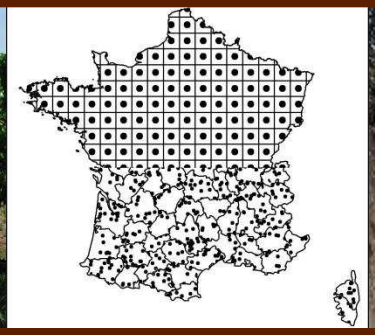
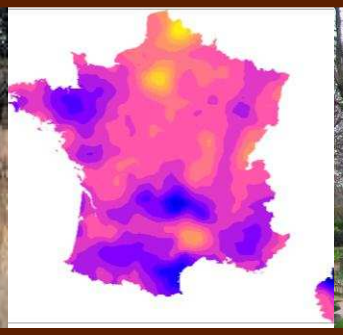
Modélisation statistique de la distribution spatio-temporelle des propriétés du sol pour la surveillance

Infosol & coll.




INRA

Groupement d'intérêt scientifique sol

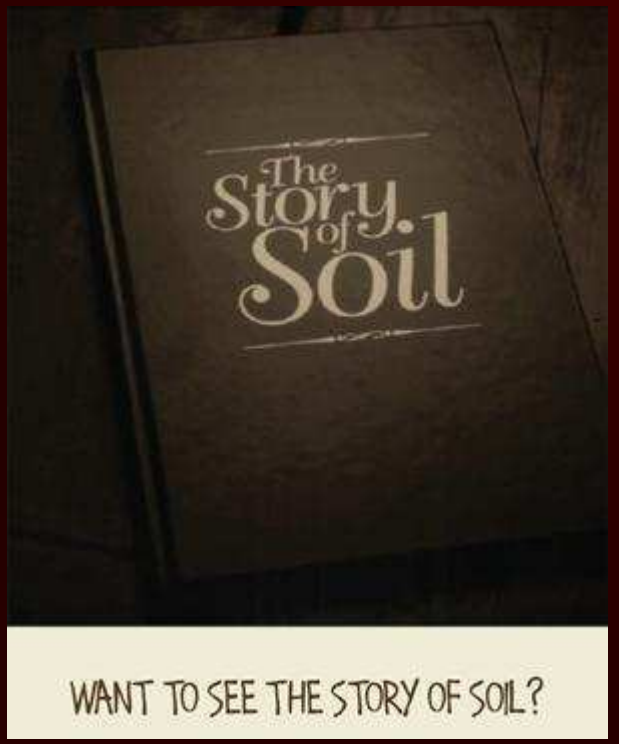
Les auteurs

- Nicolas Saby, Dominique Arrouays, Claudy Jolivet, Florent Millet, Estelle Villanneau : *Infosol, US 1106, Orléans, France*
- Ben Marchant, Murray Lark : *BGS , Nottingham, UK*
- Thomas Orton: *Université of Sydney, Australia*
- Didier Chauveau: *MAPMO Orléans; France*

Le sol

- Interface dans l'environnement
- Soumis à de fortes pressions
- Besoin de surveiller sa qualité
  - État ?
  - Évolution ?

http://www.iheartsoil.org/



La surveillance des sols en Europe

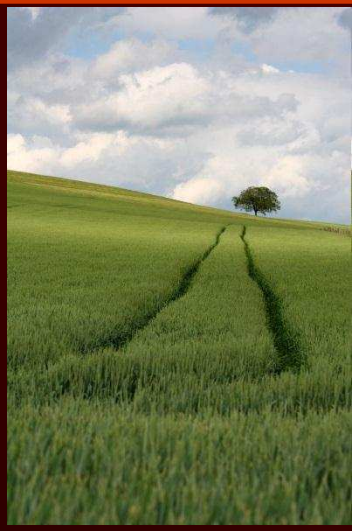
- Réutiliser (ex : France, Pays Bas)
- Re-prélever des sites (ex : Belgique)
- Développer des programmes de surveillance (tous les pays)
- Peu ou pas de recul temporel au niveau européen (exceptés l'Angleterre, le Pays de Galle et la Suisse)

La surveillance en France

LES DONNÉES

Base de Données d'Analyses de Terre

Réutiliser 

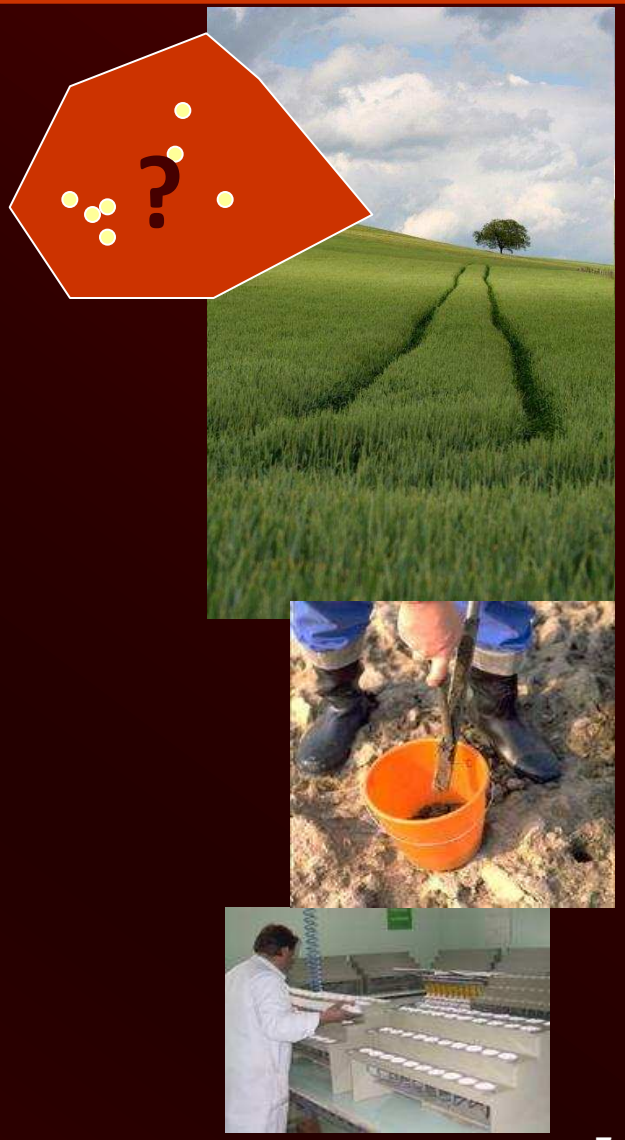


Analyses d'échantillon de surface d'une parcelle agricole	
Individus: Isolés, dispersés, atemporels, peu réutilisés	Base de données : Riche, temporelle, Spatialement distribuée
2 000 000 échantillons **25 000 000** déterminations élémentaires De 1990 à 2009	

Saby et al. 2004. Etude et Gestion des Sols.

La « stratégie » d'échantillonnage de la BDAT

- **Géoréférencement :** origine communale du prélèvement
- **Echantillonnage :** parcelles agricoles
- **Date :** année de prélèvement
- **variables :**
  - BDAT = pédologiques
  - BDETM = 8 ETMs

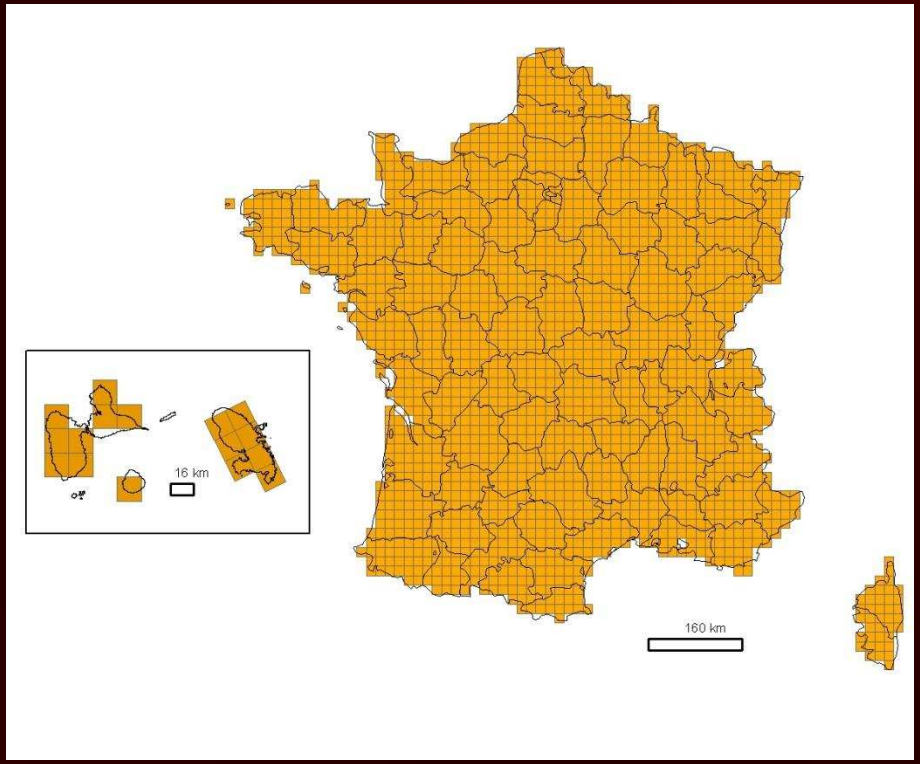


Saby et al. 2004. Etude et Gestion des Sols.

