# Exploitation de l'enquête TERLAB pour l'estimation du rendement des cultures à la parcelle à partir de séries temporelles Sentinel-2

Jordi Inglada

# Exploitation de l'enquête TERLAB pour l'estimation du rendement des cultures à la parcelle à partir de séries temporelles Sentinel-2

Jordi Inglada

9 septembre 2020

# Table des matières

**Résumé**

Ce document synthétise les résultats d'une étude menée au CESBIO visant à explorer la possibilité d'estimer le rendement des grandes cultures agricoles à partir d'imagerie satellitaire d'observation de la Terre.

Des séries temporelles d'imagerie satellitaire optique à haute résolution (10 et 20 m) acquises avec une fréquence temporelle de 5 jours sur l'ensemble du territoire sont utilisées pour estimer le rendement au niveau de la parcelle. Un modèle de régression non-linéaire est obtenu par apprentissage automatique en utilisant les données de l'enquête TERLAB (déclarations de rendement moyen des exploitations sur un échantillon). Le parcellaire agricole et la culture de chaque parcelle sont supposés connus.

Dans un premier temps, nous montrons qu'il est possible de construire des estimateurs communs à toutes les cultures, mais locaux et spécifiques à des zones géographiques de la taille d'un carré de 100 km de côté. Les performances de ces modèles se dégradent quand ils sont appliqués à d'autres zones géographiques que celles où ils ont été appris. Ceci limite leur intérêt dans un objectif de statistique nationale sur les productions.

Dans un deuxième temps, nous construisons un modèle applicable à tout le territoire métropolitain mais spécialisé pour une seule culture (le blé tendre d'hiver dans notre exemple). Nous constatons de très bonnes performances dans l'estimation des rendements avec des erreurs inférieures à 10 q/ha. Les résultats de ce modèle sont comparés à la statistique agricole annuelle consolidée montrant une très bonne cohérence.

Nous terminons le document avec des propositions d'amélioration de la méthodologie et des recommandations pour un passage à l'opérationnel.

# 1 Introduction

## 1.1 Demande

Le Service de la statistique et la prospective (SSP) du Ministère de l'agriculture et de l'alimentation (MAA) s'est adressé au CESBIO pour avoir des informations sur la possibilité de disposer d'estimations exhaustives, très fines au niveau géographique, sur divers indicateurs concernant les productions végétales.

Parmi ceux-ci, les indicateurs de rendement ou production sont ceux qui sont les moins facilement accessibles dans l'ensemble des données à disposition actuellement par le SSP, bien que, pour beaucoup de postes des grandes cultures, leur connaissance par les professionnels soit parfaitement maîtrisée (transmission de la récolte en direct par GPS par les appareils de récolte par exemple).

Le SSP considère que trois buts généraux peuvent être poursuivis. Les deux premiers se situent dans le cadre des opérations existantes du SSP : la statistique agricole annuelle (SAA) et la conjoncture végétale, grandes cultures en priorité, avec pour but d'apporter des informations qui peuvent fiabiliser le travail de production de synthèses.

1. Conjoncture :
   - établir de manière *précoce* (c'est à dire en cours de saison) des estimations de surfaces et productions ;
   - concerne la majorité des grandes cultures (et quelques légumes et fruits).
2. SAA :
   - bilan d'une année, réalisé l'année suivante, sur une nomenclature de production fine, en surface et production.
3. Créer un ensemble d'informations très fines géographiquement en surface et production :
   - restitution en terme de nomenclatures et de mesures : correspondre à la différentiation visible et fiable ;
   - source nouvelle, indépendante des sources et opérations existantes ;

— elle peut devenir à terme, si la nomenclature des cultures a pu s'affiner et les indicateurs se fiabiliser, la source de référence des productions végétales.

Il pourrait être adjoint d'autres buts :

1. Fournir des renseignements territoriaux, hors agriculture (connus par le RPG), la déprise agricole en particulier ; il s'agirait en fait d'une forme de spécialisation de la couche OSO [1] fournie par le CESBIO ;
2. Fournir des indications de pratiques culturales (détection des irrigations, travail du sol) ;
3. Fournir des indications sur les stades culturaux ;
   — en particulier, la conjoncture Grandes cultures (GCMENS) produit un indicateur pour chaque poste de la nomenclature de pourcentage semé pour les premiers mois de l'année ;
   — cet indicateur pourrait être complété d'informations sur certains stades culturaux qui seraient discernables par satellites.

Ceci sera traité ultérieurement dans une suite éventuelle de l'étude.
Ces buts généraux conduisent à examiner quatre points :

1. les indicateurs à relever,
2. leur échéancier,
3. les nomenclatures supports de ces indicateurs,
4. la granularité géographique de collecte.

## 1.2 Proposition d'étude de la part du CESBIO

N'ayant pas de cadre contractuel ni de moyens spécifiques pour répondre complètement à la demande du SSP, le CESBIO n'est pas en mesure de répondre à l'ensemble des points soulevés ci-dessus. Par ailleurs, le CESBIO n'a pas vocation à se substituer à des prestataires pouvant faire des tâches de production et capables de mettre en œuvre les ressources d'ingénierie sans composante de recherche nécessaires pour répondre à une grande partie des demandes du SSP.

Les travaux du CESBIO dans des projets comme *Sentinel-2 Agriculture* et *SENSAGRI* ont déjà démontré la faisabilité de la cartographie des grandes cultures. Des sociétés comme *OneSoil* fournissent déjà ce service.

Parmi les demandes du SSP, celle qui semblait nécessiter un travail de recherche et pouvoir être réalisée sans ressources spécifiques, est celle de l'estimation du rendement.

1. https://www.theia-land.fr/product/carte-doccupation-des-sols-de-la-france-metropolitaine/

Il a donc été proposé de faire une étude concernant la possibilité d'estimer le rendement des grandes cultures à l'échelle de la parcelle en fin de saison en utilisant l'imagerie satellitaire. Ce document synthétise les travaux réalisés et les résultats obtenus. Il s'agit de 250 h de travail réalisé entre janvier 2019 et juillet 2020.

## 1.3 L'enquête TERLAB

L'enquête a pour objectif premier d'estimer les rendements d'une trentaine de cultures issues de terres labourables (dites aussi « grandes cultures ») aux niveaux départemental, régional et national. Elle peut également être utilisée pour l'estimation précoce de l'évolution des surfaces cultivées.

L'interrogation porte sur la surface principale et le rendement de l'ensemble des cultures présentes sur l'exploitation, ainsi que sur les prévisions de semis pour l'année suivante.

Sont sélectionnés 66 départements français, respectant deux critères :

— la superficie en terres labourables du département est supérieure à 50 000 ha,

— et la superficie d'au moins une culture dépasse 20 000 ha.

Au total, on interroge 13 000 exploitations réparties sur les 66 départements. La précision attendue des rendements des principales cultures répond aux exigences européennes de précision fixée à un maximum de 3% au niveau national.

En croisant les données TERLAB et les parcelles issues du RPG, le SSP a mis à disposition du CESBIO une base de données où chaque parcelle du RPG est caractérisée par un ensemble de variables dont le rendement de l'exploitation pour la culture concernée par cette parcelle. D'autres informations comme la surface totale de cette culture dans l'exploitation sont aussi disponibles.

Le tableau 1.1 liste, pour chaque département, le nombre de parcelles pour lesquelles le rendement de l'exploitation est disponible, toutes cultures confondues. Il y a en tout 225367 parcelles avec données de rendement sur 70 départements en 2017.

## 1.4 L'utilisation de l'imagerie satellitaire comme source supplémentaire

Il est proposé d'évaluer la possibilité d'estimer le rendement à la parcelle en utilisant des séries d'images Sentinel-2 (voir 2.3 pour une description des données satellite). L'estimation à la parcelle permet ensuite de faire des statistiques à différents niveaux d'agrégation spatiale en fonction des besoins.

Il ne s'agit pas de remplacer l'enquête TERLAB, mais de proposer une méthode basée sur l'imagerie satellitaire pour :

TABLE 1.1 – Parcelles par département

| | Département | Parcelles | | Département | Parcelles |
|---|---|---|---|---|---|
| 1 | Charente-Maritime (17) | 5833 | 36 | Ille-et-Vilaine (35) | 3102 |
| 2 | Côte-d'Or (21) | 5760 | 37 | Nièvre (58) | 3077 |
| 3 | Vienne (86) | 5616 | 38 | Haut-Rhin (68) | 3033 |
| 4 | Yonne (89) | 5547 | 39 | Loir-et-Cher (41) | 3000 |
| 5 | Deux-Sèvres (79) | 5373 | 40 | Vendée (85) | 2960 |
| 6 | Bas-Rhin (67) | 5036 | 41 | Haute-Vienne (87) | 2878 |
| 7 | Loiret (45) | 4948 | 42 | Pyrénées-Atlantiques (64) | 2859 |
| 8 | Somme (80) | 4878 | 43 | Landes (40) | 2818 |
| 9 | Pas-de-Calais (62) | 4746 | 44 | Haute-Marne (52) | 2747 |
| 10 | Haute-Garonne (31) | 4713 | 45 | Allier (03) | 2597 |
| 11 | Aube (10) | 4705 | 46 | Calvados (14) | 2567 |
| 12 | Oise (60) | 4618 | 47 | Loire-Atlantique (44) | 2562 |
| 13 | Aisne (02) | 4467 | 48 | Vosges (88) | 2549 |
| 14 | Saône-et-Loire (71) | 4435 | 49 | Tarn-et-Garonne (82) | 2529 |
| 15 | Isère (38) | 4287 | 50 | Nord (59) | 2495 |
| 16 | Indre-et-Loire (37) | 4230 | 51 | Meuse (55) | 2475 |
| 17 | Charente (16) | 4215 | 52 | Manche (50) | 2423 |
| 18 | Gers (32) | 4194 | 53 | Haute-Saône (70) | 2366 |
| 19 | Cher (18) | 4162 | 54 | Creuse (23) | 2345 |
| 20 | Seine-et-Marne (77) | 4099 | 55 | Sarthe (72) | 2333 |
| 21 | Tarn (81) | 3877 | 56 | Jura (39) | 2284 |
| 22 | Eure (27) | 3876 | 57 | Meurthe-et-Moselle (54) | 2274 |
| 23 | Eure-et-Loir (28) | 3827 | 58 | Ardennes (08) | 2261 |
| 24 | Côtes-d'Armor (22) | 3709 | 59 | Orne (61) | 2215 |
| 25 | Marne (51) | 3695 | 60 | Mayenne (53) | 2162 |
| 26 | Lot-et-Garonne (47) | 3648 | 61 | Morbihan (56) | 2088 |
| 27 | Ain (01) | 3474 | 62 | Seine-Maritime (76) | 1978 |
| 28 | Puy-de-Dôme (63) | 3437 | 63 | Hautes-Pyrénées (65) | 1804 |
| 29 | Aude (11) | 3421 | 64 | Aveyron (12) | 1604 |
| 30 | Drôme (26) | 3354 | 65 | Gironde (33) | 1267 |
| 31 | Maine-et-Loire (49) | 3298 | 66 | Yvelines (78) | 1233 |
| 32 | Dordogne (24) | 3296 | 67 | Essonne (91) | 834 |
| 33 | Gard (30) | 3218 | 68 | Bouches-du-Rhône (13) | 812 |
| 34 | Finistère (29) | 3197 | 69 | Val-d'Oise (95) | 464 |
| 35 | Indre (36) | 3168 | 70 | Vaucluse (84) | 15 |

1. transformer l'information de rendement au niveau de l'exploitation en une information au niveau de la parcelle ;
2. estimer le rendement sur les parcelles des exploitations non enquêtées et ceci sur tous les départements.

