La qualité des données

Line Boulonne, Bertrand Laroche, Céline Ratié, Nicolas N. Saby

La qualité des données

Line Boulonne, Bertrand Laroche, Céline Ratié, Nicolas Saby

NOM DE L'AUTEUR

JOUR / MOIS / ANNEE

La qualité des données

Les missions du GIS Sol sont :
d'améliorer la connaissance et la surveillance des sols de France
et de capitaliser les analyses de sols

La capitalisation de données sur les sols ne doit pas se faire au détriment de leur qualité

→ mise en place d'un certain nombre d'étapes ou de stratégie pour s'assurer de la qualité des données stockées dans la base de données

Avec plusieurs objectifs :

- Mettre à disposition une donnée de qualité
- Les utilisateurs disposent d'une information sur le contexte d'acquisition et d'intégration au SI.
- Crédibilité des données fournies

Les contrôles sont effectués :

- A toutes les étapes; de l'acquisition à l'intégration des informations au SI,
- Ces étapes sont différentes suivant le programme dans lequel les données ont été acquises.

Contrôle dans la collecte des données

IGCS	RMQS
Rédaction de documents (cahiers des charges, documents qualités, protocoles, documents explicatifs …)	
Choix des partenaires (compétences, crédibilité locale …)	
Aide au montage du programme (calendrier, objectifs, ..)	Mise en place de conventions
Formation des partenaires (formation à l'utilisation de DoneSol) http://www.gissol.fr/actualite/formation.php	
	Formation sur le terrain (Protocole RMQS (échantillonnage, observations)
Suivis techniques (comités de pilotage, échanges réguliers, …)	Un contrôle du travail de l'équipe locale sur site par un opérateur d'InfoSol, effectué tous les 5 à 10 sites

Collecte des données : les données analytiques

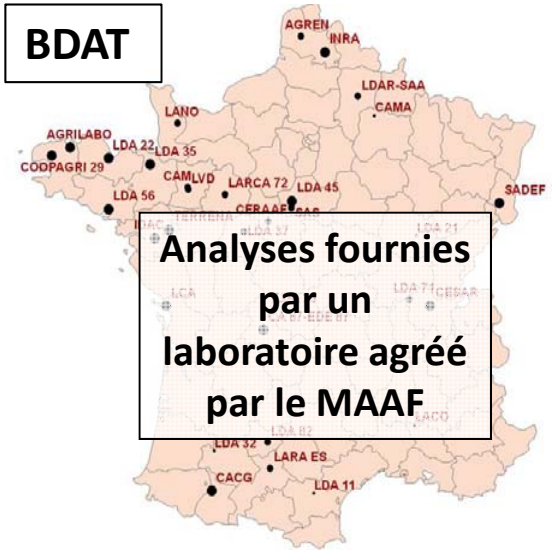
RMQS

Un seul laboratoire : LAS (Laboratoire d'Analyses des Sols (LAS) : laboratoire central pour les analyses de sol de l'INRA)

IGCS

Chaque région collabore avec son laboratoire (appel d'offre, ..) -> seule contrainte laboratoire agréé

BDAT



Validation et expertise des données

Harmonisation des formats de stockages
Harmonisation des unités
Vérification des méthodes de mesure
Présence de doublons
Absence/présence des données
Identification correcte d'une commune

Intégration dans la BDAT

Les laboratoires effectuent des tests de cohérences sur les résultats d'un même échantillon
3,5 < **pH eau** < 9 ou encore **$CaCO_3$ total** < 1,5 g/kg *si* pH eau < 6,5

BDAT **RMQS** **IGCS**

Qualité des résultats analytiques (tous programmes)

Les laboratoires agréés

Dans le cahier des charges, les analyses des échantillons doivent être faites par des laboratoires agréés.
Un laboratoire est agréé par le Ministère de l'Agriculture :

- sous la condition d'utiliser des méthodes normalisées
- de se soumettre à un circuit de contrôle national

-> Ceci garantit de pouvoir comparer les résultats d'un laboratoire à l'autre.

