Position du problème

Nous allons analyser dans ce chapitre les différentes questions soulevées par les objectifs poursuivis, faire un rapide état de l'art et préciser les questions qui sont prises en charge dans ce travail.

Position du problème

L'estimation des propriétés du sol par inversion d'un modèle de culture avec des observations a déjà été testée par quelques auteurs (Irmak et al., 2001; Timlin et al., 2001; Braga and Jones, 2004; Guérif et al., 2006). Par exemple, Timlin et al. (2001) cherchent à retrouver l'humidité à la capacité au champ (*HCC*) de trois couches de sol en utilisant un modèle simple et des observations de cartes de rendement du maïs ; Braga et Jones (2004) estiment cinq paramètres de neuf couches de sol en inversant le modèle de culture CERES avec des observations de rendement du maïs et de teneur en eau du sol ; Guérif et al. (2006) estiment douze paramètres du sol en inversant le modèle de culture STICS avec des observations d'indice foliaire (*LAI*) et de teneur en azote de la plante (*QN*). Cependant, les résultats de l'estimation des paramètres du sol montrent qu'il est difficile d'obtenir une bonne qualité d'estimation, impliquant de ce fait une difficulté d'utilisation des valeurs estimées pour l'amélioration des prédictions. La question est en effet délicate pour plusieurs raisons que nous présenterons ici.

1.1.1. Les modèles de culture sont des modèles dynamiques complexes avec un grand nombre de paramètres

Les modèles de culture sont des outils particulièrement intéressants pour valoriser les informations recueillies sur les cultures, en estimant les paramètres d'entrée du modèle par inversion. Ces observations représentent en effet des observations de variables d'état simulées par ces modèles. Il est donc possible de

les inverser pour retrouver les valeurs de leurs paramètres d'entrée tels que les caractéristiques du sol. Par ailleurs, ces modèles simulent des variables de sortie pertinentes pour la prévision des consommations en eau et en azote de la plante, du rendement et de la qualité de la récolte, lesquelles variables peuvent être optimisées pour la mise au point d'itinéraires techniques.

Ce sont des modèles complexes qui impliquent le sol, la plante et l'atmosphère. Ces trois compartiments sont reliés entre eux par des flux de matières qui sont pilotés par des bilans énergétiques. Par exemple, le flux de carbone dans le système de culture est piloté par un bilan d'énergie du rayonnement (Brisson et al., 2006). Ces modèles décrivent, sous forme d'équations mathématiques, les processus physiques et biologiques qui caractérisent le fonctionnement d'une culture, en interaction avec son environnement. Les simulations issues de ces modèles sont réalisées à un pas de temps journalier. Ces modèles comportent deux types de variables : des variables d'entrée et des variables de sortie. Les variables d'entrée sont représentées par les variables climatiques (pluies, températures, rayonnement global,…) qui sont généralement mesurées chaque jour et qui sont imposées au modèle (sous la forme d'un forçage climatique). Les variables de sortie sont calculées par le modèle et sont des variables agroenvironnementales qui décrivent l'état du système sol-plante. Leur nombre diffère selon les modèles, mais les principales variables généralement modélisées sont l'indice foliaire (*LAI*), la biomasse aérienne, la quantité d'azote contenu dans la plante, la quantité et la qualité de la récolte ainsi que le contenu en eau et en azote du sol.

Ces modèles peuvent comporter un grand nombre de paramètres qui règlent les différentes lois d'action et équations du modèle. Ils concernent les paramètres généraux, les caractéristiques de la plante cultivée, les techniques agricoles employées et les propriétés du sol. Le modèle STICS (Brisson et al., 2008), avec lequel nous travaillons, considère 227 paramètres pour le cas du blé tendre qui sont répartis en trois groupes : 129 paramètres liés aux caractéristiques de la plante, 23 paramètres des techniques agricoles et 75 paramètres de propriétés des sols. Les valeurs des paramètres liés aux caractéristiques de la plante proviennent d'études antérieures décrites dans la littérature, issues de mesures expérimentales précises

ou bien d'une calibration préalable sur une base de données (Flenet *et al.*, 2003; Hadria *et al.*, 2007; Singh *et al.*, 2008). L'incertitude liée à ce premier groupe de paramètres peut être importante car les valeurs issues des mesures ou de la calibration sont généralement entachées d'erreurs. Les paramètres des techniques agricoles appliquées au système cultural renseignent, entre autres, la date de semis, la nature des résidus de la culture précédente et le type de travail du sol, la date et la dose de fertilisant (ou d'eau) apportée au système. Les valeurs de ces paramètres sont généralement bien renseignées au niveau de la parcelle car elles correspondent aux décisions techniques prises par l'agriculteur. Dans le cas de l'agriculture de précision, la variabilité spatiale des modalités techniques est cependant difficile à appréhender et peut également être entachée d'erreurs. Les propriétés des sols, qui font l'objet central de ce travail, peuvent être déterminées à partir d'analyses ou de cartes de sols, mais cela n'est pas adapté, nous le rappelons, au contexte de l'agriculture de précision au niveau intra-parcellaire. Les paramètres du sol sont les plus difficiles à connaitre en chaque point de l'espace et nous proposons dans cette étude de les estimer par inversion du modèle STICS.

1.1.2. Le problème posé par le grand nombre de paramètres à estimer

Nous visons dans ce travail l'estimation des paramètres descriptifs du sol, en considérant que les autres paramètres sont connus. Même avec cette restriction, cela n'est pas chose aisée (Tremblay and Wallach, 2004; Launay and Guérif, 2005). La principale raison, indépendamment de la méthode d'estimation choisie, est qu'il est impossible d'estimer simultanément tous les paramètres du sol de STICS car une grande partie des paramètres n'est pas *identifiable* ; ceci est dû à la structure des équations du modèle (Niu and Fisher, 1997; Makowski et al., 2006a). Un manque d'identifiabilité apparaît lorsque plusieurs valeurs de paramètres aboutissent aux mêmes valeurs des variables observées : il est difficile d'estimer correctement ces paramètres à partir de ces observations. Par exemple, une des équations du modèle STICS (voir Brisson et al., 2008) permet de calculer la quantité d'azote organique actif dans le sol (*NHUM, t ha*$^{-1}$), provenant de la minéralisation de la matière organique du sol, de la façon suivante :

$$NHUM = Norg \times PROFHUM \times DA(1) \times (1 - FINERT) \quad (1\text{-}1)$$

où *Norg (%)* est le contenu en azote organique du sol, *profhum (cm)* est la profondeur de minéralisation, *DA(1) (*en *g cm^{-3})* est la densité volumique de la première couche de sol et *FINERT* représente la proportion d'azote inactif (fixée à 0.65). Ainsi, lorsque des observations reliées à la quantité en azote organique actif dans le sol sont disponibles, l'équation ci-dessus ne permet pas d'estimer simultanément *Norg*, *profhum* et *DA(1)* : il y a donc un problème d'identifiabilité. Pour cette raison, il est conseillé de sélectionner un sous-groupe de paramètres à estimer et de fixer les autres à une valeur convenable (appelée valeur nominale).