Les données TERLAB sont nécessaires pour construire le modèle d'estimation reliant les observations satellitaires et le rendement. La culture de chaque parcelle est supposée connue, ce qui ne pose pas de difficulté majeure si les déclarations des agriculteurs pour le RPG sont rendues disponibles.

# 2 | Méthode d'estimation du rendement

## 2.1 Hypothèses et position du problème

Voici l'ensemble des hypothèses sur le comportement des cultures et les informations disponibles :

— La dynamique temporelle de l'état de la végétation permet de prédire le rendement.

— La prédiction est faite à la fin de la saison agricole. Les images de l'année civile complète sont utilisées.

— Pour chaque parcelle, la culture est connue. On dispose du parcellaire agricole.

Le problème d'estimation du rendement est posé comme une régression où la variable prédite est le rendement de chaque parcelle et les prédicteurs sont des indices de végétation calculés à partir des observations satellitaires le long de la saison agricole.

Le modèle de régression est calibré en utilisant les données TERLAB, même si seulement les rendements moyens des exploitations et non pas ceux de chaque parcelle sont disponibles.

## 2.2 Préparation des données TERLAB

Les données TERLAB pour la campagne 2017 ont été fournies au CESBIO sous forme d'un fichier au format ESRI Shapefile pour chaque département. Des fichiers pour tous les départements, y compris ceux qui ne sont pas enquêtés ont été fournis. Les parcelles RPG appartenant à des exploitations non enquêtées ont une valeur nulle dans la colonne du rendement (attribut `RENDNORME`).

Les images Sentinel-2 sont fournies selon un découpage en tuiles tandis que les données TERLAB sont fournies par département (figure 2.1). Les données TERLAB ont été réorganisées par tuile Sentinel-2 pour faciliter la suite des traitements.

Les colonnes suivantes sont éliminées : `PRECISION`, `SEMENCE`, `DEST_ICHN`, `CULTURE_D1`, `CULTURE_D2`, `ENGAGEMENT`, `MARAICHAGE`, `AGROFOREST`, `TLENQ`, `CODUTISOL`. Les valeurs man-

quantes pour les colonnes BIO, SURUTISOL, RENDNORME, MMEAU son mises à zéro, car certains des formats de données qui seront utilisés ne supportent pas les valeurs manquantes.

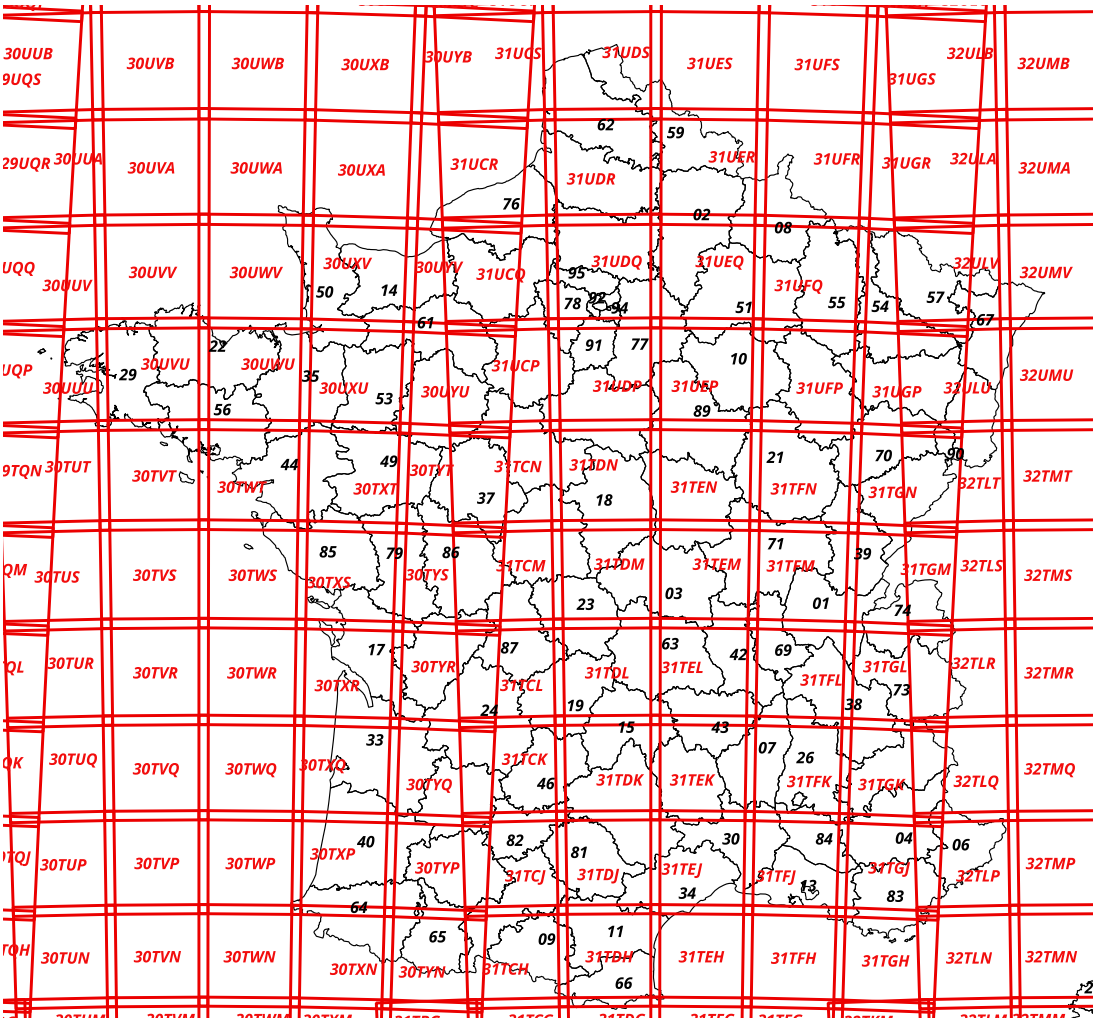


FIGURE 2.1 – Tuiles Sentinel-2 sur la France métropolitaine hors Corse

## 2.3 Préparation des données S2

Pour chaque tuile Sentinel-2, toutes les acquisitions disponibles sur l'année 2017 ont été utilisées. La première moitié de l'année est couverte par un seul satellite, ce qui permet d'avoir une acquisition tous les 10 jours. La deuxième moitié de l'année est couverte par 2 satellites et la revisite est de 5 jours [1]. Cependant, la présence de nuages, le recouvrement des orbites et d'autres éléments font que le nombre d'acquisitions disponibles est variable sur le

1. À partir de 2018, deux satellites sont toujours disponibles ce qui présente un potentiel d'amélioration des résultats par rapport à 2017.

territoire et peut aller de moins de 10 sur la pointe du Finistère jusqu'à 70 dans les Bouches du Rhône (figure 2.2).

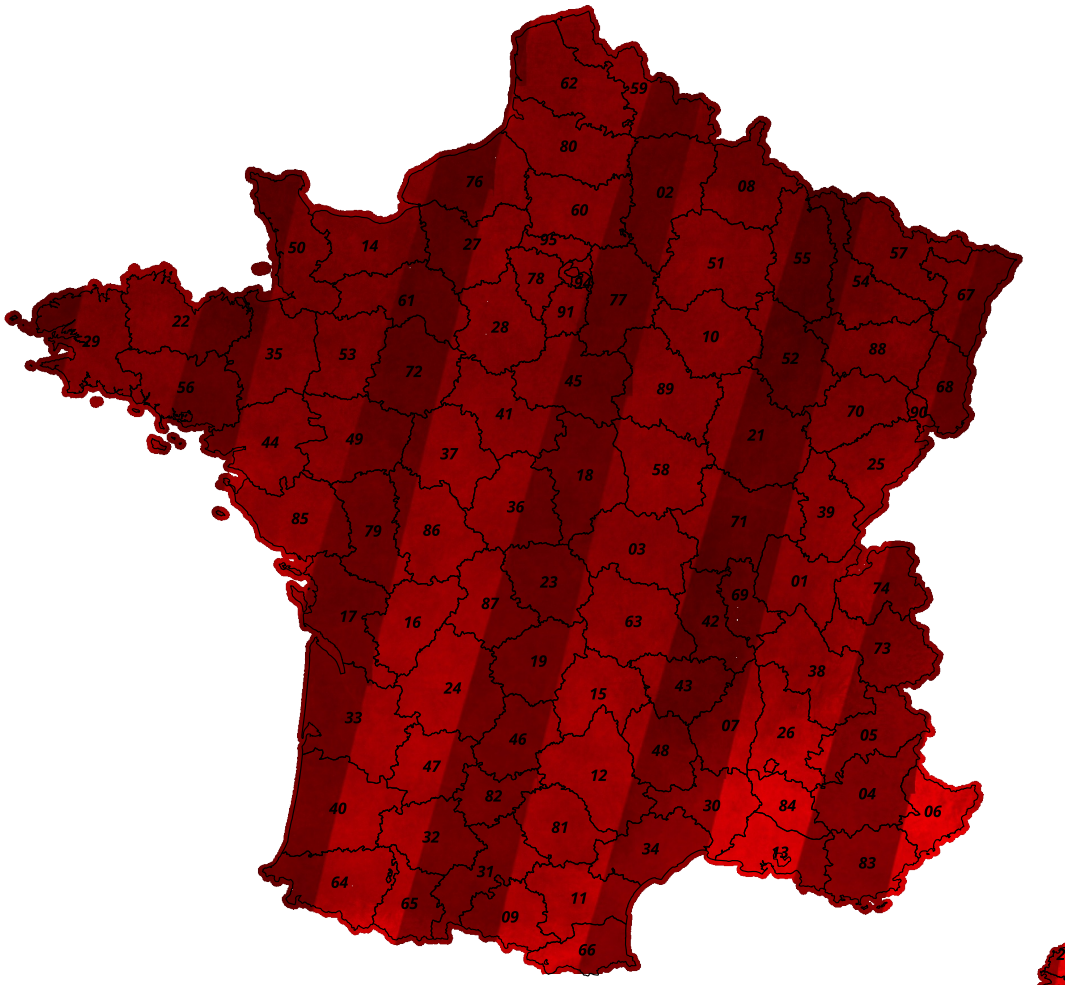


FIGURE 2.2 – Le nombre d'acquisitions exploitables sur l'année civile s'étale entre 0 et 73 selon les points du territoire.

