Huitièmes Rencontres de Théo Quant
Besançon, 10-12 janvier 2007
http://thema.univ-fcomte.fr/

Probabilité d'apparition d'un phénomène parasitaire et choix de modèles de régression logistique

Florian Tolle, François Pierre Tourneux

Laboratoire ThéMA UMR 6049, CNRS/Université de Franche-Comté
32, rue Mégevand – 25000 Besançon, France
florian.tolle@univ-fcomte.fr francois.tourneux@univ-fcomte.fr

MOTS-CLÉS

Épidémiologie spatiale
Régression logistique binaire
Courbes ROC
Modélisation prédictive

RÉSUMÉ

Les processus épidémiologiques sont de plus en plus fréquemment abordés à l'aide des outils de la statistique spatiale et de la modélisation. Ces travaux ont généralement pour but de mettre en évidence des foyers de contamination et d'identifier des variables pouvant expliquer leur présence. Le caractère binaire des données sanitaires nécessite la mise en œuvre de méthodes spécifiques comme la régression logistique binaire, qui permet d'attribuer à chaque échantillon une probabilité de présence de l'agent pathogène d'intérêt. La classification postérieure de chaque échantillon nécessite le choix d'un seuil de probabilité pour définir le caractère à risque ou sans risque des observations. La méthode ROC permet de définir ce seuil et de comprendre précisément les conséquences de ce choix. À partir d'une base de données épidémiologique ponctuelle, plusieurs modèles ont été générés et testés. Des indices paysagers ont été dérivés dans l'environnement des points à trois échelles d'analyse. Les probabilités attribuées à chaque échantillon par les modèles ont été représentées spatialement ce qui donne un outil d'interprétation de la répartition attendue des échantillons contaminés. Les variables identifiées à ces trois niveaux d'échelle ont conduit à ébaucher des hypothèses quant aux facteurs paysagers qui entrent en jeu dans les processus épidémiologiques. L'apparition de foyers de risque potentiel constitue un premier résultat.

KEY WORDS

Spatial epidemiology
Binary logistic regression
ROC curves
Predictive modelling

ABSTRACT

Binary logistic regression for predictive modelling in epidemiology

Epidemiological processes are now using spatial statistics and modelling tools. The main objective of most health risks studies consists in identifying potential contamination sources and factors capable of explaining their localization. Health data often prove binary (typically presence/absence) and specific methods such as binary logistic regression have to be used. This method's output consists in a probability for the pathogen of interest. A posterior classification of each sample is then conducted using a probability threshold. The method used to maximize this threshold is called the ROC curve which consists in giving a representation of the behaviour of the model and then to choose the optimal value to discriminate between negative and positive predictions. Using a point data epidemiological base, several models were generated and tested. Landscape indices have been derived in the environment of the points at three scale levels. Probability values allocated to each sample were then spatially represented, giving an insight on the expected geographical dispersion of contaminated samples. Variables identified in the models were then used to establish hypothesis as to which landscape factors might play a role in epidemiological processes. The outlining of potential high risk areas is a result of first importance in the geography of health.

Tolle F., Tourneux F. P., 2008, Probabilité d'apparition d'un phénomène parasitaire et choix de modèle de régression logistique, in : Foltête J.-C. (dir.), *Actes des Huitièmes Rencontres de Théo Quant*, Besançon.
ISSN 1769-6895. Article mis en ligne le 17 mars 2008.

1. Introduction

Les maladies à vecteurs sont au cœur des problématiques de santé publique. Ces vecteurs sont généralement des animaux qui, en fonction de leurs caractéristiques étiologiques, se contaminent, transportent et transmettent les agents pathogènes. De diverses natures (virus, parasites,...), ces agents engendrent le développement de maladies infectieuses au sein de la population humaine. Ces processus épidémiologiques font l'objet de nombreux travaux ayant recours aux outils de la statistique spatiale et de la modélisation. Le but principal de ces recherches est de mettre en évidence des foyers de contamination importants et d'identifier des variables pouvant expliquer la présence de ces foyers. La connaissance de ces paramètres présente un double enjeu : tout d'abord, dans le temps court, un besoin d'information de la population et la mise en place de mesures de protection visant à minimiser les contaminations humaines ; puis, dans le temps plus long, la proposition de mesures de gestion du risque sanitaire et de contrôle des sources propres de l'épidémie. Le caractère binaire des données sanitaires (sain / malade, présence ou absence d'un parasite) ainsi que les difficultés d'échantillonnage et de récolte des données nécessitent la mise en œuvre de méthodes spécifiques. Brooker *et al.* (2002) indiquent que les modèles de prévision d'occurrence et de distribution d'un agent pathogène sont souvent issus de modèles de régression logistique binaire. Dans leurs travaux sur la schistosomiase en Afrique, Brooker *et al.* (2001) mettent en œuvre ces modèles. De même, Wint *et al.* (2002) ont utilisé cette méthode pour identifier les variables ayant le plus fort potentiel de discrimination entre les échantillons positifs et les échantillons négatifs dans le cas de la tuberculose bovine au Royaume-Uni.

