HAL
open science

Éléments d'écologie de la transmission d'Echinococcus multilocularis en Chine (Sichuan) Modélisation des distributions spatiales des communautés et populations des hôtes : des données de terrain aux prédictions

Amélie Vaniscotte

To cite this version:

Amélie Vaniscotte. Éléments d'écologie de la transmission d'Echinococcus multilocularis en Chine (Sichuan) Modélisation des distributions spatiales des communautés et populations des hôtes : des données de terrain aux prédictions. Ecologie, Environnement. Université de Franche-Comté, 2009. Français. NNT : . tel-00483842

HAL Id: tel-00483842

https://theses.hal.science/tel-00483842v1

Submitted on 17 May 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Éléments d'écologie de la transmission d'*Echinococcus multilocularis* en Chine (Sichuan)

Modélisation des distributions spatiales des communautés et populations des hôtes : des données de terrain aux prédictions

par Amélie Vaniscotte
Laboratoire Chrono-environnement,UMR UFC/CNRS 6249 aff. INRA

Présentée à
L'UFR Des sciences et techniques
De L'université de Franche-Comté

pour obtenir le grade de
docteur de l'université de Franche-Comté
Spécialité : Sciences de la vie
École doctorale "Homme, Environnement, Santé"

Soutenue le 30 novembre 2009 devant la commission d'examen :

Rapporteurs	TATONI Thierry	Professeur, Université Cezanne, Marseille
	BUREL Françoise	Directrice de recherche, Université de Rennes 1, Rennes
Examinateurs	CALENGE Clément	Docteur, ONCFS, Saint Benoist-Auffargis
	HEGGLIN Daniel	Docteur, Université de Zurich, Zurich
Directeur de thèse	GIRAUDOUX Patrick	Professeur, Université de Franche-Comté, Besançon
Co-directeur de thèse	RAOUL Francis	Docteur, Université de Franche-Comté, Besançon
Invité	CHRETIEN Stéphane	Docteur, Université de Franche-Comté, Besançon

Remerciements

Je vous remercie, Monsieur et Madame les professeurs Thierry Tatoni et Françoise Burel, pour avoir accepté d'être rapporteurs de ma thèse, pour vos lectures et avis critiques sur mon travail. Je remercie également les docteurs Daniel Hegglin, Clément Calenge et Stéphane Chrétien d'avoir accepté d'être examinateurs et membres du jury de cette thèse.

Mes remerciements vont naturellement vers ceux qui m'ont permis de me lancer dans cette aventure de recherche et qui m'y ont accompagné, guidée et soutenue lorsque cela était nécessaire. Á Patrick Giraudoux pour son impulsion, nos échanges d'idées, son amitié, et finalement la confiance avec laquelle il m'a laissé travailler. Á Francis Raoul pour les longues discussions critiques de notre travail, pour ses efforts de compréhension, pour son calme et son organisation exemplaire. Á David Pleydell, pour ses folles idées et son amitié sincère. Á Stéphane Chrétien pour son optimisme et ses encouragements dans mes recherches, les réponses à mes questions et nos conversations teintées entre autres de statistiques et d'écologie. Je souhaite que nos recherches convergent à nouveau.

Cette thèse n'aurait pas pu être réalisée sans le soutien financier de l'Institut National de la Santé des Etats-Unis (NIH) que je tiens à remerçier. Je remercie également tous les collaborateurs de notre équipe internationale, et particulièrement ceux qui étaient là sur le terrain et après : nos amis de Chengdu pour l'organisation des campagnes de terrain, Nima de Shiqu pour ses traductions, les thésards du NIH, Jasmin Moss et Christopher Marston et leurs encadrants Phil Craig et Mark Danson, Thomas Romig pour nos rencontres et collaborations.

Je remercie également Marie Lazarine Poulle et Marie Hélène Guislain pour leur aide et leurs astuces précieuses en compagnie des chiens. Enfin, un grand merci à Emile Carry et Pierre Yves Bourgeois pour leur aide concrète dans les tréfonds de LaTeX.

Ce travail n'aurait pas pu être réalisé sans l'aide des "habitants" de la place Leclerc, je pense particulièrement à Françoise et Brigitte, à Dom, pour leurs aides concrètes, à Micka pour son amitié, aux copains les thésards pour leur complicité, à Mickael et Renaud et aux autres collègues maîtres et professeurs pour leurs conseils avertis et leur solide compagnie.

Mon travail a été généreusement encouragé par mes amis. Clairette, Stéphane, Rémo, Guislaine et toute la petite famille des francs comtois d'adoption, je vous dis merci et espère vous le dire encore ! Un grand merci à toute l'équipe du Buc alpin, sans qui je n'aurais sans doute pas réalisé que "la thèse est une course de fond, plutôt qu'un sprint", ni relativisé du haut de nos escapades et rigolades.

Merci aussi à ceux qui sont loin et qui ont été là quand je ne l'étais pas.

Un merci singulier à Didier, pour le courage que sa présence m'a insufflé jusqu'à l'écriture de ces lignes et pour, malgré tout et entre tout, nos petits et grands sommets.

Enfin, à ma mère, mon père et ma soeur pour m'avoir soutenue depuis mes premiers rêves d'être "zoologiste" dans la garrigue tavelloise, avoir encouragé et laissé libre cours à ma passion et à la réalisation de ce devenir, mais aussi et surtout pour avoir compris mes absences.

LISTE DES PUBLICATIONS ÉCRITES ET ORALES

Publications écrites

Vaniscotte A, Pleydell D, Raoul F, Quéré JP, Bernard N, Coeurdassier M, Delattre P, Takahashi K,Tiaoying L, Qiu, J, Wang Q, Weidmann JC, Giraudoux P (2009) Modelling and spatial discrimination of small mammal assemblages : an example from western Sichuan (China). Ecological Modelling 220 : 1218–1231.

Raoul, F., Quere, J., Pleydell, D., Vaniscotte, A., Giraudoux, P., 2008. Small mammals assemblage response to deforestation and afforestation in central china : a multinomial based modelling approach. Mammalia 72.

Pleydell D, Raoul F, Vaniscotte A, Craig P. Towards understanding the impacts of environmental variation on Echinococcus multilocularis transmission, Morand S, Krasnov BR, Poulin R, editors. Micromammals and Macroparasites. NewYork : Springer, 2006 : 545-64.

Giraudoux P, Pleydell D, Raoul F, Vaniscotte A, Ito A and Craig P. 2007. Echinococcus Multilocularis : Why are multidisciplinary and multiscale approaches essential in infectious disease ecology ? Tropical Medicine and Health, 35 (4) : 293-299.

Communications orales

Amélie Vaniscotte, Francis Raoul and Patrick Giraudoux. Habitat modelling of small mammal assemblages in Western Sichuan (China) :from field trapping data to a regional predictive mapping. IALE. 14-17 July 2009, Salzsburg.

Amélie Vaniscotte, Francis Raoul, David R. J. Pleydell and Patrick Giraudoux. Habitats des assemblages de micromammifères dans l'ouest du Sichuan (Chine) : des modèles locaux à l'échelle paysagère à une cartographie régionale des prédictions. Neuvième de rencontres de Théoquant. Nouvelles approches en géographie théorique et quantitative. 4-6 Mars 2009, Besançon.

Amélie Vaniscotte, Francis Raoul, David R. J. Pleydell and Patrick Giraudoux. Predictive mapping of host assemblages : an example with small mammals of western China (Sichuan). Xth European Multicolloquium of Parasitology. 24-28 Août 2008, Paris.

Amélie Vaniscotte, Francis Raoul, David R. J. Pleydell, Jean P. Quéré et Patrick Giraudoux. Ecologie de la transmission de l'échinocoque alvéolaire en Chine : modélisation des distributions spatiales des communautés d'hôtes intermédiaires dans la région du Plateau Tibétain (Sichuan). XIIIème Forum des Jeunes Chercheurs. 14-15 juin 2007, Dijon.

Amélie Vaniscotte, Francis Raoul, David R. J. Pleydell and Patrick Giraudoux. Small mammal assemblages response to deforestation and afforestation gradients in Central China : a multinomial-based modelling approach. Huitième rencontres de Théoquant. Nouvelles approches en géographie théorique et quantitative. 10-12 Janvier 2007, Besançon.

Table des matières

Avant propos **1**

L'éco-épidémiologie : des nouveaux défis pour des enjeux de société 3

L'échinococcose alvéolaire (EA) : une zoonose émergente 4

Les facteurs de risque éco-épidémiologiques pour EA 6

Objectifs généraux 7

Plan du mémoire 8

Introduction **9**

1 Écologie et modélisation de la transmission d'*Echinococcus multilocularis* en Chine **11**

1.1 Écologie et modélisation de la transmission des macro-parasites 11

1.1.1 Écologie de la transmission 11

1.1.2 Importance des distributions des hôtes 12

1.1.3 Modéliser les distributions spatiales des hôtes (modélisation statique) 13

1.1.4 La rencontre dans les modèles déterministes (modélisation dynamique) 14

1.2 Écologie de la transmission d'*Echinococcus multilocularis* 15

1.2.1 Paramètres écologiques de la rencontre 15

1.2.2 Variations spatiales et multi-scalaires des paramètres écologiques de la transmission 16

1.2.2.1 Échelle locale (n x 10 m) 16

1.2.2.2 Échelle régionale (n x 10 km) 17

1.2.2.3 Échelle continentale (n x 100 km) 17

1.2.3 Bilan des connaissances écologiques et épidémiologiques en Chine . . 18

1.2.3.1 Situation épidémiologique en Chine 18

1.2.3.2 Paysages et distributions spatiales des populations d'hôtes intermédiaires 20

1.2.3.3 Le rôle du chien dans la transmission 22

1.2.3.4 Hétérogénéité des systèmes de transmission 23

1.3 Modélisation de la transmission d'*Em* 24

1.3.1 Modélisation des distributions spatiales des populations d'hôtes . . . 24
1.3.2 Modélisation dynamique de la rencontre 25
1.3.3 Modélisation spatio-déterministe 26
1.4 Nouvelles questions et objectifs 26
1.4.1 Écueils méthodologiques et nouvelles questions 27
1.4.1.1 Définir des assemblages d'hôtes intermédiaires 27
1.4.1.2 Modéliser les distributions spatiales des assemblages d'hôtes intermédiaires 28
1.4.1.3 L'hétérogénéité spatiale multi-scalaire 28
1.4.1.4 Le rôle du chien domestique dans la transmission 29
1.4.2 Objectifs détaillés de la thèse 30

2 Méthodes pour la modélisation et l'écologie comportementale des hôtes 31

2.1 Modélisation des distributions spatiales des hôtes intermédiaires 31
2.1.1 Les données 31
2.1.1.1 Sur les espèces 31
2.1.1.2 Sur l'environnement 32
2.1.2 Modéliser les distributions spatiales des assemblages 32
2.1.2.1 Niche écologique et habitat : éléments de définition 32
2.1.2.2 Modéliser les distributions spatiales des espèces 34
2.1.2.3 Un domaine de recherche interdisciplinaire en pleine effervescence 35
2.1.3 Une large diversité d'outils de modélisation 36
2.1.3.1 De la distribution des espèces à celle des assemblages : les stratégies de modélisation 37
2.1.3.2 Les modèles paramétriques *versus* non-paramétriques . . . 38
2.1.4 Construire un modèle prédictif 41
2.1.4.1 Les principales étapes 41
2.1.4.2 L'erreur de prédiction : une définition inter-disciplinaire . . 42
2.2 Écologie comportementale du chien domestique 45
2.2.1 Outils moléculaires et comportement de défécation 45
2.2.2 Analyse du mouvement et comportement de prédation 45
2.2.2.1 Utilisation de l'espace par les chiens domestiques 45
2.2.2.2 Distribution des hôtes définitifs et transmission 46
2.3 Plan méthodologique du mémoire 47
2.3.1 Modélisation des distributions spatiales de micro-mammifères 47
2.3.1.1 Axe 1 : Construction d'un modèle explicatif 48
2.3.1.2 Axe 2 : Vers un modèle prédictif des assemblages de micro-mammifères 48

2.3.2 Axe 3 : Le rôle du chien domestique (*Canis lupus familiaris*) dans la transmission d'*Em* 48

Travaux de recherche **51**

3 Axe 1 : Construction d'un modèle explicatif - Modéliser la distribution spatiale des assemblages de micro-mammifères **53**

3.1 Introduction 53

3.1.1 Contexte environnemental et épidémiologique 53

3.1.2 Rappel des objectifs et des questions 54

3.2 Axe 1a : Modélisation et discrimination spatiale des assemblages 55

3.2.1 Précisions méthodologiques 55

3.2.1.1 Étape 1 - Classifier : définition des assemblages 55

3.2.1.2 Étape 2 - Modéliser les distributions des assemblages : un contexte de classification 56

3.2.1.3 Estimer l'erreur de prédiction : la validation croisée et le bootstrap 632+ 58

3.2.2 Article - Modélisation et discrimination spatiales 60

3.3 Axe 1b : Classifier et prédire simultanément 75

3.3.1 Objectifs et méthodologie 75

3.3.2 Résultats et discussion 75

3.4 Principaux résultats du chapitre 77

4 Axe 2 : Vers un modèle prédictif : prédictions régionales des assemblages **79**

4.1 Rappel des objectifs 79

4.2 Article - Des données de piégeage aux prédictions régionales 80

4.3 Article - Effet de la taille de l'échantillon sur la précision des prédictions . . 100

4.4 Principaux résultats du chapitre 108

5 Axe 3 : Le rôle du chien domestique dans la transmission **109**

5.1 Rappel des objectifs 109

5.2 Article - Les fèces de canidés et le comportement des chiens 110

5.3 Principaux résultats 131

Discussion générale **133**

6 Modélisation des distributions spatiales des assemblages d'hôtes intermédiaires : apports, limites et amélioration du protocole **135**

6.1 De la définition des assemblages 135

6.1.1 Apports méthodologiques 135
6.1.2 Les assemblages comme groupes de transmission d'*Em* 136
6.2 De la modélisation des assemblages 137
6.2.1 Rôle des variables environnementales indirectes et distributions des assemblages 137
6.2.2 Améliorations du protocole 138
6.2.2.1 Description des micro-habitats 138
6.2.2.2 La dimension paysagère 139
6.2.3 La stratégie de modélisation : incorporer l'avis d'experts 141
6.3 De l'évaluation des prédictions 142
6.3.1 La stratégie d'échantillonnage 142
6.3.2 Les erreurs spatiales 143
6.3.2.1 La structure spatiale des données 143
6.3.2.2 Le contexte environnemental et biogéographique 143
6.3.3 La dynamique des populations 144
6.4 Perspectives pour la cartographie des assemblages d'hôtes intermédiaires . . 145