Quand bien même un sous-groupe de paramètres à estimer a été sélectionné, il subsiste un autre problème. Les trois groupes de paramètres (liés à la plante, aux techniques agricoles et aux propriétés des sols) dépendent les uns des autres à travers les équations de STICS. Cela a pour conséquence que les valeurs estimées des paramètres du sol dépendent des valeurs auxquelles les paramètres des deux autres groupes sont fixés. Une mauvaise valeur donnée à un paramètre technique ou plante entraîne donc un biais sur l'estimation des paramètres du sol. Ce biais est appelé biais d'omission (Miller, 2002). Voici un exemple d'équations de STICS où un paramètre plante (sensibilité de la plante à la sècheresse *SENSRSEC*) et des paramètres du sol (humidité à la capacité au champ *HCC* et au point de flétrissement *HMIN*) donnent ensemble la valeur de l'effet de la sècheresse du sol à la date *t* (*HUMIRAC$_t$*) sur la germination :

$$HUMIRAC_t = SENSRSEC + (1 - SENSRSEC)\frac{HUMSOL_t - HMIN}{HCC - HMIN},\ HUMSOL_t > HMIN \quad (1\text{-}2)$$

où *HUMSOL$_t$* est le contenu en eau du sol à la date *t*. A partir d'observations (directes ou indirectes) des variables *HUMIRAC$_t$* et *HUMSOL$_t$*, l'estimation de *HCC* et de *HMIN* peut être biaisée si *SENSRSEC* n'est pas fixé à une bonne valeur. Il est donc important dans ce cas de calibrer correctement les paramètres liés aux caractéristiques de la plante afin d'estimer les paramètres du sol en minimisant au plus le biais d'omission.

1.1.3. Les observations dont on dispose sont généralement peu nombreuses et imprécises

Nous avons vu que les observations sur les sols par des mesures indirectes n'étaient pas encore exploitables pour estimer les propriétés permanentes des sols ainsi que pour fournir des informations sur des variables d'état des sols (contenu en eau et en azote) qui pourraient elles-mêmes être utilisées en mode inverse pour accéder aux propriétés des sols. Seules sont facilement exploitables les observations sur les couverts végétaux obtenues par télédétection ou par les capteurs de rendement, que les agriculteurs capitalisent année après année dans les contextes d'agriculture de précision. Ces données ne sont toutefois pas exemptes de problèmes. Pour les images de télédétection à haute résolution il s'agit de :

(i) la faible répétitivité temporelle, croisée avec une haute probabilité de nuages, qui, dans les régions septentrionales comme la Picardie ne permet souvent pas d'avoir plus de 5 images exploitables par année,

(ii) la variabilité de la richesse spectrale, qui dépend du capteur disponible : celle des capteurs aéroportés de type CASI peut être grande (mais chère et donc rare) et donne accès au *LAI* et à la teneur en azote de la plante *QN* ; celle des capteurs satellites de type SPOT est faible et ne donne accès qu'au *LAI*,

(iii) le problème de l'inversion des mesures de réflectance pour estimer le *LAI* et le *QN* et des erreurs associées, de l'ordre de 17% pour *LAI* et 30% pour *QN* (Moulin et al., 2007). A cette erreur, s'ajoute celle liée à la mauvaise connaissance du contenu en aérosols de l'atmosphère utilisé pour obtenir les réflectances de surface à partir des mesures au niveau du satellite (Launay et al., 2000).

En ce qui concerne les cartes de rendement, établies à l'aide de capteurs embarqués sur la moissonneuse, le principal problème d'obtention réside dans la difficulté à étalonner ces capteurs de manière précise sur l'ensemble de la parcelle, ce qui engendre une erreur de l'ordre de 9% (Machet et al., 2007).

Compte tenu de la complexité du problème posé, il est nécessaire de mettre en œuvre une méthodologie adaptée, qui consiste dans une première étape à cibler les paramètres qui sont les plus pertinents à estimer et dans une seconde étape à

appliquer la méthode d'estimation la mieux adaptée à notre contexte. Nous présenterons dans la suite une manière de répondre à ces objectifs.

1.2. Les méthodes d'estimation et de sélection des paramètres à estimer

1.2.1. Les méthodes d'estimation de paramètres

Il existe un large panel de méthodes pour estimer les paramètres d'un modèle complexe. Ces méthodes peuvent être regroupées en deux principales familles : l'approche fréquentiste et l'approche Bayesienne (Makowski et al., 2006a). La mise en œuvre de l'approche fréquentiste ne nécessite qu'un jeu d'observations alors que l'approche Bayesienne utilise en plus, une information sur la distribution des paramètres à estimer. Les approches Bayesiennes sont devenues de plus en plus utilisées ces dernières années pour estimer les paramètres de modèles complexes car elles permettent de mieux prendre en compte les incertitudes, aussi bien sur les paramètres d'entrée que sur les simulations du modèle. Par ailleurs, leur utilisation a été largement facilitée par le décuplement des vitesses de calcul des ordinateurs et le développement de nouveaux algorithmes.

Les méthodes fréquentistes ne considèrent pas les paramètres du modèle comme étant des variables aléatoires, comme le font les méthodes Bayesiennes, mais plutôt comme étant des variables fixées à une certaine valeur, à estimer. L'application d'une méthode fréquentiste permet alors de déterminer une valeur particulière de chaque paramètre à partir d'un jeu d'observations, et cette valeur est appelée estimateur du paramètre. Parmi ces méthodes fréquentistes, il existe celle du maximum de vraisemblance (Aldrich, 1997; Hald, 1999), des moindres carrés (Seber and Wild, 2003), les algorithmes génétiques (Mitchell, 1998) ou les méthodes variationnelles (Bouttier and Courtier, 1999) qui permettent, en plus d'estimer les paramètres, de contrôler les variables d'état du modèle (lissage, prédiction à court terme, …).