Nous avons appliqué une approche similaire à l'analyse des facteurs de risque associés à une zoonose répandue en France. L'échinococcose alvéolaire, maladie grave causée par les œufs du ténia échinocoque *Echinococcus multilocularis*, est une maladie considérée comme émergente bien qu'elle soit connue de longue date (Eckert *et al.*, 2000). La situation épidémiologique en France n'est pour l'heure que partiellement connue. Certaines zones d'endémie du nord-est de la France et du Massif Central ont fait l'objet de multiples travaux (Pétavy *et al.*, 1984 ; Aubert *et al.*, 1987). Les efforts conjoints de l'université de Franche-Comté, de l'ERZ[1] et de l'AFSSA[2] ont permis de constituer une base de données sur la zoonose en France et notamment dans le département du Doubs, zone d'endémie et terrain d'étude retenu pour cette approche (figure 1). A partir de ces données ponctuelles renseignant sur la présence du parasite responsable de l'échinococcose alvéolaire, plusieurs modèles ont été générés et testés. Le lien présumé entre l'infestation des hôtes du parasite et les structures paysagères (Giraudoux *et al.*, 2003) nous a conduit à explorer la capacité d'explication des indices paysagers. L'hypothèse posée ici est que la persistance du cycle parasitaire dans la transmission d'*Echinococcus multilocularis* à l'homme et aux hôtes peut s'expliquer par les caractéristiques et les structures paysagères. Il serait ainsi possible d'identifier des paysages à risque sur le plan épidémiologique.

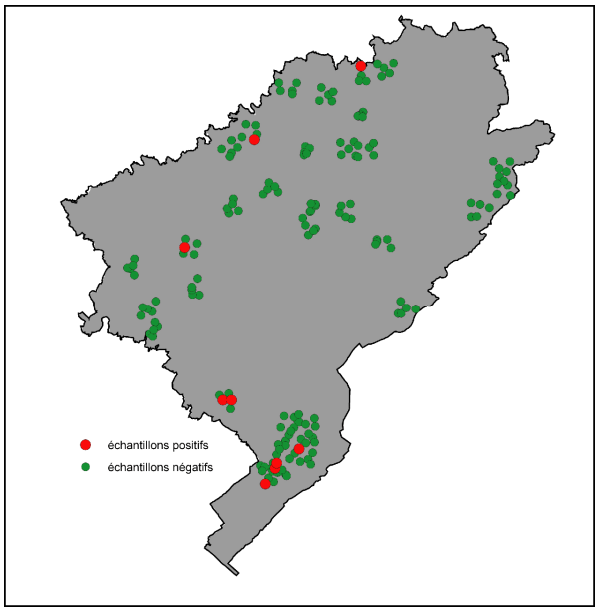

Figure 1. Répartition spatiale des 175 échantillons (dont 9 porteurs du parasite) de fèces de renard dans le département du Doubs.

2. Données et méthodes

Les 175 échantillons disponibles dans le Doubs ont servi de base au calcul d'indices de composition et de configuration paysagère. Ces calculs ont été réalisés à partir d'une image IRS classée à une résolution de 25 mètres. L'image ainsi obtenue permet de caractériser l'occupation du sol en 10 catégories. Le choix de l'échelle à laquelle dériver les indices paysagers a été effectué par analyse radiale et maximisation de l'hétérogénéité paysagère dans l'environnement des points, d'après une méthode issue de Foltête *et al.* (2002). L'objectif ici était de mettre en évidence les niveaux d'échelle auxquels les paysages des échantillons positifs se distinguent le plus nettement des paysages des échantillons négatifs. Ainsi, trois échelles d'analyse ont été retenues correspondant respectivement à un rayon de 700, 2200 et 4775 mètres autour des échantillons. Un grand nombre d'indices paysagers classiques (McGarigal *et al.*, 2002) ont été calculés et introduits comme variables dans les modèles.