7 Le rôle du chien dans la transmission 149
7.1 Comment le chien contamine l'environnement ? 150
7.1.1 Contamination de l'environnement des hommes 150
7.1.1.1 Principaux résultats 150
7.1.1.2 Applications au contrôle de la transmission 151
7.1.2 Contamination des habitats des hôtes intermédiaires 151
7.2 Comment le chien se contamine ? 151
7.2.1 Éléments d'écologie comportementale du chien domestique 151
7.2.2 La modélisation spatiale de la rencontre 153
7.2.3 Applications au contrôle des prévalences d'*Em* 154
7.3 Degré de domestication et place du chien errant dans le cycle 154

Conclusion générale 157

Bibliographie 159

Avant propos

L'éco-épidémiologie : des nouveaux défis pour des enjeux de société

Depuis le début des années soixante, les facteurs de risque des maladies sont reconnus comme multiples et comme pouvant intervenir à différents niveaux d'organisation (March and Susser, 2006). La formalisation par Pavlowski (1884-1965) de l'"épidémiologie paysagère", a rendu possible l'intégration des facteurs environnementaux en épidémiologie par la spatialisation des maladies, de leurs vecteurs et de leurs hôtes ainsi que par la prise en compte du paysage. Ainsi, le concept de foyer naturel infectieux (Pavlowski, 1964) intègre trois idées fondamentales : i) les maladies tendent à être limitées spatialement, ii) ces variations spatiales trouvent leur origine dans celles des conditions physiques et/ou biologiques nécessaires à l'existence du pathogène et de ses vecteurs et réservoirs, et iii) délimitées sur des cartes, ces conditions abiotiques et biotiques devraient être utilisables pour prédire à la fois le risque contemporain et ses changements futurs. Le **concept d'habitat** pour une maladie est alors énoncé. Parallèlement, le géographe Max Sorre pose les bases de la géographie médicale en introduisant le concept de complexe pathogène (Sorre, 1933). Il souligne l'existence d'aires de possibilités maximales et de zones marginales de distribution des maladies telles qu'exprimées pour les espèces par le **concept de niche écologique** qui se développe alors en écologie. Les nombreux outils issus de l'écologie (biogéographie, statistiques, analyses multivariées) ont rendu effectif le champ disciplinaire de l'éco-épidémiologie. Ils ont été rapidement appliqués par le professeur J.A Rioux, aux maladies vectorielles dans le cas par exemple des leishmanioses en Guadeloupe (Rioux et al., 1977).

Aujourd'hui, les maladies infectieuses, causées par la transmission d'un agent pathogène, parasite, virus ou bactérie, comptent pour 1/4 des maladies sur terre (Patz et al., 2004). La diversité des espèces dans les écosystèmes exerce un effet régulateur sur le développement de maladies si chaque espèce occupe une niche écologique définie empêchant l'invasion d'autres espèces impliquées dans la transmission de pathogènes
(Daszak et al., 2001). Or, l'avènement de l'ère industrielle puis post-industrielle, l'augmentation rapide des populations humaines ainsi que leurs mouvements ont engendré de nombreuses perturbations des écosystèmes et par conséquent des cycles de transmission des pathogènes. L'altération des habitats des vecteurs et réservoirs, l'invasion de niches ou le transfert d'hôtes, la modification de la biodiversité (perte de prédateurs, changements des densités des populations d'hôtes) sont autant de mécanismes causés par de tels changements (Hassan et al., 2005). Les effets du paysage s'exercent alors sur les distributions et les dynamiques de l'ensemble des êtres vivants constituant les complexes pathogènes (Sorre, 1933). De surcroît, les changements climatiques, et par suite les conditions environnementales abiotiques (hydrométrie, température,...) peuvent engendrer des modifications des dynamiques de population, des cycles de développement des pathogènes ou des vecteurs et par conséquent de l'aire d'extension des maladies (de La Rocque et al, 2007). Ainsi, les changements paysagers anthropiques constituent le principal facteur d'émergence des maladies infectieuses, c.a.d qu'ils contribuent à l' augmentation récente de leurs incidences ou de leur aires de distributions géographiques, à l'extension de leurs spectres d'hôtes, à la découverte de nouveaux pathogènes ou à leurs implications récentes des maladies (Daszak et al., 2001).

Les pathogènes zoonotiques, transmis des populations de vertébrés mammifères aux populations humaines, représentent 73% des maladies émergentes (Daszak et al., 2001). Ils sont à l'origine de 49% des maladies humaines et sont nécessairement liés à la faune sauvage et aux animaux domestiques (Daszak et al., 2001; Hassan et al., 2005; Taylor et al., 2001). En parallèle, les maladies de la faune sauvage introduites à travers les hôtes domestiques ou par "pollution" peuvent menacer la biodiversité et la conservation des espèces (Daszak et al., 2000). Ces pathogènes ont en général un cycle de vie indirect nécessitant alors une phase de dissémination de leur forme infectieuse et l'intervention de plusieurs hôtes. L'existence de plusieurs hôtes semble être à l'origine de la multiplication des risques pour l'émergence de pathogènes sauvages ou de zoonoses (Daszak et al., 2001). La relation pathogène-hôte ainsi que ses variations environnementales y sont en effet multipliées. Une telle complexité rend difficile les études de la transmission et la prédiction du risque. Les études concernant les effets environnementaux sont en effet moins développées pour les parasites nécessitant un contact entre ses différentes espèces hôtes, telles que ceux à l'origine de zoonoses dites "négligées" (la Rage, l'Anthrax, la Cysticercose,...) (World Health Organization , 2006), que ceux à transmission vectorielle (Maladie de Lyme, Malaria,...) (Hassan et al., 2005).

L'éco-épidémiologie s'est développée pour comprendre et prédire le lien entre les changements environnementaux, les distributions et l'émergence des maladies. Comprendre leurs distributions implique de connaître le système complexe de transmission, considérant les interactions entre chaque niveau d'organisation et chaque population impliqués (Giraudoux et al., 2008). Le pathogène étant difficilement observable, cela consiste à étudier les distributions, niches et habitats, des populations humaines, vecteurs et hôtes, leurs interactions et de ce fait à intégrer de multiples disciplines de la biologie et en particulier l'écologie. De nouveaux outils issus de l'écologie numérique, de la géographie quantitative et des statistiques modernes ont été récemment développés pour expliquer et prédire la transmission et son risque à partir de données éco-épidémiologiques récoltées à de multiples échelles et niveaux d'organisation. Face à l'urgence des situations épidémiologiques et le manque de connaissances sur les systèmes de transmission complexes, le défi consiste à trouver parmi les outils développés ceux qui vont pouvoir extraire le maximum d'informations de telles données.

L'échinococcose alvéolaire (EA) : une zoonose émergente

Ce travail de thèse s'inscrit dans l'étude de la transmission du cestode (helminthe) *Echinococcus multilocularis* (*Em*) responsable d'une zoonose, l'échinococcose alvéolaire (EA).

Le cycle naturel de transmission d'*Em* est constitué universellement de micro-mammifères comme hôtes intermédiaires et de carnivores (principalement des canidés tel que le renard ou le chien) comme hôtes définitifs. Dans ce cycle, l'homme intervient en tant qu'impasse parasitaire puisqu'il ne participe pas à la transmission. La transmission s'effectue par ingestion des oeufs par contact direct avec l'oeuf disséminé par l'hôte définitif dans l'environnement, par ingestion de nourriture contaminée, ou par contact direct avec les hôtes définitifs (Figure 1). La maladie est engendrée par la fixation sur le foie du stade métacestode du parasite. Cette forme alvéolaire composée de nombreuse vésicules peut se développer et s'infiltrer de manière prolifique à la manière d'une tumeur dans d'autres organes jusqu'à entraîner la mort si l'organisme

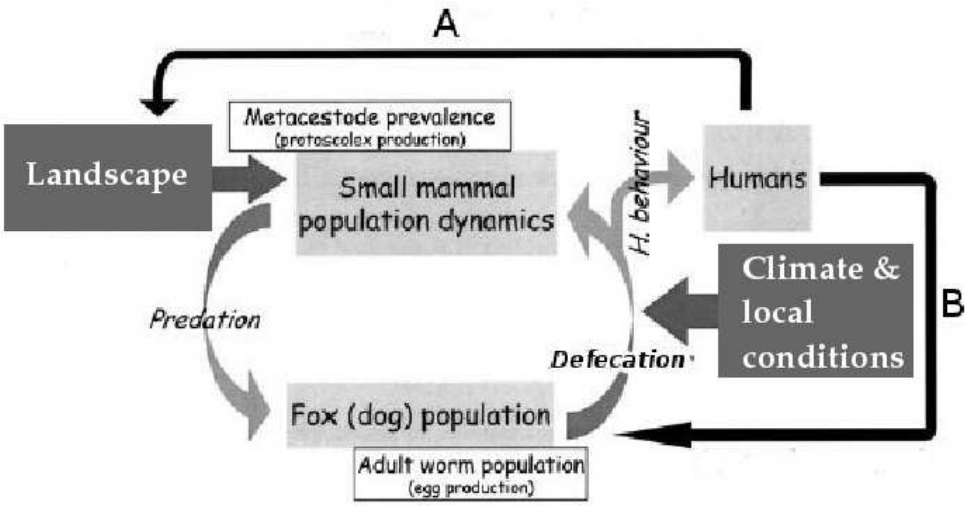

FIG. 1 – Le cycle de transmission d'*Em* et les variables comportementales des hommes (H) et des animaux (A) qui affectent la transmission. Extrait de Giraudoux et al, 2003.

n'est pas traité. Du fait de la longue période de développement a-symptomatique du parasite chez l'homme (de 5 à 20 ans) le diagnostic, effectué par échographie complétée d'un test immunologique sérologique, peut être très tardif et les traitements restent coûteux (intervention chirurgicale ou chimiothérapie) (Vuitton et al., 2003).

La maladie humaine est généralement considérée comme rare, l'incidence annuelle variant entre 0.02 et 1.4 pour 100 000 habitants dans les pays ou régions endémiques en Europe et au Japon (Eckert and Deplazes, 2004). Cependant, des prévalences très élevées ont été enregistrées dans certains endroits du globe atteignant un maximum de 14 % dans certains villages de la Chine centrale (Tiaoying et al., 2005). Aussi, du fait de son issue fatale et de la difficulté des traitements, elle est considérée comme une des zoonoses les plus dangereuses (Vuitton et al., 2003).

Indigène des zones de toundra Arctique d'Amérique du Nord depuis la période post-glaciations, le parasite s'est dispersé vers le Sud et dans les régions tempérées et arctiques de l'hémisphère Nord et à des altitudes élevées. Actuellement on trouve des foyers endémiques d'AE en Amérique du Nord, Asie centrale, Chine, Japon, Europe centrale et du Nord, ainsi que dans quelques états de l'ancienne URSS. On ne la rencontre pas en dessous du 28 ° de latitude Nord c'est à dire au sud de la Tunisie, la Turquie, l'Iran et le Nord de l'Inde où quelques cas ont été enregistrés.

Les données d'incidences récentes permettent de constater un risque persistant et considérable dans des aires d'endémie, en Europe (Eckert et al., 2000) comme en Chine (Craig and The Echinococcosis Working Group in China, 2006). Elles montrent également l'extension de la maladie au Japon (Eckert et al., 2000). Cependant, l'analyse des incidences ou des prévalences humaines ne suffit pas pour mettre en évidence ou prédire une émergence du fait des disparités spatiales et temporelles des suivis épidémiologiques, particulièrement dans certains endroits reculés où l'accès aux centres de soins est difficile. En outre, les données récoltées ne correspondent pas nécessairement à la situation épidémiologique actuelle puisque la période d'incubation de la maladie chez l'homme est longue.

Les facteurs de risque éco-épidémiologiques de l'EA

L'étude de facteurs éco-épidémiologiques à l'origine des prévalences humaines rend possible la détection de l'émergence de la maladie (Eckert et al., 2000). Bien que négligées dans le rapport du "Millenium Ecosystems Assessment" de l'OMS (Hassan et al., 2005), les actions de l'homme sur les écosystèmes et sur les populations d'hôtes qui influencent la transmission de la zoonose sont multiples et constituent des facteurs d'émergence de la maladie en Europe (Eckert et al., 2000). Elles peuvent modifier indirectement, par leurs actions sur le paysage (Figure 1, A) ou directement, par le contrôle des populations (Figure 1, B), les habitats, les densités et les prévalences des populations d'hôtes intermédiaires et définitifs.

Action sur les habitats et populations d'hôtes intermédiaires

Contrairement aux paysages dits naturels des steppes de l'Amérique du Nord où des cycles de transmission intenses ont été mis en évidence, les aires endémiques Eurasiennes seraient naturellement occupées par des taïgas ou des forêts caducifoliées peu propices à la transmission d'*Em* (Romig et al., 2006). Dans certaines de ces zones occupées et exploitées par l'homme, la transmission est favorisée par des changements paysagers tels que les pratiques agricoles (prairies de fauche en France, surpâturage en Chine) ou forestières (déforestation en Chine) (Giraudoux et al., 2002). Ces changements influencent alors indirectement les prévalences en agissant sur les paramètres de la transmission (Romig et al., 2006) et, en l'occurrence, sur les distributions spatiales des populations d'hôtes intermédiaires (micro-mammifères), augmentant leurs populations, les prévalences des hôtes définitifs et par conséquent les prévalences humaines (Giraudoux et al., 2002) (Figure 1).

Action sur les populations des hôtes définitifs et cycles synanthropiques

L'homme intervient également dans le cycle en modifiant les densités et les habitats des populations d'hôtes définitifs. En Europe, dans les années 90, le contrôle de la rage, le changement de pratiques de chasse ou de perception de la faune sauvage par les populations et l'augmentation de la disponibilité de ressources anthropiques, sont autant de causes à l'origine de l'augmentation des effectifs des populations de renards (Romig et al., 2006). Ces augmentations étaient alors corrélées à l'augmentation simultanée des prévalences d'*Em* vulpines (Romig, 2002). En parallèle, la présence de renards est reconnue dans de nombreuses villes (Deplazes et al., 2004) et l'augmentation de leurs densités favorisée par la disponibilité de ressources anthropiques (Comtesse, 2004). Aussi les prévalences élevées observées chez le renard urbain, bien qu'inférieures à celles des renards ruraux, et la disponibilité de populations d'hôtes intermédiaires potentiellement infectées à proximité des villes (zones périphériques) prouvent l'existence de cycles de transmission se réalisant à proximité des habitations humaines (Stieger et al., 2002). L'existence de tels cycles dits synanthropiques constitue un risque de contamination de l'environnement des hommes et par conséquent de transmission (Romig et al., 2006; Deplazes et al., 2004; Robardet et al., 2008). De surcroît l'existence d'hôtes intermédiaires contaminés constitue un facteur de risque de transmission aux hôtes domestiques.

