O.D.R. Observatoire Du Développement Rural	NOTE METHODOLOGIQUE **L'indicateur d'intensité agricole *Coûts Intrants par hectare* : méthode de calcul et d'estimation à l'échelle de la petite région agricole**	**Auteurs :** F. Teillard[1] G. Allaire E. Cahuzac F. Léger E. Maigné M. Tichit Version du 01/12/2011

1 L'indicateur *Coûts Intrants par hectare*

L'indicateur d'intensité agricole *"Coût Intrants par hectares"* décrit ici est exprimé en €/ha, on l'abrège *"IC / ha"* (*Input Cost / ha* en anglais). Il se définit comme le ratio entre la somme des coûts dépensés pour diverses catégories d'intrants - engrais et amendements, aliments pour le bétail, pesticides, semences, carburants, produits vétérinaires et eau d'irrigation - et la surface agricole utile de l'exploitation.

L'utilisation de coûts permet d'agréger différentes catégories d'intrants et d'obtenir un indicateur continu. L'indicateur *IC / ha* est calculé pour 5 types de productions agricoles : cultures industrielles, céréales oléagineux protéagineux, mixte culture / élevage, élevage bovin lait et élevage bovin viande. 67% des exploitations professionnelles françaises appartiennent à ces types de production, elles occupent environ 80% de la surface agricole du pays. Les types de production ayant une faible empreinte territoriale (maraîchage, horticulture) ou étant principalement hors-sols (élevage de granivores) ont été exclus du calcul de l'indicateur. Ont également été exclues les cultures pérennes (vignes, vergers) dans lesquelles les niveaux d'intrants trop élevés pour être comparables à ceux des types de productions pris en compte.

La base de données utilisée pour le calcul de l'*IC / ha* pour l'année 2006 est le Réseau d'Information Comptable Agricole (RICA). Le RICA contient spécifiquement en plus d'autres variables, le détail des coûts dépensés pour les catégories d'intrants prises en compte dans l'indicateur, la surface agricole utile des exploitations (utilisée au dénominateur de l'indicateur *IC / ha*). Les informations sont fournies au niveau d'exploitations individuelles, enquêtées chaque année. Cependant, l'échantillon d'exploitations est limité (7361 exploitations en 2006 soit seulement 2% environ des exploitations professionnelles française) et ne confère de représentativité qu'au niveau régional. D'autres bases de données plus exhaustives on ainsi été utilisées pour estimer la valeur de l'*IC / ha* à l'échelle de la petite région agricole (PRA).

2 Estimation à la petite région agricole

2.1 Données

L'ensemble des variables enquêtées par le RICA est plus large que celles décrites dans la section 1 et utilisées pour le calcul de l'indicateur d'intensité *IC / ha* . Un second groupe de variables concernant les caractéristique socio-économiques, topo-climatiques, d'utilisation des sols et des primes perçues est aussi disponible. On retrouve les variables de ce second groupe dans quatre autres jeux de données (voir le tableau A.1 en annexes) qui présentent l'avantage de fournir leur valeur pour un échantillon exhaustif et à un niveau fin, celui de l'exploitation ou de la commune. Comme pour le RICA, l'année 2006 de ces différentes bases a été utilisée. La base de données de la mutuelle sociale agricole (MSA) collecte un ensemble de variables socio-économiques à l'échelle de l'exploitation pour la quasi totalité des exploitations françaises (n = 306773 exploitations). La représentation étant exhaustives, les valeurs peuvent être agrégées pour obtenir une moyenne au niveau de la commune. La base de données des déclarations PAC (Politique Agricole Commune) fournit une information sur l'usage des sols et les primes perçues, à l'échelle de chaque commune française. Seules les communes dominées par l'urbain ne sont pas concernées (n = 34609 communes disponible sur un total de n = 36584 en France métropolitaine). Pour chaque commune, la surface des différents usages des sols est connue, ainsi que la somme des primes de différentes catégories, perçues par toutes les exploitations situées dans la commune. La base de données nationale d'identification des bovins suit chaque animal individuellement et a donc permis d'obtenir une valeur du chargement (unités gros bétail par unité de surface et de temps) dans chaque commune où des bovins ont été présents (en 2006, n = 27399 communes). Enfin, les données topo-climatiques utilisées sont celles de la base "Météo France", couvrant l'ensemble des communes françaises.