Réseau de Mesures de la Qualité des Sols

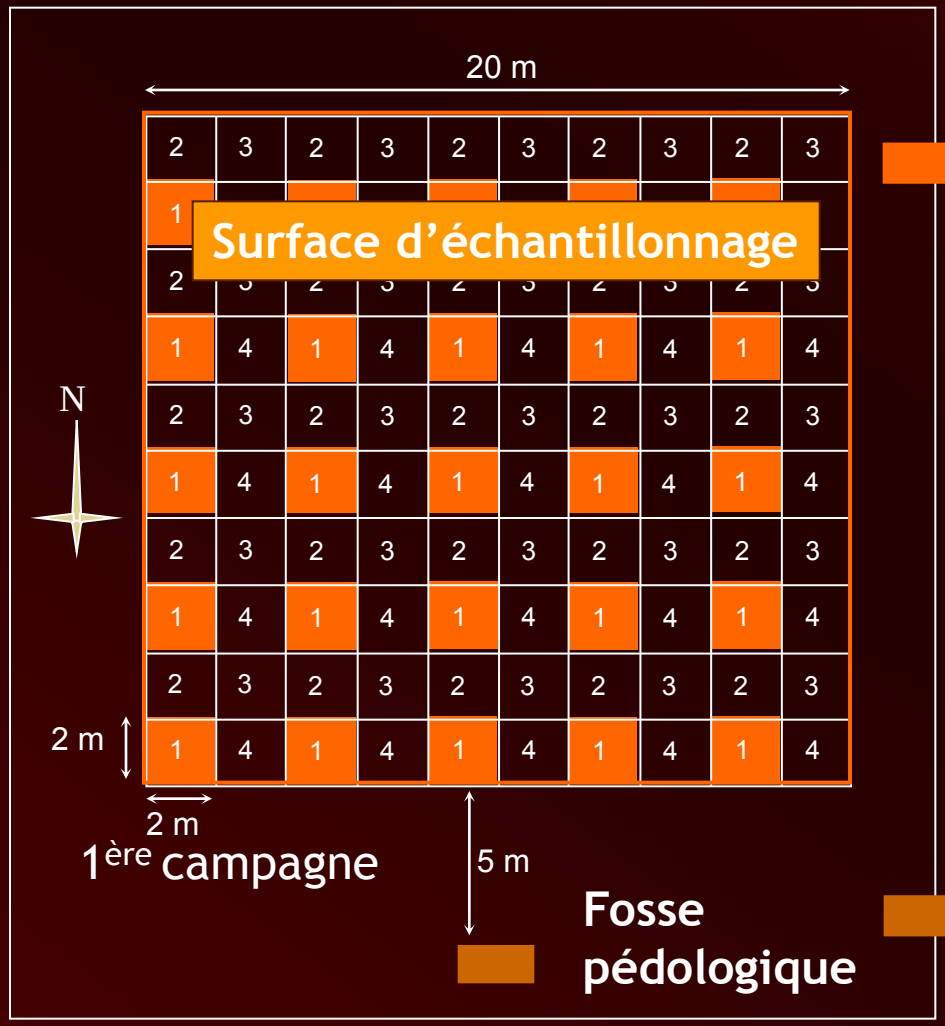
Développer des programmes dédiés

- Un réseau de sites répartis selon une grille de 16 km x 16 km
- Un réseau « pluri-occupation »
- 2200 sites échantillonnés tous les 10 ans
- Plus de 100 paramètres biologiques, agronomiques, polluants…



Jolivet et al. 2006. Etude et Gestion des Sols.

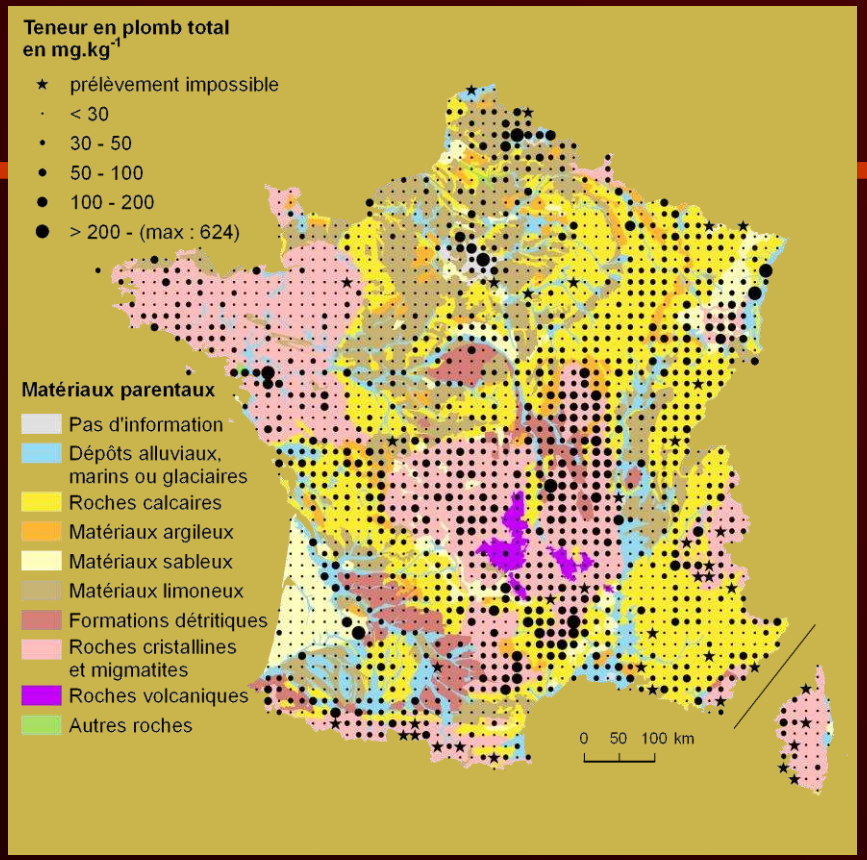
La stratégie d'échantillonnage du RMQS



Jolivet et al. 2006. Etude et Gestion des Sols.

Prélèvement d'échantillons

Description d'un profil



La surveillance et le RMQS

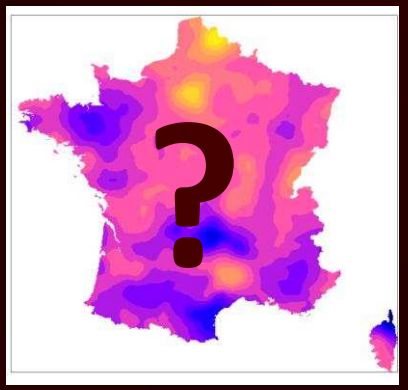
MODÈLES STATISTIQUES

INRA ALIMENTATION AGRICULTURE ENVIRONNEMENT

17e journée CASCIMODOT 6 décembre 2012

Questions liées à la surveillance

- Quelles **cartographies** des propriétés du sol sommes-nous en mesure de fournir ?
- Quelles **évolutions ?** pouvons-nous en donner des ordres de grandeurs?
- Peut-on en identifier les **déterminants** ?
- Et dispositifs de surveillance de demain ?



Questions

- Quelles **cartographies** des propriétés du sol sommes-nous en mesure de fournir ?
  - Quels sont les processus de distribution spatiale des propriétés du sol?
  - Peut-on cartographier des informations fiables sur la distribution spatiale des propriétés du sol?

Cadre statistique

- Géostatistique = $\mathbf{z}$ est une V.A.
- Modèle linéaire mixte pour des données avec ou sans transformation (log, BoxCox)

$$\mathbf{z} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{u}$$

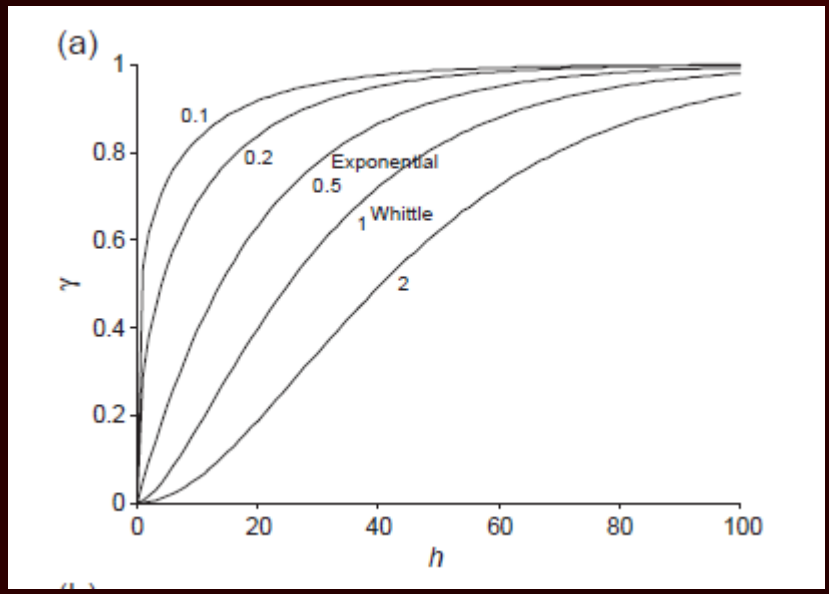
- $\mathbf{X}$ représente les covariables, $\boldsymbol{\beta}$ les paramètres associés (brutes ou data mining)
- $\boldsymbol{u}$ représente les résidus spatialement corrélés
- BLUP ou krigeage pour l'interpollation

R. M. Lark & B. R. Cullis, European Journal of Soil Science, December 2004, 55, 799–813

Le variogramme Matérn

- Modéliser la covariance spatiale
- Modèle Matérn pour décrire la covariance
- flexibilité

$$\gamma(h) = c_0 + c_1\left(1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{h}{r}\right)^{\nu} K_{\nu}\left(\frac{h}{r}\right)\right).$$

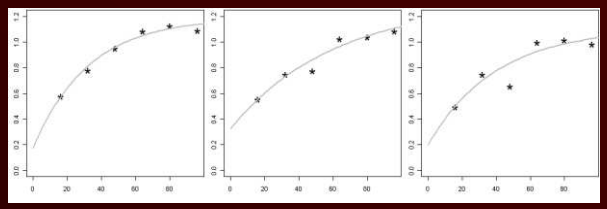


B. Minasny, A.B. McBratney / Geoderma 128 (2005) 192–207

Estimations des paramètres

- Moindres carrés
  Regression kriging
- Model-baed
  - Maximum
    de vraisemblance
  - Covariable = REML

$$\widehat{\mathbf{u}} = \mathbf{z}^* - \mathbf{M}\widehat{\boldsymbol{\beta}}.$$



Variogrammes expérimentals

$$\ell(\theta,\boldsymbol{\beta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}(\mathbf{z}-\mathbf{M}\boldsymbol{\beta})^T\mathbf{C}^{-1}(\mathbf{z}-\mathbf{M}\boldsymbol{\beta})$$

$$\mathbf{y} = \mathbf{T}\mathbf{z}$$

$$\ell(\theta,\mathbf{y}) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\log|\mathbf{W}| - \frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{Q}\mathbf{y}$$

R. M. Lark & B. R. Cullis, European Journal of Soil Science, December 2004, 55, 799–813

Statistiques de validation croisée

1. $$\theta_i = \frac{\text{Squared validation error}}{\text{Prediction variance}}$$

$$\mathrm{E}\left[\bar{\boldsymbol{\theta}}\right] = 1$$

$$\mathrm{E}\left[\tilde{\boldsymbol{\theta}}\right] = 0.455$$

2. **QQ plots des résidus (prédits – observés)**

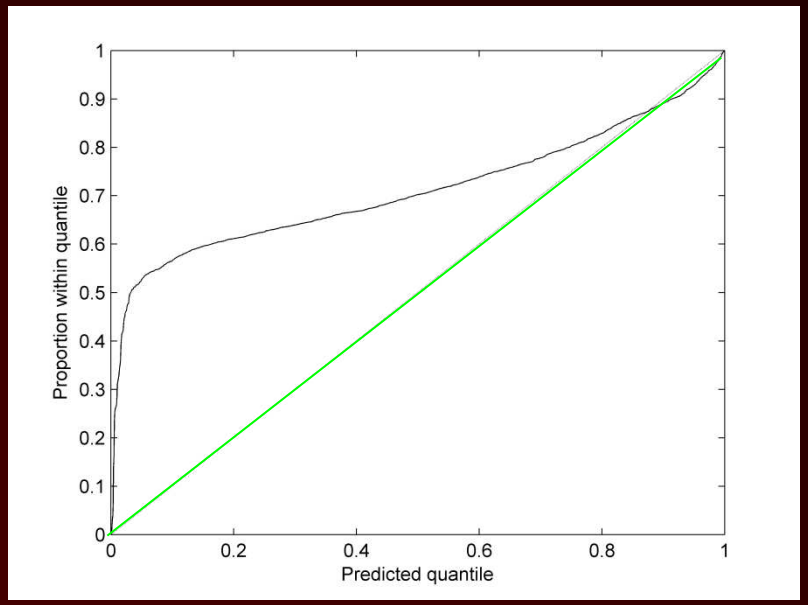
Lark, R.M. (2000) EJSS, 51, 137-157

Statistiques de validation croisée (Cd)

