**Hugo SENGES**

**SOAD**  
Variables numériques en appui aux métiers de L’OFB

Vincennes, le 25 Octobre 2021

Josselin Barry ,

Technicien supérieur à l’OFB

Constitution d’une donnée d’entraînement basée sur l’observation terrain à destination de la télédétection des CIPAN

Il est question dans cette note du cadrage et de la définition d’une mission d’acquisition de données de vérité-terrain, dans le cadre d’un projet plus large de développement de nouvelles chaînes algorithmiques en télédétection sur la thématique des CIPAN. L’objet principal est la capitalisation des expériences précédentes, l’identification des contraintes liées à l’environnement de développement et les bonnes pratiques correspondantes à adopter.

1. Contexte et objectif

Dans le cadre de l’activité d’ajout d’une chaîne de traitement dans le package INRAE SISPPEO (dédié à la télédétection des CIPAN[[1]](#footnote-1)), il est nécessaire de disposer d’une vérité-terrain actualisée. Celle-ci prendra la forme d’une donnée sémantique, issue d’un repérage sur le terrain dans les départements choisis (Ille-et-Vilaine et Morbihan). Elle devra caractériser le type de couverture du sol ainsi que la qualité de celle-ci, et être rattachable à la maille d’entraînement, donnée vectorielle dérivée des données annualisées « Registre Parcellaire Graphique » mis à disposition publiquement par l’IGN[[2]](#footnote-2) (GeoServices). Se basant sur l’AMI précédent (Intelligence Artificielle en appui à l’élaboration des plans de contrôle Eau & Nature), l’AMI « Variables Numériques » s’inscrit dans un environnement de développement établi. Ainsi les méthodes de collecte, tuilage et traitement des images Sentinel sont déjà implémentées et validées au sein du package python SISSPEO, hébergé sur les serveurs de l’INRAE. C’est dans ce cadre qu’il s’agit de développer une nouvelle branche de ce projet, qui permette l’élaboration d’un outil statistique d’aide à la décision sur la thématique des CIPAN. Cette activité se rapporte à l’objectif 1.1, inclue dans le document Descriptif de phase 1 du projet :

**Objectif 1.1. Mise à disposition d’une API de résultats statistiques aux mailles de référence INPN, sur la thématique du type de couverture de sol sur les parcelles agricoles mentionnées au RPG.**

**1.1.1. Accès aux images satellitaires et préparation des données issues du registre parcellaire graphique** :

Correspond à la première étape, mais aussi à la préparation des données nécessaires aux étapes suivantes (notamment RPG). Ce sous-objectif poursuit la collecte des données initiée lors de la *phase 0*. *Cette phase est d’autant plus importante que les sources d’images satellitaires sont appelées à évoluer, en prenant en compte les améliorations attendues (nouveaux programmes satellitaires) mais aussi la durée de vie (7 ans et 3 mois) attendue des satellites Sentinel.*

**1.1.2. Consolidation des chaînes de traitement depuis les images satellitaires vers des résultats de classification et intégration à SISPPEO** :

Correspond à l’ensemble des groupes d’étapes ‘ Préparation des données ’ (hors première étape), ‘ Composition des bandes ’ et ‘ Classification et scores ’. Les classes de qualité de couverture de sol visées sont :

* Sol nu
* Végétation sèche
* Végétation mixte basse (sol sec / végétation peu couvrante)
* Végétation peu couvrante
* Végétation mixte couvrante (végétation peu couvrante / végétation couvrante)
* Végétation couvrante
* Végétation hétérogène
* Couverture inconnue

En plus de cette légende prévisionnelle, il sera également nécessaire de renseigner une seconde légende qui renvoie au type de couverture du sol observé. Dès lors, et pour des besoins d’inter-comparabilité, il semble pertinent de reprendre tout ou partie de la légende correspondante, implémentée dans le cadre du projet de télédétection via images Sentinel des CIPAN d’hiver, mené par l’Irstea (aujourd’hui INRAE) en 2019, sur la base de relevés de la DRAAF Occitanie et de la DDT du Gers.

Il s’agit donc, dans un premier temps, de renseigner une vérité terrain sur la base de cette typologie. Deux difficultés se présentent alors :

D’une part il faut pouvoir atteindre un niveau de représentativité géostatistique correct. D’autre part il est nécessaire de bien définir en amont les ‘seuils’ discriminant un poste de la légende choisie d’un autre. A ce titre il sera intéressant de se baser sur les expériences précédentes, menées depuis 2014 par différentes équipes sur la thématique de la télédétection des CIPAN d’hiver.