Les méthodes Bayesiennes, quant à elles, utilisent une information supplémentaire sur la distribution des paramètres, dite information *a priori*. Les paramètres étant ici considérés comme des variables aléatoires définies par une densité de probabilité *a priori*, le résultat de l'application d'une approche Bayesienne sur un jeu de données est une nouvelle densité de probabilité appelée densité *a posteriori* des paramètres. L'application d'une méthode Bayesienne peut être abordée en deux étapes. La première consiste à déterminer l'information *a priori* des paramètres à estimer, à partir de différentes sources qui peuvent être constituées par des mesures, la littérature ou bien des dires d'experts. Cette information *a priori* peut se limiter à de simples bornes sur les valeurs des paramètres sans être plus informative sur la distribution de ces valeurs (densité uniforme) ; elle peut aussi être plus précise quant à cette distribution (densité normale, de Poisson, Gamma, …). La seconde étape consiste à déterminer la densité *a posteriori* à partir de la densité *a priori* et du jeu d'observations, en utilisant le théorème de Bayes (Makowski et al., 2006a). Cette densité *a posteriori* peut alors être utilisée à diverses fins comme estimer la valeur la plus probable de chaque paramètre (en considérant le mode ou la moyenne), calculer l'incertitude sur l'estimation des paramètres ou encore calculer l'incertitude sur les variables simulées par le modèle à partir de l'incertitude sur l'estimation des paramètres. Parmi les méthodes Bayesiennes, on peut citer MCMC (Metropolis et al., 1953; Hastings, 1970), Importance Sampling (Beven and Binley, 1992; Beven and Freer, 2001) ou bien certaines méthodes de filtrage (Hilgert et al., 2005; Rossi and Vila, 2005) qui permettent en plus, comme les méthodes fréquentistes variationnelles, de contrôler les variables d'état du modèle.

En ce qui concerne les paramètres du sol, on considère qu'il est possible d'obtenir de l'information *a priori* sur ces paramètres à l'échelle de la parcelle. Cette information est disponible à partir de différentes sources et sont associées à différentes précisions. La première source d'information provient des cartes de sols couplées à des fonctions de pédotransfert qui permettent de fournir, pour une parcelle donnée, des valeurs de paramètres. Cette première source est, nous le rappelons, associée à une faible précision car les cartes sont généralement établies à une échelle plus grande que celle de la parcelle agricole. Cependant, il est par exemple possible de définir l'information *a priori* au niveau intra-parcellaire comme

étant une distribution statistique centrée sur ces valeurs et avec une certaine variance. La seconde source d'information provient de mesures de paramètres issues d'analyses d'échantillons de sols pouvant être réalisées de manière plus ou moins fréquentes dans l'espace. A partir de ces mesures, une distribution statistique peut ensuite être proposée pour chaque paramètre, définissant ainsi l'information *a priori*. Cette seconde source peut amener à déterminer une information *a priori* très précise lorsque les mesures issues des analyses de sols deviennent très fréquentes dans l'espace. L'approche Bayesienne est donc préférée à l'approche fréquentiste dans notre cas. Comme les observations dont on dispose sont généralement peu nombreuses, les méthodes de type filtrage, gourmandes en nombre d'observations, sont donc épargnées. Du point de vue du temps de calcul de la distribution *a posteriori*, la méthode Importance Sampling est une méthode très économique lorsqu'il s'agit d'estimer spatialement (i.e. sur un grand nombre de points) les paramètres du sol dans un domaine comme la parcelle agricole, ce qui n'est pas le cas des méthodes MCMC ou de filtrage. De plus, Importance Sampling s'avère être d'une performance comparable à MCMC (Makowski et al., 2002). Cette méthode est donc celle que nous avons retenue pour faire l'estimation des paramètres sol dans toute notre étude.

Nous avons déjà évoqué au Chapitre 1.1.2, que le principal problème posé par le grand nombre de paramètres à estimer, indépendamment de la méthode choisie, était le problème d'identifiabilité. A cela il faut ajouter d'autres problèmes liés au choix de la méthode fréquentiste ou Bayesienne. Pour les méthodes fréquentistes, il faut aussi compter sur le problème du modèle *sur-paramétré* (l'estimation d'un grand nombre de paramètres conduit à une grande variance des estimateurs) ainsi que sur le problème de la divergence des estimateurs par rapport à la solution optimale (le nombre de minimas locaux est accru par le nombre de paramètres à estimer). Concernant les méthodes Bayesiennes, l'autre principal problème est que l'estimation d'un grand nombre de paramètres est synonyme d'autant de définition d'informations *a priori* que de paramètres à estimer. Ce travail constitue une étude très lourde.

1.2.2. Les méthodes de sélection des paramètres à estimer

Compte tenu de la quantité de problèmes causés par l'estimation d'un grand nombre de paramètres, il est recommandé de sélectionner un sous-groupe de paramètres à estimer et de fixer les autres à une valeur nominale. Pour cela, Makowski et al. (2006a) proposent quatre méthodes pour sélectionner ces paramètres :

- sélection basée sur la littérature,
- sélection pour éviter les problèmes d'identifiabilité,
- analyse de sensibilité,
- choix statistique des paramètres à estimer.

Le principe de la première méthode est de sélectionner les paramètres pour lesquels aucune valeur n'est fournie par la littérature. Par exemple, Bonesmo et Bélanger (2002) fixent 4 paramètres, parmi ceux liés aux caractéristiques de la plante, à des valeurs définies par la littérature et en estiment 17 autres. Cependant, cette méthode n'est pas valable pour l'estimation des paramètres du sol car leurs valeurs ne peuvent pas être définies par la littérature en tout point de l'espace.

L'analyse des équations du modèle STICS permet d'éviter en partie des problèmes d'identifiabilité. Reprenons l'exemple de l'Equation (1-1) du Chapitre 1.1.2. Nous avons vu pour cette équation qu'il n'était pas possible d'estimer simultanément les paramètres *Norg*, *profhum* et *DA(1)*. La méthode proposée ici suggère alors d'estimer soit le produit *Norg*×*profhum*×*DA(1)* dans sa globalité, soit un seul des trois paramètres du produit en fixant les autres à une valeur nominale.