Gaussian

$\bar{\theta} = 1.00$

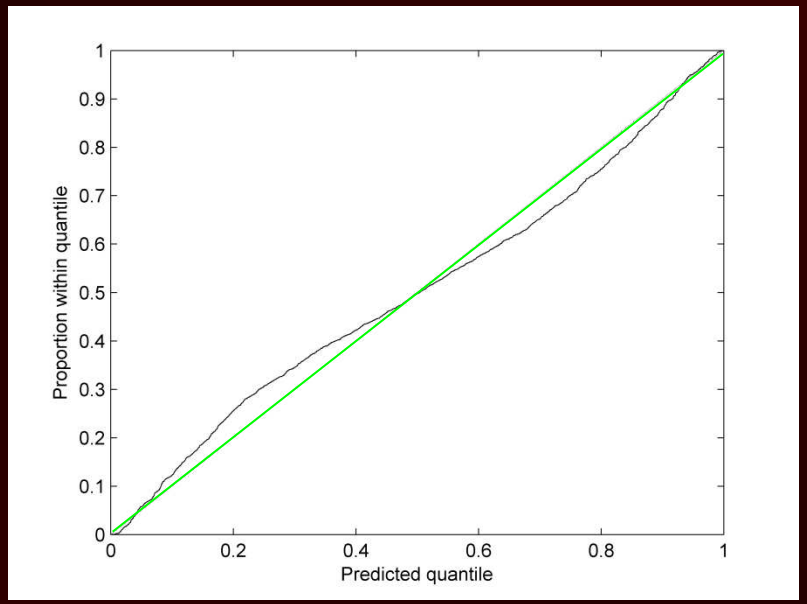
$\tilde{\theta} = 0.09$



Box-Cox Transformed

$\bar{\theta} = 1.00$

$\tilde{\theta} = 0.27$



Questions

- Quels sont les processus de distribution spatiale des propriétés du sol?

  Méthodes robustes

- Peut-on cartographier des informations fiables sur la distribution spatiale des propriétés du sol?

Questions

- Quels sont les processus de distribution spatiale des propriétés du sol?

  Méthodes robustes

- Peux-ton cartographier des informations fiables sur la distribution spatiale des propriétés du sol?

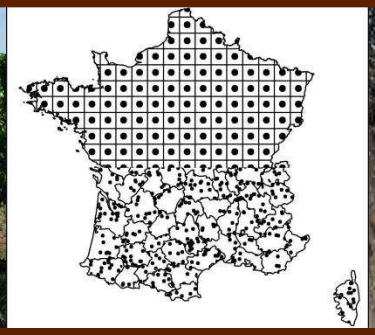
  Les copules

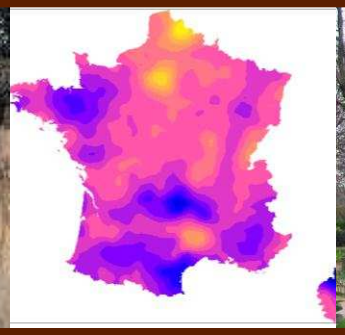
Bárdossy & Li (2008) Water Resources Research, 44, W07412

Kazianka & Pilz (2010) Stoch. Environ. Res. Risk Assess., 24, 661-673

Méthodes robustes

Variographie & krigeage

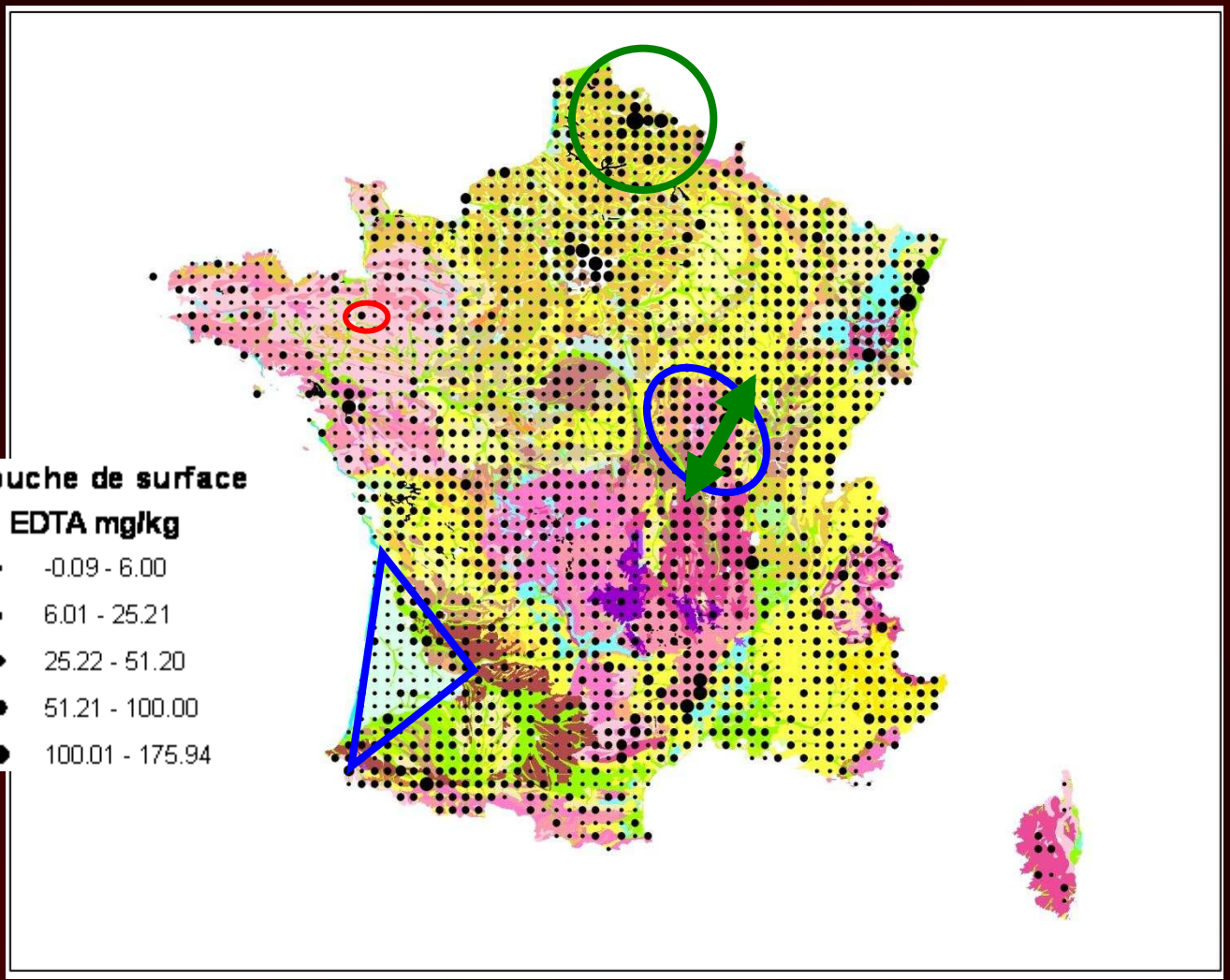






INRA

Groupement d'intérêt scientifique sol

3 sources de variation spatiale



Pédologie (Matériau Parental)

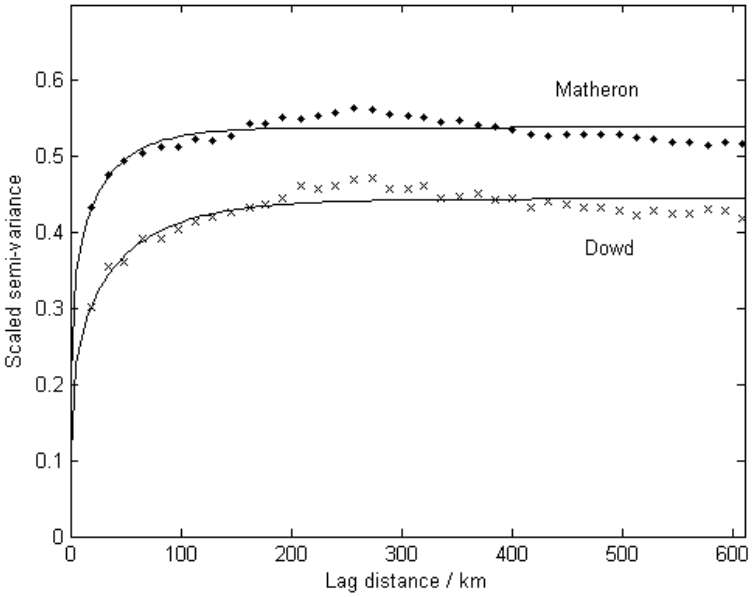
Variations continues

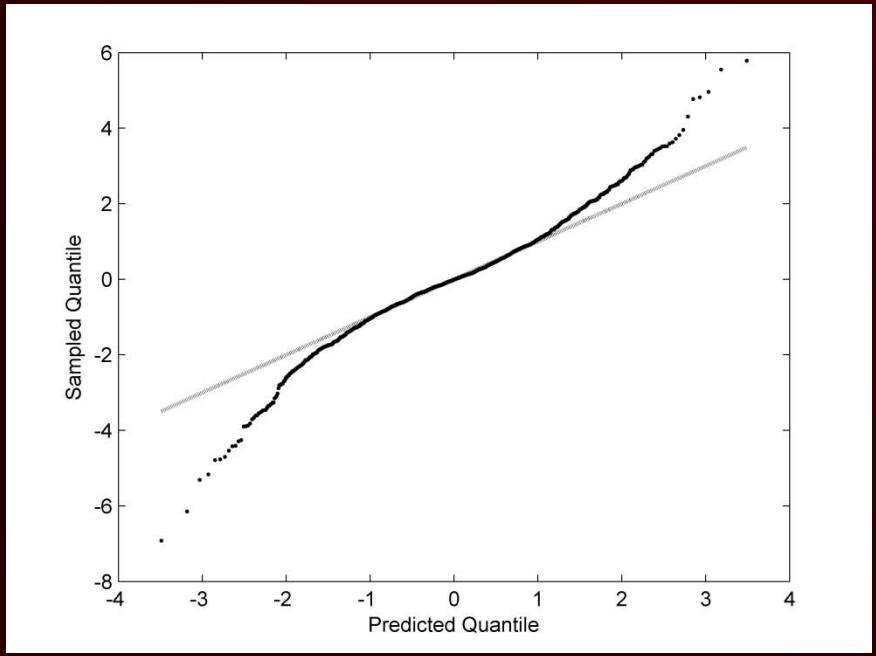
Outliers spatiaux

Marchant et al. / EJSS (2010) 2010, **6**
Saby et al. / Geoderma 162 (2011) 30
Lacarce et al. / Geoderma 170 (2012