Les données en entrée de la procédure sont des images Sentinel-2 corrigées des effets de l'atmosphère (réflectances de surface) accompagnées de masques de validité. Ces masques indiquent si la surface est visible à une date donnée en fonction de la présence de nuages, ombres de nuages ou d'autres raisons comme des saturations ou des zones hors champ d'acquisition du capteur. Les données sont produites par le Pôle Theia et disponibles gratuitement et avec une licence ouverte [2].

La procédure de préparation des données Sentinel-2 est la suivante :

1. Empilement des réflectances de surface et des masques de validité associés par ordre

2. https://www.theia-land.fr/product/reflectance-sentinelle-2/

chronologique. Les bandes à 20 m sont ré-échantillonnées à 10 m. Il n'y a pas de ré-échantillonnage temporel ni d'imputation des données manquantes.

2. Pour chaque parcelle RPG, on calcule la moyenne, l'écart-type, le minimum et le maximum des variables suivantes :
   - les 10 réflectances de surface Sentinel-2 ;
   - le masque binaire de validité par pixel ;
   - des informations topographiques issues du MNT SRTM
     - altitude ;
     - pente ;
     - exposition,
3. Pour chaque parcelle et à partir de la moyenne des réflectances de surface on calcule les indices spectraux NDVI, NDWI, NDRE1, NDRE2 (voir 2.3.1).
4. Pour chaque indice, on calcule le minimum et le maximum par trimestre et sur l'année.

Dans les sections suivantes, des détails sur les différentes étapes de la procédure son donnés.

### 2.3.1 Les réflectances de surface Sentinel-2 et les indices dérivés

Les satellites Sentinel-2 ont 13 bandes spectrales (voir tableau 2.1). Les bandes 1, 9 et 10 ont une résolution de 60 m et fournissent des informations qui ne sont pas utiles sur les surfaces continentales. Elles ne sont pas utilisées dans cette étude. Comme énoncé ci-dessus, les bandes qui ont une résolution de 20 m (5, 6, 7, 8A, 11 et 12) sont ré-échantillonnées à 10 m.

Nous travaillons avec des images corrigées des effets atmosphériques, ce qui permet d'avoir des séries temporelles cohérentes et avec très peu de bruit lié à l'état de l'atmosphère au moment du passage du satellite.

À partir des réflectances de surface, il est possible de calculer des indices qui renseignent sur l'activité photo-synthétique de la végétation et sur le taux d'humidité [1]. Les indices choisis sont les suivants :

- $NDVI = \frac{R_{833}-R_{665}}{R_{833}+R_{665}}$,
- $NDWI = \frac{R_{833}-R_{1614}}{R_{833}+R_{1614}}$,
- $NDRE1 = \frac{R_{740}-R_{705}}{R_{740}+R_{705}}$,
- $NDRE2 = \frac{R_{783}-R_{705}}{R_{783}+R_{705}}$,

où $R_\lambda$ est la réflectances de surface de la bande dont la longueur centrale est $\lambda$.

Il faut noter que ces indices ne sont pas calculés au pixel, mais sur la moyenne des réflectances à la parcelle. Les indices étant non-linéaires, ce calcul n'est pas exact, mais nous considérons que l'erreur introduite ne pénalise pas les estimations.

TABLE 2.1 – Bandes spectrales des satellites Sentinel-2

| Bande | Longueur d'onde centrale (nm) |
|---|---|
| 1 – Coastal aerosol | 442.7 |
| 2 – Blue | 492.4 |
| 3 – Green | 559.8 |
| 4 – Red | 664.6 |
| 5 – Vegetation red edge | 704.1 |
| 6 – Vegetation red edge | 740.5 |
| 7 – Vegetation red edge | 782.8 |
| 8 – NIR | 832.8 |
| 8A – Narrow NIR | 864.7 |
| 9 – Water vapour | 945.1 |
| 10 – SWIR – Cirrus | 1373.5 |
| 11 – SWIR | 1613.7 |
| 12 – SWIR | 2202.4 |

### 2.3.2 L'utilité des masques

Pour chaque acquisition, nous disposons aussi de masques de nuages et de leurs ombres, ce qui permet d'invalider les pixels pour lesquels la surface n'est pas visible.

Une parcelle dont plus de 10% de la surface est couverte par des nuages ou des ombres pour une date donnée, est considérée comme non observée à cette date.

### 2.3.3 Les statistiques temporelles

La méthode de régression nécessite d'avoir les mêmes prédicteurs pour tous les pixels. Le fait que le territoire observé est couvert par 5 orbites différentes, fait que les dates d'acquisition ne sont pas les mêmes pour toutes les parcelles. La présence de nuages et d'ombres rend l'échantillonnage temporel hétérogène au sein d'une même orbite.

Dans les problèmes de classification de l'occupation des sols, la méthode habituellement utilisée pour avoir les mêmes prédicteurs pour tous les pixels utilise une interpolation temporelle et un ré-échantillonnage sur une grille temporelle commune [2]. Dans le cas présent, le profil temporel complet n'est pas forcément utile, car nous faisons l'hypothèse simplificatrice que le rendement est lié à un cumul d'activité végétative plutôt qu'à des évènements à des dates précises.

Nous choisissons donc de calculer des minima et des maxima des indices radiométriques présentés dans 2.3.1 et ceci par trimestre et sur l'année complète. Nous y associons aussi les dates auxquelles ces extréma sont observés.

### 2.3.4 Les prédicteurs retenus

Des études préliminaires de dimensionnement qui ne sont pas présentées dans le présent document ont permis d'éliminer des variables qui avaient été identifiées comme potentiellement utiles, comme par exemple celles dérivées du MNT.

Les prédicteurs finalement retenus sont au nombre de 80. Nous calculons en effet 4 indices (NDVI, NDWI, NDRE1, NDRE2) et pour chaque indice nous calculons 2 statistiques (minimum et maximum) et leurs dates et ceci sur 5 périodes (4 trimestres et l'année entière).

Les prédicteurs ont des noms du type `max_ndvi_date` pour la date du maximum de NDVI sur l'année ou `min_ndwi_q2` pour la valeur du minimum de NDWI sur le 2è trimestre.

## 2.4 Méthode d'estimation

Comme indiqué dans 2.1, l'estimation du rendement est faite par régression. Étant donné que l'objectif n'est pas d'expliquer les liens entre rendement et observations satellite, mais d'avoir une estimation précise des rendements, nous choisissons un algorithme d'apprentissage capable d'ingérer des volumes de données tabulaires importants (des dizaines de prédicteurs pour des dizaines de milliers d'individus).

Parmi les multiples possibilités disponibles dans la boîte à outils de l'apprentissage automatique, nous choisissons l'algorithme des Forêts aléatoires (Random Forests, [3]) en raison du compromis entre précision et coût de calcul ainsi que la simplicité du paramétrage.

Cet algorithme de régression permet de modéliser des dépendances non-linéaires et fournit une estimation de l'importance des différents prédicteurs. Cette importance n'est pas analogue aux coefficients d'une régression linéaire, car elle n'indique pas le signe de la corrélation, mais simplement la contribution relative à la réduction de l'erreur de prédiction.

C'est l'étude de l'importance des variables qui a permis de ne garder que les prédicteurs listés dans 2.3.4.

# 3 | Résultats

Dans ce chapitre, 2 analyses sont présentées. La première porte sur l'ensemble des cultures, mais sur un nombre limité de tuiles Sentinel-2. La deuxième se concentre sur une seule culture, mais sur l'ensemble du territoire.

## 3.1 Estimation du rendement toutes cultures confondues

La figure 3.1 montre les 7 tuiles sélectionnées pour cette analyse. Le choix des tuiles a été guidé par le nombre de parcelles TERLAB avec des informations de rendement et la diversité géographique et climatique.

La figure 3.2 montre la distribution des cultures (en surface) pour chacune des tuiles. On y retrouve évidemment les grandes cultures :

— colza d'hiver (CZH),

— tournesol (TRN),

— blé dur/tendre d'hiver (BDH, BTH),

— orge d'hiver et de printemps (ORH, ORP),

— maïs et maïs d'ensilage (MIS, MIE),

— soja (SOJ).

Les proportions des classes sont différentes entre les tuiles, notamment pour ce qui concerne le tournesol qui est surtout présent dans le sud. Le blé tendre d'hiver est la culture majoritaire dans toutes les tuiles.

Nous construisons ici un modèle de régression pour chaque tuile. Les prédicteurs sont ceux introduits dans 2.3.4 enrichis avec des prédicteurs binaires qui encodent la classe de la parcelle. Cela veut dire que la prédiction pour une parcelle utilise la connaissance de la culture de la parcelle, mais un seul modèle de régression est utilisé pour toutes les cultures. Cette approche peut poser question du point de vue agronomique. Elle est proposée dans un objectif de simplification de déploiement.