Il est possible de sélectionner les paramètres à estimer, pour un jeu d'observations donné, en appliquant des méthodes statistiques. Pour ce faire, il faut proposer à la méthode une liste de sous-groupes candidats de paramètres à estimer (soit en scrutant toutes les possibilités du groupe de paramètres, soit en faisant une proposition raisonnée de liste) afin de fournir des critères permettant de décider du sous-groupe optimal de paramètres à estimer, compte tenu du jeu d'observations.

Parmi les nombreux critères qui existent, le critère *BIC* est un des plus performants (Tremblay and Wallach, 2004) :

$$BIC = -2\log(Lik) + P\log(N)$$

où $\log(Lik)$ est le logarithme de la fonction vraisemblance, *N* est le nombre d'observations et *P* est le nombre de paramètres du sous-groupe candidat. La fonction vraisemblance (Makowski et al., 2006a) doit être calculée à partir des estimations des paramètres du sous-groupe, ce qui sous-entend le choix préalable d'une méthode d'estimation de paramètres adaptée. Cependant, l'application de cette méthode pour sélectionner les paramètres à estimer peut être très coûteuse en temps de calcul. Par exemple, la sélection optimale d'un sous-groupe de paramètres à estimer parmi un groupe de 13 paramètres mène à proposer 2^{13}=8192 sous-groupes candidats et à effectuer autant de procédures d'estimation de paramètres.

L'analyse de sensibilité est une des méthodes les plus utilisées pour la sélection des paramètres, car elle permet en même temps d'accéder à la compréhension du fonctionnement du modèle (Saltelli et al., 2000b). Elle consiste en effet à détecter les paramètres dont les incertitudes ont un effet significatif sur les variables observables. En quantifiant l'effet de l'incertitude, associée à chaque paramètre, sur les variables observables (sous la forme d'un indice par exemple) et en définissant un seuil en deçà duquel l'effet est considéré comme étant insignifiant, il est ainsi possible de restreindre le groupe de paramètres analysés à une sélection des principaux paramètres à estimer.

Afin de procéder à la sélection des paramètres du sol à estimer, nous avons choisi dans notre étude : (i) de réduire les problèmes d'identifiabilité, en analysant les équations du modèle et en fixant certain paramètres sol à une valeur nominale, (ii) d'utiliser l'analyse de sensibilité des variables observables du modèle STICS aux paramètres. Au détriment des méthodes statistiques telles que *BIC*, l'analyse de sensibilité permet de quantifier la quantité d'information disponible dans les observations pour estimer les paramètres (voir Chapitre 1.3.1) et permet ainsi de répondre à un de nos objectifs. De plus, le nombre de procédures d'estimation, que

la méthode *BIC* suggère d'effectuer pour la sélection, implique un temps de calcul trop important lorsque cette méthode est appliquée au modèle *STICS*. De ce fait, l'analyse de sensibilité est ici préférée aux méthodes statistiques pour sélectionner les paramètres à estimer. Ce travail de sélection de paramètres, à travers les points (i) et (ii), est décrit de manière plus précise dans le Chapitre 3, chapitre dans lequel la question de l'analyse de sensibilité est traitée sous la forme d'un article. Avant cela, il est utile de présenter les différentes méthodes d'analyse de sensibilité d'un modèle complexe.

1.2.3. Les méthodes d'analyse de sensibilité

Les modèles de simulations deviennent de plus en plus complexes et la compréhension du fonctionnement du modèle et du comportement des sorties en fonction des entrées devient de plus en plus difficile. L'analyse de sensibilité du modèle est un moyen intéressant pour aider à cette compréhension. Un certain nombre d'auteurs (dont Campolongo and Saltelli, 1997; Brun *et al.*, 2002; Cariboni *et al.*, 2004; Ratto *et al.*, 2007; Manache and Melching, 2008) qualifient l'objectif de l'analyse de sensibilité comme étant une réponse aux questions du type :

1. Quels sont les paramètres d'entrée θ_i, *i=1,…,P*, dont l'incertitude influence le plus celle sur la sortie $Y = f(\theta)$ du modèle *f* ?
2. Quels paramètres ont un effet négligeable sur l'incertitude de la sortie de sorte qu'il est possible de les fixer à une valeur nominale ?

Répondre à ces questions permet de connaitre, parmi un sous-groupe de paramètres, lesquels sont à estimer en priorité (1) et lesquels sont à fixer à une valeur nominale (2). Pour répondre à ces questions, diverses méthodes d'analyse de sensibilité existent et le choix de l'une d'entre elles doit être fait en fonction du cas d'étude et des propriétés du modèle. Cariboni (2004) dresse un schéma décisionnel concernant le choix de la méthode (voir Figure 1-1) et permet de discriminer les différentes méthodes d'analyse de sensibilité entre elles suivant 3 principaux critères :

- le nombre *P* de paramètres d'entrée considérés,
- le temps d'exécution du modèle *f*,
- la linéarité du modèle.