Les méthodes robustes

1. Estimation robuste du variogramme



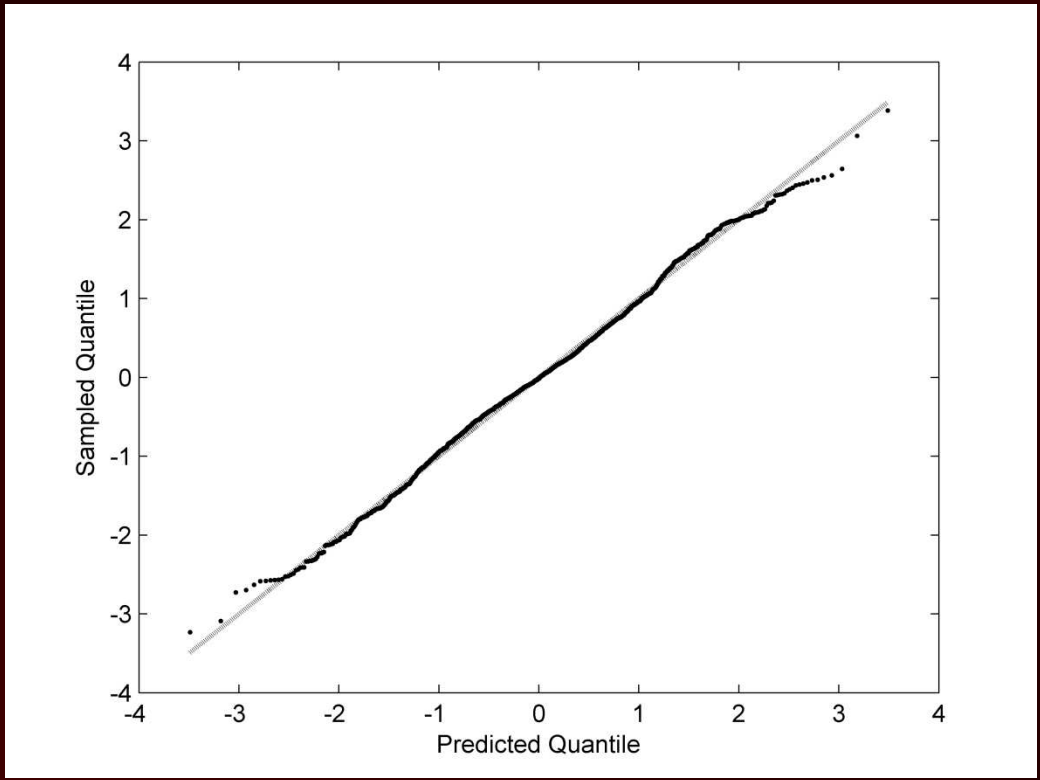


Marchant et al. / EJSS (2010) 2010, **61**, 144–152
Saby et al. / Geoderma 162 (2011) 303–311 ;
Lacarce et al. / Geoderma 170 (2012) 359–368

$$\bar{\theta} = 1.55$$

$$\tilde{\theta} = 0.44$$

(Winsorizing) Equeutage des données



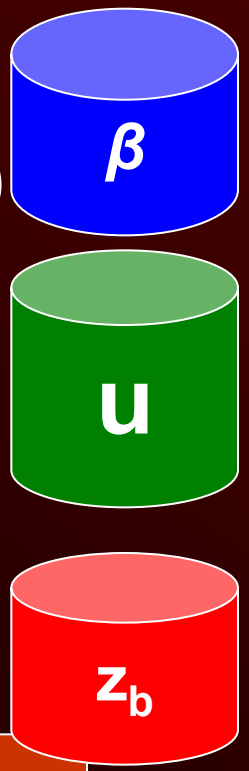
Krigeage robuste (médiane) lors de la validation croisée = Équeutage ou séparation de la composante contaminée

Marchant et al. / EJSS (2010) 2010, **61**, 144–152
Saby et al. / Geoderma 162 (2011) 303–311 ;
Lacarce et al. / Geoderma 170 (2012) 359–368

Modèle linéaire mixte robuste

Effets fixes = matériau parental (**Mβ**)

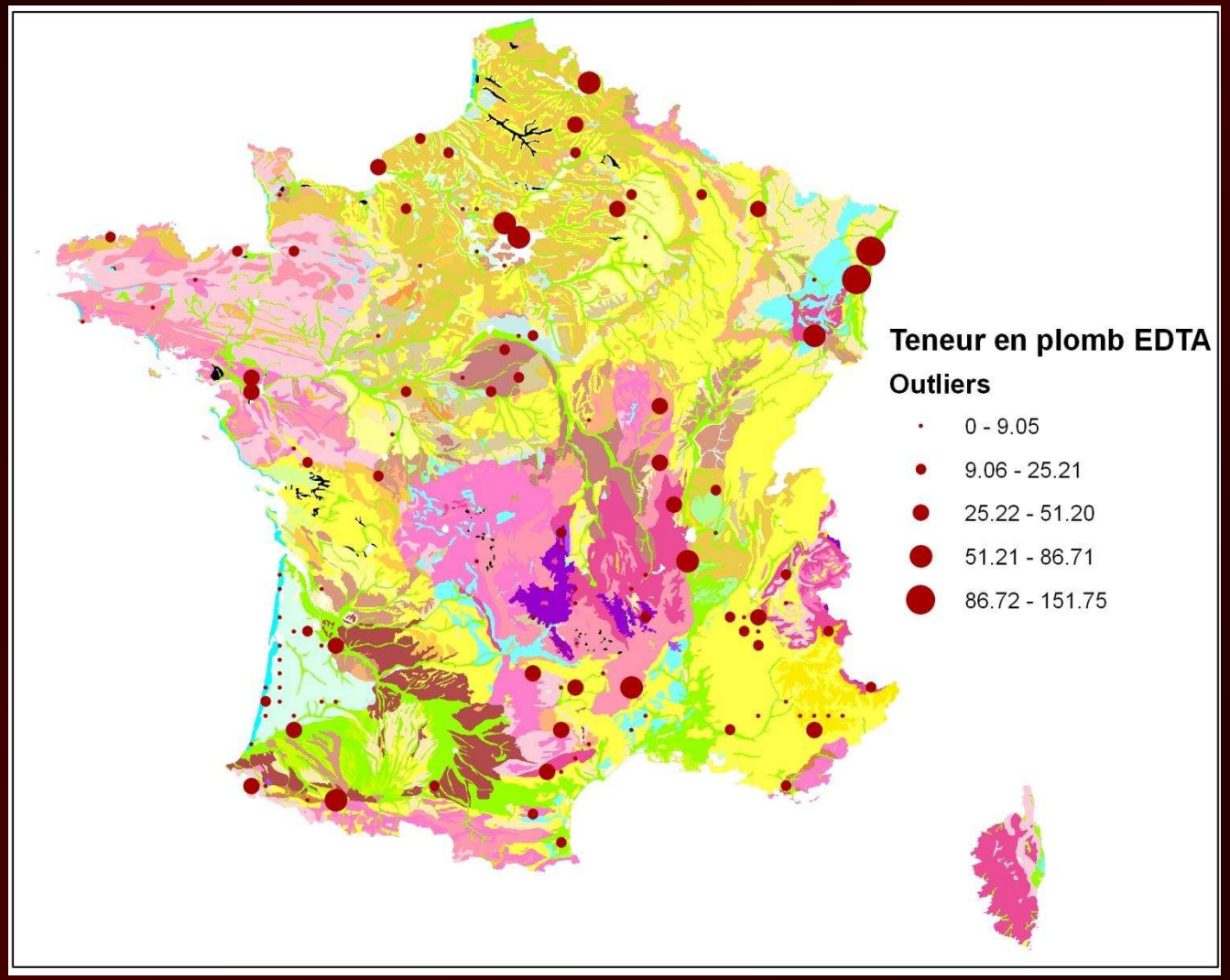
Effets continus = géostatistique (**u**)