Au premier abord, il pourrait paraître difficile d'obtenir un seul modèle de régression pour toutes les cultures. En effet, le rendement pour différentes cultures peut correspondre à

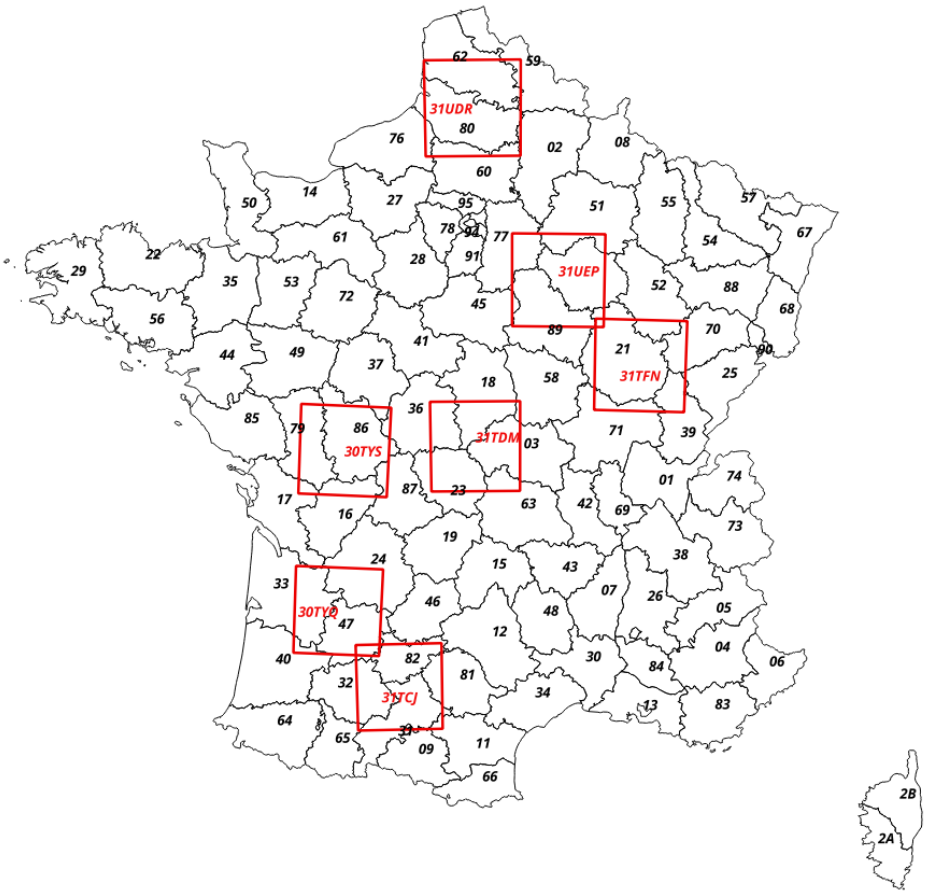


FIGURE 3.1 – Les 7 tuiles sélectionnées pour l'étude avec toutes les cultures

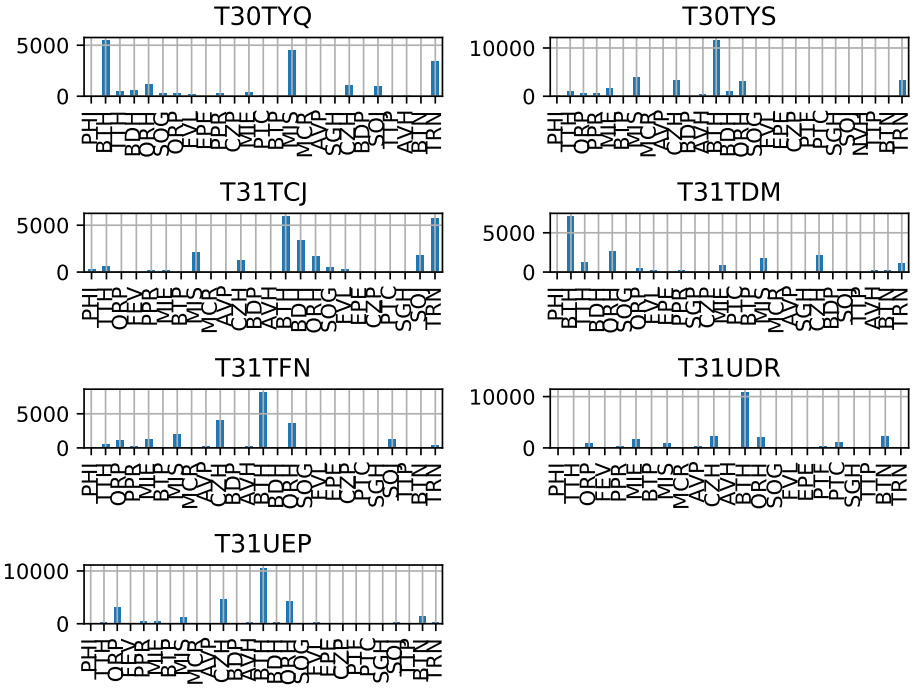


FIGURE 3.2 – Répartition des cultures dans les 7 tuiles sélectionnées

des grandeurs différentes (grain presque sec, biomasse totale aérienne avec beaucoup d'eau) et donc avec des dynamiques de valeurs très différentes.

Il faut noter que l'introduction du type de culture permet au modèle de se rapprocher de la valeur moyenne du rendement pour une culture donnée et que les autres prédicteurs servent à réduire la dispersion des estimations.

La figure 3.3 présente les résultats de la régression pour chaque tuile sur des données n'ayant pas été utilisées pour l'apprentissage du modèle. L'axe des abscisses correspond au rendement TERLAB et l'axe des ordonnées correspond au rendement estimé par chaque modèle de régression. On observe des estimations bien alignées avec la droite $y = x$ avec des surestimations pour les faibles rendements et des sousestimations pour les rendements élevés. On observe aussi des points isolés loin de la droite de régression. Ils pourraient correspondre à des parcelles atypiques au sein d'une exploitation (rappelons que les données TERLAB fournissent une valeur unique par exploitation).

Sur la figure, notamment pour les tuiles T31UDR et T31UEP, on observe des amas de points centrés sur des plages de rendements distinctes et qui correspondent à des cultures différentes.

Le tableau 3.1 liste les erreurs pour chacune des tuiles. RMSE est la racine carrée de l'erreur quadratique moyenne et MAE est l'erreur absolue moyenne. Les deux valeurs sont donc exprimées dans les mêmes unités que le rendement (quintaux par hectare). Si la RMSE est

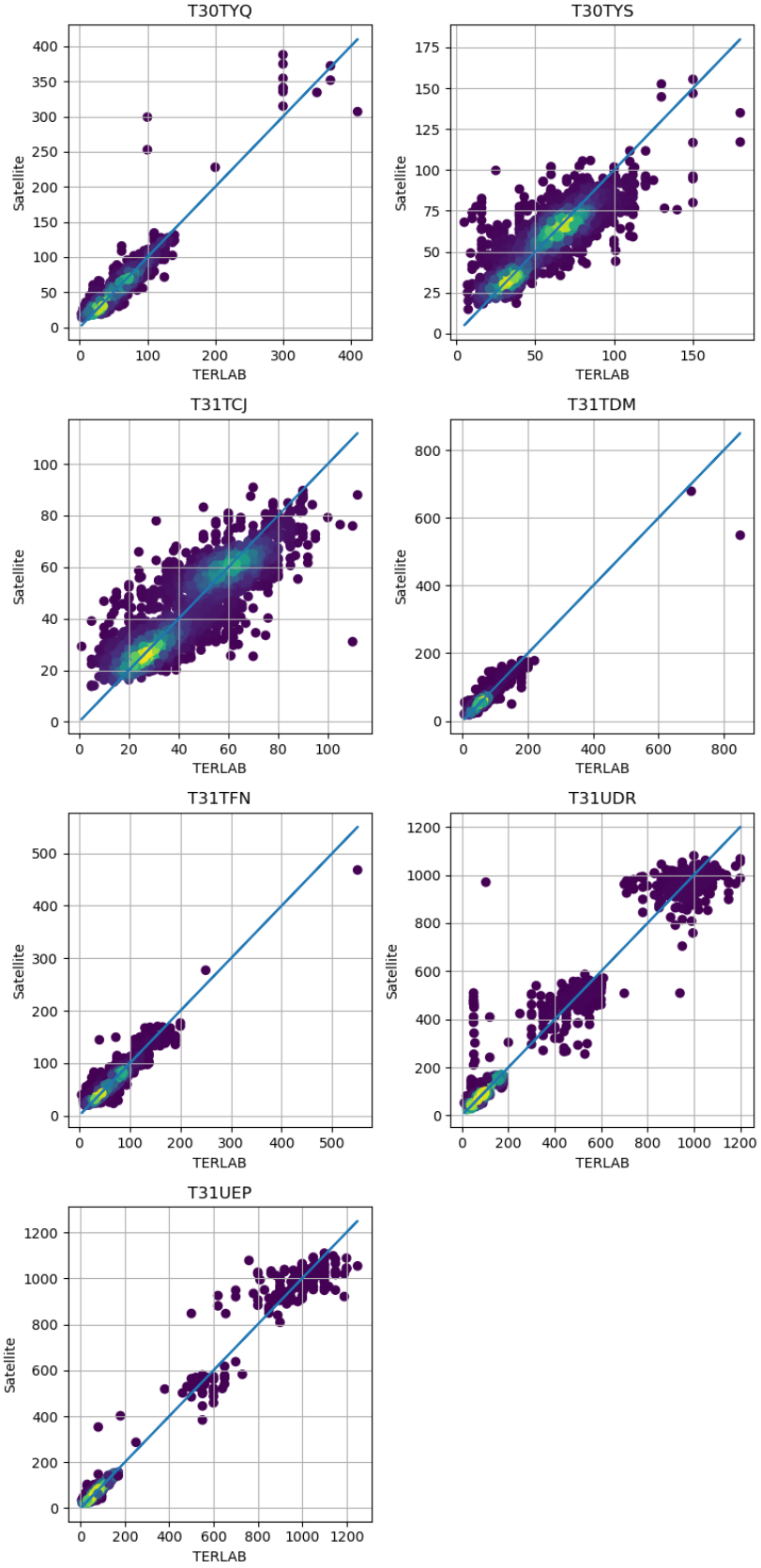


FIGURE 3.3 – Régression du rendement par tuile, toutes cultures.

une métrique plus fréquente, la MAE est jugée préférable ici, car moins sensible aux fortes erreurs sur un nombre réduit de parcelles.

Excepté pour les 2 tuiles les plus septentrionales (T31UEP et T31UDR), la MAE oscille entre 6 et 8. Cette valeur est à interpréter en tenant compte que les valeurs obtenues par enquête ce sont des moyennes à l'exploitation et arrondies à des valeurs entières.

Pour les tuiles septentrionales, on constate que le fait d'avoir des groupes de cultures avec des rendements très différents ne permet pas au modèle de réduire la dispersion des estimations. Ceci invite à construire des modèles par culture.