C’est sur la base de cette donnée que seront ensuite développées les chaînes de traitement en télédétection. En outre, l’avantage des données vectorielles RPG liées consiste en la mise à disposition d’un profil de culture, pour l’année du millésime et pour chaque parcelle (entité). La table attributaire correspondante est la suivante :

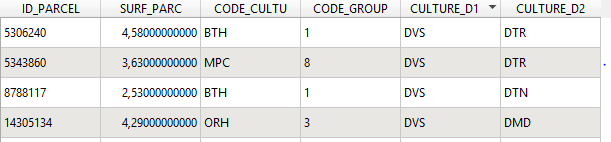


Figure 1: table attributaire d'un jeu de données RPG 2.0

Le code culture ainsi disponible permet de constituer une ressource de variables auxiliaires. Celles-ci ne peuvent se substituer à l’acquisition d’une « vérité-terrain » en bonne et due forme. Mais, dans une optique d’entraînement d’un programme de télédétection des CIPAN, elles fournissent des informations (sur une base déclarative) qui peuvent être confrontées dans un second temps avec les prévisions du modèle de classification des surfaces. De plus, la manipulation de données vectorielles ainsi renseignées permet une appréhension facilité des zones de cultures (possibilité de filtrer, trier, masquer les zones d’intérêt).

1. **Environnement de développement :**

Le développement des chaînes algorithmiques en télédétection devant se faire par enrichissement du package python dédié SISSPEO (développé par Arthur Coqué de l’INRAE en 2020), le format des données en entrée, ainsi que certains traitement prévisionnels doivent être réfléchis en amont, au regard des incompatibilités potentielles.

Le package se présente sous la forme d’un repository gitHub accessible en téléchargement et en écriture[[3]](#footnote-3). Développé en « orienté-objet », il se compose de plusieurs classes, une par algorithme dédié qui comporte elle-même deux méthodes obligatoires (\_\_init\_\_ et \_\_call\_\_). On retrouve également des classes dédiées à une tâche transversale particulière (‘reader’ pour la lecture des ressources satellites ; ‘L3AlgoProduct’ pour la composition d’images optiques, ‘Mask’ pour le masquage sélectif et dynamique des ressources…). Ainsi, l’ensemble de la chaîne d’acquisition-calibration-extraction-composition des ressources satellites est pris en charge au sein du package.

L’un des intérêts de ce package est la possibilité de recourir à une analyse inter-temporelle (très utile pour les cultures annuelles) via les objets ‘timeseries’ qui structurent des données satellites de dates différentes sur une même emprise.

Un bémol concerne l’absence (à tout le moins dans la documentation) de traitements dédiés aux ressources radar, notamment Sentinel-1, dans les fonctions de calibration et composition des rasters d’entraînement. Le développement de fonctionnalités dédiées pourrait alors varier fortement suivant les étapes considérées. Concrètement, l’algorithmie de télédétection proprement dite (entraînement et expansion d’un modèle de classification sur base de données (entre autres) radar) ne pose aps de difficultés particulières. En revanche cela demande d’intégrer de nouveaux outils (via des bibliothèques déjà existantes) pour la calibration et le *de-speckle* des données natives.

En somme, on peut envisager les traitements à venir sur la base d’une analyse préférentiellement optique et intégrant une forte dimension d’inter-temporalité à l’échelle du mois (temps de revisite des Sentinel 2 de 5 jours et de 12 jours pour les Sentinel 1). L’ajout de la ressource Radar, pourrait se faire dans un second temps, suite à la classification des types de couverts. Etant donné l’apport majeur de cette donnée sur la texture des surface, cela pourrait permettre d’accroître la précision du modèle de classification après disjonction des surfaces monitorées sur la base de leur ‘famille’ de type de couvert. Une telle analyse en deux temps et en disjonction des cas est présente dans la démarche de l’expérimentation TETIS/ONEMA (GOEURY & BARRY) menée sur l’hiver 2014-2015.