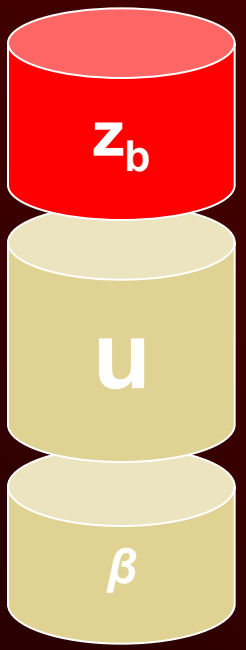


$$\mathbf{z} = \mathbf{M}\beta + \mathbf{u}$$

Krigeage robuste élimine l'effet des variations à courte distance

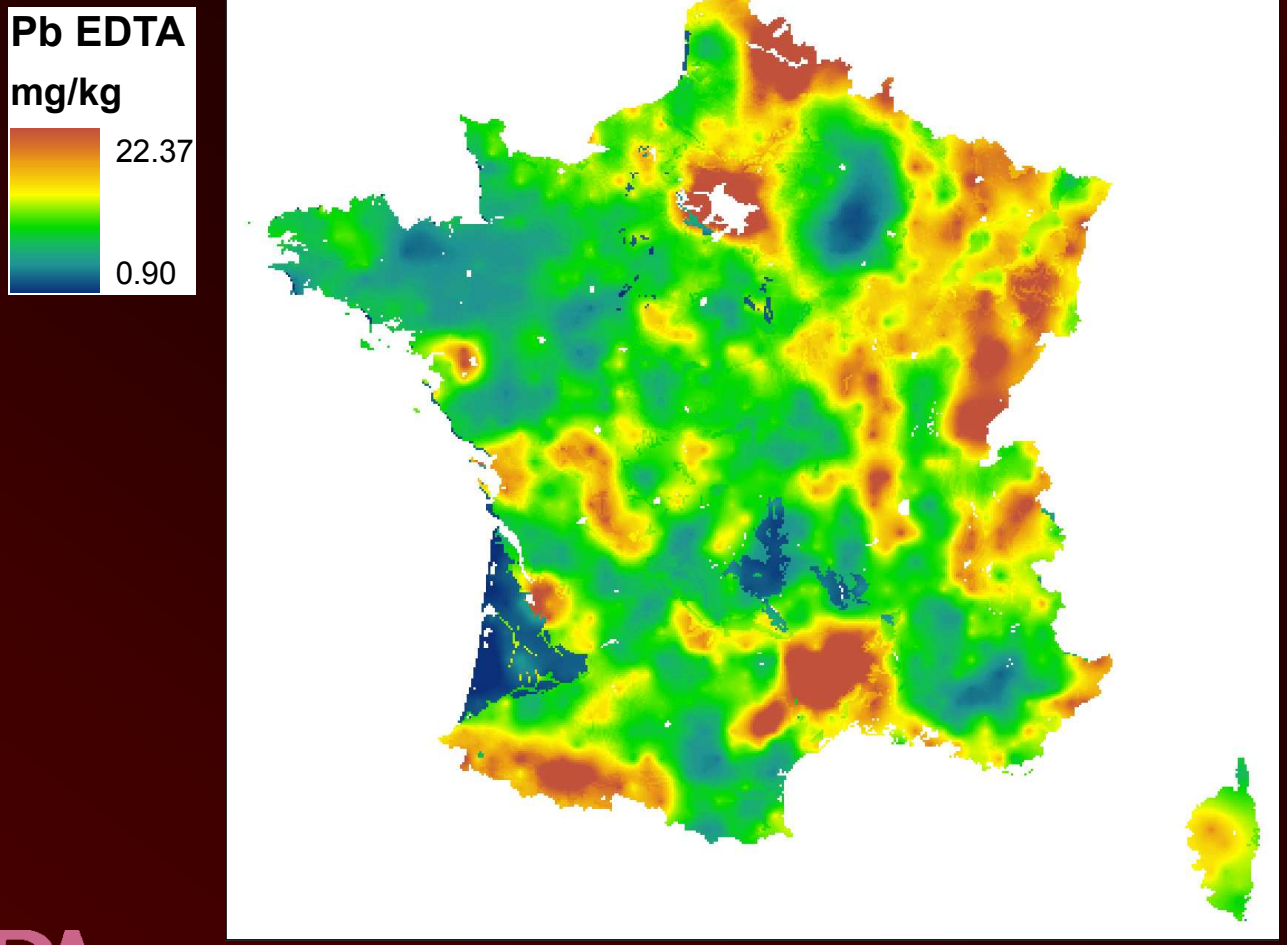
Les outliers

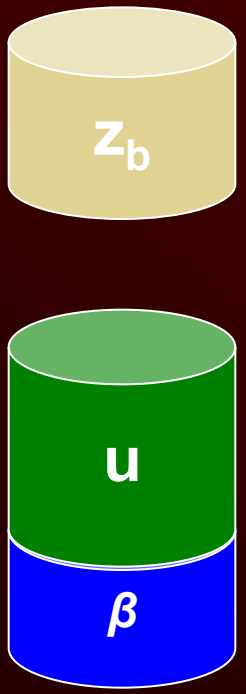




Cartographie

$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \mathbf{u}$$

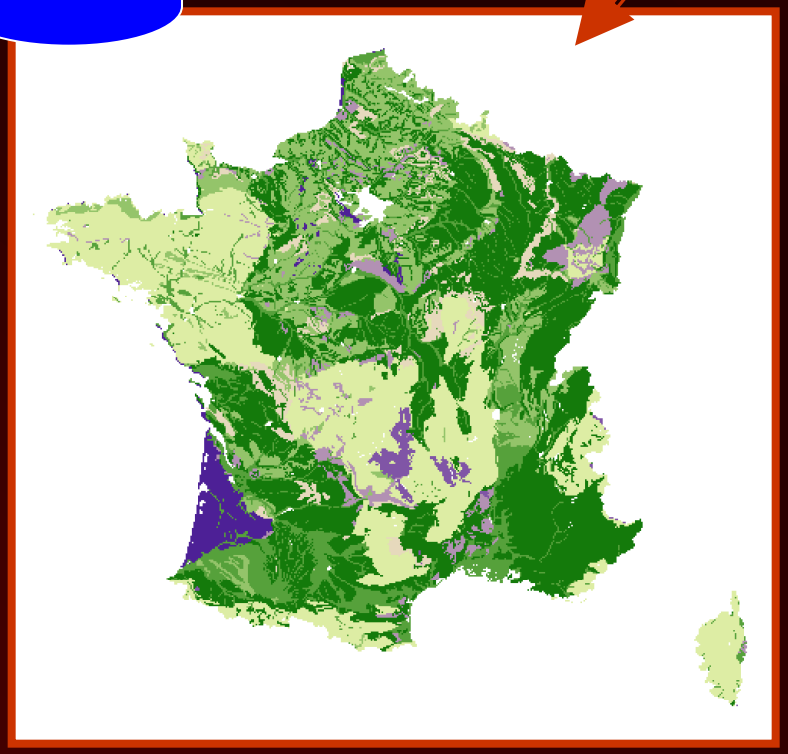




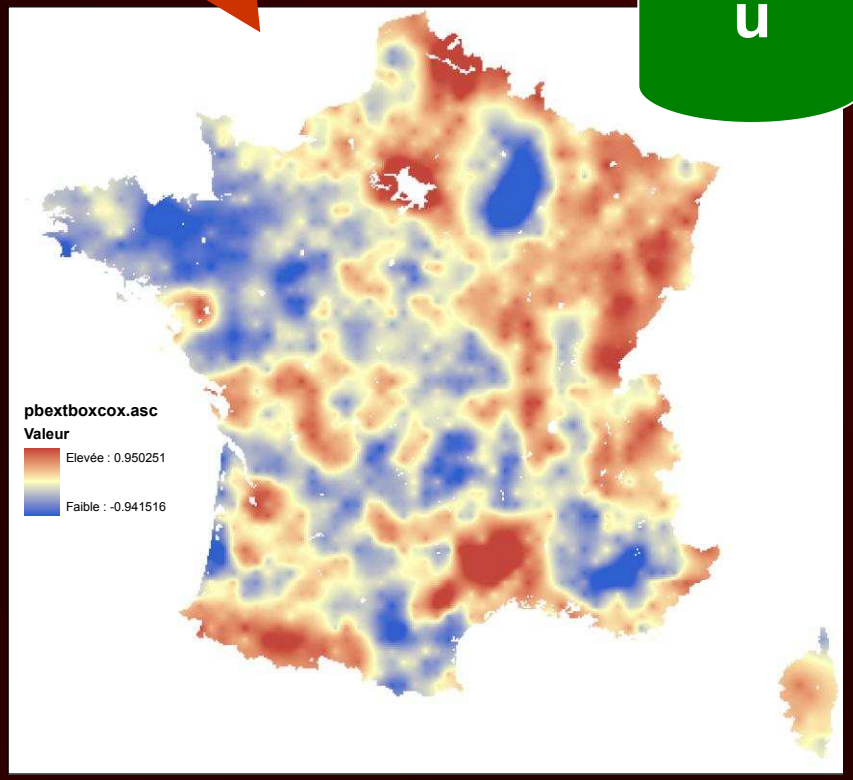
Séparer les effets

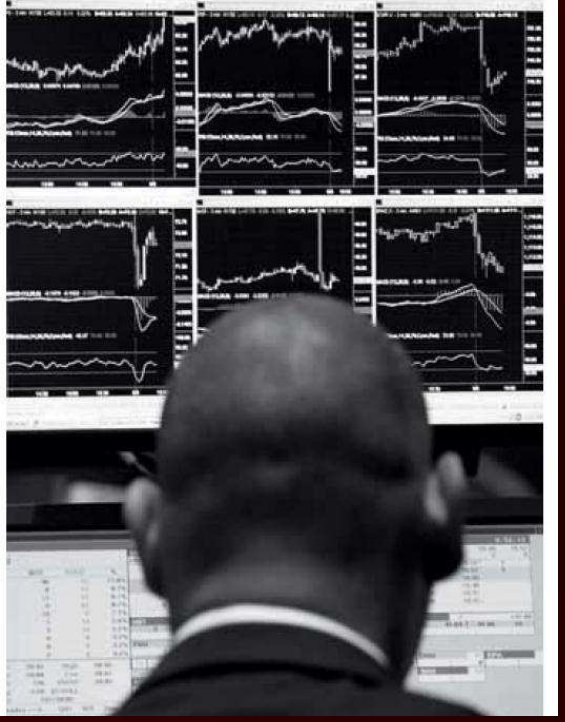
$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \mathbf{u}$$

β



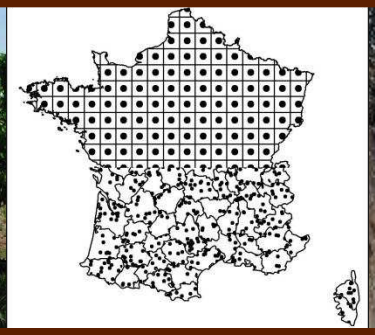
u

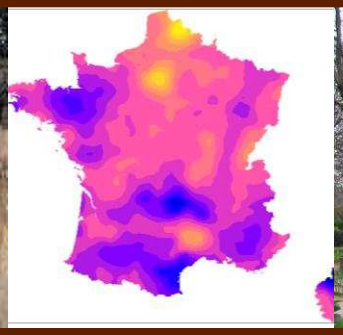




Les Copules

Recipe for Disaster: The Formula That Killed Wall Street



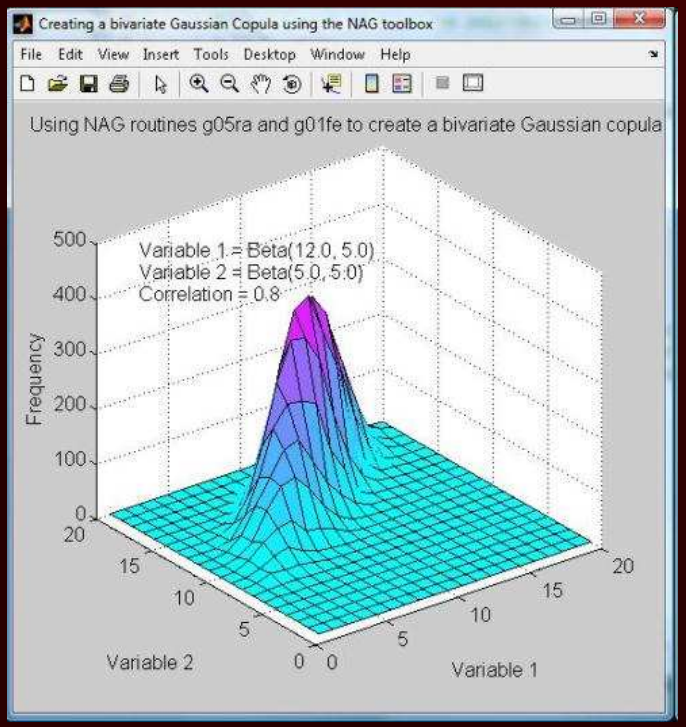

INRA

Groupement d'intérêt scientifique sol

Une copule ?