La liste des laboratoires agréés est disponible via les liens suivants :

http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000028316901&dateTexte&categorieLien=id

http://www.gemas.asso.fr/?accueil=laboratoires (site du GEMAS : Groupement d'Etudes Méthodologiques pour l'Analyse des Sols)

Chaque laboratoire fournit une incertitude sur les analyses -> prise en compte dans les calculs

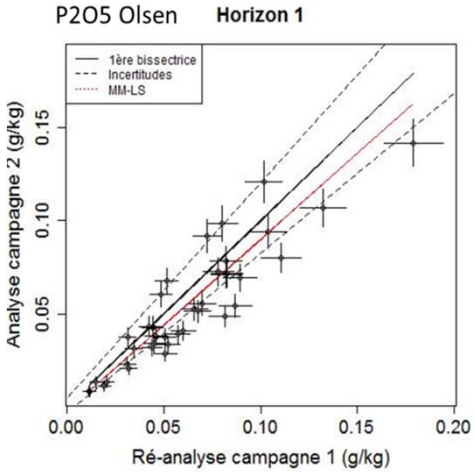
Aldana Jague, E. (2011). *Estimation des sources de variabilité du stock de carbone dans les sols du RMQS.* (Rapport de stage, Université de Tours).
Martin, M., Boulonne, L., Saby, N., Jolivet, C., Arrouays, D., Aldana Jague, E. (2012). *Estimating Sources Of Uncertainties Of Soil Organic Carbon Stock Assessments Within A Soil Monitoring Network*. Presented at 4th International Congress of the European Soil Science Societies Eurosoil 2012, Bari, ITA (2012-07-02 - 2012-07-06).
BOUKIR, H. (2012). *Estimation des sources d'erreurs sur la mesure des teneurs en contaminants dans les sols du Réseau de Mesure de Qualité des Sols par analyse de variance hiérarchique* (Rapport de stage, Université d'Orléans).

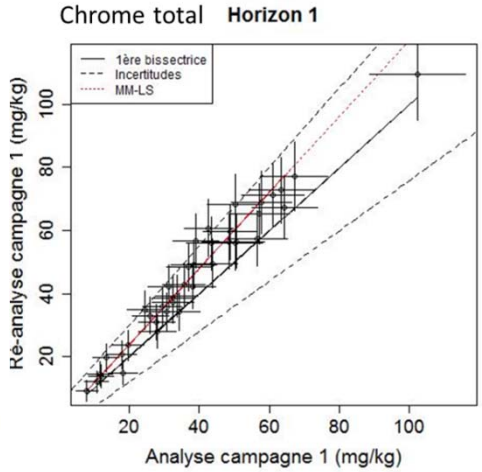
Qualité des résultats analytiques

Evolution des données analytiques avec le temps

RMQS

- Des travaux sont conduits avant d'identifier s'il y a une variation significative ou l'absence de variation des propriétés du sol dans le temps -> campagne test en région Centre 2001/2004 - 2010





La détection d'une variation temporelle est indissociable de la prise en compte de l'incertitude analytique (voir ici l'exemple du P2O5 Olsen)

un biais systématique a été détecté sur le Chrome, lié à un léger changement dans la méthode non signalé par le LAS

Boulonne, L., et al (2014). Enseignements tirés de la comparaison des résultats analytiques entre deux campagnes du RMQS, en Région Centre. In: *12. Journées d'Etude des Sols : Le sol en héritage*. Présenté aux 12ème. Journées d'Etudes des Sols , Le Bourget du Lac, FRA (2014-06-30 - 2014-07-04).

BDAT

Les méthodes d'analyse sont normalisées par l'AFNOR et garantissent la répétabilité des mesures. Ainsi les résultats obtenus à cinq ans de distance sont comparables dans la même parcelle à condition d'utiliser la même méthode d'analyse.

Contrôle des données saisies

IGCS	RMQS
Rédaction de documents (cahiers des charges, documents qualités, protocoles, documents explicatifs …)	
Choix des partenaires (compétences, crédibilité locale …)	
Aide au montage du programme (calendrier, objectifs, ..)	Mise en place de conventions
Formation des partenaires (formation à l'utilisation de DoneSol) http://www.gissol.fr/actualite/formation.php	
	Formation sur le terrain (Protocole RMQS (échantillonnage, observations)
Suivis techniques (comités de pilotage, échanges réguliers, …)	Un contrôle du travail de l'équipe locale sur site par un opérateur d'InfoSol, effectué tous les 5 à 10 sites
Vérification des couches graphiques (topographie, ….)	Contrôle manuel du rapport d'échantillonnage de site afin de corriger des dérives éventuelles.
	Contrôle manuel de données hors Donesol (environnement végétation, pratiques, …)
DonesolWeb : contrôle à la saisie des données	
Vérifications sémantiques des données (manuellement et à l'aide de sivercoh)	
Corrections par les partenaires IGCS	Corrections InfoSol

Qualité des données saisies

Des contrôles manuels

Des tests supplémentaires en tenant compte d'autres variables ou entre plusieurs échantillons d'un même profil

Exemple : taux de carbone de la couche de surface est supérieur au taux de carbone de la couche de sub-surface

Des contrôles à l'interface, toutes les données saisies via DonesolWeb (RMQS, IGCS)

Avec des contrôles sur l'interface de saisie :

sur la typologie des donnés rentrées : numérique ou non

sur l'amplitude des valeurs pour certaines données

(ex pH, les valeurs doivent être comprises entre 1 et 14)

Sur la dépendance de champs par rapport aux autres

(ex une taille d'éléments grossiers ne peut être renseignée si l'abondance de ces éléments est de 0%)

Le traitement des données permet aussi d'identifier des anomalies à posteriori.