1. **Expériences précédentes :**

L’on peut se fonder dans notre développement sur les approches similaires précédentes qui commence à constituer, sur la thématique de la télédétection des CIPAN, un corpus robuste :

En 2012 ont lieux les premiers travaux de l’UMR TETIS sur le projet  ‘CIPAN / Télédétection’, à la demande de services de l’Etat (DDT des Deux-Sèvres) à des fins de contrôle liés à la Directive Nitrate.

Une nouvelle expérimentation réunissant TETIS et ONEMA (GOEURY & BARRY) est menée sur l’hiver 2014-2015.   
Le cadre conceptuel est en grande partie transposable, bien que les données actuelles permettent de meilleures précisions qu’à l’époque, avec des données Landsat 8 (30m d’emprise) et des données RPG 1.0[[4]](#footnote-4). Comme cela est nécessaire aujourd’hui, les vérités-terrain avaient été acquise sur le moment.   
Le projet devait répondre à un besoin d’aide aux contrôles sur la directive nitrates, et il proposait une approche par facteurs de risques (renseigné sur le type de culture en T et en T-1, la surface de la parcelle, reliquats d’azotes liés aux pratiques culturales, proximité aux masses d’eau sensibles…). L’étude de la corrélation de cet indicateur avec le type de couvert d’une part et l’homogénéité de celui-ci d’autre part (5 classes ‘type’ et 7 classes d’homogénéité) offrait la perspective d’une dimension prédictive (ou probabiliste) dont il n’est pas encore question dans le projet actuel. L’objectif étant de discriminer les seuils de risques par type de couvert détecté , cela permettait d’optimiser l’efficacité du ciblage pour les contrôles.

En 2017, suite à la sollicitation d’Envylis pour un projet de recherche collaborative et pluridisciplinaire (PollDiff Captage) centré sur l’appui aux acteurs locaux engagés dans la préservation des eaux souterraines dans les périmètres de captage, TETIS est chargé du volet cartographie des couverts hivernaux par télédétection

De ce projet on peut retenir la proposition qui y est faite d’indicateurs dérivés des données optiques, à composer dans la ressource de classification :

Bandes et indices à concaténer dans l’image :

* NDVI (B08 -B04) / (B08 + B04)
* NDWI (B08 -B11) / (B08 + B11)
* EVI (2.5 \* (B08 -B04) / ((B08 + 6.0 \* B04 -7.5 \* B02) + 1.0)
* NDTI (B11 -B12) / (B11 + B12)

Il serait à ce titre intéressant de tester la transposabilité de cette méthode "couverts hivernaux par télédétection".

Enfin, l’Irstea (aujourd’hui INRAE) a mené en 2020 un projet de classification des séries temporelles Sentinel S1 (radar) et S2 (optique) en utilisation les relevés terrain pour calibrer les algorithmes et évaluer la précision des résultats.

Réalisé grâce à des données vérité-terrain de la DDT (Gers) et de la DRAAF (Occitanie), auxquelles nous avons accès, ces résultats sont donc reproductibles. Le projet se proposait la détection de type de couvert (sol nu, maïs, CIPAN, colza, prairie, cultures permanentes, repousse, tournesol) et de la qualité du couvert (terre, couvert faible, moyen, fort, très fort), elle–même distinguée selon le type de couverture précédente (sols nus / autres couverts).

L’environnement technique était le suivant: Apprentissage supervisé (Algorithme Random Forest) pour classification au niveau pixellaire et vectoriel (basé sur le parcellaire), via le package Orfeo ToolBox.