Dans leurs travaux sur la tuberculose bovine, Wint *et al.* (2002) ont utilisé la régression logistique binaire. Cette

[1] Entente interdépartementale de lutte contre la Rage et autres Zoonoses.
[2] Agence Française de Sécurité Sanitaire des Aliments.

méthode leur permet d'identifier les variables ayant le plus fort potentiel de discrimination entre les échantillons positifs et les échantillons négatifs. Le modèle de régression logistique qui en est issu attribue à chaque échantillon un score de probabilité de présence de l'agent pathogène d'intérêt. L'application des paramètres de ce modèle à la totalité de l'espace d'étude permet d'obtenir une image globale du risque de présence attendu. Un seuil de probabilité doit être choisi pour définir le caractère à risque ou sans risque des différentes observations. Une méthode existe pour définir ce seuil, c'est la méthode ROC (Receiver-Operator Characteristic) (Metz, 1978), qui consiste à obtenir une visualisation du comportement du modèle, et permet de comprendre précisément les conséquences du choix du seuil. La courbe ROC s'obtient en mettant en regard, pour chaque valeur de seuil, la sensibilité (le pourcentage d'échantillons positifs correctement classés) et la spécificité (le pourcentage d'échantillons négatifs correctement classés) (Greiner *et al.*, 2000). L'aire sous la courbe, définie comme l'AUC (Area Under Curve), donne une estimation de la performance du modèle. Plus cette valeur est proche de 1, plus le modèle discrimine nettement et de manière satisfaisante entre les positifs et les négatifs, entre l'occurrence et l'absence du phénomène étudié (Pearce et Ferrier, 2000).

L'analyse statistique par régression logistique binaire a été conduite aux trois niveaux d'échelle retenus. Les variables sont toutes introduites à la première itération. À chaque étape, la variable la moins significative est retirée du modèle et le modèle est testé à nouveau jusqu'à ce que toutes les variables restantes soient significatives ou, le cas échéant, jusqu'à l'élimination de toutes les variables. La matrice de confusion propre à chaque modèle est générée simultanément en fonction de la valeur de césure précisée par l'utilisateur. Cette matrice renseigne sur la sensibilité et la spécificité des modèles et sur leur performance globale. Il est possible, pour un modèle donné, d'enregistrer les valeurs de la probabilité calculée pour chaque individu. La prédiction de l'individu en positif ou en négatif dépend ensuite du choix d'une valeur de seuil séparant les deux groupes d'observations. Pour chaque modèle, le coefficient de probabilité noté e^B nous renseigne sur le rôle joué par chaque variable dans la prédiction. Des valeurs de e^B supérieures à 1 indiquent qu'une augmentation des valeurs de la variable entraîne une augmentation de la probabilité de prédiction en positif des échantillons. Au contraire, des valeurs de e^B inférieures à 1 indiquent des variables qui augmentent la probabilité de prédiction en positif de l'échantillon quand leurs valeurs baissent. Dans l'exemple du modèle à 2200 mètres du tableau 1, deux variables ont ainsi une influence positive et deux ont une influence négative.

variables	B	σ	Wald	p.	e^B
indice d'auto-adjacence des zones mixtes	-0.2213	0.0823	7.2352	0.0071	0.8015
indice d'auto-adjacence des forêt de feuillus	-0.0777	0.0289	7.2322	0.0072	0.9252
fréquence de contact entre prairies et zones mixtes	0.2294	0.0756	9.2103	0.0024	1.2579
fréquence de contact entre zones mixtes et forêts	0.1954	0.0631	9.5913	0.0020	1.2158
constante	-3.8289	2.3429	2.6708	0.1022	0.0217

Tableau 1. Modèle de régression logistique binaire à un rayon d'analyse de 2200m.

Lorsque ces modèles sont établis, il s'agit d'évaluer leurs performances de manière à effectuer un choix raisonné et à donner une interprétation juste des résultats qui en sont issus. Plusieurs objectifs peuvent être envisagés, nécessitant une approche différente des critères de sensibilité et de spécificité. Dans un premier cas de figure, l'accent peut être mis sur les échantillons négatifs, et il peut être impératif de ne pas obtenir de faux positifs. Dans ce cas, la spécificité sera privilégiée. Dans le cas de figure inverse, on peut chercher à prédire la totalité des échantillons positifs et ce, quel que soit le nombre de faux positifs que cela engendre. C'est la sensibilité qui sera alors favorisée. Dans la problématique qui nous intéresse ici, cette approche peut sembler intéressante. En effet, la variabilité temporelle du parasite, ainsi que les méthodes d'analyse utilisées en laboratoire, ont pu conduire à ne pas repérer le parasite dans des zones où il est, ou a été, présent, et où les caractéristiques paysagères et écologiques sont réunies pour que le cycle parasitaire soit actif.