[1]Coordonnées
INRA UMR 1048 SAD APT
16, rue Claude Bernard
75231 Paris Cedex 05, France
Tel: +33144081775
Email: teillard@agroparistech.fr

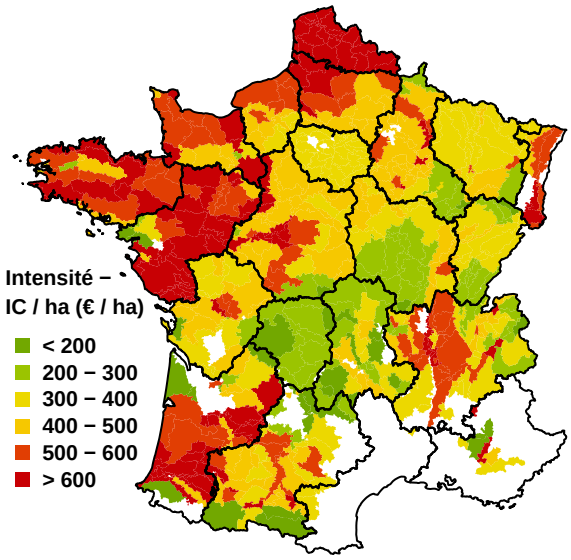


Figure 1: Valeur de *Coûts Intrants par hectare* (*IC / ha*) des petites régions agricoles (PRA) françaises dominées par la production de cultures industrielle, céréales oléagineux protéagineux, mixte bovin cultures et bovin lait et viande. Ces 5 types de productions doivent occuper plus de 60% de la surface agricole utile de la PRA. Les autres PRA apparaissent en blanc. La valeur continue de l'*IC / ha* est ici représentée en 6 classes, de la plus extensive (vert) à la plus intensive (rouge) (voir légende). Les frontières des régions administratives (NUTS 2) sont tracées en noir.

2.2 Méthode d'estimation

Nous avons développé une méthode d'estimation de l'indicateur *IC / ha* à l'échelle de la PRA. Le zonage PRA est conçu spécifiquement pour définir des unités homogènes pour leur systèmes agricoles et conditions topo-climatiques (Klatzmann, 1955). La largeur moyenne (± écart type) d'une PRA est de 22.4 ± 13km. Le niveau d'agrégation PRA permet d'obtenir le meilleur compromis entre la précision de l'estimation, et sa finesse géographique. Les valeurs de l'*IC / ha* à la PRA sont obtenues en agrégeant des valeurs d'abord calculées au niveau de la la commune. Afin de calculer ces valeurs communales, les variables communes entre le RICA et les autres bases de données ont été utilisées comme prédicteurs de l'*IC / ha* dans une régression multinomiale calibrée sur les exploitations RICA. Les modèles de régression ainsi calibrés on permis, pour chaque commune, d'estimer la valeur de l'*IC / ha* à partir de la valeur des prédicteurs disponible via les autres bases de données (section 2.1).

La première étape de l'estimation est donc la calibration d'un modèle de régression multinomiale sur les exploitations du RICA. La valeur de l'*IC / ha* de chaque exploitation i (Y_i^{FADN}) est estimée à partir d'une combinaison linéaire de p prédicteurs :

$$\mathrm{Y_i^{FADN}} = \alpha + \beta_1 \mathrm{X_{1i}^{FADN}} + ... + \beta_\mathrm{p} \mathrm{X_{pi}^{FADN}} + \epsilon_\mathrm{i} \quad (1)$$

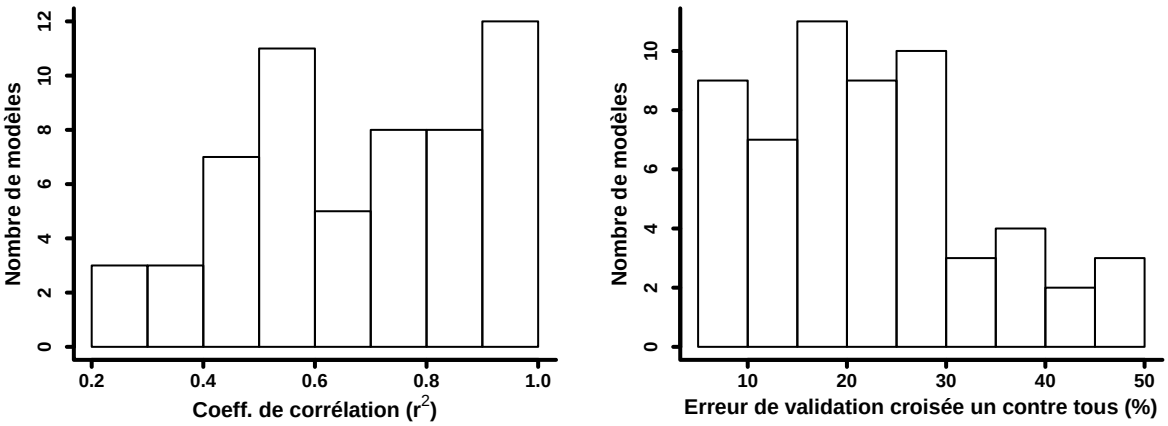
Le modèle calcule l'intercept (α), et les pentes (β) associés à chacun des p predicteurs. On suppose une distribution normale des résidus ϵ. A partir du modèle maximal comprenant tous les prédicteurs comme variables explicatives, on procède à une simplification jusqu'au modèle minimal adéquat ne comprenant que les p' prédicteurs ayant le meilleur pouvoir explicatif (Crawley, 2007). Cette procédure de simplification par étapes est effectuée sur la base du critère AIC, en utilisant la fonction *mixAIC* du package R *MMIX*.