Le chien et, dans une moindre mesure, le chat domestique peuvent intervenir dans les cycles de transmission (Eckert and Deplazes, 2004). La présence d'hôtes définitifs domestiques dans l'environnement des hommes tend à augmenter les prévalences humaines (Gottstein et al., 2001) et participe à l'émergence de la maladie (Eckert et al., 2000). L'implication du chien

domestique dans le cycle d'*Em* a été mise en évidence dans toutes les aires d'endémie mondiale. En Europe, dans le canton de Zurich, les contaminations portées par les chiens et les chats, bien que plus nombreux que les renards (estimée à environ 48000, 145000 et 4700 respectivement), ne représentent que 9.3% et 35.5% de la "capacité de portage" du renard. Au Japon, la contribution du chien reste également faible comparée à celle du renard (Macpherson and Craig, 2005). En revanche, en Alaska (île St Lawrence) et en Chine de fortes prévalences ont été enregistrées chez le chien domestique (jusqu'à 12 %) (Schantz et al., 1995; Craig et al., 2000; Budke et al., 2005a) et de nombreux facteurs de risque associés aux chiens ont été identifiés (Stehr-Green et al., 1988; Craig and The Echinococcosis Working Group in China, 2006). Le chien domestique est alors considéré comme l'hôte définitif d'un cycle péri-domestique, où il constituerait le lien entre le cycle sauvage et l'homme, en consommant des hôtes intermédiaires et constituant la principale source de contamination de l'environnement des hommes (cf paragraphe 2.2.2).

Objectifs généraux

Les recherches en écologie présentées ici s'inscrivent dans des projets de recherche internationaux (Transech) financés par l'Institut National de la Santé des Etats-Unis (NIH) et le Fond National pour la Science (NSF), dans le cadre du programme "Ecology of infectious diseases". L'objectif en est de comprendre et prédire la dynamique de la transmission d'*Em* en Chine. Dans cette région du monde la maladie est reconnue comme un problème de santé publique dans de nombreux foyers endémiques où les plus fortes prévalences au monde ont été enregistrées dans des populations importantes pour le tiers constituées de jeunes. Le projet de recherches s'est tourné vers des analyses spatiales et de modélisation de données écologiques et épidémiologiques collectées à de multiples échelles et dans différents sites d'étude. Ces recherches ont impliqué un travail d'équipe en collaboration avec des médecins (Centre for Disease Control en Chine), des géographes (Université de Salford), des vétérinaires (Université de Zurich) et des parasitologues (Université de Salford).

Dans ce cadre général, ce travail de thèse a pour objectif général d'expliquer de manière quantitative et de prédire les paramètres écologiques des populations d'hôtes intervenant dans la transmission d'*Em*, sur différentes aires d'études, tels que :

1) les distributions spatiales des communautés d'hôtes intermédiaires (micro-mammifères)

2) les comportements spatiaux "à risque" des populations de chiens domestiques (*Canis familiaris*) intervenant dans la transmission du parasite à l'homme (utilisation de l'espace et défécation).

Pour ce faire, nous avons eu recours à des méthodes empruntées à l'écologie comportementale, des populations et des communautés. Dans ces champs disciplinaires, notre démarche générale a été de construire des modèles ou d'utiliser des outils de modélisation appropriés à l'analyse de données éco-épidémiologiques.

Plan du mémoire

Le mémoire est composé de huit chapitres regroupés en trois grandes parties : l'introduction (chapitres 1 et 2), les travaux de recherches (chapitres 3 à 6) et la discussion générale (chapitre 7 et 8).

L'introduction qui suit expliquera dans un premier chapitre l'importance de l'étude de l'écologie des populations d'hôtes dans l'étude et la modélisation du risque de transmission des macro-parasites. Après avoir exposé les paramètres écologiques intervenant dans la transmission, nous passerons en revue les méthodes existant pour modéliser le risque de transmission du parasite. Puis, l'état des connaissances de l'écologie de la transmission d'*Em* sera présenté avant un bilan détaillé de la situation éco-épidémiologique en Chine. Nous exposerons les méthodes de modélisation appliquées à la transmission d'*Em*. Leur synthèse nous permettra de mettre en relief les principales limites et écueils méthodologiques que les chercheurs rencontrent pour expliquer et prédire (c'est-à-dire modéliser) le risque de transmission d'*Em* dans cette région du monde. Dans ce contexte nous présenterons alors les questions éco-épidémiologiques auxquelles ce travail de thèse a tenté de répondre.

Le deuxième chapitre sera consacré à la présentation des principaux pré-requis méthodologiques nécessaires pour répondre aux questions éco-épidémiologiques. Nous présenterons les connaissances dans le domaine de l'écologie concernant la modélisation des distributions spatiales des espèces et des groupes d'espèces. Puis, nous exposerons les méthodes d'écologie comportementale disponibles pour l'étude des comportements à risque des hôtes définitifs. Ces connaissances nous permettront d'expliquer la démarche et les questions méthodologiques de ce travail.

Les résultats des recherches réalisées seront présentés dans les chapitres 3 à 5 de ce mémoire. Le troisième chapitre du mémoire exposera, après avoir expliqué les principales méthodes, les résultats de la construction d'un modèle explicatif des assemblages de micro-mammifères en Chine, sous la forme d'une publication scientifique et de résultats complémentaires. Suivra, en quatrième chapitre, l'évaluation des prédictions d'un modèle régional des distributions des assemblages présentées sous la forme d'une publication scientifique en cours de préparation ainsi que d'une expérience de simulation complémentaire. Le cinquième chapitre sera consacrée à l'évaluation du rôle du chien dans la transmission du parasite sur le plateau tibétain, en exposant les résultats sous la forme d'une publication scientifique soumise à publication.

Enfin, les sixième et septième chapitres du mémoire seront consacrés à la discussion générale de nos travaux de recherche.

Introduction

Chapitre 1

Écologie et modélisation de la transmission d'*Echinococcus multilocularis* en Chine

1.1 Écologie et modélisation de la transmission des macro-parasites

1.1.1 Écologie de la transmission

Les parasites ont pour habitat tout ou partie de l'organisme de leurs hôtes, excepté pendant leur phase libre durant laquelle ils sont soumis aux conditions environnementales (Combes, 2001). Aussi leur survie dépend-elle de l'accomplissement d'un cycle constitué de plusieurs stades de l'oeuf à l'adulte, chacun correspondant à un habitat particulier (hôte et environnement). La dispersion de leurs propagules ainsi que leur passage d'un habitat à l'autre constituent le processus de transmission. Celle-ci pourra s'effectuer par contact direct (inhalation, contacts sexuels,...), par consommation (d'un hôte contaminé ou du stade libre du parasite), ou encore par l'intermédiaire d'un vecteur arthropodes piqueur (Woolhouse, 2002).

Les macro-parasites, tels que les helminthes comme *Em*, sont des parasites multi-cellulaires qui réalisent des cycles hétéroxènes impliquant l'existence de plusieurs espèces hôtes, chacune correspondant à une étape de leur cycle. Pour de tels parasites on distinguera alors les hôtes intermédiaires, à l'intérieur desquels le développement et la reproduction asexuée du parasite a lieu, des hôtes définitifs, à l'intérieur desquels le parasite réalisera sa reproduction sexuelle.

Les paramètres écologiques de la transmission

L'existence d'une interaction durable avec ses hôtes conditionne la survie à long terme du parasite et nécessite la réalisation de deux processus : leur rencontre et leur compatibilité (Combes, 2001). Ces processus ont été conceptualisés sous la forme de filtres qui doivent être franchis par le parasite pour qu'il soit transmis. Les degrés d'ouverture de ces filtres détermineront l'intensité de la transmission. Alors que la rencontre détermine leur probabilité de contact et dépend de la distribution et des comportements des hôtes, leur compatibilité constitue la probabilité qu'ils ont de vivre ensemble durablement et dépend des ressources (métabolites) et systèmes immunitaires de l'hôte. Les épidémiologistes ont concentré leurs efforts sur l'étude des interactions parasite/hôtes autour du filtre de compatibilité qui relève de la biologie cellulaire,

de l'immunologie et de la parasitologie. Dans notre champ disciplinaire nous nous concentrerons sur les variables déterminant le filtre de rencontre.

Le complexe parasite-hôtes ne peut fonctionner durablement que dans les zones géographiques et pendant la période où la conjonction des deux conditions du filtre de rencontre, cohabiter puis se rencontrer, a lieu. La première condition est la cohabitation des hôtes et du parasite dans le même écosystème et la même localisation. Cela nécessite une superposition spatiale à différentes échelles concernant i) les distributions spatiales (aires géographiques et habitats) du parasite et des hôtes, ii) les conditions locales pour la survie du parasite ainsi que la superposition locale des hôtes et de la forme infectieuse du parasite (oeufs), de première importance lorsque la transmission est directe. L'analyse de la cohabitation inclut donc l'étude des variables écologiques telles que les conditions de l'environnement ainsi que les distributions spatiales (habitats), les densités et la dynamique des hôtes.

Si les acteurs cohabitent, il faudra qu'ils puissent ensuite se rencontrer. La transmission dépend donc de la présence des hôtes au moment et à l'endroit où le parasite est accessible (Morgan et al., 2004). Elle dépendra des comportements de l'hôte et du parasite qui favoriseront leur rencontre et la transmission. L'hôte définitif étant responsable de la dispersion des propagules du parasite, il parcourt en général de plus grandes distances que les hôtes intermédiaires (Morgan et al., 2004). Par exemple, le raton laveur disperse les oeufs du nématode *Baylisascaris procyonis* dans l'environnement des micro-mammifères. Ainsi, les mouvements individuels des hôtes définitifs pourront être également considérés pour expliquer la rencontre entre les propagules et hôtes intermédiaires.

La transmission peut varier dans le temps et dans l'espace, elle est donc dynamique. L'étude de la probabilité de se rencontrer suppose la prise en compte de la variabilité de ces paramètres et nécessitera l'intégration de plusieurs disciplines telles que l'écologie des communautés, la dynamique des populations et l'éthologie. Cela constitue alors le champ de recherche de l'écologie de la transmission du parasite, défini comme la détermination et l'analyse des "facteurs qui modulent la transmission d'un parasite au sein d'un Système Parasite Hôte (SHP), tant dans sa composante temporelle que spatiale" (Raoul, 2001). Elle portera donc essentiellement sur les facteurs déterminant les distributions spatiales et temporelles des hôtes du parasite.

1.1.2 Importance des distributions des hôtes

On ne peut expliquer, prédire (modéliser) et cartographier la distribution spatiale de la forme infectieuse d'un parasite, d'une zoonose, uniquement par l'analyse des prévalences humaines. Cette démarche n'offre qu'une perception réduite de la transmission du pathogène et de la distribution des prévalences elles-mêmes puisque (Peterson, 2006) :

- il peut exister une non correspondance entre le risque et l'incidence

- les données de prévalences humaines peuvent être rares ou relevées à de faibles résolutions spatiales ou encore sous-estimées

- elles sont souvent peu précises quant à la localisation et à la période de la contamination.

Il s'agit plutôt, pour comprendre et prédire leurs distributions, d'analyser les variations du processus de transmission dans son ensemble et de considérer chacun des paramètres de l'écologie de la transmission, en faisant appel à une recherche multidisciplinaire (Giraudoux

et al., 2007). Dans cet objectif, les paramètres de l'écologie des hôtes, définissant la cohabitation et la rencontre (distributions spatiales et comportements), pourront être utilisés pour modéliser la transmission de manière statique ou dynamique.

1.1.3 Modéliser les distributions spatiales des hôtes (modélisation statique)

La probabilité de transmission des pathogènes à un hôte susceptible dépend de sa distance à un hôte infecté (Ostfeld et al., 2005). Les facteurs environnementaux et paysagers influençant les distributions spatiales des hôtes au sein du système écologique qu'il intègre, ainsi que leurs probabilités de rencontre sont donc d'une importance essentielle pour comprendre les dynamiques des maladies.

La survie, la reproduction et la dispersion des espèces animales sont largement déterminées par les attributs biotiques et abiotiques définissant leurs habitats. Il en est de même pour le parasite, pour qui l'habitat, excepté pendant sa phase libre, demeure l'organisme hôte. La superposition spatiale des habitats de chacun de ses hôtes permettra de mieux comprendre et prédire la distribution spatiale de cette zoonose.

Des modèles statistiques des distributions spatiales ont été développés pour définir les habitats des hôtes et de ce fait expliquer les relations entre des données spatiales environnementales et épidémiologiques. La modélisation des niches écologiques est un outil emprunté à l'écologie des populations récemment utilisé en épidémiologie (Peterson, 2006). Du fait de l'avancée méthodologique dans ce domaine, une grande diversité des méthodes de modélisation est accessible aux épidémiologistes pour expliquer et prédire les distributions spatiales des espèces. Ils ont été développés pour s'adapter aux problèmes rencontrés dans les jeux de données éco-épidémiologiques : auto-corrélation spatiale, taille faible d'échantillonnage et événements rares. Nous développerons cet aspect méthodologique dans le paragraphe 2.3.

Parallèlement, la disponibilité de nouveaux Systèmes d'Information Géographique (SIG) offre de multiples outils pour le développement des cartographies du risque pour de nombreuses maladies (Hay (2000); Pffeifer and Hugh-Jones (2002) pour une revue des outils disponibles). Les multiples utilisations de l'imagerie satellitale tiennent de la disponibilité de telles données (Danson et al., 2008) : i) pour la quasi totalité de la surface du globe, permettant la cartographie dans les zones reculées où aucune autre donnée paysagère n'est disponible, ii) en tant qu'archives historiques pour comprendre les changements paysagers des 30 dernières années, iii) pour une large gamme de résolutions permettant des analyses à de multiples échelles. Des données satellites et topographiques sont en accès libre sur le World Wide Web et servent d'outils de base pour de nombreuses recherches. Depuis peu, les données topographiques du SRTM (Shuttle Radar Topographic Mission) à 30 m de résolution sont disponibles librement pour l'ensemble du globe.