Dans cette optique, la prévision d'échantillons en faux positifs peut éventuellement mettre en évidence des foyers potentiels dont les caractéristiques paysagères sont proches des vrais positifs bien que le parasite n'y ait pas été formellement trouvé. Une troisième alternative existe, qui vise à établir une valeur de césure permettant d'obtenir le meilleur compromis entre la prédiction du maximum de vrais positifs et le nombre de faux positifs. Cette dernière option équivaut en fait à ne conserver que les vrais positifs les mieux expliqués par le modèle.

L'évolution du rapport entre la sensibilité et la spécificité, en fonction des valeurs de césure possibles, peut se représenter sous forme de courbe appelée courbe ROC. Ce mode de représentation offre un moyen d'estimation des performances d'un modèle de régression logistique binaire, ainsi qu'un outil d'aide au choix de valeurs de césure offrant un compromis raisonnable entre sensibilité et spécificité.

3. Résultats

Pour chacun des trois modèles élaborés à partir des données du Doubs, les valeurs de spécificité et de sensibilité ont été calculées selon des valeurs de césure croissantes avec, pour limite, la valeur n'offrant plus de prédiction en positif. Ces tables nous ont permis de tracer, pour chaque modèle, la courbe ROC et d'en calculer l'« aire sous la courbe » (AUC) (figure 2). Les valeurs d'AUC obtenues dans la figure 2 indiquent une capacité de discrimination raisonnable pour les modèles à 700 et 4775 mètres (AUC comprise entre 0,7 et 0,9) (Brooker *et al.*, 2002) et une bonne capacité de discrimination pour le modèle à 2200 mètres (AUC supérieure à 0,9). À titre d'exemple, la valeur d'AUC de 0,94 du modèle à 2200 mètres indique qu'un échantillon positif tiré au hasard aura 94 % de chances d'être prédit comme tel par le modèle.