TABLE 3.1 – Erreurs d'estimation des rendements à la parcelle exprimées en q/ha

| Tuile | RMSE | MAE |
|---|---|---|
| T30TYQ | 18.79 | 8.66 |
| T30TYS | 11.05 | 7.33 |
| T31TCJ | 8.50 | 6.02 |
| T31TDM | 13.37 | 7.66 |
| T31TFN | 16.60 | 7.71 |
| T31UDR | 55.53 | 23.39 |
| T31UEP | 43.81 | 15.69 |

Il est aussi intéressant d'étudier la transférabilité des modèles de régression entre les tuiles. La figure 3.4 montre les scatterplots résultants de l'application des modèles appris sur chaque tuile à toutes les autres. En ligne, on trouve la tuile où le modèle a été appris, et en colonne la tuile sur laquelle il est appliqué. En diagonale, on retrouve la même information que sur la figure 3.3. On constate que, sauf dans quelques exceptions, les modèles ne sont pas applicables à des zones géographiques différentes de celles où ils ont été calibrés. Cet effet n'est pas souhaitable si on vise à estimer les rendements dans les départements non sélectionnés par l'enquête TERLAB.

Les raisons de la non transférabilité des modèles n'ont pas été analysées, mais on peut faire l'hypothèse que la variabilité climatique joue moins que la diversité des cultures ou des pratiques entre les zones. Ayant constaté que la connaissance de la culture pour chaque parcelle sert à régler le biais du modèle, si les rendements moyens pour une même culture diffèrent entre les zones, le modèle est incapable de les restituer.

## 3.2 Rendement à la parcelle au niveau national pour le blé tendre d'hiver

Nous analysons ici la possibilité de produire un seul modèle de régression pour l'ensemble du territoire pour une seule culture.

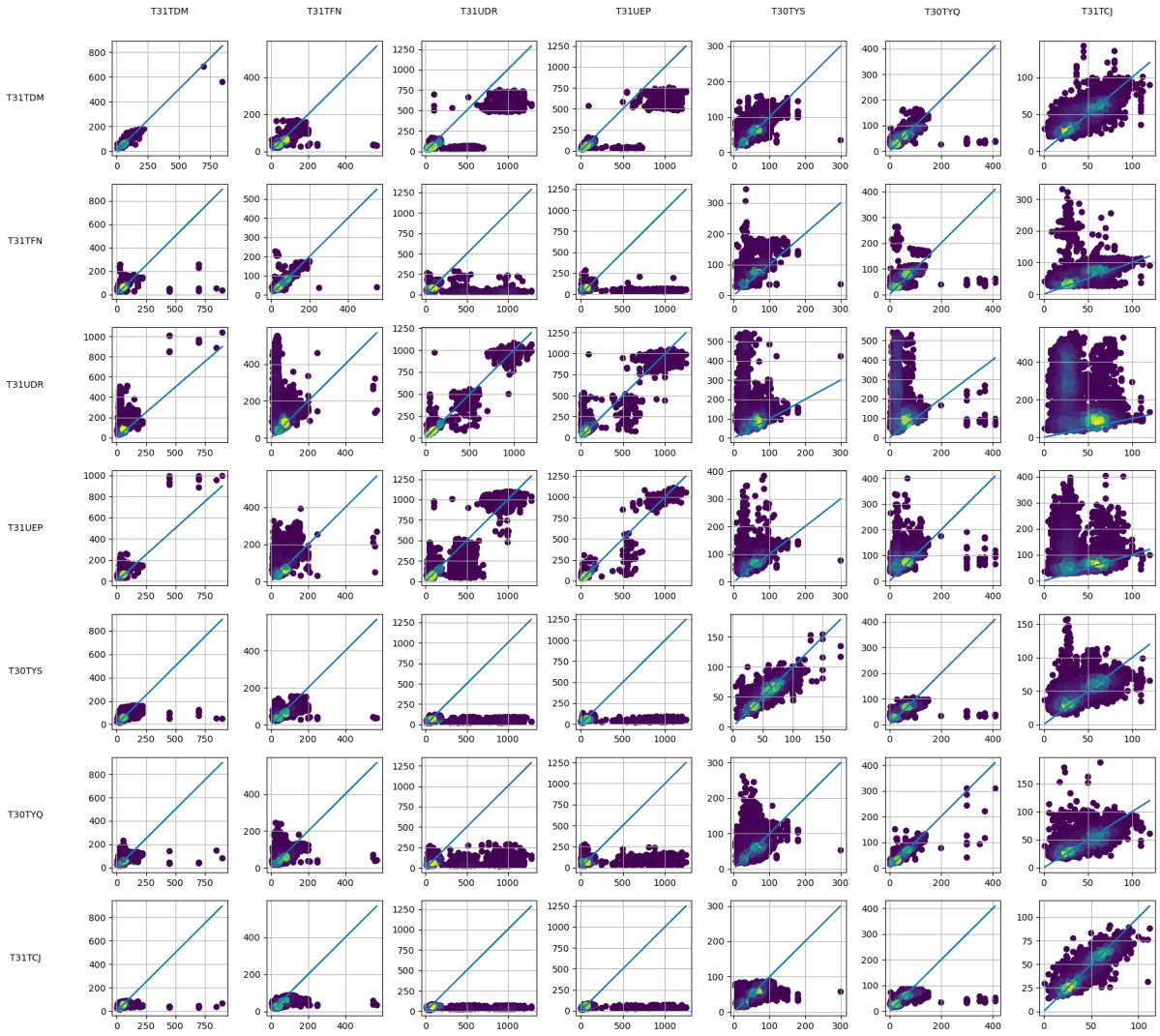


FIGURE 3.4 – Croisement des modèles de régression entre les tuiles. En ligne, la tuile où le modèle a été appris, et en colonne la tuile sur laquelle il est appliqué.

Le blé tendre d'hiver (BTH) est choisi car il s'agit d'une des cultures les plus présentes.

### 3.2.1 Modèle de régression

Ici, nous utiliserons toutes les tuiles Sentinel-2 couvrant tous les départements de la métropole excepté la Corse. Il s'agit de 77 tuiles au lieu des 84 strictement nécessaires, car nous avons éliminé celles dont l'étendue utile était trop faible. Pour chacune de ces tuiles, toutes les parcelles étiquettées comme BTH sont utilisées. Parmi celles contenant des informations de rendement (95550), 2/3 (63700) sont utilisées pour l'apprentissage et 1/3 (31850) sont mises de côté pour réaliser une validation.

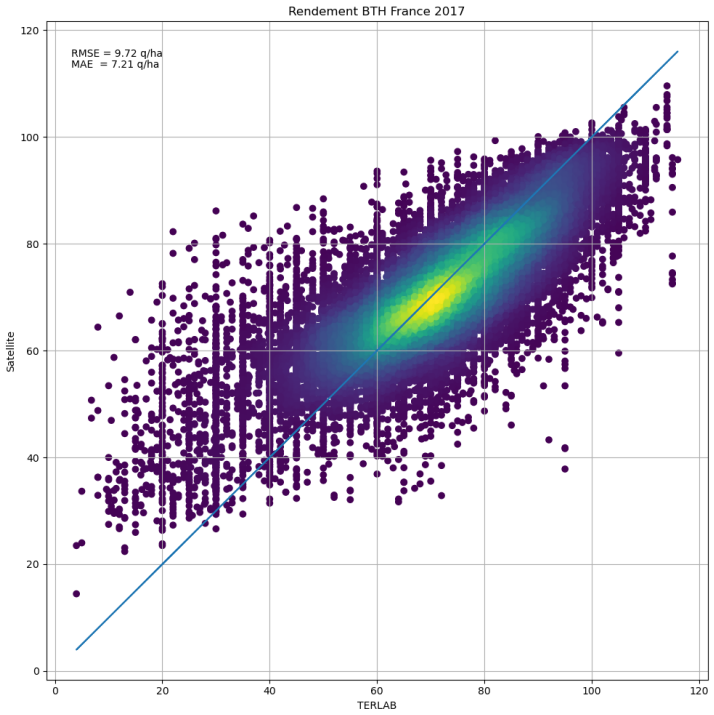


FIGURE 3.5 – Validation de l'estimation du rendement du BTH sur l'ensemble des départements

La figure 3.5 montre les résultats de la validation. Les erreurs mesurées sont de 9.72 q/ha en RMSE et de 7.21 q/ha en MAE. La figure montre une bonne corrélation entre les données TERLAB et les estimations issues de la donnée satellite. Des écarts importants existent pour les faibles rendements (inférieurs à 40 q/ha) pour lesquels la régression surestime. Il s'agit cependant de très peu de parcelles par rapport à l'ensemble de la validation (3.4 % pour un rendement TERLAB inférieur à 40 q/h).

Le rendement de toutes les parcelles, y compris pour celles des exploitations non enquêtées, est estimé à partir des données issues des images Sentinel-2.

Malheureusement, nous ne disposons pas de données précises de rendement à la parcelle pour pouvoir mener une validation plus fine. Il est donc difficile de conclure sur la nature de la dispersion des estimations obtenues à partir des données satellite. On peut admettre qu'une partie des différences avec les données TERLAB est due à la quantification des données issues de l'enquête.

### 3.2.2 Comparaison avec la SAA

Dans les opérations menant à la Statistique agricole annuelle (SAA) l'enquête TERLAB est la source principale pour le calcul du rendement. Les services statistiques du ministère chargé de l'agriculture établissent, dans le cadre de la SAA, des données de production agricole cohérentes aux niveaux départemental, régional et national. Ces données sont publiés sur le site Agreste.

Nous avons donc utilisé ces données consolidées pour faire une évaluation des estimations satellite par régression calibrée sur les données TERLAB pour le cas du blé tendre d'hiver. La base de données utilisée est nommée *Cultures développées (hors fourrage, prairies, fruits, fleurs et vigne)* pour l'année 2017 [1].

L'idée est de voir si l'utilisation conjointe de l'enquête TERLAB et des données satellite permet de fournir des estimations de rendement et de production par département similaires à celles de la SAA consolidée.

La figure 3.6 montre le résultat de la comparaison des estimations satellite et les données Agreste. La colonne de gauche présente les départements pour lesquels l'enquête TERLAB fournit des données pour le BTH, et la colonne de droite contient les départements pour lesquels aucune exploitation de BTH n'a été enquêtée.

Nous observons que pour les *départements TERLAB* (colonne de gauche) les rendements sont proches de ceux de la SAA avec une légère sous-estimation pour les rendements élevés et une sur-estimation pour les faibles. C'est le même type de comportement observé pour le rendement à la parcelle. Les productions montrent la même sous-estimation pour les valeurs élevées, mais il ne semble pas y avoir de problème pour les faibles productions, ce qui est logique, car la pondération par les surfaces faibles fait chuter les erreurs.