La modélisation des distributions spatiales des différentes espèces hôtes peut être réalisée indépendamment. Les cartes de prédictions peuvent ensuite être superposées (Ostfeld et al., 2005). On peut alors modéliser le risque pour les populations humaines d'être en contact avec le parasite, risque qui conditionne les prévalences observées (Peterson, 2006). Cette méthodologie en plusieurs étapes a par exemple apporté des éléments de connaissance sur la dispersion régionale de la grippe aviaire par certaines espèces d'oiseaux afin de prévenir l'émergence de la

maladie en Amérique du Nord (Peterson, 2007). La cartographie des habitats des arthropodes vecteurs de maladies est aujourd'hui largement répandue du fait de la sensibilité de leurs paramètres physiologiques intervenant dans la transmission aux conditions de température et d'humidité, c'est-à-dire à des variables abiotiques accessibles en données satellitales. La reproduction des tiques et des moustiques est en effet conditionnée par les pluies. Par exemple dans le cas de la maladie de Lyme, la définition des habitats des tiques a permis d'expliquer les facteurs environnementaux à risque et de prédire les aires potentiellement à risque (Guerra, 2002). En revanche, la cartographie des hôtes vertébrés est plus rare. Elle s'est avérée prédictive des cas humains pour les parasites à pathogène à transmission directe tel qu'un hantavirus transmis par des espèces de rongeurs (Glass et al., 2000; Langlois et al., 2001; Suzan et al., 2006). Cependant, l'inconvénient majeur de ces approches est qu'elles n'intègrent pas la dynamique de la transmission.

1.1.4 La rencontre dans les modèles déterministes (modélisation dynamique)

Les modèles déterministes

Les modèles déterministes permettent d'étudier la dynamique de la transmission des populations de parasites. Contrairement aux modèles statistiques, leurs paramètres correspondent à des paramètres biologiques de la transmission (taux de natalité, de mortalité, d'infection, densité des populations,...).

Les différentes équations qui composent de tels modèles gouvernent la transmission du parasite entre différents hôtes et/ou leurs différents états infectieux. Ils incorporent alors le taux de transmission (ou de contact) et les densités de chaque état/hôtes, comme par exemple la simple formule du modèle βSI (MacCallum,2001), où β représente le taux de transmission et S et I les densités des hôtes susceptibles et infectés respectivement, qui suppose que la transmission est seulement un phénomène densité-dépendant. Ces modèles sont largement développés dans le domaine vétérinaire, tel que le "susceptible-infectious-removed" (SIR) qui considère le nombre d'animaux immunisés (Anderson and Clements, 2000) et sont souvent développés pour une seule espèce hôte. La transmission entre espèces hôtes, engendrée par leur rencontre, peut donc être résumée par le paramètre β, c'est-à-dire le taux de transmission entre individus infectés (Dobson and Foufopoulos, 2001).

Bien que simplificateurs de la réalité, ces modèles basés sur les processus de la rencontre peuvent aider à tester des hypothèses et construire des théories, ou à estimer les facteurs qui affectent la transmission pouvant être utilisés dans le contrôle (de Jong, 1995). Cependant leur développement se heurte à deux inconvénients majeurs :

i) les paramètres concernant les populations d'hôtes et de pathogènes ne peuvent pas toujours être estimés de manière précise et le sont par simulation ou sur des bases théoriques ; l'estimation de β peut s'avérer particulièrement difficile (Haydon, 2008),

ii) ils supposent un équilibre de la situation épidémiologique et une transmission homogène dans l'espace et le temps, les densités et le taux de transmission étant considérés comme constants.

De nouveaux outils

Concernant l'estimation des paramètres, de récents travaux suggèrent que des données

écologiques empiriques peuvent aider à estimer le taux de contact entre 2 espèces hôtes (β) (Kauhala and Holmala, 2006). Des recherches sur la transmission de la maladie de carré (Canine Distemper Virus) dans une communauté de carnivores du parc national de Serengeti (Tanzanie) ont mis en exergue la possibilité d'approcher l'estimation du paramètre par une recherche interdisciplinaire réunissant des résultats de l'écologie comportementale et de l'épidémiologie (Craft et al., 2008). Les comportements spatiaux (obtenus par suivis télémétriques par radio pistage), ainsi que la structure sociale des populations ont permis par exemple d'observer qu'un seul assemblage d'espèces hôtes pouvait expliquer les prévalences sporadiques et discontinues observées chez les lions.

Aussi, des outils de statistiques modernes offrent la possibilité de dépasser les limites des hypothèses d'homogénéité des paramètres des modèles déterministes et permettent l'exploration de l'hétérogénéité des facteurs environnementaux susceptibles d'influencer la transmission. Les modèles non paramétriques (Modèles Additifs Généralisés, Réseaux de neurones,...) ou les régressions spatialement explicites (auto-régression) permettent de modéliser des processus non-linéaires, variables spatialement et peuvent être incorporés dans des modèles de transmission traditionnels et déterministes, par exemple l'utilisation de Réseaux de neurones dans un modèle SIR (Keeling et al, 1999). De tels modèles dits spatio-déterministiques permettent alors de cartographier le risque de manière dynamique (Ostfeld et al., 2005; Kitron, 2000).

1.2 Écologie de la transmission d'*Echinococcus multilocularis*

1.2.1 Paramètres écologiques de la rencontre

Nous nous concentrerons à décrire à travers le cycle de vie du parasite (Figure 1, p 3), les paramètres de l'écologie des hôtes et du parasite qui interviennent dans l'ouverture/fermeture du filtre de rencontre et donc ceux ayant trait aux individus et populations des hôtes intermédiaires ou définitifs et du parasite. Les paramètres immunologiques soumis à des variations sont décrits dans Pleydell et al. (2006).

Le cycle de vie du parasite consiste en trois états : l'oeuf, la larve et l'adulte. L'oeuf est pondu par des vers adultes fixés sur la paroi intestinale de l'hôte définitif. Les oeufs sont excrétés dans l'environnement *via* les fèces de l'hôte définitif. La distribution de la contamination est donc principalement gouvernée par le comportement de défécation de l'hôte définitif.

Les oeufs pourront alors être ingérés par les hôtes intermédiaires au sein desquels le parasite réalise la seconde étape de son cycle : le stade larvaire et la production de protoscolex (reproduction asexuée). La probabilité pour un hôte intermédiaire d'ingérer un oeuf dépend de la juxtaposition de la distribution spatiale des oeufs par l'hôte définitif avec celle des micro-mammifères.

Le passage du stade protoscolex au stade adulte se réalisera par le passage de ces derniers dans l'intestin de l'hôte définitif et dépend donc du comportement de prédation de l'hôte définitif. La prédation des hôtes intermédiaires sera fonction de leurs distributions spatiales, de leurs densités et de leurs accessibilités.

1.2.2 Variations spatiales et multi-scalaires des paramètres écologiques de la transmission

Les variations spatiales et temporelles des paramètres écologiques des hôtes intervenant dans l'ouverture du filtre de rencontre entre les hôtes et le parasite (distributions, comportements) ont constitué l'objet des travaux antérieurs en éco-épidémiologie d'*Em* à trois échelles spatiales principalement : continentale, régionale et locale (Giraudoux et al., 2002). Ci-après, les variations temporelles du cycle seront évoquées parallèlement aux variations spatiales par des processus de dynamique des populations.

1.2.2.1 Échelle locale (n x 10 m)

De manière générale une distribution agrégative spatiale et temporelle du parasite est constatée. Les résultats des études concernant les charges parasitaires mettent en évidence une distribution agrégative du parasite au sein des populations d'hôtes définitifs de renards en Europe (Hofer et al., 2000; Raoul et al., 2001b; Guislain, 2006) et de chiens en Chine (Budke et al., 2005a; Torgerson, 2003). Il est possible que cette agrégation résulte de la résistance individuelle de l'hôte définitif. Aussi la densité élevée (suite à la reproduction asexuelle) du parasite au sein des micro-mammifères (plusieurs centaines de protoscolex par individus) (Galvani, 2003) peut favoriser de fortes concentrations du parasite au sein de l'hôte définitif après ingestion de telles proies. Le comportement de défécation de certains individus aura davantage de poids dans la transmission que d'autres et la contamination pourra être attribuée à quelques individus plutôt qu'à l'ensemble de la population. De plus, la survie des oeufs dans l'environnement dépend des conditions micro-climatiques, la longévité des oeufs (d'environ un an dans les conditions optimales) étant fortement réduite dans des conditions sèches et chaudes (Veit et al., 1995). Ainsi, les prévalences des renards roux sont corrélées à la moyenne annuelle des précipitations (Miterpakova et al., 2003) et aux habitats à sols humides en Allemagne (Staubach et al., 2001). Le comportement de défécation chez le renard roux varie également en fonction du milieu et de la saison et ceci ne semble pas être corrélé aux densités de ses populations (Giraudoux et al., 2002; Guislain, 2006) mais davantage à la proportion des habitats optimaux et des densités de leurs proies (Robardet et al., 2008; Guislain, 2006). Les distributions spatiales et les densités des hôtes intermédiaires varient également localement, en fonction de la disponibilité locale de leurs habitats et des saisons (liées à la reproduction, la mortalité ou la dispersion). Par voie de conséquences le comportement de prédation de l'hôte définitif varie localement et saisonnièrement avec les densités d'hôtes intermédiaires. Le renard roux va modifier son régime alimentaire avec la densité de proies : spécialiste lorsqu'elles sont importantes et généraliste lorsqu'elles déclinent (Giraudoux, 1991; Dupuy et al., 2009). Ainsi, la diminution des prévalences vulpines du milieu rural vers le milieu urbain a été observée parallèlement à une diminution de consommation de campagnols terrestres (*Arvicola sherman*) selon ce même gradient et à la capacité pour le renard de diversifier ses habitudes alimentaires (Hegglin et al., 2007; Robardet et al., 2008).

La rencontre parasite/hôtes intermédiaires est donc favorisée dans des micro-foyers (n x 10m) ou des macro-foyers (n x 1km) où la contamination est agrégée dans les habitats optimaux des hôtes intermédiaires fréquentés par des hôtes définitifs. Un tel type d'habitat où l'ouverture du filtre de rencontre est maximale a été identifié en France par exemple : dans les Ardennes comme

les bordures de prairies à végétation de taille moyenne (Guislain, 2006), en Franche-Comté comme les lisières de cultures et les banquettes herbeuses en bordure de chemins (Delattre et al., 1990) et dans l'agglomération de Nancy comme la zone péri-urbaine (Robardet et al., 2008).

1.2.2.2 Échelle régionale (n x 10 km)

La composition du paysage influence la dynamique des populations d'hôtes intermédiaires (Lidiker 95) et par conséquence la transmission d'*Em* (Pavlowski, 1964; Giraudoux et al., 2003). Ainsi l'hypothèse du ROMPA (Ratio of Optimal to Marginal Patch Area) a été développée pour mettre en évidence les effets de la composition du paysage sur la dynamique des populations de rongeurs (Lidicker 1995, 2000). La proportion que représente l'habitat optimal pour l'espèce considérée sur l'aire totale de la zone d'étude permet d'estimer le ROMPA. Sa valeur influence alors la probabilité pour l'espèce de suivre une dynamique cyclique (fluctuations de densités multi-annuelles ou vague voyageuse) par un effet combiné de la prédation et de la disponibilité d'habitat optimal sur la dispersion.

Les habitats des espèces de micro-mammifères sont définis localement à partir de données de piégeage (lignes de pièges) ou de transects (relevés d'indices de présence de surface de micro-mammifères à intervalles réguliers) menés en parallèle à des analyses paysagères. L'habitat optimal est alors caractérisé par la classe paysagère présentant les plus fortes densités relatives pour l'espèce considérée, ceci en tenant compte de la dynamique des populations dans la zone géographique étudiée lorsque des données pluriannuelles sont disponibles.

En France, dans le département du Doubs, l'augmentation de la surface en herbe pour l'élevage des vaches laitières depuis la fin des années cinquante offre des habitats optimaux pour deux espèces principales de rongeurs *Microtus arvalis* et *Arvicola terrestris*. Ainsi, de fréquents pics de densités apparaissent lorsque la proportion des prairies permanentes en herbe excèdent un certain seuil (Giraudoux et al., 1997). Le paysage influence alors indirectement la relation prédateur-proie et par conséquent la transmission d'*Em*. En effet, il a été démontré que la dynamique de la transmission du parasite est plus active dans différentes aires endémiques où les densités de proies pour l'hôte définitif sont élevées pendant quelques mois ou années (Giraudoux et al. 1996, Saitoh et Takahashi 1998). Aussi, dans les mêmes zones où les densités sont élevées, les prévalences humaines et vulpines pour EA étaient élevées dans le département du Doubs (Viel et al., 1999; Raoul et al., 2001b; Pleydell et al., 2004).

Les effets des perturbations paysagères sur la dynamique de la transmission ont été particulièrement étudiés en Chine et seront détaillés dans la partie 1.2.3 de l'introduction.

1.2.2.3 Échelle continentale (n x 100 km)

Á l'échelle continentale, la sensibilité des oeufs aux conditions climatiques (température et humidité) influence la répartition géographique continentale du parasite qui se limite à l'hémisphère nord (Rausch, 1995).

La transmission dépend de la présence d'habitats favorables pour les hôtes susceptibles et de la juxtaposition de leurs aires de distributions avec celle du parasite. Étant donné le caractère cosmopolite des aires de distribution des principaux hôtes définitifs (renards et chiens), ceci

tient essentiellement aux distributions des micro-mammifères. Ces dernières dépendent alors, à cette échelle, des facteurs environnementaux mais aussi biogéographiques. Ainsi peut être expliquée la large diversité globale observée dans le spectre des hôtes d'*Em* et de ce fait la large diversité de complexes hôtes/pathogènes. Alors que dans les toundras des régions arctiques (Alaska), le cycle sauvage implique un seul hôte définitif, le renard arctique, se nourrissant principalement du campagnol nordique, les communautés d'espèces hôtes sont plus diversifiées dans les régions sub-arctiques (Amérique du Nord et Eurasie) du fait de l'hétérogénéité des conditions climatiques et des paysages observés.

1.2.3 Bilan des connaissances écologiques et épidémiologiques en Chine

Des études environnementales de terrain (campagnes de piégeages ou de transects), des analyses paysagères ainsi que des suivis épidémiologiques des populations de chiens domestiques et d'hommes ont été menées en parallèle dans différentes provinces chinoises reconnues comme endémiques. Ces recherches éco-épidémiologiques ont apporté des éléments en faveur des deux principales thèses expliquant l'importance de la zoonose dans ces régions :

i) L'influence des changements paysagers d'origine anthropique sur les distributions spatiales des populations d'hôtes intermédiaires, et de ce fait sur les prévalences humaines.

ii) Le rôle principal du chien domestique en tant qu'hôte définitif d'un cycle péri-domestique impliquant l'homme comme hôte accidentel.