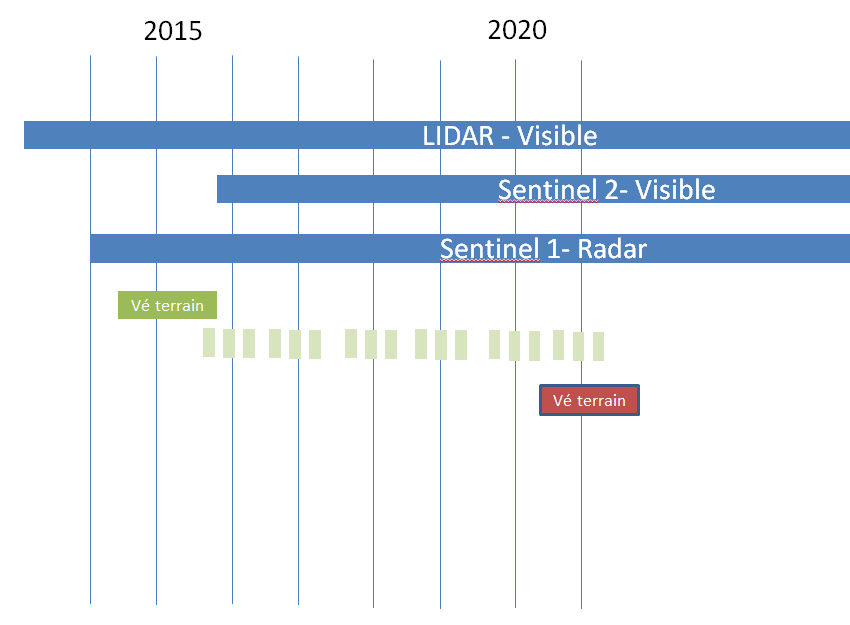
1. **Pistes de réflexion:**

* Application de la méthode à des années antérieures pour évaluer la progression des couverts hivernaux dans les pratiques agricoles, travail sur l’évolution du NDVI (indice de végétation) en lien avec l’activité photosynthétique de la végétation (suivi sécheresse).
* Récolte de la vérité-terrain via Qfield (basés sur le RPG et des images SENTINEL- 2 de 2021) : à 2 périodes (de la mi à la fin octobre et de la mi-décembre à fin décembre ou début janvier).
  + Récolte de données de terrain ponctuelles dont on se sert pour enrichir la table attributaire des parcelles RPG par jointure spatiale.
* On procède ensuite, sur les parcelles renseignées, à une érosion des parcelles sur 10m (pour supprimer les effets de bords).
* Puis on partitionne les polygones de vérités terrains en différents jeux, dédiés à l’entraînement et à l’expansion (de façon mutuellement exclusive). On proposera ici une répartition des polygones a priori : jeux d’entrainement (70%), jeux d’expansion (30%)
* Pour les besoins de l’allègement du temps de calcul, on peut envisager une démarche exploratoire, visant à limiter le transfert des informations pixellaires au segment selon une sélection aléatoire stratifiée de 10000 pixels max. Si les temps de calculs ne sont pas prohibitifs, l’on peut penser à effectuer l’entraînement sur la totalité des pixels contenus dans les segments test.
* L’appariement de la classe prédite aux pixels au parcellaire RPG par vote majoritaire permet la régularisation des prédictions du modèle, mais induit une perte d’information potentielle intra-parcellaire.
* Production des indices de qualification (confusion matrix, precision, recall, f-score et overall accuracy) et comparaison avec les résultats obtenus par les projets précédents. En moyenne une OA comprise entre 0.7 et 0.8 est attendue pour les types de couvertures, et une OA supérieur à 0.8 est attendue pour la qualité des couverts.
* Il faudra également procéder à l’examen des successions de couverture entre les 2 classifications, notamment au vu de la faible amplitude entre les deux dates estimées de récolte de la vérité-terrain, via l’appréciation des répartitions de transition maximales de classes entre les deux dates. Soit un tableau à double entrée donnant pour type de couvert détecté en T1 le type de couvert le plus détecté sur les mêmes polygones en T2, et leur proportion dans cette population.
* On pourra enfin croiser avec les données de contrôle de la directive nitrates effectués par la police de l’environnement  (contrôles OSCEAN et OPALE)

1. **Axes de travail :**

On peut donc se proposer comme ordre du jour, pour la réunion du 28 Novembre 2021, l’évaluation des aspects suivants :

* Critères de représentativité géostatistique de la vérité-terrain, et méthodologie à implémentée en ce sens
* Définition des seuils entre les postes de qualité de couverture du sol
* Préparation logistique d’une mission de collecte de la vérité-terrain entre Novembre et Décembre de l’année 2021, sur les départements cibles (Morbihan et Ille-et-Vilaine)
* Réflexion sur la façon d’intégrer les données auxiliaires de vérités terrains (contrôles police de l’environnement) dans la démarche. A priori plutôt comme variable auxiliaire que comme substitut de la vérité-terrain.
* Discuter de l’opportunité de d’intégrer des données radar dans la démarche, au vu des apports, des contraintes de développement et de l’inter comparabilité induite d’avec les précédentes vérités-terrains (tel que représenté sur le schéma suivant)



1. **Protocole de récolte de vérité-terrain :**
   1. **Calendrier et Zonage**

La récolte d’un premier jeu de vérité-terrain est prévue pour la semaine 48, du 29 Novembre au 3 Décembre 2021. Elle sera assurée par Josselin Barry et Hugo Senges.