-> report des anomalies dans un outil central (bugzilla) pour corrections

Qualité des données saisies

SIVERCOH : un outil de vérification de la cohérence des données

Lancement de jeux de requêtes

Cohérence des données

- par rapport CCTG
- au sein d'une même table
  - si il y a un excès d'eau -> les champs concernant l'hydormophie doivent être renseignés
- Entre les tables
- par rapport à la classification choisie

Qualité des données saisies

SIVERCOH : un outil de vérification de la cohérence des données

Rapport n°211 25/08/2014 - 11:01:26
exécution du jeu Surfaciques composé de 178 requête(s)
par LAROCHE Bertrand - bertrand.laroche@orleans.inra.fr
avec les paramètres : no_etude = 32360
47 requête(s) en anomalie 434 ligne(s) anomalique(s)

n° Anomalie	id_ucs	no_ucs	id_uts	no_uts	id_strate	no_strate	Anomalie
1300379	16008	3108					Le nombre d'UTS annoncé dans UCS est # du nombre d'enregistrements dans L_UCS_UTS pour cette même UCS
1300661			22040				PROF_APPAR_MIN, MOY OU MAX de la strate 2 ne correspond pas a EPAIS_MIN, MOY ou MAX de la strate 1
1300662			22052				PROF_APPAR_MIN, MOY OU MAX de la strate 2 ne correspond pas a EPAIS_MIN, MOY ou MAX de la strate 1
1300663			25577				PROF_APPAR_MIN, MOY OU MAX de la strate 2 ne correspond pas a EPAIS_MIN, MOY ou MAX de la strate 1
1300664			25567				PROF_APPAR_MIN, MOY OU MAX de la strate 2 ne correspond pas a EPAIS_MIN, MOY ou MAX de la strate 1
1300665			22082			2	prof_appar_min de la strate N+1 ne peut etre < à prof_appar_min+epais_min de la strate N
1300666			22090			3	prof_appar_min de la strate N+1 ne peut etre < à prof_appar_min+epais_min de la strate N
1300667			22019			3	prof_appar_min de la strate N+1 ne peut etre < à prof_appar_min+epais_min de la strate N
1300668			22045			2	prof_appar_min de la strate N+1 ne peut etre < à prof_appar_min+epais_min de la strate N
1300669			22041			3	prof_appar_min de la strate N+1 ne peut etre < à prof_appar_min+epais_min de la strate N
1300670			22044			3	prof_appar_min de la strate N+1 ne peut etre < à prof_appar_min+epais_min de la strate N
1300713			25554	118		3	Cette strate ne présente pas de données quantitatives
1300714			25510	112		2	Cette strate ne présente pas de données quantitatives
1300724			22080	62			Pour chaque UTS, les strates doivent être numérotées de 1 à n avec pour la strate de surface le n° 1.
1300732			22019	1			Les UTS doivent être caractérisées par un profil (affectation dans l_profil_etude)
1300733			22020	2			Les UTS doivent être caractérisées par un profil (affectation dans l_profil_etude)
1300734			22021	3			Les UTS doivent être caractérisées par un profil (affectation dans l_profil_etude)
1300735			22022	4			Les UTS doivent être caractérisées par un profil (affectation dans l_profil_etude)
1300736			22024	6			Les UTS doivent être caractérisées par un profil (affectation dans l_profil_etude)

Gains de temps importants
Adaptable à d'autres bases de données
Suivi des justifications

Conclusions

L'unité est très impliquée dans la démarche qualité -> référentiel V2 INRA

- ✓ Mise en place d'un groupe qualité,
- ✓ Rédaction de documents qualité,

…

Un souci permanent

- ✓ De travailler en amont sur la qualité des informations : dérives analytiques, méthodologiques par des formations des personnels (à l'INRA et hors INRA), des documents explicatifs (notice, CC…),
- ✓ De prendre en compte les incertitudes sur les données (métadonnées),
- ✓ De tracer nos activités,
- ✓ D'insérer des informations de qualité,
- ✓ Et de fournir de la donnée de qualité.

Avec des points faibles

- ✓ Quantité et diversité de l'information
- ✓ Historique DoneSol important,
- ✓ Réseau de partenaires -> un contrôle impossible sur toutes les activités,
- ✓ nombreux opérateurs pour l'acquisition et la saisie,
- ✓ Une charge de travail très importante.