La colonne de droite montre les résultats pour les départements sans donnée TERLAB pour le BTH. On constate une très bonne cohérence avec la SAA pour les productions. Il est intéressant de noter que la Moselle (57) constitue un cas atypique (production très élevée) dont l'estimation est très bonne. Le problème de la surestimation des faibles rendements est ici accentuée par le fait que beaucoup de départements non enquêtés ont de faibles rende-

1. Elle a été téléchargée à partir de `https://stats.agriculture.gouv.fr/disar-web/disaron/SAANR_DEVELOPPE_2/detail.disar#` avant la refonte du site Agreste. Elle est maintenant disponible ici `https://agreste.agriculture.gouv.fr/agreste-web/disaron/SAANR_DEVELOPPE_2/detail/`

ments.

Ces interprétations pour les départements hors enquête TERLAB sont à prendre avec précautions, car sans connaître la procédure utilisée pour la SAA pour ces départements, il est difficile de savoir si l'estimation par satellite est vraiment loin de la réalité. Ceci ne remet pas en cause la qualité de la SAA, car les conséquences sur les estimations de productions sont minimes pour ces départements avec peu de surfaces associées à cette culture.

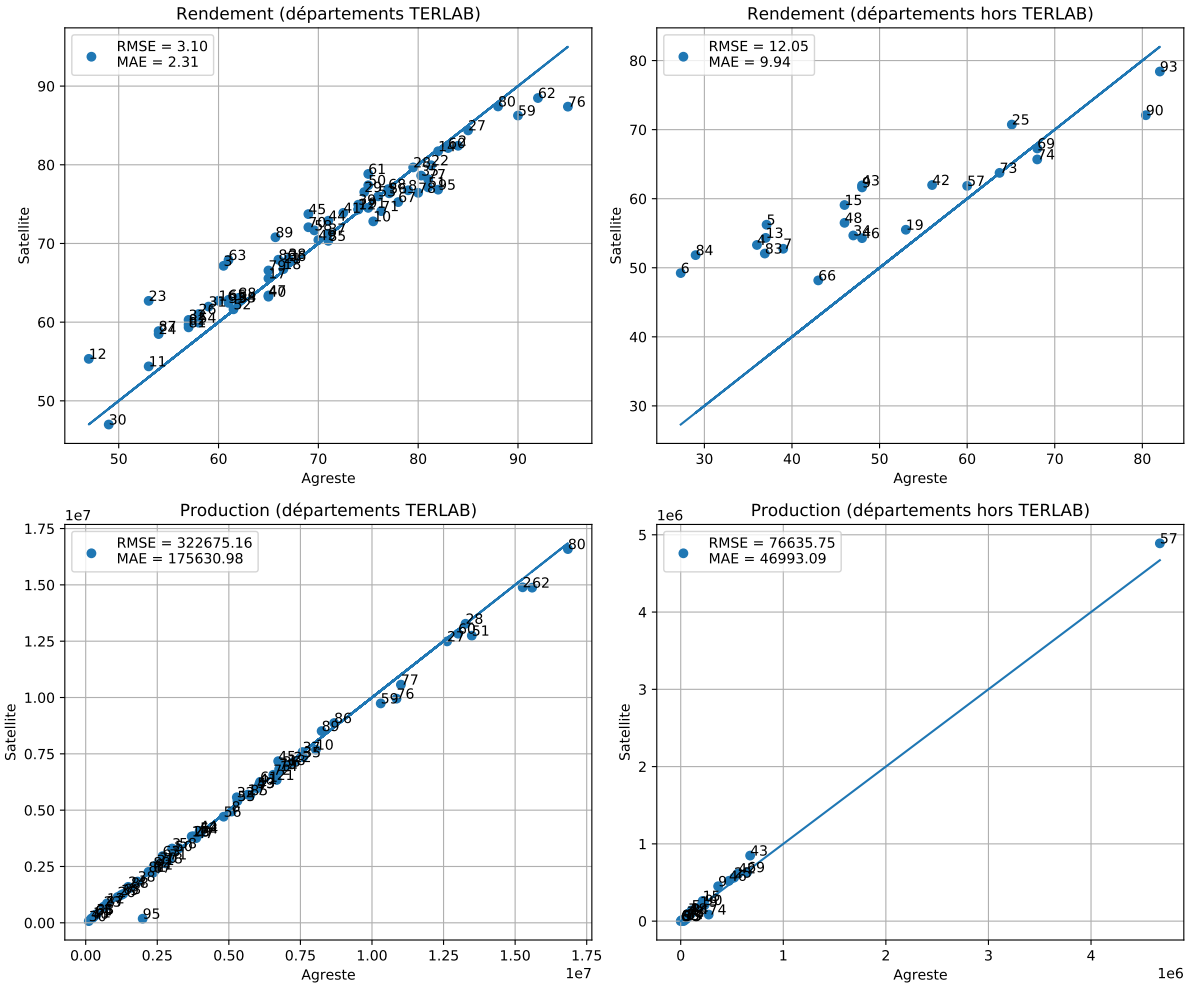


FIGURE 3.6 – Comparaison entre les estimations satellite et les données Agreste

Nous pouvons nous intéresser à la dépendance des erreurs avec le nombre de parcelles enquêtées par département. La figure 3.7 illustre ce lien. Le graphe de gauche montre la MAE (calculée à la parcelle) en fonction du nombre de parcelles dans la base de données TERLAB. On observe que, sauf pour 5 départements, la MAE est inférieure à 9 q/ha. Si on élimine ces départements de l'analyse, il ne semble pas y avoir de forte corrélation entre erreur d'estimation et taille de l'échantillon. Le graphe de droite illustre l'erreur en fonction de la production. On note le même type de comportement.

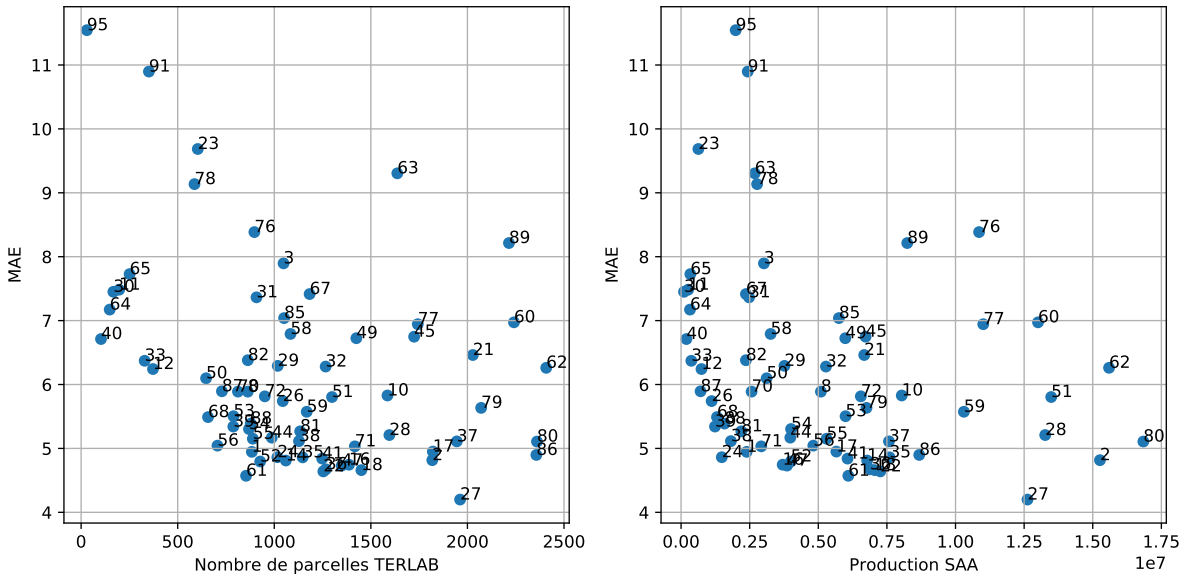


FIGURE 3.7 – Correlation des erreurs d'estimation avec la taille de l'échantillon TERLAB et les productions

Nous terminons cette analyse par une représentation spatialisée des résultats. La figure 3.8 montre des cartes de rendement et de production. Nous y associons aussi une carte d'erreurs (moyenne départementale des erreurs calculés sur les parcelles comme pour la figure 3.5) et aussi une carte du nombre de parcelles TERLAB disponibles pour l'étude.

### 3.2.3 L'utilité du satellite par rapport à l'enquête TERLAB

Dans un objectif de statistique au niveau départemental, l'utilisation du satellite peut être redondante avec l'enquête TERLAB. Afin de prendre du recul, nous avons réalisé une estimation directe des rendements et productions de chaque département à partir des parcelles TERLAB.

Pour ce faire, nous avons calculé un rendement moyen par département à partir des données TERLAB, puis nous avons multiplié ce rendement par la surface RPG totale pour le BTH dans chaque département. Nous obtenons ainsi des valeurs de production pour chaque département.

La figure 3.9 montre la comparaison entre les données de rendements et productions départementales de la SAA issue d'Agreste et le calcul simplifié à partir des données de l'enquête TERLAB décrit ci-dessus. Nous observons une correspondance presque parfaite. Ceci peut être dû à deux raisons non exclusives :

— l'enquête TERLAB est de très bonne qualité, notamment à cause d'un échantillonnage très représentatif de la réalité ;