Dans une perspective d’analyse diachronique, il serait très utile de pouvoir ensuite retourner sur les parcelles renseignées en premier lieu à la fin du mois de Décembre, ou bien début Janvier. Toutefois, en fonction de l’opportunité d’un tel protocole, et de la disponibilité des agents (notamment, Hugo Senges dont le contrat s’achève à la fin de l’année), une adaptation de cette démarche demandera une concertation ultérieure.

Concernant le zonage, cela dépendra de la localité choisie pour le logement des agents. Toutefois une fois cette information connue, l’extraction des parcelles RPG dans un rayon de 100km autour de ce point sera aisée. Au sein de ce nouvel ensemble on peut également filtrer en fonction d’un itinéraire choisi. Il est même possible de maximiser un indicateur comme le nombre de parcelles « renseignables » par km parcouru afin de déterminer rationnellement cet itinéraire.

Visuellement on peut s’appuyer sur les éléments de références ci-dessous pour appréhender la façon avec laquelle les agents se repèreront sur le terrain.

Figure 2 : Représentation du RPG superposé à une orthoimage. Ressource type sur laquelle visualiser la saise des observations terrains sur Qfield

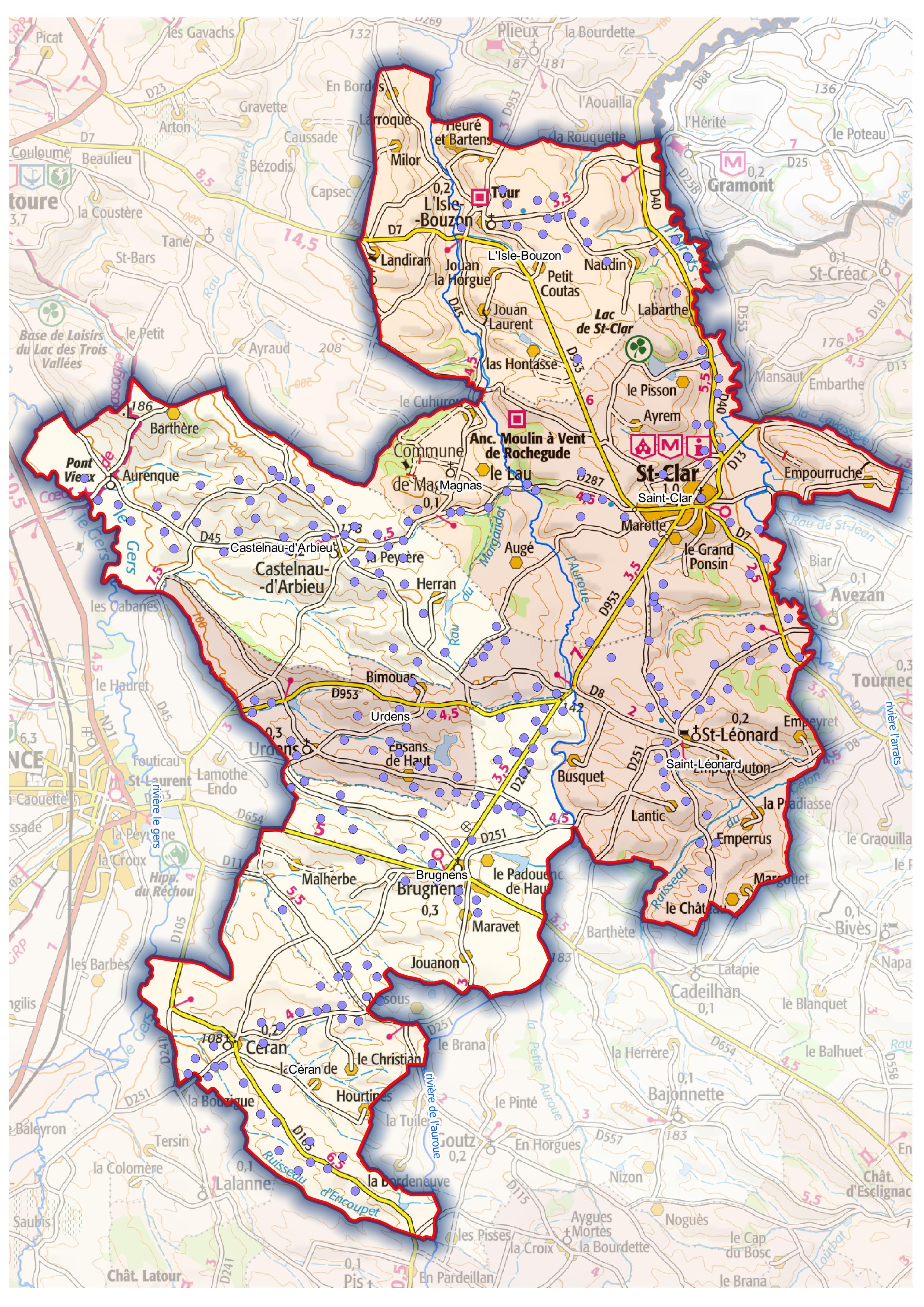


Figure 3 : Cartographie d'un itinéraire de collecte de vérité-terrain, issu du projet DRAAF Occitanie 2019

* 1. **Logistique**

A bord d’un véhicule les agents se rendront sur dans différentes communes de la zone investiguées et, sur la base d’un plan représentant les différentes parcelles RPG renseignées dans la zone, enregistreront leurs observations directement sur portable ou tablette, via l’application Qfield. Y sera chargé un projet superposant un fond de carte (OSM par exemple) et la donnée vectorielle du RPG de l’année 2019 (le 2020 n’étant toujours pas publié). Cette donnée vectorielle sera issue d’un sous-échantillonnage de la donnée départementale, et comportera des champs textes vides, dont le remplissage se fera selon les observations terrains et en accord avec la nomenclature choisie.

En fonction de la densité des champs dans les communes traversées, des conditions de trafic et de temps d’acquisition, deux ou trois jours d’observations devraient permettre d’acquérir un nombre suffisant de renseignements. Un critère devra être gardé à l’esprit qui touche à la diversité des couverts observés en vérité-terrain. A ce titre, la lecture régulière par les agents sur le terrain des statistiques d’effectifs de champs renseignés par poste de la nomenclature est importante.