Le modèle minimal est ensuite ustilisé pour estimer la valeur de l'indicateur d'intensité *IC / ha* pour chaque commune ($\hat{Y}_j^M$) (équation (2)). Cette estimation se base sur la valeur de tous les prédicteurs p' restant dans le modèle minimal, connue pour chaque commune, ainsi que sur celle de l'intercept (α) et des pentes (β) précédemment calculées.

$$\hat{\mathrm{Y}}_\mathrm{j}^\mathrm{M} = \alpha + \beta_1 \mathrm{X_{1j}^M} + ... + \beta_{\mathrm{p'}} \mathrm{X_{p'j}^M} \quad (2)$$

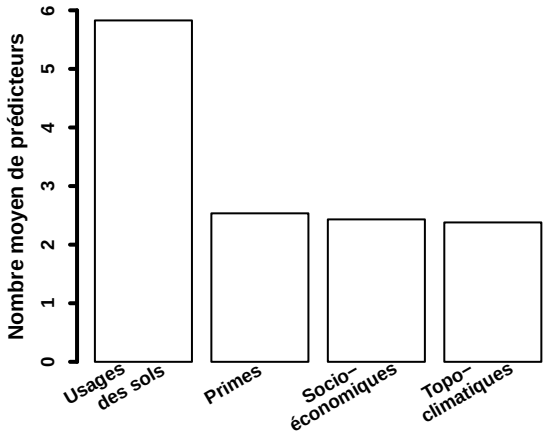
On suppose que le pouvoir explicatif des variables utilisé comme prédicteur n'est important qu'au sein de systèmes agricoles relativement homogènes. Aussi, la procédure décrite ci-dessus est effectuée séparément pour chaque type de production agricole, au sein de chaque région (84 combinaisons possibles). Afin d'éviter d'employer des modèles sur-paramétrés, ceux-ci n'ont été calibrés que sur des échantillons comprenant un nombre suffisant d'exploitations RICA (> 20 exploitations, n = 58 modèles). Ainsi, la valeur de l'*IC / ha* peut se trouver non calculé pour certains types de productions dans certaines régions.

Nous avons enfin agrégé les valeurs communales d'*IC / ha* ainsi calculées, au niveau de la PRA en effectuant une moyenne arithmétique. Ce niveau d'agrégation permet d'améliorer la précision de l'estimation (se référer à la section 3) tout en conservant une résolution spatiale satisfaisante. Les valeurs de l'indicateur d'intensité *IC / ha* pour toutes les PRA dominées par les types de production pris en compte sont présentées dans la figure 1.



(a)

(b)



(c)

Figure 2: Carractéristiques des modèles de régression utilisés pour l'estimation de l'indicateur d'intensité *Coûts Intrants par hectare*. Un modèle est utilisé pour chaque type de production et pour chaque région (n = 58 modèles au total). (2a) Distribution des valeurs des coefficients de corrélation (r^2) des modèles. (2b) Nombre moyen de prédicteurs retenus dans les modèles minimales, pour chacune des quatre catégories de variables. (2c) Distribution des valeurs d'erreurs de prédictions (validation croisée un contre tous, erreur exprimée en pourcentage de la valeur réelle).

3 Validation

Nous avons utilisé plusieurs critères pour évaluer la qualité de l'estimation de l'*IC / ha* . Le coefficient de corrélation de chaque modèle a été calculé, et le critère AIC a été utilisé pour la sélection de modèle (section 2.2). Par ailleurs, nous avons utilisé une méthode de validation croisée pour évaluer la précision des prédictions des modèles (Geisser, 1975; Hawkins et al., 2003). Les échantillons d'exploitations RICA utilisé pour la calibration étant souvent de taille modérée, nous nous sommes orienté vers une méthode de validation croisée "un contre tous" (leave-one-out cross-validation). Nous avons calculé pour chaque modèle, l'erreur moyenne de validation croisée un contre tous, exprimée en pourcentage de la valeur observée (Geisser, 1975; Zuur et al., 2007).

$$\mathrm{CV} = \frac{1}{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{100 * \left| \mathrm{Y_i} - \hat{\mathrm{f}}^{-\mathrm{i}}(\mathrm{X_i}) \right|}{\mathrm{Y_i}} \tag{3}$$

L'erreur moyenne de validation croisée un contre tous est ici notée CV. Elle correspond à la moyenne de la valeur absolue de la différence (en %) entre la valeur réelle d'*IC / ha* d'une exploitation RICA i (Y_i) et celle estimée à partir de la valeur de ses prédicteurs, en excluant cette exploitation de la calibration du modèle ($\hat{f}^{-i}(X_i)$).