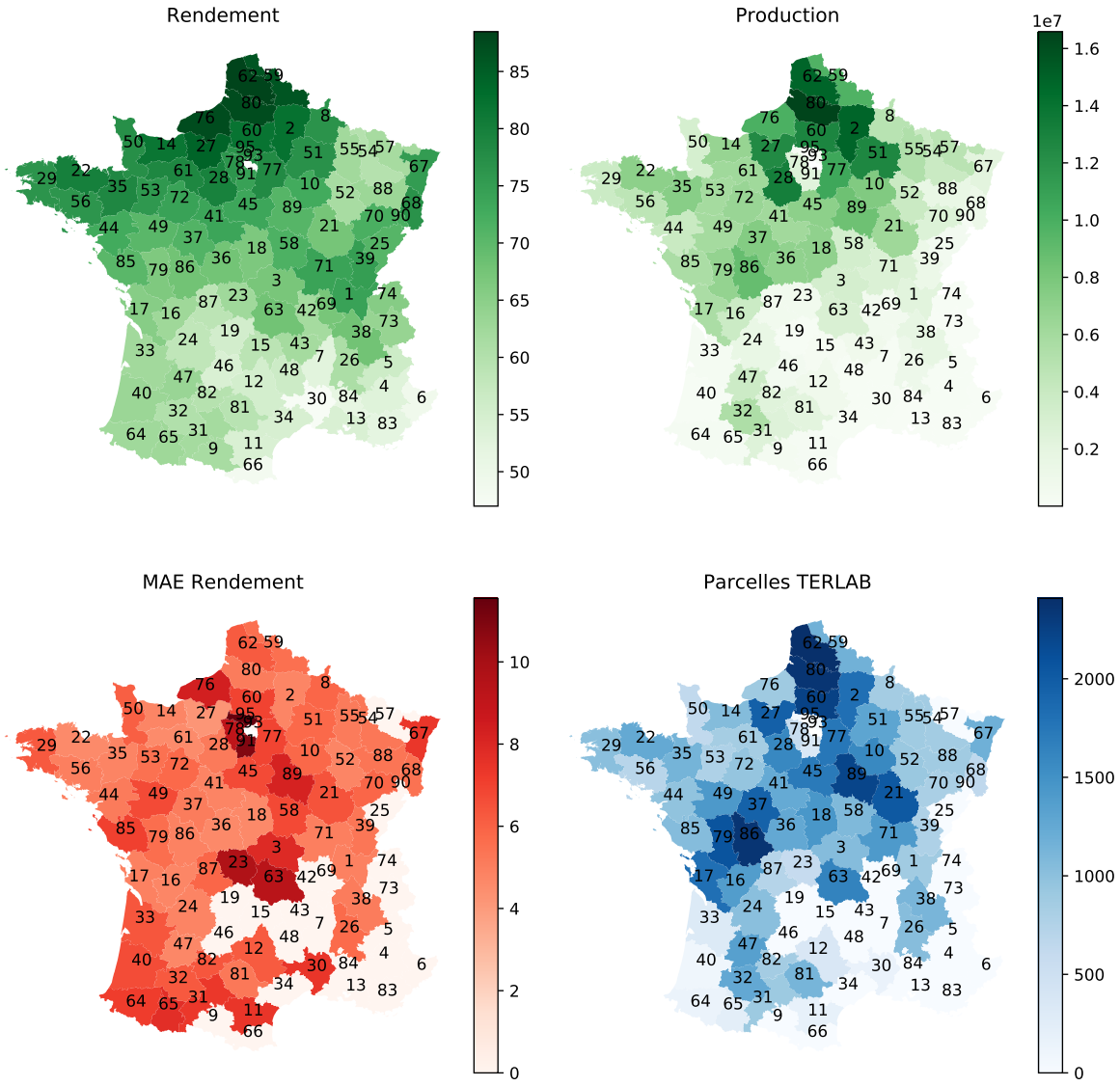


FIGURE 3.8 – Cartes des estimations de rendement et productions par département, MAE à la parcelle et nombre de parcelles TERLAB disponibles par département.

— le poids de TERLAB dans la SAA est très important.

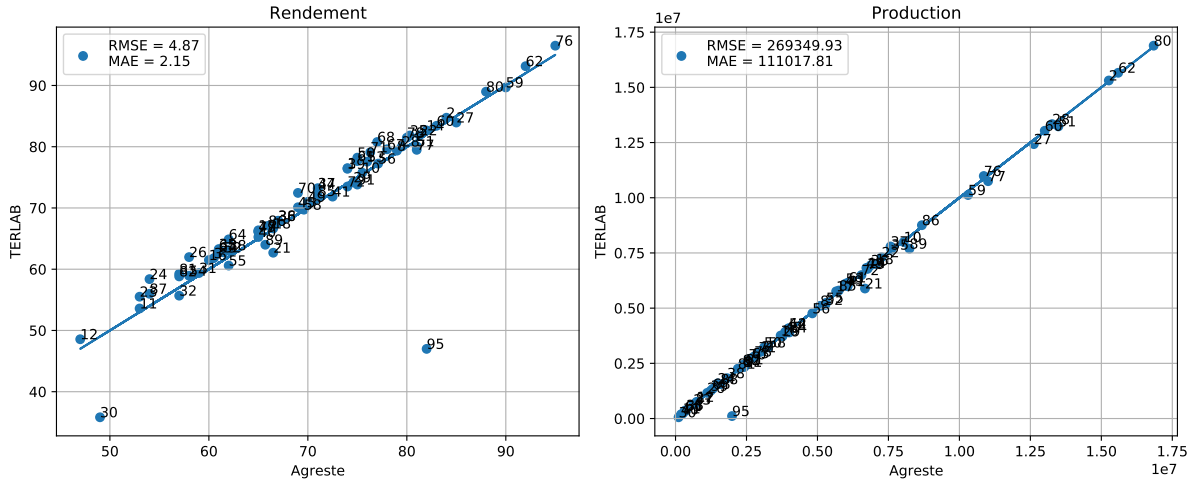


FIGURE 3.9 – Comparaison des estimations de rendements et productions entre la SAA issue d'Agreste et le calcul simplifié à partir de TERLAB.

Au regard de cette comparaison, on peut se demander l'utilité de l'estimation par satellite. Il est important de rappeler que l'estimation à partir d'imagerie spatiale permet d'avoir des rendements à la parcelle (voir illustration en figure 3.10) et donc de faire des statistiques à des niveaux plus fins que celui du département. Aussi, on peut envisager d'appliquer le modèle de régression dans des situations où l'enquête TERLAB ne serait pas encore disponible (nous abordons ce sujet dans 5.2.3).

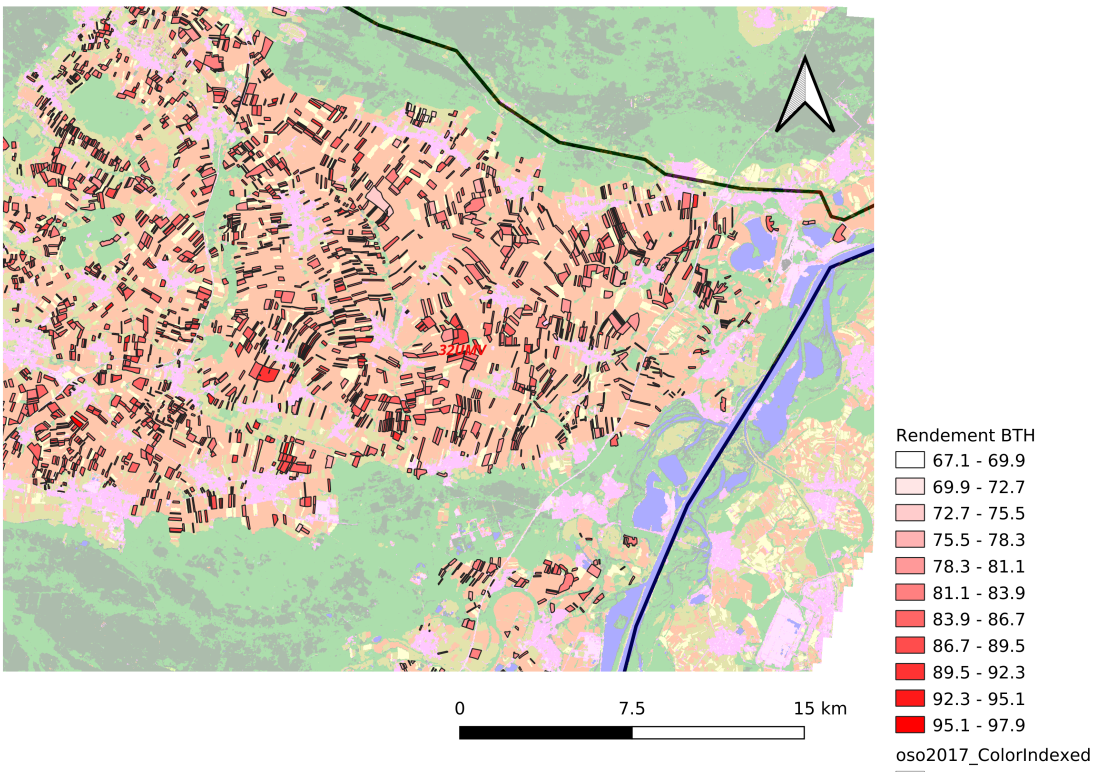


FIGURE 3.10 – Illustration de l'estimation du rendement à la parcelle.

# 4 Données et code

Pour réaliser cette étude, des données et du code ont été produits et sont mis à disposition. Les données permettront la réalisation d'analyses supplémentaires sans avoir besoin de traiter les images satellitaires. Le code permettra d'exploiter les données et pourra aussi servir dans une suite éventuelle de l'étude.

La convention de mise à disposition des données TERLAB par le MAA au CESBIO ne permet pas à ce dernier d'en faire une diffusion. Les données produites par le CESBIO seront donc fournies selon des modalités respectant ces contraintes.

De son côté, le code est une production 100% CESBIO et est distribué sous une licence de logiciel libre.

## 4.1 Données

### 4.1.1 Fichiers de statistiques zonales

Pour chacune des 77 tuiles il a été généré un fichier au format GeoJSON contenant, pour chaque parcelle de BTH, les statistiques des réflectances de surface, masques de validité, altitude, pente et exposition.

Ces fichiers permettent de réaliser la totalité de l'étude sans avoir à re-traiter les images Sentinel-2.

L'ensemble des fichiers compressés correspond à un volume de 16 GO.

### 4.1.2 Le modèle de régression

Le résultat de l'apprentissage sur 2/3 des parcelles TERLAB de BTH en utilisant les 80 prédicteurs sélectionnés est sauvegardé. Il s'agit d'un modèle Random Forest compatible avec la bibliothèque Scikit-learn. Il pèse 1.1 GO.

Ce modèle peut être utilisé pour produire des estimations sur d'autres années et ainsi évaluer la transférabilité temporelle.

Le modèle est aussi accompagné d'un fichier CSV avec les estimations de rendement sur les parcelles de validation (1/3 des parcelles TERLAB).

### 4.1.3 Fichiers de résultats

L'ensemble des estimations pour toutes les parcelles de BTH a aussi été conservé suivant 2 découpages : par tuile (la sortie des estimations) et par département (réorganisation pour les analyses de ce document).

Ces fichiers sont fournis au format ESRI Shapefile et contiennent pour chaque parcelle les attributs suivants issus des fichiers TERLAB `CODE_CULTU`, `SURF_ADM`, `SURUTISOL`, `BIO`, `MMEAU`, `RENDNORME`. Nous y ajoutons de code du département `DPT` et l'estimation du rendement `RENDPRED`.

Les fichiers par tuile sont aussi fournis au format csv et contiennent aussi les prédicteurs, ce qui permettrait de faire des apprentissages avec d'autres algorithmes de façon aisée.

Enfin, un fichier shapefile des départements avec les données nécessaires aux analyses de la section 3.2.2 (estimations de rendements et productions par département, données SAA d'Agreste, etc.) est aussi fourni.

Le volume total des fichiers de résultats est inférieur à 4 GO.