L’acquisition proprement dite, se fera préférentiellement hors du véhicule, le plu près possible du champ investigué. Cela permettra de réduire au maximum les incertitudes liées à la faible levée des couverts végétaux et de réduire les indécisions relatives au classement d’une surface entre deux postes distincts de la nomenclature (par exemple « végétation peu couvrante » et « végétation sèche »).

* 1. **Nomenclature(s)**

Dans une perspective de comparabilité de nos résultats avec les études déjà implémentées sur la question, l’adoption d’une nomenclature compatible avec ces résultats préexistants semble être un point important de notre méthode. Toutefois il ne sera pas possible de répondre exactement aux mêmes distinctions que présentées dans les méthodes de 2015 (TETIS-ONEMA) et de 2019 (IRSTEA-Envilys). La cause en est la trop grande ambiguïté introduite par le croisement de leurs nomenclatures respectives (impossible de distinguer ainsi la « végétation peu couvrante » de la méthode de 2015 et les couverts « autres » que « sols-nus » de qualité « faible » dans la méthode de 2019).

Ainsi c’est une approche proche mais propre qui sera adoptée ici, en reprenant le principe de la double qualification des surfaces. D’une part une nomenclature de couverture du sol traditionnelle. EN fonction des couverts rencontrés en vérité-terrain certains postes peuvent être amenés à évoluer mais l’on peut partir à cet endroit de la nomenclature proposée pour la couverture du sol dans la méthodologie de 2015 :

|  |  |  |
| --- | --- | --- |
| **Couverture du sol** | | |
| **Poste** | **Observation** | **Code** |
| Maïs | Probablement sous forme de végétation sèche issue de la fauche | Mai |
| Céréales | Céréales d’hiver t faiblement couvrante dans la période | Cer |
| Tournesol | Probablement relativement proche du Maïs | Tou |
| Colza | A distinguer des pois | Col |
| Pois de printemps | A distinguer du colza | Poi |
| Prairies | Poste le plus simple | Pra |
| Sol nu | Ambiguïtés à lever par approche diachronique | Sol |
| Autres | A détailler ultérieurement | Aut |