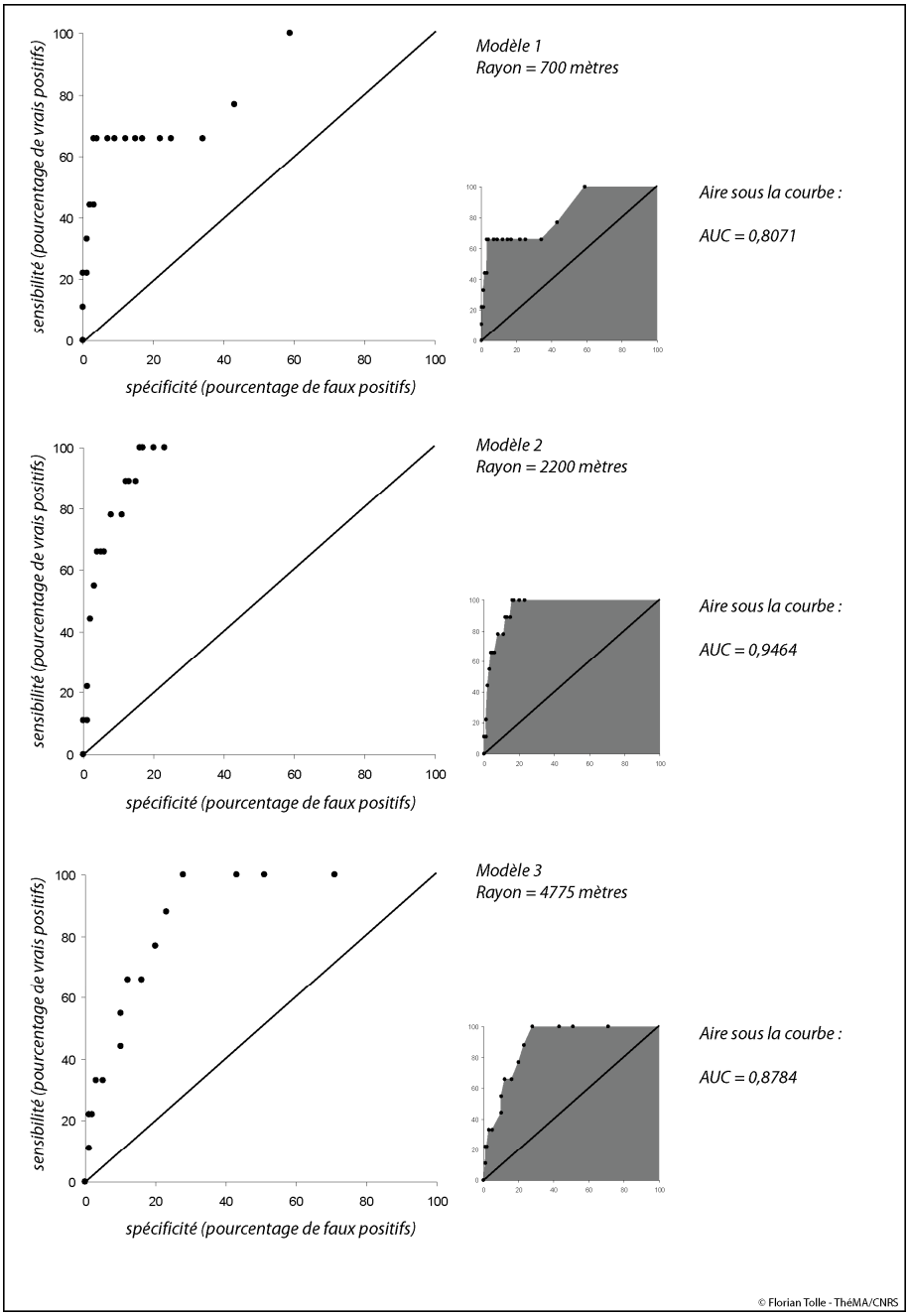

Figure 2. Courbes ROC et AUC des trois modèles du Doubs.

L'analyse des résultats du tableau 2 confirme une meilleure performance du modèle à 2200 mètres. C'est en effet ce modèle qui offre la meilleure qualité de prédiction globale pour une sensibilité de 100 %. En pratique, ce modèle permet de prédire correctement tous les échantillons positifs tout en conservant une spécificité de 84,24%. Le modèle à 4775 mètres offre également, pour une sensibilité de 100%, une performance satisfaisante mais clairement inférieure au modèle précédent. Aux valeurs de césure correspondant à un bon compromis (ici permettant de prédire 6 des 9 échantillons positifs), les deux premiers modèles sont les plus performants.

	AUC	valeur de césure correspondant à une sensibilité de 100 %	spécificité associée	performance globale du modèle	valeur de césure correspondant à un bon compromis entre sensibilité et spécificité	sensibilité associée	spécificité associée	performance globale du modèle
modèle 1 (700m)	0,8071	0,01	41,21 %	44,25 %	0,2	66,66 %	96,97 %	95,40 %
modèle 2 (2200m)	0,9464	0,05	84,24 %	85,05 %	0,3	66,66 %	96,36 %	94,83 %
modèle 3 (4775m)	0,8784	0,04	71,51 %	72,98 %	0,08	66,66%	87,88 %	86,78 %

Tableau 2. AUC, valeurs de césure et performance des 3 modèles.

Les valeurs de probabilité attribuées à chaque échantillon par les modèles peuvent être représentées spatialement. De même, la prédiction des échantillons en fonction des différentes valeurs de césure retenues offre un outil d'interprétation spatiale de la répartition attendue des échantillons contaminés.