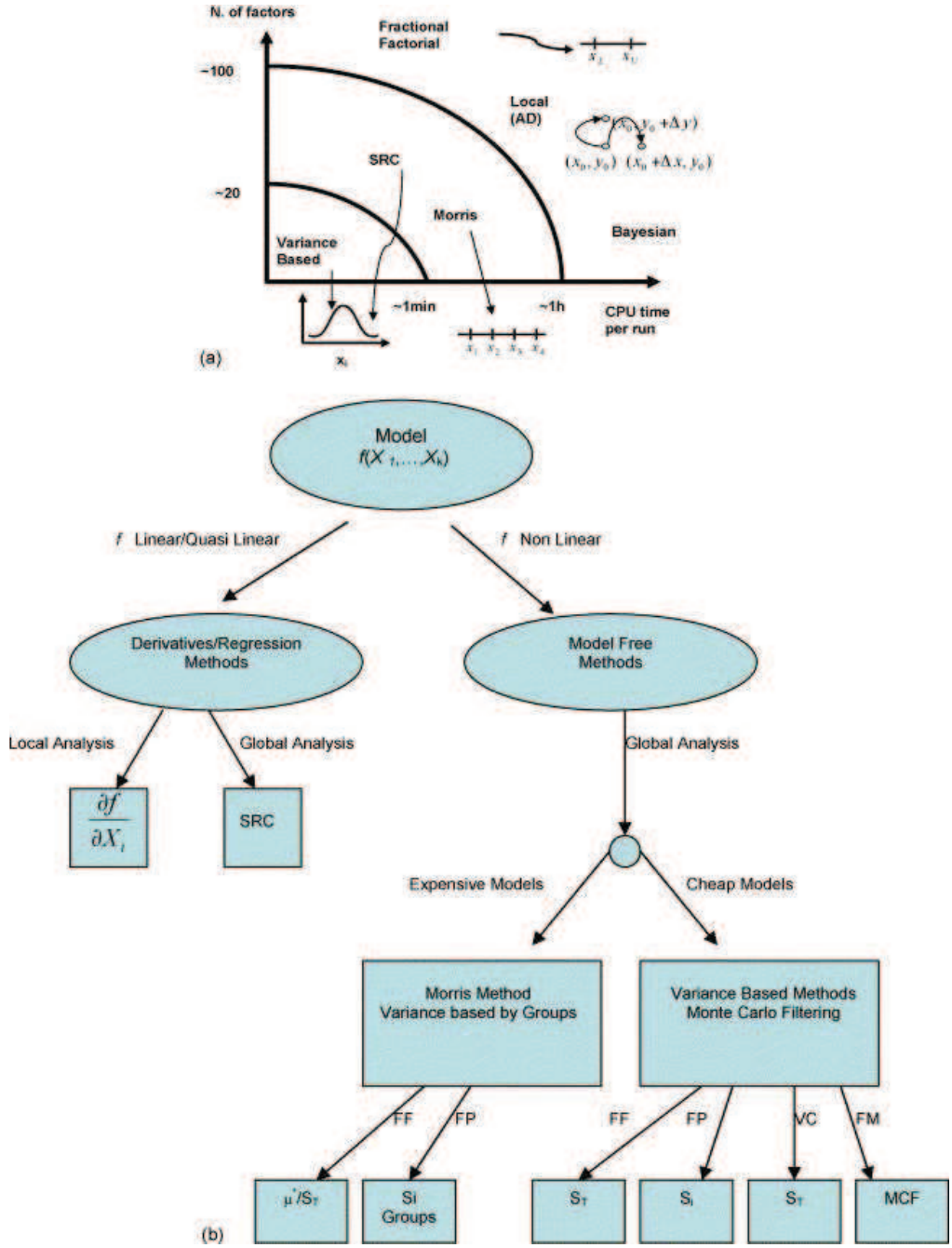

Figure 1-1. Les différents types d'analyses de sensibilité (Cariboni *et al.*, 2004).

Pour un nombre *P* inférieur à 20 et un temps d'exécution inférieur à 1 minute, il est conseillé d'utiliser une méthode basée sur la variance ("Variance Based") (Chan et al., 2000; Makowski et al., 2006b) ; pour un nombre *P* compris entre 20 et 100 et

un temps d'exécution compris entre 1 minute et 1 heure, une méthode "Standardized Regression Coefficients" (SRCs) ou de Morris (Morris, 1991; Campolongo et al., 2006) serait plus adaptée ; lorsque *P* est supérieur à 100 ou bien que le temps d'exécution est supérieur à 1 heure, des méthodes locales type dérivées (Varma et al., 1999; Grievank, 2000), "Fractional Fractorial" ou encore Bayesiennes (Saltelli et al., 2000a) seraient plutôt à envisager. La linéarité du modèle *f* est le dernier critère permettant de discriminer les méthodes entre elles, comme le montre la Figure 1-1b. Lorsque le modèle a un comportement non-linéaire par rapport à l'incertitude des paramètres d'entrée, seules des méthodes de type globales peuvent être utilisées parmi lesquelles on retrouve la méthode de Morris et les méthodes basées sur la variance. Par contre, si le modèle se comporte linéairement, le choix se partage entre les méthodes locales, de type dérivées, et les méthodes globales SRCs ; choix qui doit être affiné en fonction des deux autres critères. En réponse à la question (1), l'utilisation de l'une de ces méthodes permet de quantifier l'influence de l'incertitude des paramètres θ_i, *i=1,…,P*, sur celle de la sortie $Y = f(\theta)$ et de connaître ceux dont il est important d'éliminer l'incertitude pour réduire au minimum celle sur la sortie. Les méthodes locales de type dérivées donnent des résultats informatifs seulement lorsque le modèle est linéaire ou si la gamme d'incertitude est petite ; si la gamme est grande et le modèle proche du linéaire, les méthodes SRCs sont des outils intéressants pour répondre à ces questions ; sinon d'autres types de méthodes globales doivent être utilisés, comme la méthode de Morris ou les méthodes basées sur la variance. Ces deux dernières méthodes permettent également de quantifier l'importance que les paramètres peuvent avoir sur la sortie lorsqu'ils agissent en interaction entre eux. Pour un paramètre donné, il est possible que son incertitude n'influence celle sur la sortie qu'à travers son interaction avec un ou plusieurs autres paramètres. Si l'incertitude de ce paramètre n'influence pas celle sur la sortie, même à travers son interaction, il est donc possible de le fixer à une valeur nominale. En ce sens, la méthode de Morris ou les méthodes basées sur la variance permettent de répondre à la question (2).

Dans notre cas d'étude appliqué aux paramètres sol du modèle de culture STICS, on ne s'intéressera qu'à un nombre *P* de paramètre toujours inférieur à 20. De plus, le modèle STICS est un modèle complexe non linéaire et une exécution du

modèle est de l'ordre de la seconde. Les méthodes basées sur la variance semblent donc être les plus adaptées à notre problème. Parmi elles, deux méthodes sont couramment utilisées : la méthode de Sobol' (Sobol, 1993) et Extended FAST (Saltelli et al., 1999). Makowski (2006b) montre sur un modèle peu coûteux, non linéaire et sur 13 paramètres que les résultats fournis par la méthode Extended FAST convergent plus rapidement vers des solutions stables que ceux fournis par la méthode de Sobol'. Ce comportement a également été constaté par d'autres auteurs (Saltelli and Bolado, 1998; Saltelli et al., 1999). La méthode globale Extended FAST sera donc celle considérée dans toute notre étude.