Après une description de la situation épidémiologique des populations humaines et animales dans les principaux foyers endémiques chinois, nous détaillerons les principaux résultats de ces recherches en Chine centrale. Nous dresserons enfin un bilan des systèmes de transmission existants dans cette région du monde.

1.2.3.1 Situation épidémiologique en Chine

Les prévalences humaines

Sur le territoire chinois, des données de prévalences humaines ont pu être récemment enregistrées (depuis une quinzaine d'années) suite à des campagnes de suivis épidémiologiques par imagerie médicale (échographies, Ultrasons) ou sérologiques, les relevés des cas hospitaliers n'étant souvent pas utilisables pour des analyses (non numérisés) (Vuitton et al., 2003; Craig and The Echinococcosis Working Group in China, 2006). Ces campagnes sont effectuées par les centres nationaux de santé (CDC, Center of Disease Control) dans des aires ciblées.

Trois principaux foyers et sept provinces ou régions autonomes ont été identifiées comme endémiques (Ito et al., 2003; Craig and The Echinococcosis Working Group in China, 2006) (Figure 1.1) :

- au Nord-ouest du pays : la province du Xinjiang,
- en Chine centrale : les provinces du Qinghai, du Tibet, du Gansu, du Sichuan et du Ningxia,
- au Nord Est du pays : les provinces de la Mongolie Intérieure et du Heilongjiang.

Les communautés humaines concernées sont multiples : tibétaine (provinces du Qinghai et du Tibet), Han (province du Gansu), Hui (province du Ningxia), Uygur et Kazak (province du

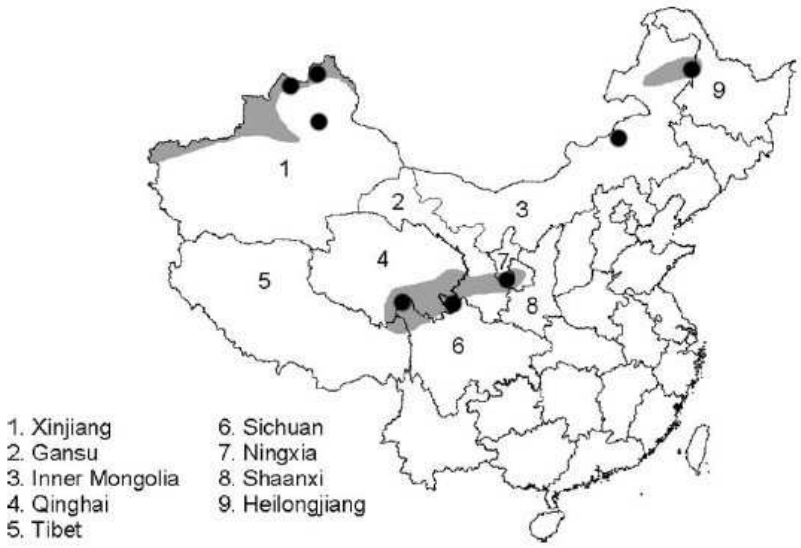

FIG. 1.1 – Carte de la Chine et de ses provinces. Les points noirs représentent les sites pour lesquels des données éco-épidémiologiques sont disponibles, les zones grisées correspondent aux aires endémiques pour EA. Figure modifiée de Craig and The Echinococcosis Working Group in China (2006)

Xinjiang) et Mongoles (provinces de la Mongolie Intérieure et du Heilongjiang). Ces communautés partagent des conditions économiques et sociales : elles sont pauvres, constituent souvent mais pas seulement des groupes ethniques minoritaires de religions musulmane ou bouddhiste et leurs principaux revenus sont issus de l'agriculture et/ou de l'élevage de bétail (Vuitton et al., 2003).

Des campagnes de diagnostic récentes et en cours permettent d'actualiser les prévalences (Vuitton et al., 2003). Sur le plateau tibétain, la majorité des cas d'EA sont concentrés dans les communautés nomades et pastorales tibétaines dans la Préfecture Autonome Tibétaine sur l'est du plateau et dans l'ouest du Sichuan (Tiaoying et al., 2005). De ce fait, la zoonose est considérée dans cette région comme un défi important de santé publique (Craig and The Echinococcosis Working Group in China, 2006). Aussi, de fortes prévalences (jusqu'à 15.8 %) ont été enregistrées dans les communautés de fermiers Han et Hui habitant les hautes vallées du Sud du Gansu (Craig et al., 2000) ou dans le Sud du Ningxia (Yang et al., 2005) (jusqu'à 8%). Enfin, les foyers du Xinjiang ont été redéfinis récemment ; la zoonose est assez fréquente dans les montagnes de l'Altai, de l'ouest du Junggar et des Tianshan (jusqu'à 4 % dans l'Altai) (Zhou et al., 2000). De manière générale, sur l'ensemble du continent chinois les distributions des cas humains semblent être corrélées aux variations régionales du climat et du paysage (Zhou et al., 2000).

Les hôtes

Une grande diversité d'hôtes potentiels définitifs ou intermédiaires est observée au sein des foyers chinois. La diversité des complexes pathogènes intervenant dans les cycles sauvages ou

domestiques d'*Em* est résumée dans le tableau 1.

Au regard des prévalences relevées, le renard roux (*Vulpes vulpes*), le renard corsac (*V. corsac*) et le renard tibétain (*V. ferrilata*) apparaissent comme les principaux hôtes définitifs sauvages dans les aires chinoises, le renard roux (*Vulpes vulpes*) étant commun à toutes les aires endémiques (Table 1). L'implication du chien domestique *Canis familiaris* dans le cycle d'*Em* a été mise en évidence dans le Sichuan, le Qinghai le Gansu et le Xinjiang (Table 1).

Les espèces d'hôtes intermédiaires susceptibles à *Em* sont nombreuses et hétérogènes d'une région à l'autre. Elles incluent des espèces des régions biogéographiques du Palearctique et de l'Indo-Malasien. Les espèces du genre *Microtus* (*Microtus limnophilus/oeconomus*, *Microtus irene*, *Microtus brandti*) sont impliquées dans la plupart des foyers. Cependant, des espèces de gerbilles (*Meriones spp.*), de hamsters (*Cricetulus spp.*) et de lagomorphes (*Ochotona spp.*, *Lepus spp.*) peuvent tenir une place importante dans le cycle.

1.2.3.2 Paysages et distributions spatiales des populations d'hôtes intermédiaires

La définition locale des habitats à partir de données de piégeage a permis de mettre en évidence les distributions des espèces d'hôtes intermédiaires et les systèmes de transmission qui en découlent dans la plupart des régions endémiques de la Chine centrale (Giraudoux et al., 2006).

Depuis les années 80, dans le comté de Serxu, sur le plateau tibétain, des parcelles protégées du pâturage (clôturées) ont été attribuées aux familles pastorales autour des villages. Ce changement dans la gestion pastorale a entraîné une réduction des pâtures communales et une augmentation de la pression de pâturage dans ces zones. Des différences significatives de densités des espèces hôtes potentielles d'*Em* et potentiellement cycliques ont été observées de part et d'autre du gradient environnemental engendré par les fortes pressions de pâturage (Raoul et al., 2006). Alors que les habitats optimaux d'*Ochotona curzoniae* et *Cricetulus kamensis* étaient des prairies sur-pâturées, ceux de *Microtus limnophilus* étaient des prairies clôturées et ceux de *Microtus leucurus* des buissons de Potentille. Parallèlement, les prévalences humaines et la proportion de prairies encloses dans les villages de cette région étaient corrélées (Wang et al., 2004).

Dans la province du Gansu (comtés de Zhang et de Puma), une déforestation intensive a transformé durant les cinquante dernières années, une forêt de 400 km2 en un paysage composé principalement de buissons et de prairies, stade intermédiaire dans la succession végétale après déforestation. La disponibilité d'une carte d'occupation du sol a permis de déterminer, sur la base des résultats de piégeages (Giraudoux et al., 1998), les ROMPA des deux espèces dominantes et reconnues comme cycliques *Microtus limnophilus* et *Cricetulus longicaudatus* ; *Microtus limnophilus* étant une espèce susceptible à *Em*. Ils ont été définis comme la proportion de la classe paysagère prairies/buissons sur l'aire totale telle que définie sur une carte d'occupation du sol du gouvernement chinois. Ainsi une représentation cartographique qualitative des ROMPA a été mise en corrélation avec les prévalences humaines regroupées en zones (Giraudoux et al., 1996; Craig et al., 2000; Giraudoux et al., 2003). Ces dernières se sont révélées être influencées par la proportion de l'habitat optimal de *Microtus limnophilus*.

Dans la province du Ningxia, l'impact de la gestion forestière sur les distributions des es-

TAB. 1.1 – Les hôtes des foyers chinois. Les prévalences (entre parenthèses) sont reportés. Pour les micro-mammifères, les présences des espèces identifiées sont également rapportées (1).

Famille	Espèces hôtes	Sichuan	Qinghai	Ningxia	Gansu	Xinjiang	Inner Mongolia
	Vulpes vulpes	19/32 (59.4)[1]		3/20 (15)[3]		16/48 (33)[9]	
	Vulpes ferrilata	89/193 (46)[1]	3/9 (33.3) [2]				
	Vulpes corsac						2/33 (6.1) [3]
							12.6 (19/151) [3]
	Canis lupus					1/2 (50) [39]	
	Canis familiaris	40/242 (16.7) [1]	2/15 (13.3)[11]		6/58 (10.3) [10]	3/30 (10) [9]	
		46/295 (15.6) [4]					
		44/371 (12) [5]					
	Microtus brandti						NA [6]
	Microtus irene	3/12 (25) [1]					
	Microtus limnophilus/oeconomus* (C)	1	1		NA [7]	NA [9]	
	Meriones spp.						NA [6]
	Mus musculus				1		
	Spermophilus (dauricus, alashanicus)			NA [6]			
	Citellus spp.					1	
	Citellus dauricus						
	Cricetulus kamensis (C)	NA [8]					
	Eoospalax fontanierii			NA [6]	1		
	Ochotona curzoniae (C)	13/322 (5.6)[1]	11/319 (3.5)				
	Ochotona daurica					1	
	Lepus oiostolus	NA [1]					
	Marmota					1	

[1]Qiu et al. (1995) [2]Wang et al. (1999) dans Zhou (2001) [3]dans Craig and The Echinococcosis Working Group in China (2006) [4]He et al. (2000) dans Craig and The Echinococcosis Working Group in China (2006) [5]Budke et al. (2005a) [6]dans Vuitton et al. (2003) [7]Giraudoux et al. (2003) [8]Raoul, non publié [9]dans Zhou (2001) [10]Craig et al. (1992) [11]Schantz (1998) [12]Zhang et al. (2006)

(*)M. limnophilus and M. oeconomus sont des espèces différenciables uniquement par caryotypage.

C : espèces reconnues à dynamiques cycliques.

NA : l'hôte a été reconnu susceptible à *Em* mais aucune prévalence précise n'est disponible.

pèces a été explorée sur le plateau de Loess (Raoul et al., 2006). Les assemblages de micro-mammifères étaient variables le long du gradient de déforestation et de reforestation. Cependant, beaucoup d'espèces piégées étaient différentes de celles piégées dans le Gansu telles que certaines espèces semi-désertiques *Allacta sibirica* et *Dipus sagitta*. Parallèlement, les prévalences humaines étaient négativement corrélées au *ratio* de prairies permanentes, habitat optimal des *Arvicolinaes* de cette région. Ceci laisse supposer que ces espèces, contrairement à ce qui est observé ailleurs ne sont pas les hôtes principalement impliqués dans le cycle d'*Em* dans le comté de Xiji où les hamsters, gerbilles, lagomorphes, zokors et souris seraient davantage impliqués (Pleydell et al., 2008).

Enfin, les distributions spatiales des habitats et les dynamiques des populations ont été également considérées à l'échelle continentale. Ainsi, des fluctuations temporelles et régionales des densités du genre *Ochotona* ont été mises en évidence le long de transects réalisés entre les villes de Rangtang et de Xining (Lai and Smith, 2003; Giraudoux et al., 2006). De surcroît une corrélation spatiale entre les distributions des cas humains et la proportion de la classe de prairie issue d'une classification non supervisée de l'image satellite a été mise en évidence sur l'ensemble du continent (Danson et al., 2003).

1.2.3.3 Le rôle du chien dans la transmission

En Chine, les populations de chiens sont importantes du fait de l'utilité qu'en ont les hommes pour les activités pastorales (garde des troupeaux en pâtures d'été) ou pour la garde des habitations. Aussi, leur caractère sacré conféré par la religion bouddhique empêche tout contrôle des populations. De fortes densités de populations associées à une grande proximité des chiens aux populations humaines laissent à penser que les chiens constituent la principale source de contamination de l'environnement des villages par la dissémination de fèces contaminés et qu'ils se contaminent en prédatant des micro-mammifères présents dans l'environnement des villages. Des suivis épidémiologiques dans certains villages ont permis l'estimation des prévalences et des facteurs de risques en faveur de ces hypothèses.

Les prévalences

La présence et le nombre (charge parasitaire) de cestodes présents dans les contenus intestinaux des chiens ont été estimés après nécropsie ou purge suite à l'administration d'arecoline hydrobromide.

Sur le plateau tibétain les résultats d'analyse des contenus intestinaux après nécropsies (Qiu et al., 1999) et purges (Budke et al., 2005a) convergent vers une estimation de la prévalence comprise entre 12 et 17 % (Table 1). Dans les provinces du Qinghai et du Gansu, les prévalences ont été estimées après nécropsies à 13.3 et 10 % respectivement (Schantz, 1998; Craig et al., 1992). Cependant, les faibles tailles des échantillons pour ces deux dernières régions compliquent la comparaison de leurs prévalences à celle estimée dans le Sichuan. Dans le sud du Ningxia, le rôle du chien dans le passé a été attesté (Rausch, 1995) mais aucune infection n'a pu être montrée actuellement. Enfin, dans le Xinjiang aucune infection n'avait pu être identifiée sur 27 186 chiens entre 1957 et 1991 (Liu, 1993; Vuitton et al., 2003), mais la découverte plus récente d'une double infection par *E. granulosus* et *Em* suggère la possible intervention du chien en tant qu'hôte définitif dans cette zone géographique (Zhang et al., 2006).