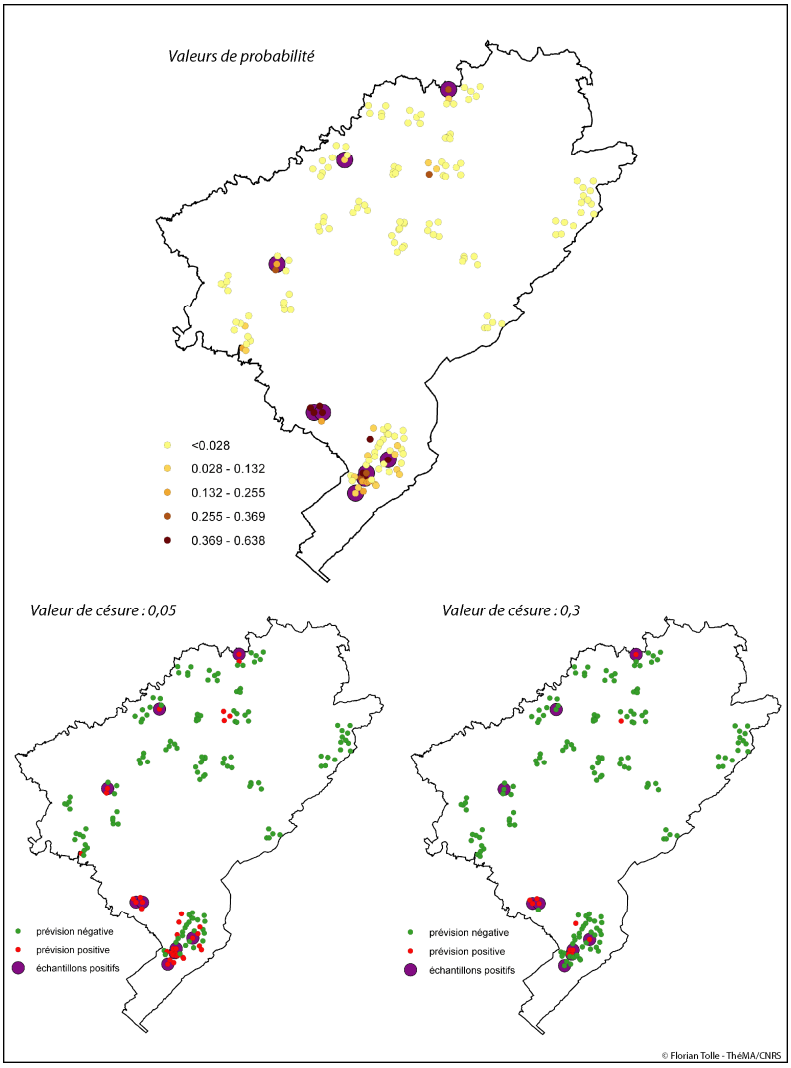

Figure 3. Probabilités associées aux échantillons et classement en fonction des valeurs de césure pour le modèle à 2200 mètres.

La figure 3 illustre les valeurs de probabilité issues du modèle au second niveau d'échelle (2200 mètres) ainsi que la prédiction des échantillons en fonction de la valeur de césure retenue. À cette échelle d'analyse, la valeur de césure de 0,05 permet de prévoir tous les vrais positifs avec un nombre relativement restreint de faux positifs. Ces derniers sont, en outre, spatialement circonscrits à des zones nettement définies. Ainsi, un foyer principal semble apparaître dans le sud du département. Les échantillons positifs du nord du département sont soit seuls prédits comme tels, soit ils s'inscrivent dans un foyer local. Il est à noter également, dans le nord du département, un petit groupe de trois échantillons négatifs prédits comme positifs par le modèle. Ce type de groupe peut être assimilé à une zone où les indices paysagers associés aux échantillons positifs sont présents, mais où le parasite n'a pas été identifié. C'est ici l'idée de potentiel de risque que nous cherchons à mettre en lumière, en insistant sur le fait que le parasite a pu se trouver ponctuellement dans ce type de zone, ou qu'il est susceptible de pouvoir y être rencontré à l'avenir. Une seconde valeur de césure visant à limiter le nombre de faux positifs à cette échelle a été fixée à 0,3. Cette valeur permet encore de prédire six des neuf échantillons positifs. Le foyer du sud du Doubs reste encore visible ainsi que l'échantillon positif du nord du département qui est ici seul prédit comme tel, et un des échantillons du groupe de trois prédits comme positifs avec la valeur de césure inférieure.

4. Discussion

Les variables identifiées à ces trois niveaux d'échelle ont conduit à ébaucher des hypothèses quant aux facteurs paysagers qui entrent en jeu dans les processus épidémiologiques. Tout d'abord, la récurrence d'indices liés aux zones mixtes peut laisser penser que ces classes paysagères et leurs agencements ont une influence sur une ou plusieurs étapes du cycle parasitaire. Si leur importance est confirmée, elles pourraient être considérées comme des indicateurs de risque présumé. La fréquence de contact entre classes d'occupation du sol semble également être un type de descripteur intéressant. Ce constat peut être interprété comme le reflet de l'importance des configurations paysagères, et plus spécifiquement des zones de lisière entre certaines classes d'occupation du sol.