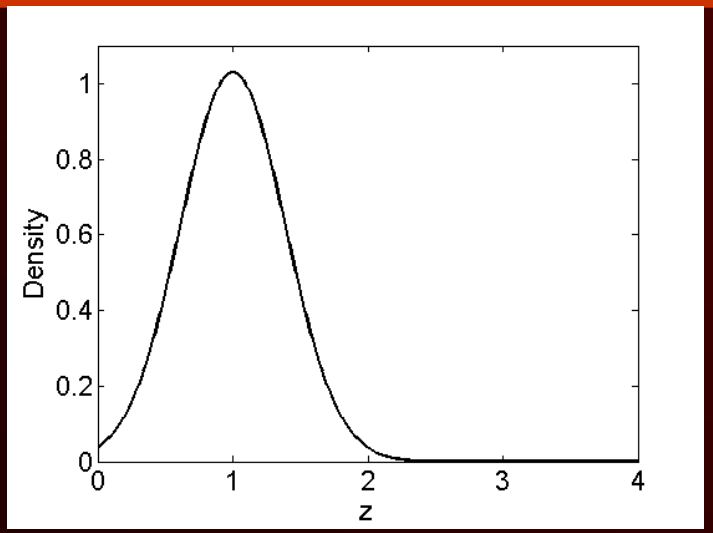
- la copule (mathématique),
  objet statistique qui lie les variations de 2 ou plusieurs variables aléatoires sans lien avec la loi marginale

- On représente avec la copule les variations entre U(**x**) et U(**x**+**h**) dans le cadre de la géostatistique

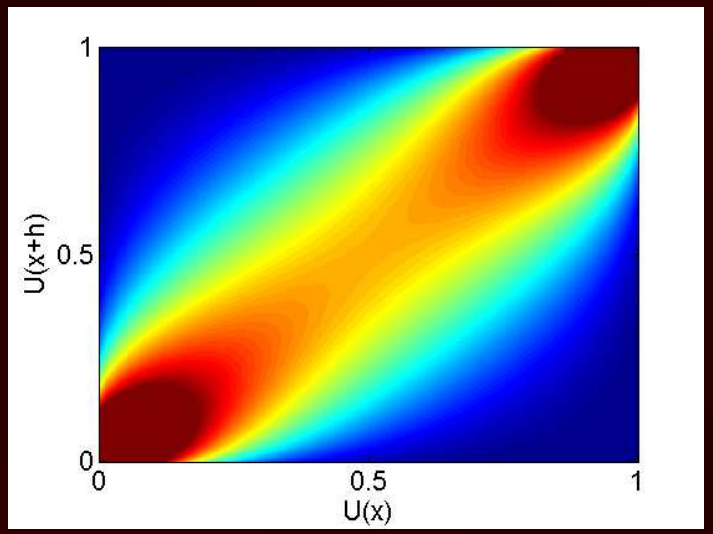


L'hypothèse gaussienne

1. Distribution marginale gaussienne

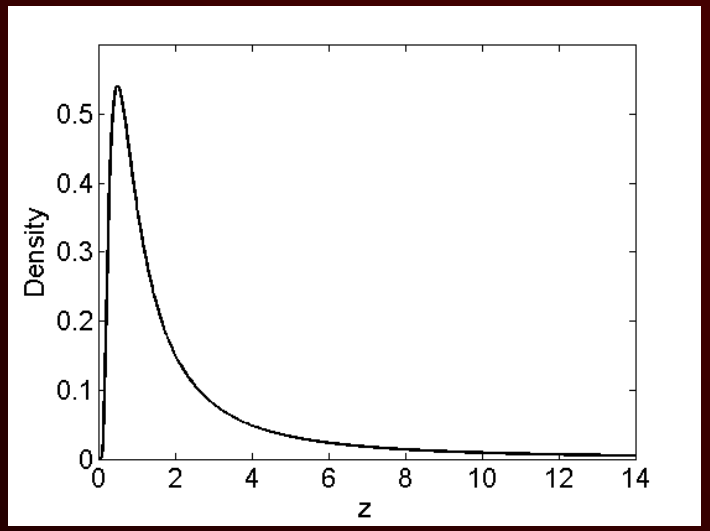


2. Structure de la dépendance gaussienne

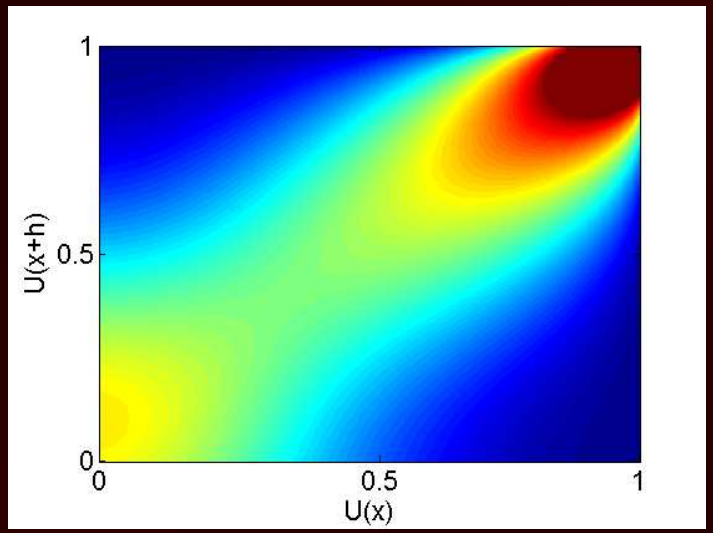


Plus d'hypothèse gaussiènne ?

1. **Distribution marginale plus générale par ex . Extreme value distribution avec effet fixe dépendant du matériau parental.**



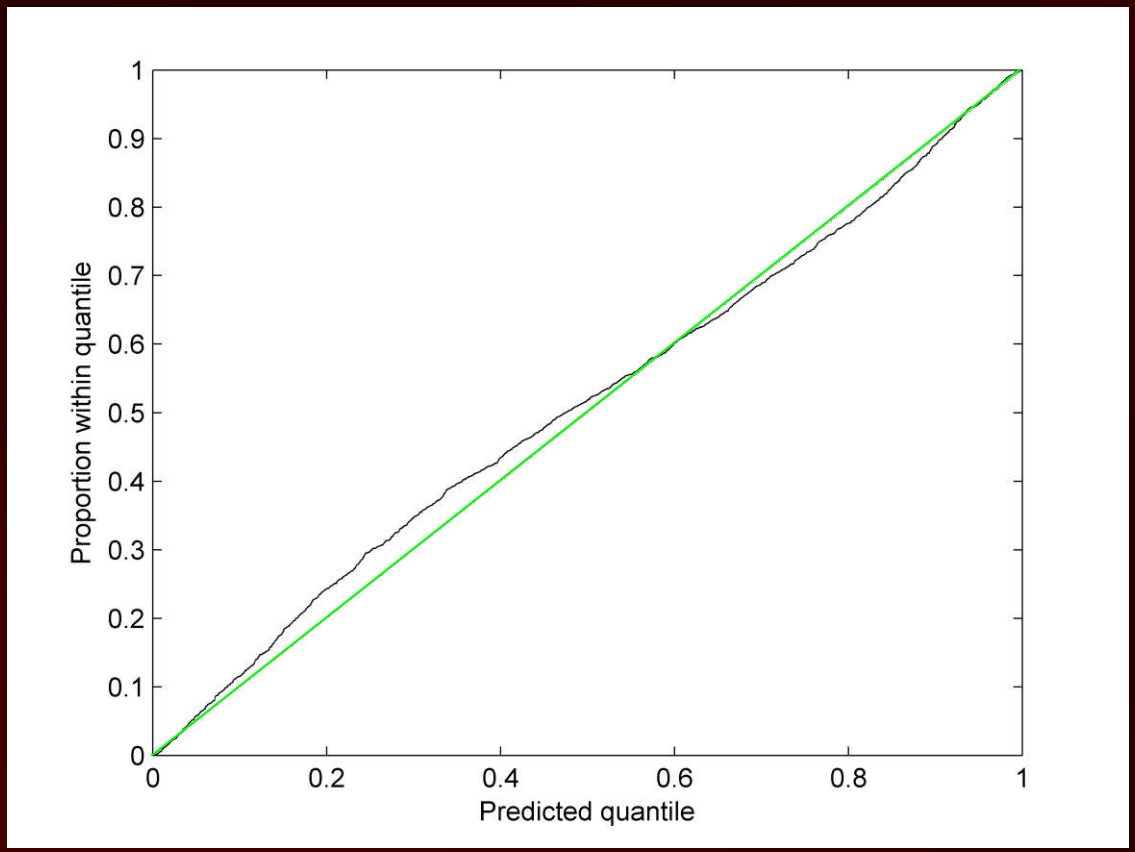
2. **Structure de dépendance asymétrique e.g. v-transformed copula**