Contamination environnementale et transmission du parasite à l'homme

Des facteurs de risque pour les populations humaines ont permis de mettre en évidence, par des questionnaires adressés aux patients, l'implication du chien domestique dans la contamination de l'environnement humain, dans les provinces du Sichuan (Serxu) et du Gansu. Le fait de posséder un chien ainsi que le nombre de chiens possédés cumulativement constituent des facteurs de risque (Tiaoying et al., 2005; Budke et al., 2005a; Craig et al., 2000).

Les comportements humains semblent jouer un rôle primordial dans la transmission. Les relations directes entretenues avec le chien, liées à son degré de domestication, telles que jouer avec lui (Wang et al., 2001) et lui permettre de dormir à l'intérieur (Schantz et al., 2003) expliquent en partie les prévalences humaines observées. De plus, le fait que les femmes soient davantage sujettes à contamination que les hommes a été interprété par leurs contacts plus fréquents avec les chiens domestiques pour leur alimentation par exemple (Tiaoying et al., 2005). Enfin, le contact aux fèces, tel que le fait d'utiliser les fèces de chiens comme fertilisant, constitue un facteur de risque (Craig et al., 2000).

Prédation et sources d'infection du chien

Les études de facteurs de risque ont également permis la mise en évidence de certains traits d'histoire de vie du chien et comportements pouvant favoriser leur contamination, c'est-à-dire la prédation d'hôtes intermédiaires. Le fait d'avoir un chien mâle est un facteur de risque supposé être engendré par la plus grande propension des mâles à prospecter et à défendre un territoire que les femelles, comme on l'observe chez les renards (Budke et al., 2005a). Aussi, les chiens ayant été observés prospectant dans les colonies de micro-mammifères et qui ne sont pas attachés en permanence constituent des facteurs explicatifs des prévalences de leurs propriétaires (Budke et al., 2005a; Craig et al., 2000).

Contrairement à l'étude des facteurs de risque, les comportements de prédation des chiens domestiques restent peu étudiés. La présence de micro-mammifères dans le régime alimentaire des chiens a été mise en évidence sur le plateau tibétain (Wang et al., 2009) : des restes osseux et des poils étaient présents dans 37 % des 60 fèces analysés. Dans le Nord du Xinjiang des macro-restes de microtinae et de marmottes ont été également retrouvés (Zhou, 2001).

1.2.3.4 Hétérogénéité des systèmes de transmission

L'hétérogénéité des systèmes de transmission en Chine centrale a été résumée par la théorie des méta-populations pour le parasite (Giraudoux et al., 2003, 2006).

Sur le plateau tibétain, la large étendue des pâtures de haute altitude, habitat optimal pour *Ochotona curzoniae* (sujet à des fluctuations multi-annuelles cycliques), les fortes densités observées pour plusieurs hôtes alternatifs du genre *Microtus*, ainsi que les fortes densités de chiens domestiques contaminés dans les villages peuvent expliquer une transmission active. Cette aire d'endémie peut être alors qualifiée de large et stable réservoir pour le parasite.

Par opposition, la transmission dans les aires périphériques du plateau (sud du Gansu et Nord du Ningxia), situées à environ 150 km du plateau, est plus sporadique. L'hétérogénéité actuelle des habitats à dominante agricole ne favorise pas les fortes densités des hôtes intermédiaires. De surcroît, des contrôles de populations de micro-mammifères (par rodenticides) effectuées dans les années quatre-vingt dix ont décimé les populations de prédateurs. En re-

vanche, dans le Gansu, durant le long processus de déforestation (trente ans), les preuves de pullulations anciennes des hôtes intermédiaires favorisées par la disponibilité de leurs habitats optimaux (prairies buissonnantes) et combinées à de fortes densités de chiens peuvent expliquer les prévalences humaines actuelles. Il est alors probable que le plateau tibétain, par le biais de dispersion des populations d'hôtes définitifs principalement (dispersion de renards et vente de chiens) pourrait alimenter les zones de contreforts en parasites et contribuer à la ré-émergence de la transmission dans les habitats favorables aux hôtes intermédiaires (Giraudoux et al., 2003, 2006).

Cependant, les diversités et compositions des espèces potentiellement hôtes pour *Em* ainsi que les facteurs influençant leurs distributions diffèrent entre les sites étudiés, suggérant une large hétérogénéité spatiale des complexes pathogènes à l'échelle continentale. L'ensemble de ces résultats soutiennent l'idée qu'*Em* peut être transmis dans une grande diversité de communautés d'hôtes et que sa dynamique méta-populationnelle dans la Chine centrale fonctionne à travers une large diversité d'éco-zones et de communautés de micro-mammifères qui demeurent encore largement méconnues (Giraudoux et al., 2006; Pleydell et al., 2008).

1.3 Modélisation de la transmission d'*Em*

1.3.1 Modélisation des distributions spatiales des populations d'hôtes

Les premières définitions des habitats des espèces étaient qualitatives. Les piégeages et transects ont été réalisés dans des habitats définis *a priori* ou décrits sur avis d'experts en fonction des compositions et structures de la végétation se succédant le long de gradients des perturbations anthropiques.

Puis, la cartographie des ROMPA pour les espèces d'hôtes intermédiaires a été développée à partir de cartes d'occupation du sol. En France, les proportions de la classe paysagère définie comme habitat optimal (généralement les prairies) (Giraudoux et al., 1997, 2002, 2003) ou des indices prairiaux réalisés à partir de plusieurs classes (Tolle, 2005) ont été estimés et cartographiés. Parallèlement, les prévalences vulpines de fèces collectées en Franche-comté ont pu être modélisées par le *ratio* de composition en prairies (Pleydell et al., 2004). En Allemagne, l'analyse des distributions spatiales des renards infectés a permis de mettre en évidence qu'ils fréquentaient davantage les paysages ouverts (prairies) à sols humides que les forêts (Staubach et al., 2001).

En Chine, les données satellitales constituent souvent les seules données paysagères disponibles. Des métriques paysagères et indices de végétation, tels que le NDVI (Normalized Differential Vegetation Index) et le EVI (Enhanced Vegetation Index) obtenus après classification et traitement d'images satellites, ont été utilisés pour redéfinir les classes définies *a priori* par les écologues et cartographier des ROMPA (Danson et al., 2003). Ainsi, la proportion des classes "prairie" et "buissons" a permis la cartographie de l'habitat optimal pour *Microtus* dans la province du Gansu. De telles variables paysagères se sont alors avérées plus explicatives que d'autres facteurs de risque (âge, genre, propriété d'un chien).

Enfin, des outils modernes de modélisation spatiale établissant des relations statistiques entre les données (prévalences, espèces) et leurs attributs paysagers (variables environnementales),

non paramétriques et spatialement explicites ont été appliquées aux données de prévalences humaines en France (Tolle et al., 2005) ou vulpines en Autriche (Duscher et al., 2006). L'incorporation explicite des variations spatiales sous la forme de "random field" dans des modèles de régression pondérés géographiquement ont permis d'incorporer l'hétérogénéité spatiale des variables environnementales et de considérer l'auto-corrélation spatiale des observations. Cependant, de tels outils ont rarement été appliqués à la définition quantitative des habitats pour les espèces d'hôtes intermédiaires. A notre connaissance, seule la définition des habitats pour le genre *Ochotona* a été réalisée par une approche de modélisation spatiale sur une aire du plateau tibétain (Marston, 2008). Les distributions des occurrences pour ce genre, collectées le long de transects, ont pu être expliquées avec succès par la composition paysagère de leurs environnements, mais aussi par des métriques paysagères (aires et tailles des tâches) et des indices de végétation (Normalized Difference Water Index principalement). Pour ce faire ce sont des modèles linéaires généralisés (GLM) ou additifs (GAM) qui ont été utilisés. Ces derniers, parce qu'ils incorporent les structures spatiales des données en modélisant les distributions par des fonctions de lissage, permettent de "coller" à la structure spatiale des données. Aussi, l 'effet de l'échelle spatiale à laquelle les relations statistiques sont établies a été testé en faisant varier le rayon de l'aire autour du point d'échantillonnage pour laquelle lax métrique paysagère est estimée ("buffer"). Ainsi, l'habitat optimal pour *Ochotona* (les prairies dégradées) étaient les plus corrélées dans un "buffer" de 300 m autour des points d'échantillonnage. En revanche, les méthodes permettant de remédier aux effets de la dépendance spatiale des données, tel que le sous-échantillonnage des données ou l'incorporation de covariables spatiales dans les GAM, se sont avérées inapplicables dans un tel contexte et les modèles utilisés incorporaient l'auto-corrélation spatiale des points d'échantillonnage. D'une part, la faible taille d'échantillon limite le sous-échantillonnage. D'autre part, l'agrégation des transects dans certaines zones en opposition à l'absence de données pour la majorité de la zone d'étude entraîne une surestimation ("over-fitting") de l'interpolation dans les zones échantillonnées. Finalement, les séries temporelles pour ces variables ont permis de prédire précisément les distributions des micromammifères. Cependant, ces prédictions se sont avérées limitées quant à leurs extrapolations sur des données indépendantes.

1.3.2 Modélisation dynamique de la rencontre

Le modèle déterministe développé par Roberts et al. (1986) pour la transmission d'*Echinococcus granulosus* permet de modéliser les prévalences des hôtes définitifs en présence ou non d'immunité par une série d'équations ordinaires différentielles. Il a été appliqué et adapté pour modéliser les prévalences de renards (Roberts and Aubert, 1995) et de chiens domestiques (Torgerson, 2003). Le taux d'infection et la présence d'immunité chez le chien domestique sur les plateaux tibétains ont été estimés (Budke et al., 2005b). Le taux d'infection (β) a été estimé à 0.52 infection par an en supposant une durée de vie du parasite à 5 mois. Aussi, contrairement à *Echinococcus granulosus*, *Echinococcus multilocularis* n'induit pas d'immunité chez le chien. La pertinence de ces modèles est contrainte par l'homogénéité du paramètre β en fonction des individus et dans l'espace. Or, ce paramètre dépendant également de la distribution des hôtes intermédiaires, son hétérogénéité spatiale doit être considérée pour expliquer et prédire les distri-

butions du parasite et les aires potentiellement à risque de transmission. Une telle hétérogénéité spatiale liée aux facteurs environnementaux, pour être décelée, devra alors être étudiée à une échelle plus large que celle à laquelle le parasite s'agrège dans les populations (Torgerson, 2003; Morgan et al., 2004). Enfin, le modèle de Takumi and Van Der Giessen (2005) a permis d'estimer la biomasse du parasite chez le renard et a été utilisé pour estimer l'efficacité des campagnes de contrôle des prévalences.

1.3.3 Modélisation spatio-déterministe

Des modèles spatialement explicites incluant une composante stochastique et basés sur les populations de parasites et d'hôtes plutôt que sur celles des hôtes définitifs uniquement ont été développés pour expliquer les distributions spatiales des prévalences d'*Em*.

La technique de modélisation RAMAS GIS (Milner-Gulland et al., 2004), basée sur un modèle méta-populationnel, a été appliquée dans une région semi-aride du Kazakhstan pour modéliser directement les populations de parasites (metacestodes), considérés comme un ensemble d'individus, ayant pour habitat les populations d'hôtes intermédiaires et le renard comme mode de dispersion de ses vers. Ce modèle nécessite alors l'estimation de paramètres écologiques tels que la végétation (carte des communautés), la biologie des rongeurs (densités et distributions spatiales) et des renards (densités, distances des mouvements) ainsi que la biologie du parasite.

Echi (Hansen et al., 2003) est un modèle basé sur une grille, discrétisée temporellement et structurée par un jeu de lois entre les sous-populations d'hôtes intermédiaires, les renards (individus) et les trois stades du parasite. Les valeurs des paramètres tels que la pression de prédation du renard, le nombre de fèces qu'il dépose et la proportion de cellules de la grille offrant des conditions favorables à la survie des oeufs, ont été simulées pour prédire les prévalences vulpines observées après contrôle dans le nord de l'Allemagne. Différents scénarii ont été testés pour expliquer l'écart entre les prédictions du modèle et les données. Parmi eux, celui considérant un effet du paysage sur la concentration locale de l'infection et en particulier quelques unes des propriétés des micro-habitats des hôtes intermédiaires (température, humidité), s'est avéré le plus explicatif.

Ces modèles requièrent donc l'estimation de certains paramètres concernant le paysage, les distributions spatiales des hôtes intermédiaires (leurs habitats) ainsi que les comportements spatiaux de l'hôte définitif (taille des domaines vitaux). De plus, les comportement "à risque" des hôtes définitifs ont été ignorés (ou supposés constants) ou simulés. L'étude de l'écologie des populations des hôtes intermédiaires et définitifs basée sur des relevés de terrain est appelée à apporter de telles informations pour prédire le risque.

1.4 Nouvelles questions et objectifs

D'une part, les recherches éco-épidémiologiques menées jusqu'ici ont permis d'élucider en partie la "black box" chinoise en ce qui concerne les paramètres écologiques des hôtes : les distributions spatiales des communautés d'hôtes et le rôle du chien domestique dans la contamination de l'environnement humain. Ces paramètres se sont révélés être en partie explicatifs des distributions spatiales des prévalences humaines. D'autre part, les modèles de la transmission d'*Em*

développés jusqu'ici nécessitent l'estimation des paramètres de l'écologie des hôtes (distribution spatiales des hôtes intermédiaires, domaines vitaux des hôtes définitifs,...) pour être réalistes. Ces travaux ouvrent de nouvelles perspectives de recherches mais soulèvent aussi des écueils méthodologiques ainsi que de nouvelles questions à résoudre pour la modélisation spatiale de la transmission d'*Em* en Chine.

1.4.1 Écueils méthodologiques et nouvelles questions

1.4.1.1 Définir des assemblages d'hôtes intermédiaires

Les résultats de captures issues de campagnes de piégeage ont mis en évidence une grande richesse spécifique (comprise entre 6 et 19 espèces) dans chaque aire et une grande diversité d'habitats échantillonnés (comprise entre 6 et 15 habitats) (Giraudoux et al., 1998; Raoul et al., 2006, 2008; Vaniscotte et al., 2009). Certaines espèces dominantes (au regard des densités relatives), susceptibles à *Em* et ayant tendance à une dynamique de population cyclique donnant lieu à de fortes densités ponctuelles ont été considérées comme des hôtes clefs dans le cycle de transmission (par exemple *Ochotona sp* dans le Sichuan). Cependant, une telle identification n'a pas toujours été possible du fait de la diversité du spectre d'hôtes intermédiaires observées dans certaines aires (cas du Ningxia). En effet, parmi les espèces piégées, nombreuses étaient celles susceptibles à *Em* (Table 1) partageant (*O. curzoniae* et *C. kamensis*) ou pas les mêmes habitats. L'éco-épidémiologiste se retrouve alors face à des systèmes complexes de transmission définis dans des habitats discrets et qui contiennent des espèces hôtes alternatives pouvant contribuer chacune à la transmission du parasite ainsi qu'à la stabilité de sa population (ou méta-populations) (Giraudoux et al., 2006). L'intensité de la transmission dépend alors des susceptibilités au parasite des différentes espèces d'hôtes intermédiaires alternatives et de leur propension à présenter des fluctuations cycliques accompagnée de fortes densités de populations.