Dans un second temps, un autre attribut sera donné à la surface en fonction de sa qualité de couvert. Cette seconde nomenclature s’appuiera plutôt sur la méthodologie de 2019, et sera donc elle-même différenciée en fonction de la couverture préalablement identifiée. La qualification du couvert se fera donc de la façon suivante :

|  |  |  |
| --- | --- | --- |
| **Qualité du couvert – Sol Nu** | | |
| **Poste** | **Commentaire** | **Code** |
| Faible | Presque aucune végétation active.  Ambiguïté vis-à-vis du SN\_6 | SN\_1 |
| Moyen | Repousses spontanées éparses | SN\_2 |
| Fort | Repousses spontanées couvrantes. Poste très ambigu. | SN\_3 |
| Résidus enfouis | Ambiguïté vis-à-vis du SN\_6 | SN\_4 |
| Résidus non-enfouis | Relativement aisée à identifier | SN\_5 |
| Terre | Aucune végétation visible et terre labourée laissée à nue | SN\_6 |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_

|  |  |  |
| --- | --- | --- |
| **Qualité du couvert – Autres surfaces** | | |
| **Poste** | **Commentaire** | **Code** |
| Faible | Presque aucune végétation active.  Ambiguïté vis-à-vis du CO\_4 | CO\_1 |
| Moyen | Végétation peu couvrante | CO\_2 |
| Fort | A distinguer du poste précédent | CO\_3 |
| Résidus enfouis | Ambiguïté vis-à-vis du SN\_4 | CO\_4 |
| Résidus non-enfouis | Relativement aisée à identifier | CO\_5 |

La dimension diachronique de l’approche envisagée en classification et entraînement du modèle étant déterminante, il est envisageable que les ambiguïtés et indécisions dans la nature et la qualité d’un couvert observé soient levées en deux temps, de la façon suivante. Les postes objets d’ambigüités sont tous renseignés pour la surface considérée (soit par concaténation des codes dans la même colonne, soit par ajout de colonnes auxiliaires). L’arbitrage se fera rétrospectivement lors de la seconde acquisition en vérité-terrain.

* 1. **Critères de réussite**

Considérant les méthodologies préexistantes, deux critères de réussite de la collecte de vérité-terrain peuvent être ici présentés.

Premièrement, la représentativité statistique, obtenue à travers la l’extension dans l’espace et le nombre des parcelles renseignées. En croisant les données raffinées du RPG avec les variables auxiliaires issues des contrôles de police OSCEAN, on peut espérer pouvoir faciliter l’étape de généralisation (ou expansion du modèle) et donc abaisser en entrée le nombre de parcelles renseignées en vérité-terrain.

En tablant sur une observation collectée toutes les 5min en moyenne, sur 8 heures ouvrées et 2 jours pleins de collectes, 250 enregistrements semblent être un objectif atteignable.

En deuxième lieu, on se proposera un critère de diversité du jeu de données ainsi constituées. Cette diversité est forcément pluridimensionnelle. On y trouve une dimension de diversité spatiale, pour laquelle c’est la bonne répartition des parcelles renseignées au sein de leur enveloppe minimale ainsi que la maximalisation de l’aire de cette enveloppe qui comptent. On trouve également une diversité sémantique, qui renvoie à la bonne représentation de chaque poste des nomenclatures dans la donnée de vérité-terrain. Le seuil d’effectif de parcelles varie en fonction des postes et de la nomenclature considérées. Sur 8 postes de la nomenclature « couverture », il serait intéressant d’acquérir plus de points en « Sol nu » que 1/8ème des observations. En effet cela nous permettra dans un second temps de satisfaire la représentativité de la nomenclature « qualité – sol nu ». On peut donc tabler, en première approximation, sur environ une 60 d’observation en « sols nus » au minimum. Cela laisserait alors une trentaine d’observation par autres postes (en ordre de grandeur). Au sein des nomenclatures « qualité de la couverture », les seuils minimums de représentativité seront abaissé en proportion. On peut estimer qu’une dizaine d’observations par postes sur la nomenclature « sol nu » serait satisfaisante, pour une vingtaine dans la nomenclature « autres couvertures »

1. Cultures Intermédiaires Piège à Nitrate [↑](#footnote-ref-1)
2. Voir l’adresse : https://geoservices.ign.fr/services-web-experts-agriculture [↑](#footnote-ref-2)
3. Voir https://github.com/inrae/SISPPEO [↑](#footnote-ref-3)
4. soit à l’Ilôt de culture et pas à la parcelle [↑](#footnote-ref-4)