Dans le choix des valeurs de césure, la sensibilité de la prédiction a été favorisée. Cela équivaut à tolérer un plus grand nombre de faux positifs mais, dans le cas de phénomènes variables dans l'espace et dans le temps, cette propriété peut éventuellement aider à mettre en évidence des foyers potentiels dont les caractéristiques paysagères sont proches des vrais positifs bien que le parasite n'y ait pas été formellement trouvé. L'apparition de foyers de risque potentiel, à partir des données du Doubs, est un premier résultat qui nous a encouragé à aller plus loin dans l'exploitation de ces modèles. Les facteurs paysagers identifiés peuvent être calculés en tout point de l'espace. Les trois modèles élaborés précédemment ont fait l'objet d'une généralisation spatiale des facteurs les composant. Nous avons ainsi obtenu une représentation continue des différentes variables entrant dans les modèles. Les coefficients issus de la régression logistique nous renseignent à la fois sur le poids éventuel à attribuer aux variables paysagères, et sur leur influence, négative ou positive, sur la prédiction de présence du parasite. Nous avons donc procédé, pour chaque indice, à un rééchantillonnage régulier en 10 classes couvrant l'étendue des données. Ainsi, des valeurs fortes de deux indices distincts sont toutes deux recodées avec la valeur de 10. Les variables agissant négativement sur la prévision de présence du parasite ont été recodées de manière similaire mais en attribuant, cette fois, la valeur de 10 aux valeurs les plus faibles. Cette normalisation réalisée, nous avons compilé les couches de chaque modèle en leur attribuant le poids donné par le coefficient e^B calculé pour chaque variable. Cette pondération vise naturellement à inclure toute la précision du coefficient e^B dans la modélisation spatiale des indices paysagers. Le résultat de cette étape permet de visualiser plus aisément les espaces où les facteurs de risque présumé s'additionnent et donc, où la présence du parasite est attendue. La figure 4 représente les caractéristiques paysagères mises en évidence par le modèle à 2200 mètres.

Le modèle à 2200 mètres laisse apparaître plusieurs foyers relativement nets. Le sud du département en constitue un premier, relativement étendu. Une zone à l'ouest de Pontarlier apparaît également nettement, renforçant la susceptibilité présumée des paysages de cette zone à accueillir le parasite. Au nord du département, trois autres foyers se dessinent nettement. En revanche, le nord-est du département semble moins concerné par le phénomène parasitaire. La variabilité des contextes paysagers, écologiques et épidémiologiques ne peut nous inciter qu'à la prudence dans l'interprétation des résultats présentés. Néanmoins, ces résultats peuvent être considérés comme encourageants dans le sens où leur interprétation écologique est possible et relativement cohérente avec les paramètres connus comme entrant en compte dans l'établissement du cycle parasitaire, et dans le fait qu'ils laissent apparaître des foyers bien identifiés et aux caractéristiques précises. Notre démarche exploratoire nous permet de faire apparaître des tendances naturelles suggérées par les données et donc, par les processus écologiques qui les conditionnent. Pour autant, nous ne prétendons pas avoir utilisé un jeu de données exempt de problèmes d'échantillonnage ou d'autocorrélation spatiale. Un plus grand nombre de données, et surtout d'échantillons positifs, nous aurait sans doute permis d'obtenir des résultats plus robustes et plus facilement généralisables. Cela nous aurait également permis de construire nos modèles sur un sous-échantillon de la base de données que nous aurions cherché à valider avec les données restantes. Dans le cas présent, chaque point a beaucoup d'importance et influe fortement sur

les résultats de la régression. La confirmation des hypothèses testées ici nécessiterait la collecte de nouveaux échantillons dans les zones attendues comme à risques. Ces données pourraient à nouveau être introduites dans les modèles pour en préciser la portée.

