilisée (package R/python utilisé, nature du modèle, variables d'entrée, algorithmes utilisés etc.),
- Un fichier au format .txt (séparateur tabulation) incluant les prédictions d'anomalie de rendement pour le maïs dans le même ordre que les situations incluses dans le fichier test.
- Un fichier incluant le code R ou Python suffisamment documentés pour que le jury puisse utiliser la méthode de prédiction.

Le document décrivant la méthode devra s'intituler :
`mais_document_nomCandidat.pdf`.

Le fichier incluant les prédictions devra être un fichier intitulé :
`mais_prediction_nomCandidat.txt` incluant les prédictions des anomalies de rendement.

Sélection du vainqueur

Seules les candidatures comportant un document descriptif détaillé et les codes R/Python seront considérés.

Parmi ces candidatures, deux classements seront établis, un pour les étudiants, un pour les professionnels. Chaque classement sera établi à partir des RMSE calculés avec les jeux de données « test ». Le vainqueur sera celui ou celle ayant obtenu.e la valeur de RMSE la plus faible pour le maïs.

Le 1er du classement « étudiant » recevra un prix de 500 euros. Le 1er du classement « professionnel » recevra un prix offert par le Réseau Mixte Technologique « Modélisation et analyse de données » (http://www.modelia.org/moodle/).

Les meilleur.e.s candidat.e.s seront invité.e.s à venir présenter leurs méthodes de prédiction lors d'un séminaire de restitution le 7 décembre.

Dates clés

- Dates limites pour déposer le fichier des prédictions : 16 novembre 2018
- Publication des résultats : 7 décembre 2018
- Séminaire de restitution : 7 décembre 2018

Liste des variables

yield_anomaly : variable à prédire représentant l'anomalie de rendement de maïs (une valeur positive indique un rendement plus élevé qu'attendu, une valeur négative indique une valeur perte de rendement par rapport à la valeur attendue), exprimée en tonne par ha.

year_harvest : année (anonyme) de récolte (1 à 57)

NUMD : numéro (anonyme) indiquant le département (de 1 à 94).

La variable yield_anomaly doit être prédite uniquement à l'aide des variables suivantes (ou d'une partie de ces variables) :

- ETP_1… ETP_9 : Evapotranspiration potentielle moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- PR_1… PR_9 : Précipitation cumulée mensuelle par année et par département (1= janvier, 9=septembre)
- RV_1… RV_9 : Rayonnement moyen mensuel par année et par département (1= janvier, 9=septembre)
- SeqPR1…SeqPR9 : Nombre de jours de pluie mensuel par année et par département (1= janvier, 9=septembre)
- Tn_1…Tn_9 : Température minimale journalière moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- Tx_1…Tx_9 : Température maximale journalière moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- IRR : variable comprise entre 1 et 5 liée à la fraction de surface agricole irriguée dans chaque département. La valeur 1 indique une fraction faible, la valeur 5 indique une fraction élevée de surface irriguée. Ces valeurs sont indicatives car établies sur la base d'information collectée pendant une seule année.

Important : Le maïs est généralement semé au printemps et est récolté à l'automne. Les valeurs des variables climatiques pour les mois 1 à 9 correspondent aux valeurs obtenues l'année de récolte. Elles sont disponibles avant la récolte et peuvent donc être utilisées directement pour prédire le rendement. Les valeurs des variables climatiques des mois 10 à 12 sont absentes des fichiers.