Suivant un "effet de dilution", la diversité des espèces hôtes accessibles aux parasites peut diminuer l'intensité de la transmission (Keesing et al., 2006). Cela a été observé par exemple dans le cas de la maladie de Lyme, où une corrélation négative a été mise en évidence entre la diversité des espèces hôtes de micro-mammifères terrestres et le taux *per capita* de cas de maladie de Lyme chez l'homme en Amérique du Nord (Ostfeld and Keesing, 2000). Ainsi, la présence d'hôtes alternatifs a tendance à faire diminuer les densités de populations d'autres espèces hôtes (par effet de compétition) davantage susceptibles (Begon, 2008). Pour le macro-parasite *Ribeiroia ondatrae*, touchant des espèces d'amphibiens, l'augmentation des densités et la diminution de la diversité tendent à diminuer les abondances du parasite (Johnson et al., 2008).

Dans le cas d'*Em*, la diversité de proies accessibles pour l'hôte définitif peut lui permettre de diversifier son régime alimentaire (dans le cas de prédateurs non spécialistes) avec des espèces aux charges parasitaires variables. Par opposition, l'existence d'hôtes alternatifs au parasite peut favoriser la transmission comme par exemple les coexistences de *Microtus arvalis* et *Arvicola terrestris* en France ou de *Ochotona sp*, *Microtus irene*,*Microtus leucurus* ou *Cricetulus kamensis* sur le plateau tibétain. Il apparaît alors important d'intégrer la diversité spécifique des hôtes intermédiaires dans l'étude de la transmission de ce parasite. L'écologie des communautés permet l'étude de la richesse spécifique des hôtes et l'analyse des distributions spatiales de

leurs densités relatives ainsi que de leurs interactions.

Concernant la définition des habitats pour ces espèces, plusieurs d'entre elles peuvent partager certains attributs environnementaux et être présentes dans les mêmes habitats échantillonnés. Symétriquement, des habitats peuvent être similaires en termes de composition et de densités relatives en espèces et par conséquent redondants dans l'information paysagère qu'ils fournissent pour expliquer la variabilité des densités relatives des espèces.

Face à cette diversité d'hôtes et d'habitats échantillonnés, la définition d'assemblages d'espèces, c'est-à-dire le regroupement des habitats partageant les mêmes compositions et densités relatives en espèces est apparue indispensable pour l'interprétation écologique mais aussi épidémiologique des résultats des campagnes de piégeage. Dans les travaux précédents, les assemblages d'espèces ont été définis de manière subjective ou par des statistiques uni-variées, c'est-à-dire par l'usage de tests de comparaison des densités relatives entre les différents habitats définis *a priori* et espèce par espèce. La définition d'assemblages d'espèces hôtes par l'analyse paysagère de leurs distributions spatiales a été déjà effectuée en France (Tolle, 2005) et devrait être considérée dans les sites d'études chinois.

1.4.1.2 Modéliser les distributions spatiales des assemblages d'hôtes intermédiaires

L'utilisation d'outils modernes de modélisation spatiale a permis une définition quantitative des habitats ouvrant des perspectives de prédictions précises du risque de transmission, dans le cas de données éco-épidémiologiques à faible taille d'échantillon et dispersées spatialement. Cependant, cette avancée a été réalisée seulement pour une seule espèce de micro-mammifères dont les données de présence/absence d'indices de présence collectées le long de transects (Marston, 2008). La définition des habitats pour l'ensemble des espèces de micro-mammifères issues de campagnes de piégeage, c'est-à-dire pour des assemblages, a été effectuée jusqu'ici par des regroupements d'habitats définis *a priori* sur le terrain (variables qualitatives). Or, une telle définition, bien que nécessaire et inévitable dans le contexte d'études exploratoires de terrain dans des régions inconnues tant au point de vue richesse spécifique que paysagère (et qui plus est dans un contexte éco-épidémiologique), apparaît limitée dans une optique prédictive, puisque non quantitative.

1.4.1.3 L'hétérogénéité spatiale multi-scalaire

Une hétérogénéité spatiale des distributions des espèces a été mise en évidence au sein de chaque site d'étude mais aussi entre les sites. L'intégration de l'ensemble des données collectées sur les différents sites d'études amène à considérer une échelle plus large de modélisation, celle continentale (Giraudoux et al., 2003). Modéliser la transmission sur l'ensemble du continent au delà des sites d'études reste un problème ouvert. En effet, le ROMPA pour les populations de *Microtus* estimé dans l'est de la France n'a pu être utilisé pour expliquer les prévalences vulpines dans un site de l'Allemagne du sud (Pleydell, non publié). Les effets des facteurs environnementaux et les habitats définis sont spécifiques des systèmes de transmission et des aires sur lesquelles ils ont été définis. De ce fait, le ROMPA peut difficilement être considéré comme une variable globale incorporable dans les modèles de transmission sans précautions.

1.4.1.4 Le rôle du chien domestique dans la transmission

Le rôle du chien a été mis en évidence par une approche épidémiologique classique consistant à estimer des prévalences de certains échantillons puis étudier des facteurs de risque de transmission à l'homme. Bien que ces recherches apportent les preuves de l'existence d'un cycle péri-domestique où le chien joue le rôle de l'hôte définitif, l'intensité de la transmission ainsi que ses modalités, telles que les comportements "à risque" des chiens, restent méconnues et sont d'une importance particulière pour modéliser la rencontre et aider à la décision des contrôles d'*Em* dans les populations.

1.4.2 Objectifs détaillés de la thèse

Dans ce travail de thèse, je me suis intéressé à modéliser les distributions spatiales des hôtes intermédiaires dans deux nouveaux sites d'étude situés dans l'Ouest du Sichuan, espacés d'une centaine de kilomètres. Au regard des connaissances et écueils méthodologiques des précédents travaux, ce travail a répondu à plusieurs objectifs :

1. définir localement des assemblages d'hôtes intermédiaires de manière quantitative à partir de données de piégeage réalisée dans des habitats définis *a priori* sur le terrain
2. quantifier localement les interactions des assemblages avec leurs environnements, c.a.d modéliser leurs habitats dans l'espace multivarié de variables environnementales extraites de données satellitales
3. investir les possibilités d'extrapolation des distributions des assemblages sur une étendue régionale.

Pour ce faire, nous avons eu recours à des outils de modélisation de distributions spatiales des espèces hôtes.

Puis, le rôle du chien dans la transmission d'*Em* a été exploré, à une échelle locale, dans des villages situés sur le plateau tibétain (province du Sichuan). Nous avons cherché à :

1. caractériser et quantifier leurs comportements "à risque" tels que leurs comportements de défécation et leurs utilisations de l'environnement des hommes et des micro-mammifères
2. évaluer le rôle relatif du chien dans la contamination de l'environnement humain par rapport à celui du renard, hôte définitif du cycle sauvage.

Des outils d'écologie comportementale et de biologie moléculaire ont été utilisés pour répondre à ces objectifs.

Chapitre 2

Méthodes pour la modélisation et l'écologie comportementale des hôtes

2.1 Modélisation des distributions spatiales des hôtes intermédiaires

Les données d'occurrences des espèces animales constituent un ensemble de points (semis) qui peut être analysé dans deux espaces, l'espace environnemental et l'espace géographique, ainsi qu'à différents niveaux d'organisation biologique, l'individu et la population (Callenge 2005).

Dans ce contexte méthodologique, cette thèse comporte deux entrées pour l'étude des distributions spatiales des hôtes pour *Em* :

- un niveau populationnel sur des aires géographiques larges (n x 10 km^2), pour traiter des données concernant les distributions des assemblages de micro-mammifères. Á cette échelle, nous avons cherché à identifier l'influence de variables environnementales sur les distribution des populations.

- un niveau individuel : localement (village) concernant les comportements spatiaux des chiens, traduisant leurs interactions avec les hommes et les populations de micro-mammifères. Á ce grain plus fin de l'analyse, nous nous sommes concentrés sur l'analyse de la distribution des points dans l'espace géographique.

Ce chapitre constitue une revue des méthodes utilisables pour une analyse populationnelle des distributions. Les méthodologies pour les distributions des individus seront abordées au chapitre suivant. Suite à une présentation des données et de leurs limites, y seront exposés les principaux concepts et méthodologies utilisables pour les analyser.

2.1.1 Les données

2.1.1.1 Sur les espèces

Des atlas ont été réalisés sur la base de données de présence et nous informent sur la diversité taxonomique des micro-mammifères et leurs distributions à l'échelle continentale (Zhang et al., 1997; Panteleyev, 1998; Smith and Xie, 2008). Cependant, le géoréférencement de ces données n'était pas accessible au moment de l'étude et les résolutions employées étaient trop grossières.

Les données issues de campagnes de piégeage, utilisées dans cette étude nous informent à la fois sur la présence et l'absence des espèces. En effet, la probabilité de capturer une espèce supplémentaire après trois nuits de piégeage étaient nulles dans le comté de Serxu ou dans le Gansu, excepté pour quelques rares habitats (Raoul et al., 2006; Giraudoux et al., 1998). Cet effort de piégeage correspond donc à un minimum requis pour comparer les densités relatives des espèces entre les habitats. Toute proportion gardée, cette double information demeure assez rare dans le contexte de l'étude des distributions des espèces souvent uniquement basée sur les présences (données atlas ou collection *ad hoc* d'observations) et donc précieuse. Elle permet en effet de discriminer les habitats favorables mais aussi ceux défavorables à la présence des espèces. D'autre part, elle permet de considérer l'ensemble de la diversité présente en un site donné et donc des espèces pouvant y interagir. Cependant, ces données ont été collectées dans des zones géographiques inconnues *a priori* lors de suivis écologiques exploratoires dans des objectifs écologiques, tels que l'identification des espèces (taxonomie), l'inventaire de la diversité et l'étude des habitats, mais aussi épidémiologiques (estimation des prévalences). Leur qualité est donc contrainte par la durée restreinte des campagnes de terrain (environ trois semaines), ainsi que leur coût et leur logistique (déplacements limités, certaines zones inaccessibles).

Récoltées sur des périodes de temps courtes, de telles données n'intègrent donc pas les dynamiques des populations. Cela limitera l'extrapolation des distributions des espèces dans le temps. Aussi, l'absence de l'espèce peut être le résultat de sa dynamique, donnant lieu à de fausses absences. De plus, les lignes de pièges ont été agrégées dans l'espace localement puisque la pose des lignes a été regroupée pour des raisons logistiques dans quelques zones d'échantillonnage à l'intérieur des sites d'étude. Enfin, l'effort d'échantillonnage peut être considéré de manière générale comme faible pour chaque habitat sur les deux sites de cette étude.

2.1.1.2 Sur l'environnement

Sur nos deux terrains d'étude nous ne disposons pas de cartes d'occupation du sol pour rendre compte de l'hétérogénéité du paysage. Des variables issues de données satellitales seront alors utilisées pour décrire l'environnement des lignes de pièges ou les caractéristiques des micro-habitats pour les espèces de micro-mammifères. Ainsi des bandes spectrales et des modèles numériques de terrain ont été traités pour extraire des variables topographiques et des indices de végétation.

2.1.2 Modéliser les distributions spatiales des assemblages

2.1.2.1 Niche écologique et habitat : éléments de définition

Définition

Niche et habitat sont des concepts de base en écologie utilisés pour désigner la distribution d'une espèce ou d'un groupe d'espèce dans deux espaces multivariés principaux : l'espace environnemental, composé des variables biotiques et abiotiques d'un écosystème et l'espace géographique. Leurs multiples utilisations en écologie ont donné lieu à diverses définitions, chacune relevant d'une manière d'appréhender l'analyse des distributions, et qui peuvent selon les écoles s'opposer (Calenge, 2005).

Souvent utilisé sur le terrain pour désigner une classe paysagère, tel que type de couvert végétal, l'habitat dans sa dimension multivariée peut être utilisé simplement en référence à l'espace géographique occupé par les espèces ou "the place where an organism lives, or the place where one would go to find it" tel que défini par Odum (1971) et peut même être défini sans référence particulière à une espèce (Kearney, 2006). Cette définition a été développée par certains pour intégrer les variables environnementales traduisant les ressources et les conditions (incluant survie et reproduction) qui produisent l'occupation de l'habitat par une espèce donnée (Hall et al., 1997).

La définition de l'habitat, et les débats qui l'entourent, s'arrête où celle de la niche écologique commence, qui relève uniquement de l'espace environnemental. Définie par Grinnel (1917) comme "l'ensemble ou l'étendue des caractéristiques environnementales qui permet aux individus d'une espèce de survivre et de se reproduire", Elton (1927) y voit plutôt l'individu en tant qu'acteur sur son écosystème (sa place) sans considérer les facteurs limitant sa distribution.

L'arrivée de la définition de Hutchinson (1957) permet de formaliser mathématiquement la niche comme un hypervolume dans l'espace des variables environnementales. Cette approche géométrique s'est largement répandue dans le domaine de la modélisation. Ceci d'autant plus que même si s'inscrivant dans la pensée de Grinnel, la distinction qu'il fait de la niche fondamentale, celle excluant les interactions biotiques, de la niche réalisée qui les incluent et qui correspond à l'espace utilisé effectivement, permet de distinguer l'étude des interactions inter et intra-spécifiques de l'espace davantage géographique.

Calenge (2005) souligne que les approches grinellienne et eltonienne diffèrent en fait davantage d'un *modus operandi*. La collecte des données sur le terrain ne peut conduire qu'à la définition de la niche réalisée, résultat de l'effet des interactions biotiques et de l'exclusion compétitive sur la niche fondamentale, la niche fondamentale recourant davantage à la modélisation des contraintes physiologiques (Guisan and Zimmerman, 2000; Calenge, 2005; Soberon and Peterson, 2005). De plus, la définition de la niche écologique peut inclure la compétition, des interactions positives (mutualisme) mais aussi la dispersion, et des ambiguïtés persistent sur la place de ces facteurs dans les niches réalisées et fondamentales (Pulliam, 2000). Ainsi, ce terme sera toujours discutable si utilisé avant la résolution de tels débats.