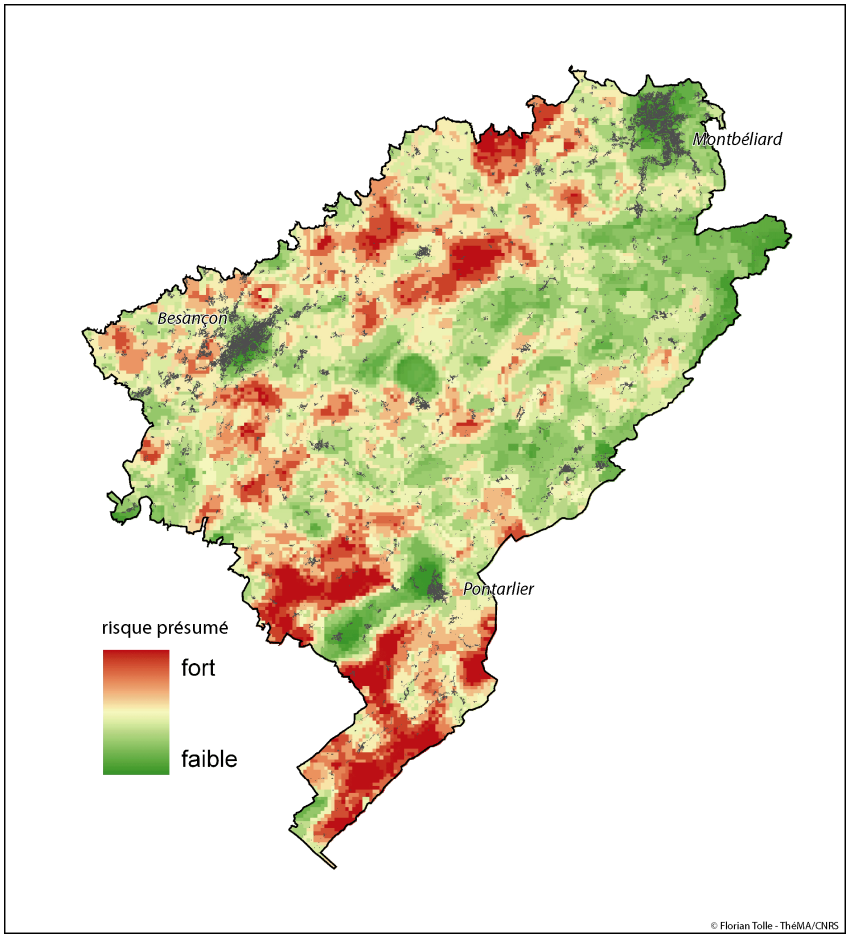

Figure 4. Représentation spatiale des indices de risque de présence du parasite pour le modèle à 2200 mètres.

Remerciements

Patrick Giraudoux du laboratoire de Biologie Environnementale ; Benoit Combes, Déborah Gottsheck, Stéphanie Favier et Frantz Catarelli de l'ERZ (Entente interdépartementale de lutte contre la Rage et autres Zoonoses) ; Denis Augot et Franck Boué de l'AFSSA (Agence Française de Sécurité Sanitaire des Aliments) ; Projet soutenu par le contrat Européen « Eurechinorisk » QLK-2-CT-2001-01995 « Risk assessment and prevention of Alveolar Echinococcosis ».

5. Références bibliographiques

Aubert M., Jacquier P., Artois M., Barrat J. M., Basile A. M., 1987, Le portage d'Echinococcus multilocularis par le renard (*Vulpes vulpes*) en Lorraine. Conséquences sur la contamination humaine, *Recueil médical et vétérinaire*, 163(10), 839-843.

Brooker S., Hay S. I., Issae W., Hall A., Kihamia C. M., Lwambo N., Wint W., Rogers D. J., Bundy D., 2001, Predicting the distribution of urinary schistosomiasis in Tanzania using satellite sensor data, *Tropical medicine and international health*, 6(12), 998-1007.

Brooker S., Hay S. I., Bundy D., 2002, Tools from ecology: useful for evaluating infection risk models? *Trends in parasitology*, 18(2), 70-74.

Eckert J., Conraths F. J., Tackmann K., 2000, Echinococcosis: an emerging or re-emerging zoonosis?, *International Journal for Parasitology*, 30, 2000, 1283-1294.

Foltête J.-C., Monteil C., Deconchat M., 2002, Habitat animal et image numérique : méthode de reconnaissance exploratoire appliquée à des occurrences d'espèces, *Photo-interprétation*, 38, 40-51.

Giraudoux P., Craig P. S., Delattre P., Bao G., Bartholomot B., Harraga S., Quéré J.-P., Raoul F., Wang Y., Shi D., Vuitton D. A., 2003, Interactions between landscape changes and host communities can regulate Echinococcus multilocularis transmission, *Parasitology*, 127, 121-131.

Greiner M., Pfeiffer D., Smith R. D., 2000, Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests, *Prev. Vet. Med.*, 45, 23-41.

McGarigal K., Cushman S. A., Neel M. C., Ene E., 2002, FRAGSTATS: Spatial pattern analysis program for categorical maps. Computer software program produced by the authors at the University of Massachusetts, Amherst. Available at the following web site: www.umass.edu/landeco/research/fragstats/fragstats.html

Metz C. E., 1978, Basic principles of ROC analysis, *Semin. Nucl. Med.*, 8, 283-298.

Pearce J., Ferrier S., 2000, Evaluating the predictive performance of habitat models developed using logistic regression, *Ecological Modelling*, 133, 225-245.

Pétavy A. F., Deblock S., Gilot B., 1984, Mise en évidence de la larve du Ténia multiloculaire chez deux campagnols (Microtus arvalis et Clethrionomys glareolus) dans le foyer d'échinococcose alvéolaire du Massif Central (France), *Académie des Sciences de Paris*, 299, 735-737.

Wint G. R., Robinson T. P., Bourn D. M., Durr P. A., Hay S. I., Randolph S. E., Rogers D. J., 2002, Mapping bovine tuberculosis in Great Britain using environmental data, *Trends in Microbiology*, 10(10), 441-444.