1.3. Les questions posées par l'estimation des paramètres sol et la prédiction de variables agroenvironnementales

1.3.1. Lien entre analyse de sensibilité et quantité d'information disponible dans les observations

Une certaine quantité d'information contenue dans les observations est nécessaire pour estimer les paramètres sélectionnés. Cependant, les observations ne fournissent pas la même quantité d'information sur tous les paramètres, ce qui implique une disparité dans les performances d'estimation des paramètres. Comme nous le verrons dans ce chapitre, la quantité d'information peut être mesurée via l'analyse de sensibilité des variables observables aux paramètres. Aussi, la linéarité du modèle par rapport à l'incertitude sur les paramètres est un aspect primordial à considérer avant de choisir une méthode d'analyse de sensibilité. L'analyse de sensibilité est classiquement utilisée dans le cas linéaire pour mesurer la quantité d'information, traduisant la performance d'estimation, alors que dans le cas non linéaires, peu de travaux ont été réalisés sur l'utilisation de l'analyse de sensibilité globale pour quantifier cette quantité. Cette question a été abordée dans ce travail de thèse. Dans ce chapitre, nous commencerons par rappeler comment la quantité d'information est mesurée dans le cas linéaire, pour ensuite présenter les bases de la mesure de cette quantité dans le cas non linéaire.

La quantité d'information au sens de Fisher

La question posée dans cette section est de savoir ce que nous apprend un jeu d'observations sur les paramètres. Plus précisément, quelle quantité d'information le jeu d'observations fournit-il sur les paramètres ? C'est ce que la théorie de l'information au sens de Fisher (Kauffmann, 1994) permet de déterminer. Ainsi, il est possible de quantifier ce que le jeu d'observations donne comme information pour estimer les paramètres, permettant de distinguer ceux qui peuvent être estimé de ceux qui ne le peuvent pas.

Soit un jeu d'observations composé de *K* observations, les sorties observées associées du modèle, f_k, *k=1,…,K*, et un sous-groupe de *P* paramètres θ_p, *p=1,…,P*. La théorie de l'information au sens de Fisher permet de construire une matrice d'information, appelée *FIM*, telle que $FIM = Q^T W Q$ où *W* est une matrice de covariance d'erreur d'observation et *Q* est de la forme suivante :

$$Q = \begin{bmatrix} \frac{\partial f_1}{\partial \theta_1} & \cdots & \frac{\partial f_1}{\partial \theta_P} \\ \vdots & \cdots & \vdots \\ \frac{\partial f_K}{\partial \theta_1} & \cdots & \frac{\partial f_K}{\partial \theta_P} \end{bmatrix}$$

$\partial f_k / \partial \theta_p$ est la dérivée de la fonction f_k par rapport au paramètre θ_p. Dès lors, deux cas de figure peuvent se présenter en ce qui concerne la linéarité des sorties :

- soit les sorties sont linéaires par rapport à l'incertitude sur les paramètres,
- soit les sorties ne sont pas linéaires par rapport à l'incertitude sur les paramètres.

Dans le cas linéaire, la matrice *Q* peut être vue comme une matrice d'indices de sensibilité où $\partial f_k / \partial \theta_p$ est l'indice de sensibilité de f_k par rapport à θ_p. Les paramètres ayant un indice de sensibilité élevé sont ceux qui peuvent être estimés avec précision, grâce au jeu d'observations. De plus, chaque composante C_{pq}, *p=1,…,P*, *q=1,…,P* de la matrice *FIM* quantifie la quantité d'information qui est fournie par le jeu d'observations sur le couple de paramètres (θ_p, θ_q). Pour finir, le calcul d'un critère d'optimalité à partir de *FIM* tel que le critère A, D ou E (Rodriguez-

Fernandez *et al.*, 2006; Pronzato, 2008) permet de quantifier la quantité d'information fournie sur l'ensemble du sous-groupe de paramètres.

Dans le cas non linéaire, la matrice *FIM* n'a de sens que si elle est calculée pour les valeurs estimées des paramètres. Dans ce cas, la matrice *FIM* permet, comme pour le cas linéaire, de quantifier la quantité d'information qui a été apportée par le jeu d'observations pour estimer les paramètres. Cependant, il n'est pas possible de calculer les indices de sensibilité des sorties du modèle par rapport à l'incertitude sur les paramètres, ainsi que de quantifier l'information apportée par le jeu d'observations pour réduire l'incertitude sur les paramètres. Pour répondre à cette question, d'autres types d'analyses que celle proposée par Fisher peuvent être entreprises. Parmi elles, les méthodes d'analyse de sensibilité globales.

Le cas non linéaire et l'utilisation de l'analyse de sensibilité globale

L'application d'une méthode d'analyse de sensibilité globale à un modèle non linéaire permet seulement d'apprécier qualitativement la quantité d'information apportée par une observation pour estimer les paramètres. Lorsqu'un paramètre agit directement sur la variable observable de sortie de manière importante et non via ses interactions avec les autres paramètres, on peut dire que l'observation de la variable fournit une grande quantité d'information pour estimer ce paramètre et qu'il est alors possible d'éliminer son incertitude. Ratto et al. (2007) qualifie de "Factor Prioritization" (FP sur la Figure 1-1b) le fait de s'intéresser à un paramètre pour en éliminer son incertitude. Lorsqu'un paramètre n'agit d'aucune manière (ni directement, ni en interaction) sur la variable observable de sortie, on peut dire que l'observation de la variable ne fournit aucune information pour estimer ce paramètre et qu'il peut être alors fixé à une valeur nominale. Ratto et al. (2007) qualifie de "Factor Fixing" (FF sur la Figure 1-1b) le fait de fixer à une valeur nominale un paramètre n'ayant aucune influence sur la sortie. Par exemple, Ruget (2002) montre sur le modèle de culture STICS qu'il est possible de choisir les principaux paramètres à estimer en étudiant la sensibilité des principales variables de sortie de STICS à 28 paramètres d'entrée (plante, sol et technique). Par conséquent, son travail montre qu'il est possible de qualifier la quantité d'information contenue dans l'observation de

chaque variable de sortie du modèle, afin de choisir les principaux paramètres à estimer.

Cependant, aucune étude n'a permis à ce jour de quantifier la quantité d'information qui est apportée par l'ensemble d'un jeu d'observations sur l'estimation des paramètres sol d'un modèle complexe comme STICS. Nous proposons alors de répondre à cette question au Chapitre 4 en proposant des critères, basés sur les indices de sensibilité globaux, qui mesurent cette quantité d'information, elle-même liée à la performance d'estimation des paramètres.