En plus de son usage pour évaluer la qualité des modèles, le coefficient CV nous a également servi à déterminer le niveau d'agrégation des valeurs d'*IC / ha* offrant le meilleur compromis entre la résolution de l'estimation et sa précision. Quatre niveaux d'agrégation étaient envisagés: la commune (non agrégé), le canton, la PRA et le département (NUTS 3). Les valeurs respective du coefficients CV pour ces niveaux d'agrégations étaient 56%, 38%,

23% and 18%. Le niveau d'agrégation retenu était la PRA : elle confère un gain de précision important par rapport au canton, en revanche elle confère une précision assez comparable à celle du département tout en représentant une bien meilleure résolution.

Les modèles de régression utilisés pour l'estimation de l'*IC / ha* présentent de bonnes performances prédictives et d'adéquations aux données (figure 2). Parmi les types de productions et régions, les coefficients de corrélations des modèles varient de 0.26 à 0.96, la moyenne ($\pm$ écart type) étant de 0.68 $\pm$ 0.22 (figure 2a). Seulement 23% des modèles présente une valeur de r^2 inférieure à 0.5 alors que cette valeur est supérieure à 0.7 pour la majorité des modèles. Le nombre moyen de prédicteurs p' retenus dans les modèles minimaux est de 13 (figure 2b, voir section 2.2). Les variables d'usages des sols sont les plus fréquemment utilisées, 6 variables de cette catégorie étant en moyenne présentes dans les modèles minimaux (voir aussi le tableau A.1 en annexes pour la liste des variables par catégories). La fréquence de sélection est similaire pour les autres catégories de variables, $\approx$ 2 à 3 de chaque catégories étant en moyenne retenues dans les modèles minimaux. Enfin, les modèles présentent de bonnes capacités prédictives (figure 2c). Sur l'ensemble des modèles, l'erreur moyenne de validation croisée un contre tous est de 23%. Une large majorité de modèles (plus de 60 %) présente une erreur de prédiction inférieure à 25% de la valeur observée.

Références

Crawley, M., 2007. The R book. Wiley, Chichester.

Geisser, S., 1975. The Predictive Sample Reuse Method with Applications. Journal of The American Statistical Association 70 (350), 320–328.

Hawkins, D. M., Basak, S. C., Mills, D., 2003. Assessing model fit by cross-validation. Journal of chemical information and computer sciences 43 (2), 579–86.

Klatzmann, J., 1955. La localisation des cultures et des productions animales en France. Tech. rep., INSEE.

Zuur, A., Ieno, E., Smith, G., 2007. Analysing ecological data. Vol. 129. Springer Verlag.

Annexes

Categorie	Source	Variable	Unité
Usage des sols	Déclarations PAC	Blé	% de la SAU
		Maïs	% de la SAU
		Protéagineux	% de la SAU
		Colza	% de la SAU
		Pommes-de-terres	% de la SAU
		Betterave	% de la SAU
		Jachère	% de la SAU
		SFP	% de la SAU
		Maïs fourrage	% de la SAU
		Cultures fourragères	% de la SAU
		Prairies permanentes	% de la SAU
	BDNI	Chargement	UGB/ha
	Corine Land Cover	Forêts	% de la surf. communale
Socio-économiques	Mutuelle Sociale Agricole	Revenu	€/an
		Salaires versés	€/an
		Employés	Equivalent temps-plein
		Taille de l'exploit.	ha
		Nb. d'exploitants	Nombre
		Année d'installation	Année
		Type de production	Catégorie
Topo-climatiques	Météo France	Altitude	m
		Température moyenne	°C
		Température min.	°C
		Température max.	°C
		Précipitations	mm
		Nb. de jours avec précipitations	Nombre
		Nb. de jours de gel	Nombre
		Humidité relative moyenne	%
Primes	Déclarations PAC	Pratiques extensives	€
		Abattage	€
		Vaches allaitantes	€
		Bovins mâle	€

Table A.1: Source, liste et unité de toutes les variables utilisées comme prédicteurs dans les modèles de régression permettant d'estimer l'indicateur d'intensité *Coûts Intrants par hectare*. La source des données est la même pour toutes les variables d'une même catégorie, sauf lorsque le contraire est spécifié. Abréviations utilisées : PAC = Politique Agricole Commune, SAU = Surface Agricole Utile, BDNI = Base de Données Nationale d'Identification individuelle des bovins, SFP = Surface Fourragère Principale, UGB = Unités Gros Bétail.