## 4.2 Code

Le code et les documents associés à l'étude sont disponibles au téléchargement [1]. Ce dépôt contient l'arborescence suivante :

- `docs` : les fichiers associés au présent document ;
- `python` : bibliothèque de traitements pour la production et les analyses des données ;
- `jobs` : fichiers spécifiques au lancement des traitements sur le cluster HPC du CNES ;
- `notebooks` : documents d'analyses exploratoires sans utilité à la fin de l'étude, mais conservés pour référence.

Des détails sur ces contenus sont en train d'être ajoutés au dépôt et pourront être enrichis par la suite.

1. `http://osr-cesbio.ups-tlse.fr/gitlab_cesbio/Jordi/rendementterlab/-/archive/public/rendementterlab-public.zip`

# 5 | Conclusions et perspectives

## 5.1 Conclusions générales

Ce travail avait comme objectif d'explorer la possibilité d'estimer le rendement des grandes cultures agricoles à partir d'imagerie satellitaire d'observation de la Terre.

Des séries temporelles d'imagerie optique à haute résolution (10 et 20 m) acquises avec une fréquence temporelle de 5 jours sur l'ensemble du territoire ont été utilisées pour estimer le rendement au niveau de la parcelle. Des modèles de régression non-linéaire peuvent être obtenus par apprentissage automatique en utilisant les données de l'enquête TERLAB (déclarations de rendement moyen des exploitations sur un échantillon) comme variable à estimer. Le parcellaire agricole et la culture de chaque parcelle sont supposés connus.

Malgré l'écart important qu'il peut exister entre le rendement moyen d'une exploitation (qui lui-même est approximatif) et le rendement réel de chaque parcelle de l'exploitation, nous avons fait l'hypothèse que la capacité de régularisation de l'algorithme d'apprentissage permet une convergence vers la valeur réelle. Ceci est possible par le fait qu'un grand nombre d'exploitations sont disponibles.

Dans un premier temps, nous avons montré qu'il est possible de construire des estimateurs locaux à des zones géographiques correspondant à un carré de 100 km de côté (découpage des données satellite) pour toutes les cultures. Les performances de ces modèles se dégradent quand ils sont appliqués à d'autres zones géographiques que celles où ils ont été appris. Ceci limite leur intérêt dans un objectif de statistique nationale sur les productions.

Dans un deuxième temps, nous avons construit un modèle applicable à tout le territoire métropolitain mais spécialisé pour une seule culture (le blé tendre d'hiver dans notre exemple). Nous avons constaté de très bonnes performances dan l'estimation des rendements avec des erreurs inférieures à 10 q/ha. Les résultats de ce modèle ont été comparés à la statistique agricole annuelle consolidée montrant une très bonne cohérence.

Les données produites et mises à disposition (rendements pour toutes les parcelles RPG) permettraient facilement de réaliser des agrégations à des niveaux géographiques plus fins. Le SSP pourrait évaluer ainsi d'autres usages de l'approche.

Cette étude ayant été réalisé sans cadre contractuel ni programmatique, elle est par na-

ture de portée limitée. Nous proposons ci dessus des perspectives de recherche pour aller plus loin dans la faisabilité du principe exposé. Nous donnons aussi des recommandations pour la mise en œuvre opérationnelle de ce type d'approches.

## 5.2 Perspectives de recherche

### 5.2.1 Extension à d'autres cultures

Nous avons montré que l'estimation à partir de données satellitaires fournit des valeurs très proches de celles de la SAA, mais l'étude s'est limitée à une seule culture qui est très présente sur le territoire.

Il faudrait consolider ces résultats en les appliquant à d'autres cultures. Il faudrait aborder dans un premier temps des cultures d'été pour lesquelles les différences entre les zones climatiques peuvent être importantes. Des différences pour une même région, notamment dans les zones du sud, peuvent aussi apparaître en raison des pratiques d'irrigation.

L'étude de cultures minoritaires serait aussi intéressante pour évaluer la capacité de la méthode à travailler avec peu de données d'apprentissage.

Enfin, réaliser un modèle par culture peut s'avérer fastidieux. La possibilité de produire des modèles de régression pour des groupes de cultures (par exemple les céréales à paille d'hiver) serait intéressante pour une simplification du passage à l'opérationnel.

### 5.2.2 Variables et algorithmes

Le choix des statistiques temporelles de quelques indices de végétation a été fait sans comparaison à d'autres possibilités. D'autres choix pourraient aboutir à des meilleures performances ou à un nombre réduit de prédicteurs. Ceci pourrait aussi permettre d'utiliser moins de données en entrée de la procédure et donc de réduire les temps de calcul et les besoins de stockage.

Sur un plan différent, des variables non issues de l'imagerie satellitaire pourraient aussi être utilisées. Des données climatiques ou météorologiques pourraient affiner les estimations (même si nous avons fait l'hypothèse que toute l'information nécessaire pour prédire le rendement est contenue dans l'évolution temporelle de l'activité de la végétation).

Pour ce qui concerne les algorithmes de régression aussi, le choix a été fait par l'expérience du CESBIO dans le domaine de la cartographie de l'occupation des sols. Le traitement des cas d'autres cultures, la possibilité de réutilisation des modèles appris sur d'autres années (voir ci dessous) ou d'autres cas d'usage pourraient nécessiter des algorithmes offrant d'autres caractéristiques.

### 5.2.3 Simplification de l'enquête TERLAB

Dans l'estimation du rendement du blé tendre d'hiver, nous avons fait le choix d'utiliser 2/3 des parcelles avec information de rendement pour calibrer le modèle.

Dans un éventuel objectif de simplification de l'enquête TERLAB, il serait utile d'évaluer des plans d'échantillonnage dégradés. On pourrait imaginer d'identifier les exploitations ou les départements les moins utiles pour la calibration du modèle.

Ce type d'échantillonnage peut être simulé avec les données déjà disponibles.

Un autre aspect intéressant est celui de savoir si pour l'apprentissage du modèle il serait préférable d'avoir des rendements réels à la parcelle (par opposition à la moyenne de l'exploitation), même si cela implique d'échantillonner moins d'exploitations. Les données TERLAB disponibles ne permettent pas de simuler cette configuration.

Une autre façon de simplifier l'enquête serait de ne pas échantillonner toutes les parcelles chaque année. Cela demande de se poser la question de la transférabilité des modèles de régression entre années. Un modèle appris sur une année pourrait avoir des performances dégradées quand il est appliqué à des images d'une autre année avec des conditions climatiques différentes. Ceci est typiquement le cas de la classification du type de culture. Dans le cas de l'estimation du rendement, il se pourrait que les modèles soient plus robustes et ce d'autant plus qu'ils sont calibrés sur l'ensemble du territoire et donc sur des régions climatiques variées.

### 5.2.4 Estimation en cours de saison

Dans ce travail, nous avons utilisé toutes les images disponibles sur l'année civile, ce qui implique que les estimations ne peuvent être réalisées que quand toutes les images ont été acquises.

Étant donné le calendrier agricole, il est sans doute possible de réduire le délai à la fin du mois d'octobre, mais on peut aussi envisager des estimations par paliers en fonction des cultures (les cultures d'hiver en juin, etc.).

Au delà de la connaissance a priori du calendrier agricole, on peut imaginer aussi que le rendement peut être estimé un certain temps avant la récolte.

Les données disponibles permettent d'étudier les incertitudes dans l'estimation du rendement en cours de saison. Cependant, étant donné que nous travaillons par apprentissage supervisé, les données TERLAB de l'année en cours ne seraient pas disponibles. Il faudrait donc d'abord avoir validé la transférabilité des modèles entre années évoquée ci-dessus.

Cette approche, permettrait de se rapprocher des informations de conjoncture présentées en 1.1.

## 5.3 Passage à l'opérationnel

La procédure mise en œuvre pour cette étude peut être déployée pour une production opérationnelle. Les données satellitaire en entrée sont accessibles gratuitement. Le code de cette étude est aussi disponible et, même si certaines améliorations sont nécessaires, il a servi à traiter un volume de données équivalent à celui nécessaire pour une production réelle dans un cadre opérationnel.

Cependant, cette étude a été rendue possible par le fait que le CESBIO a accès au centre de calcul du CNES. Deux éléments ont été déterminants pour la réalisation de l'étude :

— les ressources de calcul et de stockage ;

— la disponibilité des données satellite.

Pour ce qui concerne les ressources informatiques, seule l'étape de génération des statistiques de chaque parcelle nécessite des ressources de calcul importantes. En effet, pour chaque tuile Sentinel-2, de l'ordre 300 GO de données sont manipulés, et ceci pour 77 tuiles. Ce traitement peut être parallélisé au niveau de chaque tuile. Le traitement d'une tuile peut prendre jusqu'à 48h sur une machine avec 24 CPU et 120 GO de RAM. Sur le centre de calcul du CNES, nous avons pu traiter beaucoup de tuiles en parallèle, ce qui fait que le temps total de traitement pour cette étape est inférieur à la semaine.

Les étapes d'apprentissage et d'estimation sont beaucoup moins gourmandes et la production d'une tuile pour une culture nécessite entre quelques minutes et moins de 5h. Cette étape peut aussi être parallélisée au niveau des tuiles.

On voit donc qu'avec des ressources bien inférieures à celles du centre de calcul du CNES le traitement est possible et seulement la durée est affectée si le nombre de noeuds de calcul est limité.

D'un autre côté, les données satellite doivent être disponibles sur l'infrastructure de calcul utilisée. Le problème ne réside pas dans le stockage dont les coûts deviennent de plus en plus faibles. La difficulté vient de la nécessité de transférer les données sur l'infrastructure hôte. On parle ici de quelques dizaines de TO à transférer entre le serveur de distribution et la plate-forme de calcul. Nous avons pu bénéficier du fait que le CNES héberge un miroir des données Sentinel.

Pour un déploiement opérationnel de cette approche il faut donc envisager d'implanter les calculs sur une infrastructure hébergeant les données Sentinel. Il pourrait s'agir du centre de calcul du CNES, via la plate-forme PEPS ou bien via un des DIAS, les plate-formes Copernicus basées sur le cloud offrant un accès centralisé aux données et informations Copernicus, ainsi qu'aux outils de traitement.

# Bibliographie

[1] J.G.P.W. Clevers and A.A. Gitelson. Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on sentinel-2 and -3. *International Journal of Applied Earth Observation and Geoinformation*, 23 :344–351, Aug 2013.

[2] Jordi Inglada, Arthur Vincent, Marcela Arias, Benjamin Tardy, David Morin, and Isabel Rodes. Operational high resolution land cover map production at the country scale using satellite image time series. *Remote Sensing*, 9(1) :95, 2017.

[3] Leo Breiman. Random forests. *Machine learning*, 45(1) :5–32, 2001.