1.3.2. La prédiction des variables agroenvironnementales dépend de l'estimation des paramètres du sol

L'estimation des paramètres du sol à partir d'observations sur le couvert végétal est un objectif en soi, permettant d'accéder à la connaissance de propriétés difficilement mesurables sur des espaces étendus. Comme nous l'avons vu au chapitre précédent, cette connaissance dépend de la quantité d'information qui est fournie par les observations. En outre, l'estimation des paramètres du sol peut également être considérée comme un moyen d'affiner l'utilisation du modèle de culture, en améliorant le renseignement des paramètres du sol, visant à améliorer les prédictions des variables agroenvironnementales.

La liaison étroite qui existe entre les variables observées et les variables d'intérêt, car corrélées entre elles d'une certaine manière, implique que la sensibilité des variables observables et celle des variables à prédire aux paramètres du sol sont similaires. Dans ce cas, la méthodologie d'estimation des paramètres peut s'avérer être efficace pour améliorer les prédictions. Par exemple, la teneur en protéines est directement liée au rendement et à la quantité d'azote contenu dans la plante, ce qui signifie que leurs sensibilités respectives aux paramètres doivent être semblables. En ce sens, il est donc possible d'améliorer les prédictions des variables d'intérêt, en tout point de l'espace, à partir de l'estimation des paramètres du sol issue de l'inversion de STICS avec des observations du couvert végétal. Cette réflexion n'est pas nouvelle, elle est sous-jacente au processus connu sous le nom de

calibration/validation, où un certain nombre de paramètres sont estimés à partir d'un jeu d'observations (étape de calibration) et où l'impact de cette estimation est testé et analysé en prédiction sur un jeu d'observations indépendant (étape de validation). Un grand nombre de travaux liés à la calibration/validation existent dans la littérature (dont Hadria *et al.*, 2007; Tonitto *et al.*, 2007; Beaudoin *et al.*, 2008; Heidmann *et al.*, 2008; Dimokas *et al.*, 2009). Par exemple, Heidmann (2008) propose de calibrer des paramètres de croissance d'un modèle de culture de pomme de terre avec des observations sur deux années de matière sèche (liée à l'indice foliaire) et de quantité d'azote contenue dans la plante, pour ensuite valider cette approche sur une autre année en prédisant de la matière sèche et de la quantité d'azote contenue dans la plante. Un autre exemple, Tonitto (2007) cale des paramètres chimiques et de sol, qu'il qualifie de mal connus *a priori*, avec des observations de drainage et de lessivage des nitrates sur plusieurs années, pour ensuite améliorer la prédiction du rendement du maïs et du soja, par rapport à celle issue des valeurs *a priori* des paramètres. Dans ces travaux, il y a une réelle volonté d'améliorer les prédictions grâce à l'estimation de paramètres, ce qui est généralement soldé par un succès, car les variables prédites et les variables observées présentent des sensibilités aux paramètres estimés comparables. Dans notre cas d'étude, l'estimation des paramètres du sol pourrait donc permettre d'améliorer la prédiction des variables d'intérêt agroenvironnemental.

Dans notre étude, nous nous sommes intéressés à l'amélioration de la prédiction de variables agroenvironnementales de STICS à partir de l'estimation des paramètres du sol sous un large panel de jeux d'observations (croisant différentes configurations agropédoclimatiques), lequel n'a jamais été aussi exhaustif dans les précédentes études. Le lien existant entre les variables observables et les variables à prédire nous a donc permis d'expliquer les performances de prédiction en fonction des performances d'estimation, compte tenu d'un jeu d'observations donné. Ce travail est présenté au Chapitre 5.

1.3.3. La distribution spatiale des paramètres du sol dans la parcelle est structurée et la prise en compte de cette structure peut améliorer l'estimation des paramètres

Dans les travaux précédents, l'estimation des paramètres du sol par inversion d'un modèle de culture est entreprise sur chaque point de la parcelle agricole sans tenir compte de la structure spatiale des valeurs de paramètres (Houlès, 2004; Guérif et al., 2006). Certains auteurs travaillant à une échelle plus large (telle que le bassin de production) ont développé des approches utilisant des contraintes spatiales. Par exemple, Lauvernet et al. (2005) forcent les valeurs des paramètres liés aux caractéristiques de la plante à être égales sur chaque point de l'espace couvert par la même variété. Pour notre cas d'application, il pourrait être envisageable de forcer les valeurs des paramètres du sol à être égales sur chaque point de l'espace couvert par le même type de sol, mais les cartes de sol, nous le rappelons, sont très rares à l'échelle de la parcelle et sont très coûteuses. Il semble plus intéressant de considérer une information spatiale qui renseignerait sur la structure spatiale des propriétés des sols et que l'on pourrait introduire dans le processus d'estimation.

La disponibilité de nouveaux outils d'investigation pour l'étude du proche sous-sol s'appuyant sur des méthodes géophysiques ouvre des perspectives intéressantes dans ce domaine. Parmi elles, la mesure de la résistivité électrique des sols (Samouelian et al., 2005; Bourennane et al., 2007) qui représente la capacité d'un horizon de sol à limiter le passage du courant électrique, et est étroitement liée à ses caractéristiques intrinsèques. L'interprétation du signal fait l'objet de nombreux travaux et même s'il est encore très difficile de retrouver les propriétés des sols précisément (Samouelian *et al.*, 2005), il n'en demeure pas moins que cette méthode donne accès de manière automatisée à la structure spatiale de certaines caractéristiques du sol de la parcelle.