Le tout ajusté par maximum de vraisemblance

Validation croisée

la copule gaussienne et la distribution des valeurs extrêmes

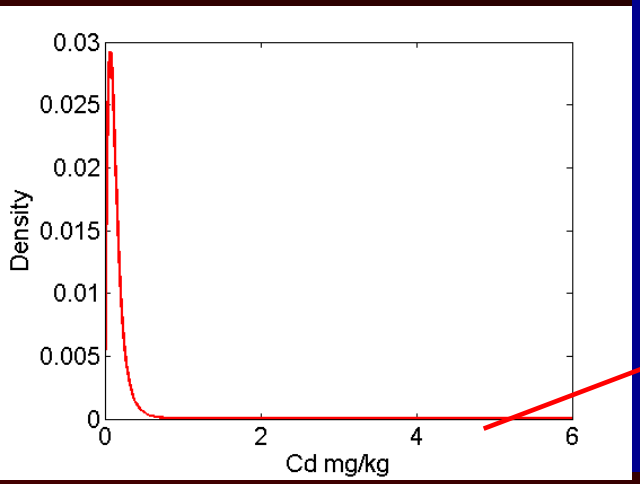


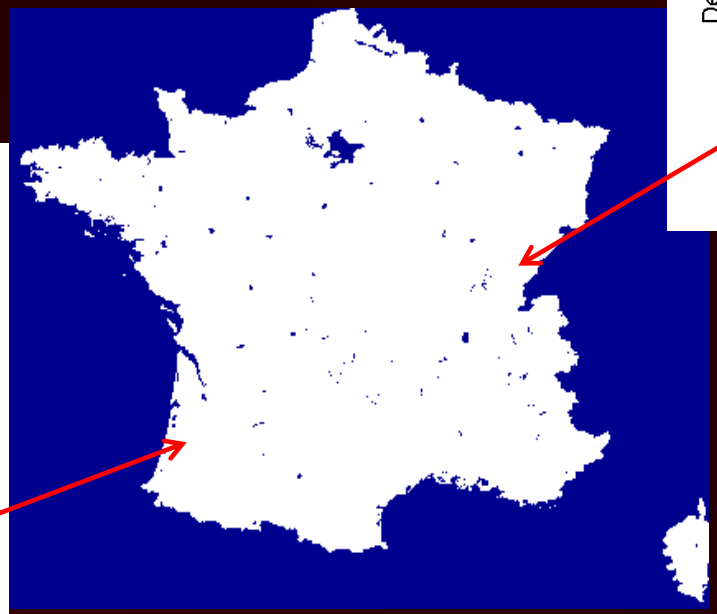
$$\bar{\theta} = 1.00$$

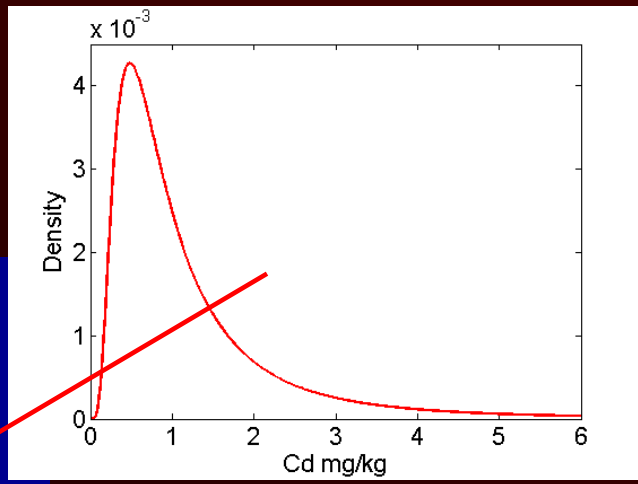
$$\tilde{\theta} = 0.34$$

Marchant et al. / Geoderma 162 (2011) 327–334

Cadmium en France





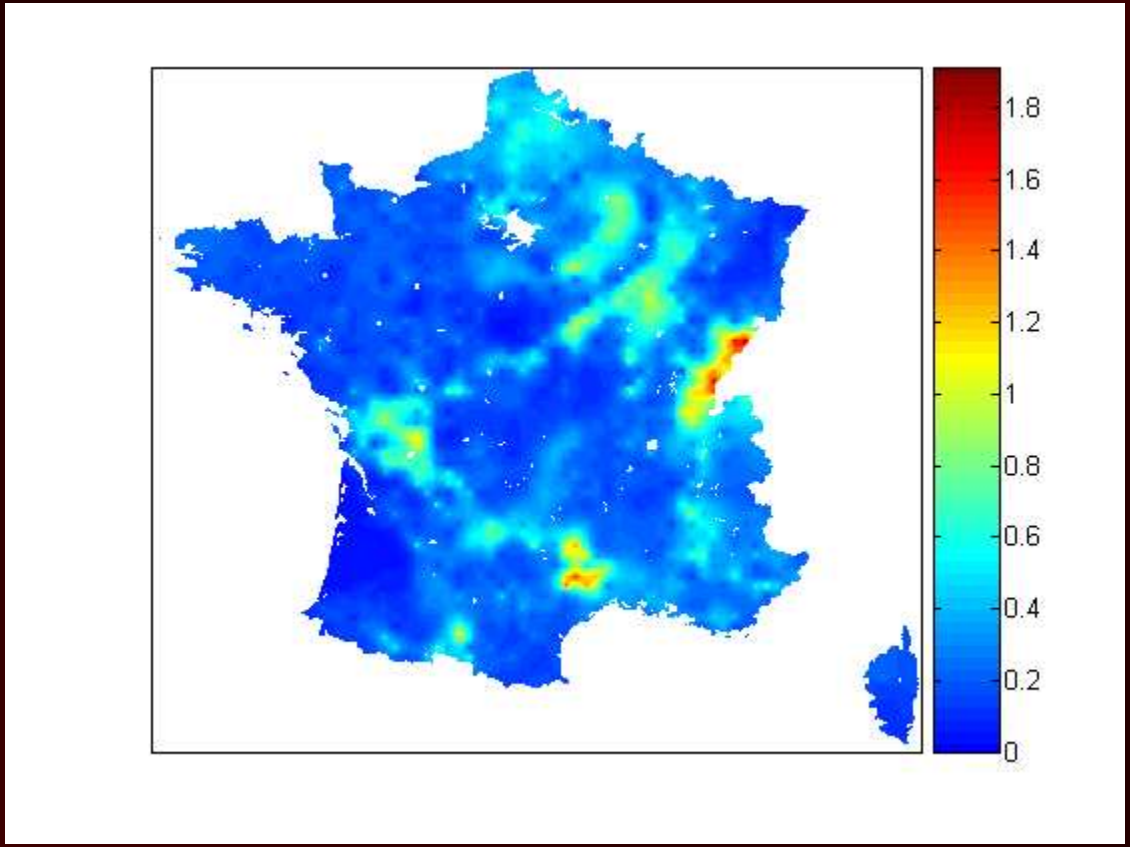


Marchant et al. / Geoderma 162 (2011) 327–334

INRA ALIMENTATION AGRICULTURE ENVIRONNEMENT

Cadmium en France

Expected concentration mg/kg

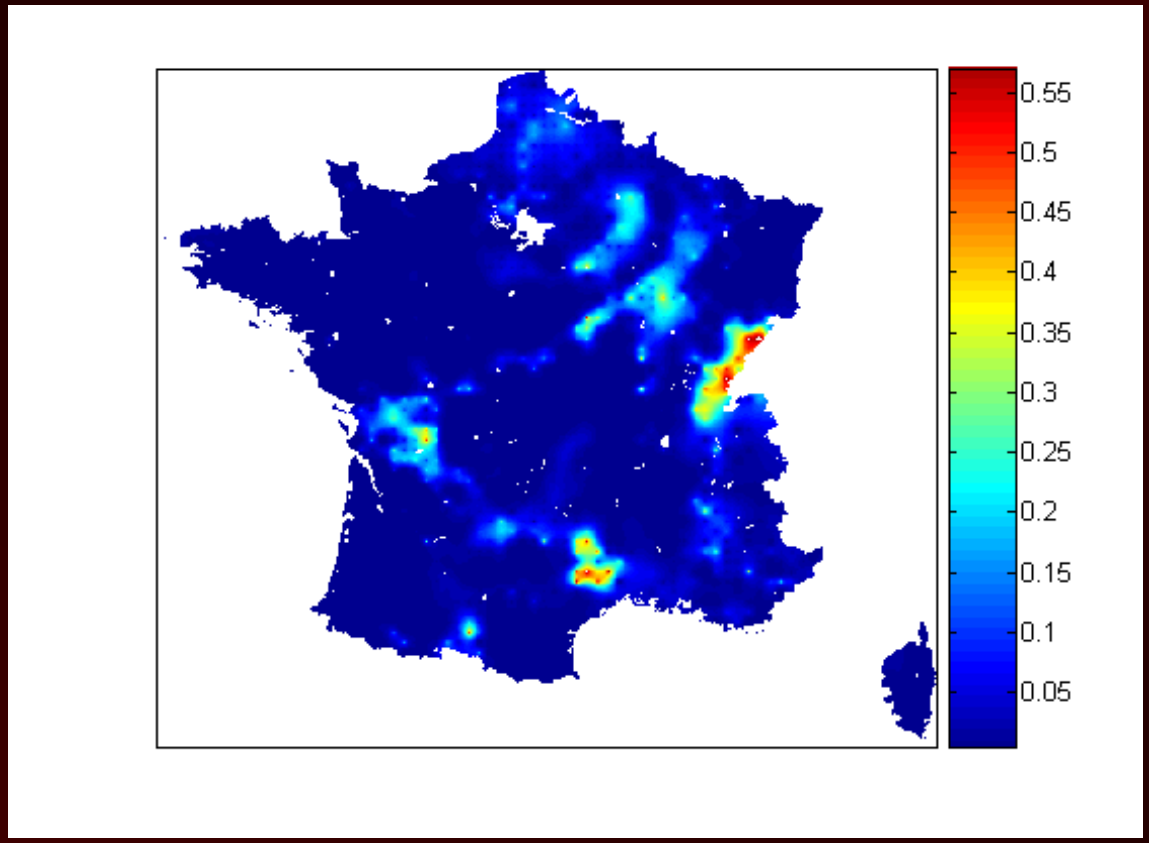


Marchant et al. / Geoderma 162 (2011) 327–334

INRA ALIMENTATION AGRICULTURE ENVIRONNEMENT

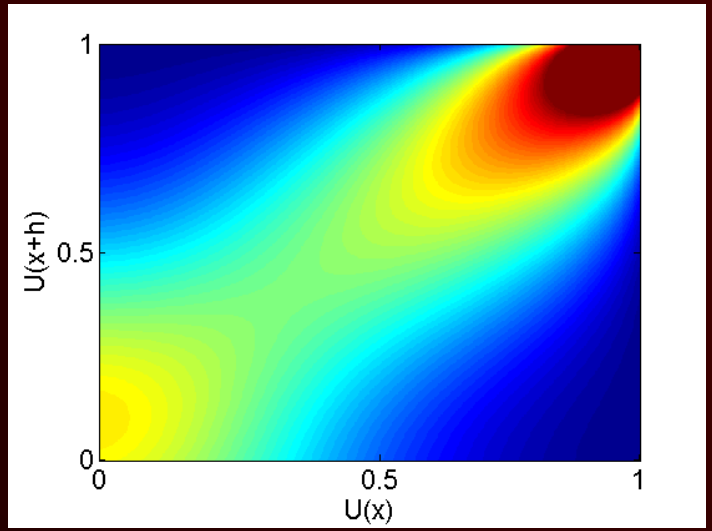
Cadmium in France

Probabilité de dépasser 1 mg/kg



Marchant et al. / Geoderma 162 (2011) 327–334

Copules Non-gaussienne



V-transformed copula

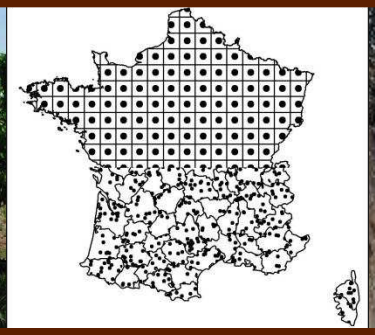
Problème:

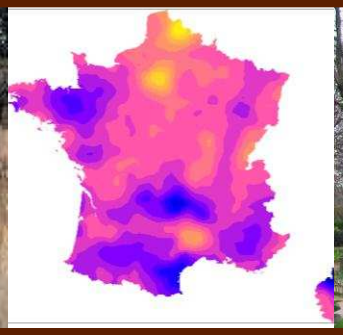
- La vraisemblance comporte un nombre de termes énorme 2^n
- 90 sites = 2^{90} = 1.23794e+27 termes

$$l(\Theta) = \frac{1}{2}\log|\mathbf{Q}| - \frac{1}{2}\mathbf{a}^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{Q}^{-1}\right)\mathbf{a} + \sum_{i=1}^{n}\log\left\{f_i(z_i)\right\}$$

Aspect temporel

L'évolution des propriétés agronomiques des sols vue par la BDAT

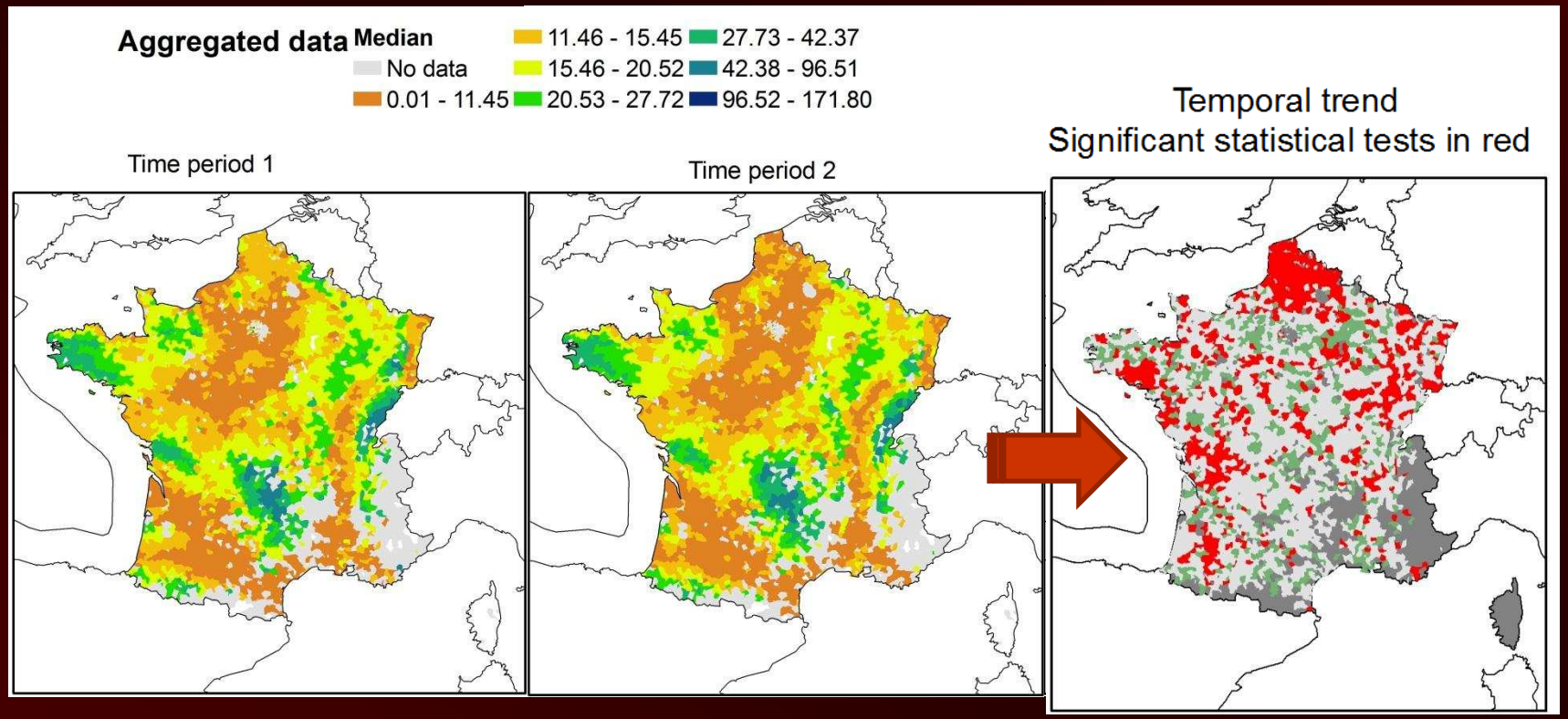



INRA

Groupement d'intérêt scientifique sol

La Base de données d'Analyses de terre

Agrégation des analyses par Canton
Détecter des évolutions = Tests d'inférence par canton



Saby et al. / Soil Use Manag, 2008

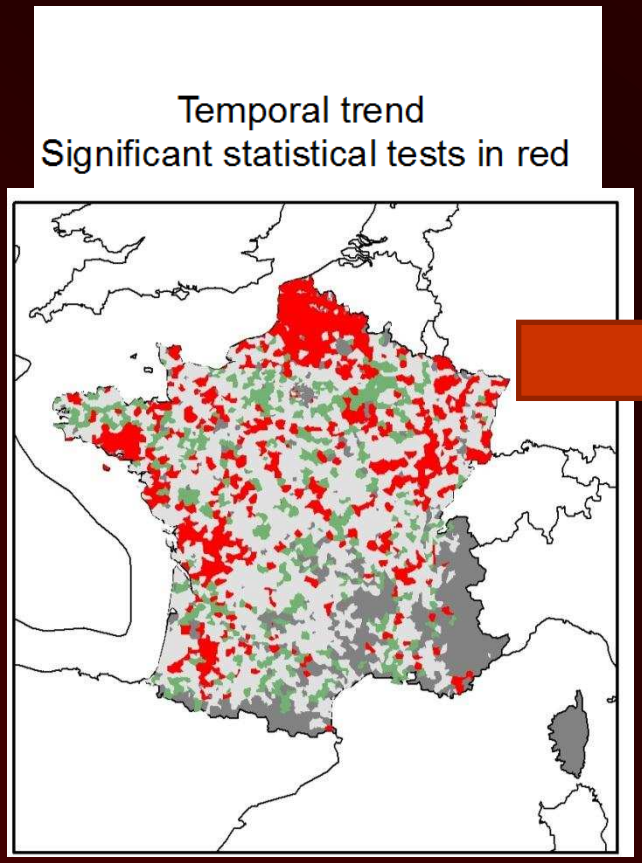
Cadre des tests multiples

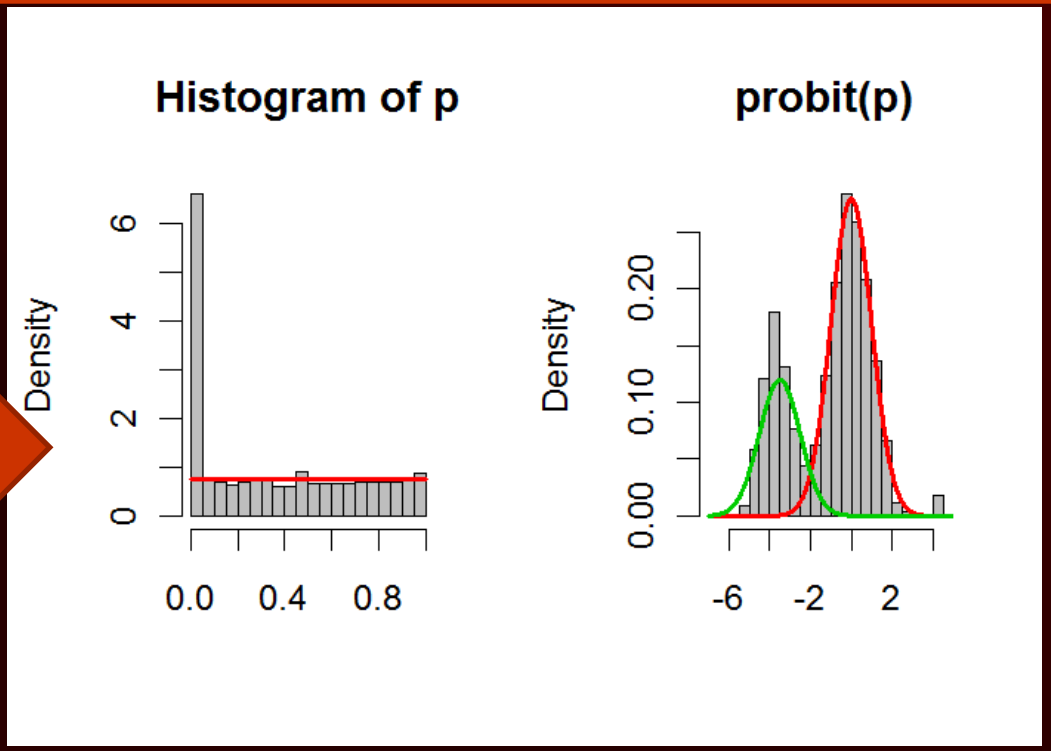
- Mélange de milliers de tests considérés simultanément => cadre des tests multiples
- FDR (Benjamini et Hochberg, 1995) : contrôle du taux de fausses découvertes
- FDR local (Robin, 2007)

$$\ell\mathrm{FDR}(p_i) = \mathbb{P}(\text{“not interesting”}|p_i)$$

Chauveau et al. / Geoderma *in prep*

Comparaison de méthodes





- Modèle de mélange
- EM semi paramétrique = > meilleure estimation de FDR local

Autres méthodes

- Méthodes pour les données censurées : estimation de la vraisemblance par MCMC et prédiction par méthodes Bayésienne

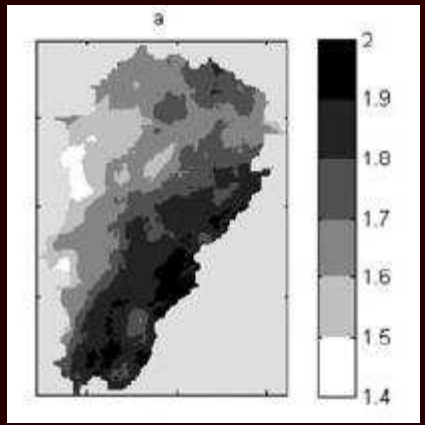
  Orton et al. Journal Environmental Quality (2013)

  Orton et al. Science of the Total Environment 443 (2013) 338–350

- Krigeage “surface vers points” en tenant compte de la densité de données

  Orton et al. Environmetrics 443 (2012a)

  Orton et al. Environmetrics 443 (2012b)



Conclusions

- Géostatistiques standards ne sont pas adaptées aux données issues des campagnes de mesures des réseaux de surveillance nationaux

- Les méthodes robustes et les copules apportent des modèles plus généraux et donc plus adaptés

Et demain, quelle stratégie ?

- Coûts prohibitifs de la densification spatiale ou temporelle
  - Recueillir des informations connexes
  - Besoin de méthodes statistiques adaptées pour mixer des observations éparses et des données connexes continues

- Vers du Digital Soil Monitoring ?

Merci de votre attention







INRA

Groupement d'intérêt scientifique sol