Prenons le cas d'une parcelle expérimentale du site agriculture de précision de Chambry (Guérif *et al.*, 2001), située dans l'Aisne et décrite par Nicoullaud et al. (2004). Cette parcelle présente une topographie relativement accentuée (7 mètres de dénivelé). Elle est située en position haute, favorisant ainsi l'érosion des matériaux. La répartition des sols dans le paysage est étroitement liée à la

topographie : le sommet de la parcelle et le versant orienté sud - sud ouest, soumis à l'érosion, sont occupés de sols calcaires développés sur la craie ou la craie sableuse magnésienne. Le versant opposé (nord-nord est) est occupé par des sols profonds moins calcaires développés sur la craie cryoturbée et de sols limoneux profonds développés sur des formations sablo-calcaires. Dans le cadre du projet agriculture de précision, un grand nombre de mesures du sol ont été faites sur cette parcelle, permettant d'accéder à sa structure spatiale. La Figure 1-2 illustre les types de structure spatiale rencontrés pour trois propriétés de l'horizon de surface de la parcelle en question, avec la carte des valeurs interpolées et le semi-variogramme associé (Chilès and Delfiner, 1999) : soit les valeurs sont corrélées quelle que soit la distance (pour la teneur en calcaire *CA* et dans une moindre mesure pour le taux d'argile *AR*), exprimant l'existence d'un gradient dans la parcelle, soit elles sont corrélées jusqu'à une certaine distance (jusqu'à environ 80 m pour le taux de sable fin *SF*), et indépendantes ensuite. La comparaison avec la Figure 1-3 montre qu'il existe une similitude de structure spatiale entre les valeurs de la teneur en calcaire *CA* et les valeurs de la résistivité électrique.

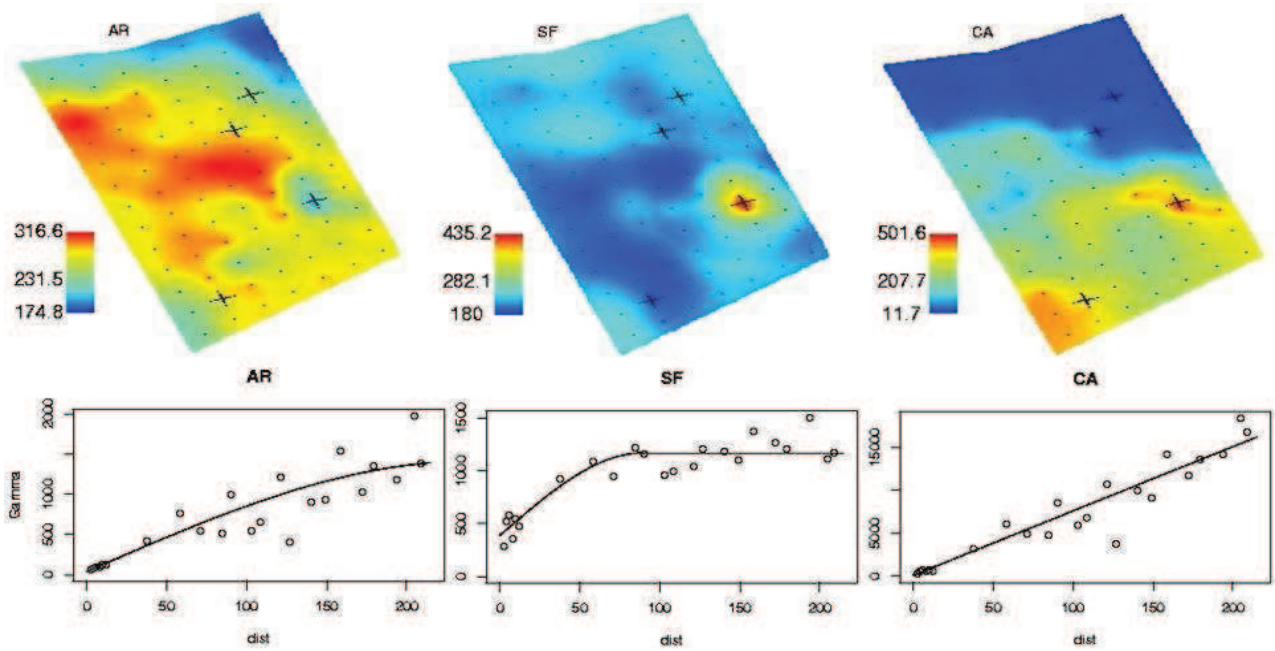

Figure 1-2. Semi-variogrammes de la teneur en argile (*AR*), en sable fin (*SF*) et en calcaire (*CA*), obtenus à partir de mesures du sol de la parcelle P2 de Chambry.

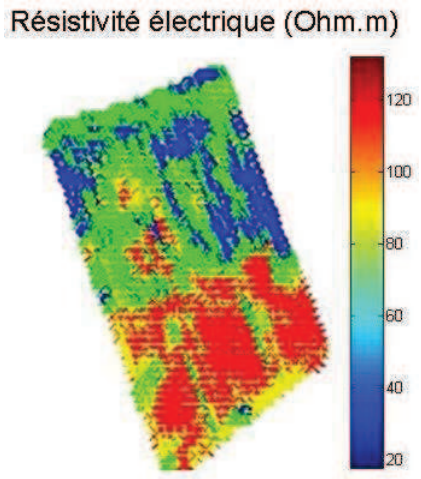

Figure 1-3. Mesures de résistivité électrique (en Ohm.m) sur la parcelle P2 de Chambry.

La résistivité électrique du sol permet donc de renseigner sur la structure spatiale de certaines de ses propriétés. Dans le Chapitre 6, nous proposons une voie d'utilisation de cette information dans le processus d'estimation des paramètres et analysons les possibilités d'amélioration de l'estimation.

1.3.4. Les questions de recherche prises en compte et organisation de la thèse

L'estimation des paramètres du sol d'un modèle de culture afin d'améliorer les prédictions de variables agroenvironnementales n'est donc pas chose aisée. Nous avons vu, à travers l'état de l'art dressé dans ce chapitre, qu'un certain nombre de questions de recherche ont été révélées et méritent notre attention. Les réponses à ces questions permettraient alors de faire un meilleur usage du modèle de culture STICS pour la prédiction des variables d'intérêt.

Les quatre questions de recherche prises en compte dans cette thèse sont alors les suivantes :

1. Quel est le sous-groupe optimal de paramètres du sol à sélectionner pouvant être estimé convenablement ?
2. L'analyse de sensibilité globale permet-elle de mesurer la quantité d'information contenue dans les observations et la qualité d'estimation des paramètres ?

3. La prédiction des variables d'intérêt est-elle améliorée à partir de l'estimation des paramètres du sol ?
4. La prise en compte d'une structure spatiale dans l'estimation des paramètres du sol permet-elle d'améliorer leur estimation ?

Après un chapitre consacré à la présentation des outils et des données utilisés dans l'ensemble des travaux (Chapitre 2), la suite de la thèse s'articule autour de ces 4 questions. Elle s'achève par une conclusion générale, incluant un bilan des résultats, une discussion et la présentation de perspectives.