Concernant la modélisation des distributions spatiales des espèces, il s'agit alors de bien distinguer les modèle de niche de ceux des variables spatialisées (Araujo and Guisan, 2006). Quoi qu'il en soit de la définition théorique de la niche ou de l'habitat, toute caractérisation de ces concepts par un jeu donné de variables environnementales est une description incomplète des facteurs biotiques et abiotiques permettant à l'espèce de satisfaire ses besoins écologiques minimum, la survie et la reproduction (Soberon and Peterson, 2005). Afin d'éviter toute ambiguïté avec la définition de la niche, nous utiliserons le terme d'habitat potentiel en ce qui concerne les prédictions des modèles développés dans nos travaux (Guisan and Zimmerman, 2000; Kearney, 2006). Nous nous attacherons alors à définir "la fonction donnant la densité de probabilité de présence de l'espèce pour une combinaison donnée des variables d'habitat" (Calenge, 2005).

Paramètre de la niche, habitat optimaux et potentiels

Le long des gradients environnementaux, les espèces ne sont pas distribuées aléatoirement. Elles opèrent au contraire une sélection dans les conditions environnementales de la région dans

laquelle elles se trouvent.

Ainsi, depuis la définition de la niche d'Hutchinson, l'existence d'un optimum des distributions des abondances est admise et dans sa transcription qualitative définit l'habitat optimal pour l'espèce. La niche est alors caractérisée par une distribution Gaussienne. Elle se caractérise par sa dimension ("niche breadth") correspondant à la portée maximale de sa distribution le long des gradients environnementaux, c'est-à-dire à l'étendue de la diversité des conditions tolérées par les organismes (Ricklefs and Miller, 1999). On distingue alors sa marginalité, écart de la moyenne de l'espèce à celle globale du site, et sa spécialité, écart entre la variance pour l'espèce et celle globale des variables environnementales du site (Hirzel et al., 2002).

Or, en réalité la niche est rarement une ellipse parfaite dans l'espace des variables environnementales. Une multitude de distributions non Gaussiennes ont été observées chez les plantes (Austin, 1999). Une espèce compétitrice peut, par exemple, déplacer l'optimum de la distribution d'une autre espèce. Du fait de la possibilité de recouvrements des niches, l'optimum de la niche réalisée peut différer de celui de la niche fondamentale, et peut prendre des formes multiples et multi-modales. Peu d'exemples concernant les distributions animales existent, mais il est difficile d'imaginer que cela diffère des distributions végétales (Austin, 1999).

De surcroît, lorsque l'on s'intéresse à modéliser les communautés, la distribution des habitats est d'autant plus non-linéaire qu'elle constitue une mixture des distributions (souvent non Gaussienne) de chaque espèce la constituant (Doledec et al., 2000).

2.1.2.2 Modéliser les distributions spatiales des espèces

Définition et objectifs

Les modèles de distribution spatiale des espèces sont développés dans l'objectif général de quantifier les relations des espèces avec leur environnement. Un modèle n'est qu'une représentation partielle d'une nature complexe et hétérogène, et de fait nécessairement faux. Les modèles peuvent être classifiés en fonction principalement de leurs pouvoir de précision, de généralisation et de réalisme (Levins, 1966). Il apparaît important de connaître les possibilités et les limites du modèle utilisé dans ces champs et de trouver le meilleur compromis pour répondre à l'objectif de l'étude (Guisan and Zimmerman, 2000).

Les modèles auxquels nous nous sommes intéressés ne sont pas ceux qui décrivent des processus de la manière la plus réaliste et précise (ex : physiologique), dits mécanistiques, ni ceux basés sur des relations mathématiques théoriques (équations de Lotka-volterra), dits analytiques, mais sont plutôt qualifiés d'empiriques ou de statistiques. En ce sens, ils permettent d'établir une relation statistique entre la réponse (l'occurrence des espèces) et les prédicteurs (les variables environnementales) et produisent des distributions de probabilités d'occurrence des espèces en fonction des valeurs des prédicteurs.

Ces modèles peuvent être utilisés pour i) expliquer des processus écologiques donnant lieu aux patrons de distribution observés (modèles explicatifs), et/ou ii) prédire la probabilité d'occurrence des espèces (modèles prédictifs) (Guisan and Zimmerman, 2000; Calenge, 2005).

Les modèles exploratoires permettent de formuler des hypothèses sur un jeu de données récolté sur avis d'experts. Dans le contexte des distributions des espèces, ils peuvent servir à examiner l'adéquation du modèle à la réalité (le "fit"), la consistance de cette relation,

et d'estimer les contributions et les rôles des différentes variables. Ils contribuent aussi à la construction de modèles prédictifs en testant les hypothèses et en estimant les paramètres en jeu (le poids de chaque variable explicative). Ils devront donc plutôt rester réalistes et précis.

Les modèles prédictifs permettent d'atteindre l'objectif ultime et le plus appliqué actuellement des modèles statistiques : prédire sur des points qui n'ont pas été échantillonnés. Les capacités de précision et de généralisation des prédictions sont requises dans cette étape. Dans un contexte de prédictions sur des échelles régionales, l'analyse peut être décomposée en deux questions distinctes : prédire dans la zone d'étude où l'on a échantillonné et extrapoler les prédictions sur de nouvelles zones. Nous verrons plus loin les multiples choix techniques auxquels doit faire face le modélisateur avant d'obtenir un modèle prédictif approprié.

Calenge (2005) souligne aussi l'importance de l'indépendance des données utilisées pour développer de tels modèles et distingue alors les études exploratoires, développant des modèles explicatifs, des études confirmatoires réalisées dans le but de prédire.

Philosophie générale de la construction

De manière générale, les modèles sont construits par un aller-retour entre la formulation statistique et les données collectées qui commence, idéalement, par la formulation avant la collecte des données (Turchin, 1998). La première étape empirique permet d'estimer les paramètres. Puis une seconde étape constitue le test rigoureux des prédictions là où les données sont indisponibles, en comparant les valeurs prédites par le modèle aux données de validation récoltées sur le terrain. Étant donné que les modèles sont toujours faux, la phase de validation ne peut se contenter de rejeter un modèle (test d'hypothèse) mais cherchera plutôt une meilleure alternative parmi différents modèles testés. Il s'agira alors d'estimer les déviations des prédictions à la réalité et de continuer la procédure tant que des alternatives sont possibles et que l'erreur minimale que l'on s'est fixée n'est pas atteinte.

2.1.2.3 Un domaine de recherche interdisciplinaire en pleine effervescence

Motivations

Cette démarche trouve de nombreuses applications dans des domaines qui connaissent un intérêt grandissant dans des écosystèmes changeants et perturbés tels que la biologie de conservation (invasions, sélection de réserves, espèces rares), la biologie de l'évolution (Kozak et al. 2008), la gestion paysagère et l'aménagement (évaluation des effets de perturbation, fragmentation de l'habitat) et l' éco-epidemiologie (voir Guisan and Thuiller (2005) pour une revue des applications). Il apparaît de plus en plus important d'obtenir des explications et des prédictions fiables et directement utilisables par les gestionnaires. Ainsi, les modèles prédictifs sont davantage développés.

L'activité de recherche utilisant des modèles de distribution des espèces est aussi variée que les termes pour la désigner sont multiples : on rencontrera dans la littérature les termes de "ecological niche models", "bioclimatic envelopes", "habitat models" et "resource selection functions". Tous ont pour perspective d'expliquer et/ou de prédire les occurrences des espèces étudiées en fonction des attributs environnementaux et géographiques.

Á l'origine de l'effervescence : une interdisciplinarité

Plusieurs étapes conceptuelles se sont succédées pour aboutir aux méthodes actuelles : des

modèles spatialement explicites, statistiques et empiriques (Guisan and Thuiller, 2005). Les formulations statistiques se sont développées rapidement avec la modélisation numérique au milieu des années 70. Les expériences de simulations, les ouvrages et les conférences ont donné lieu dans les années 80 à une multitude de modèles. Enfin, le développement dans le siècle dernier de méthodes statistiques dites "modernes" (résumée dans Hastie et al. (2001)) a ouvert un nouveau champ des possibles pour comprendre les patrons de distributions observés. Le nombre important d'articles de synthèse, théoriques ou concernant des études comparatives de méthodes, souligne des besoins de synthèse des connaissances et de clarification des concepts (niche, habitat) et des méthodes dans ce champs disciplinaire en pleine expansion.

Ce développement est la conséquence d'une forte collaboration entre la biologie des populations, des communautés, l'écophysiologie, les observations de terrain, l'écologie théorique et les statistiques. Une telle interdisciplinarité ne s'est pas accomplie sans confusions. D'une part, les méthodes sont parfois mal utilisées faute de "background" solide en statistiques et de l'attitude "presse-bouton" facilitée par les logiciels, essentiellement commerciaux. D'autre part les théories et les processus bien connus de l'écologie sont parfois oubliés des formulations statistiques et éloignent parfois le modélisateur des objectifs appliqués de l'analyse (Austin, 2002).

Ainsi, les travaux fructueux, innovants et aux nombreuses applications alliant des modèles et des méthodes statistiques (validation et sélection) et une forte connaissance du terrain et des processus biologiques émanent souvent de réelles collaborations entre chercheurs des disciplines en jeu (Elith et al., 2008) ou des sciences intégratives telles que la biométrie ou les biostatistiques.

2.1.3 Une large diversité d'outils de modélisation

L'effet retour d'un tel engouement scientifique est le développement d'une multitude d'outils et de problèmes méthodologiques auxquels l'écologiste doit se heurter et choisir avant d'atteindre son objectif.

D'une manière générale, les modèles augmentent en complexité depuis la délimitation des limites d'une enveloppe dans l'espace environnemental au début des années quatre-vingt dix, jusqu'à l'ajustement de relations statistiques non-linéaires entre l'occurrence des espèces et les variables environnementales. Encore plus récentes sont les méthodes basées sur l'intelligence artificielle et sur les statistiques bayésiennes.

Quelques grandes familles de méthodes statistiques peuvent être répertoriées :

- la régression
- l'ordination
- la classification
- les enveloppes climatiques
- l'intelligence artificielle
- les méthodes bayésiennes

Bien que partageant des suppositions quant à la théorie de la niche écologique (équilibre avec l'environnement et niche réalisée), la multiplicité des méthodes développées s'explique par la diversité concernant principalement :

- l'objectif de la modélisation qui peut être descriptif, explicatif ou prédictif,
- le type de la réponse pouvant être quantitative (densités) ou qualitative (présence/absence

ou présence seulement),
- la distribution ("shape") de cette réponse le long des gradients environnementaux, c'est-à-dire Gaussienne (et uni-modale), mixture de Gaussienne (et multi-modale), libre (non-paramétrisée). La combinaison des variables explicatives peut être définie linéaire ou non linéaire.

Nous nous concentrerons sur la description et la discussion des méthodes utilisées dans le cadre de la modélisation des assemblages d'espèces.

2.1.3.1 De la distribution des espèces à celle des assemblages : les stratégies de modélisation

Lorsque les données d'occurrence des espèces sont peu nombreuses et que beaucoup d'espèces sont rares il devient difficile de modéliser chaque espèce individuellement. Ainsi, modéliser les habitats partagés par un ensemble d'espèces se présente comme une alternative. Plusieurs outils ont été développés pour définir des groupes d'espèces et modéliser leurs réponses.

Les stratégies

Trois stratégies principales de modélisation existent selon Ferrier et Guisan (2006) :

1. "Classifier puis modéliser et prédire", qui consiste à définir des groupes d'espèces partageant les mêmes habitats (ou des groupes de points d'échantillonnage partageant les mêmes espèces), puis à modéliser leurs habitats de manière individuelle (pour chaque assemblage) ou multiple (pour tous les assemblages simultanément).
2. "Modéliser puis classifier" qui consiste à modéliser chaque espèce puis à regrouper ensuite leurs réponses.
3. "Modéliser et classifier simultanément" où les étapes sont réalisées simultanément.

Étant donné les limitations de la stratégie 2 dans le cas de faible échantillon (Ferrier and Guisan, 2006; Olden, 2003), nous nous concentrerons sur la description des stratégies 1 et 3 uniquement.

Le choix de la stratégie dépendra de l'objectif de l'étude, du type (présence/absence), de la quantité (taille échantillonnage) et de la qualité (échantillonnage) des données. Les deux stratégies (1 et 3) permettent de prendre en compte la distribution des espèces rares ainsi que les interactions entre espèces. Une contrainte majeure de la stratégie 1 est qu'elle suppose que toutes les espèces présentes dans le site d'étude aient été considérées. De ce fait, elle considère les groupes d'espèces comme des entités fixes (Olden, 2003). Les prédictions sur des points non échantillonnés seront donc la présence ou l'absence de ces classes et la prédiction de combinaisons différentes d'espèces n'est alors pas possible. Cette approche demeure cependant appropriée dans le cas de la classification des assemblages. La stratégie 3 modélisant les réponses des espèces de manières simultanée permet d'obtenir les réponses individuelles pour chaque espèce. Ceci est impossible par la stratégie 1 car elle permet de modéliser la distribution de groupes d'espèces uniquement. De surcroît, la stratégie 1 permet d'extrapoler les prédictions au delà des groupes définis sur la zone échantillonnée, ce qui peut être très utile pour prédire dans des contextes environnementaux différents où la composition *a priori* en espèces est inconnue.

La modélisation statistique

Pour la stratégie 1, la classification des réponses des espèces est souvent réalisée par des méthodes d'ordination (Ferrier et al., 2002b; Ferrier and Guisan, 2006). La modélisation des assemblages prédéfinis peut être réalisée de manière individuelle pour chaque assemblage par des méthodes de modélisation développées pour des réponses binomiales et pour construire des cartes de probabilités de présence/absence de l'assemblage. Les assemblages peuvent être également modélisés simultanément par des méthodes considérant des réponses multinomiales et dans ce cas donner lieu à des cartes de classification des assemblages dans la zone étudiée.

Pour la stratégie 3, les méthodes de modélisation statistique utilisées sont celles qui permettent de modéliser des réponses multinomiales des espèces (de manière jointive). Á l'issue de tels modèles, les cartes de prédiction pour chaque espèce sont obtenues et peuvent être combinées dans l'optique par exemple de classifier la distribution des assemblages.

On distinguera parmi les méthodes de modélisation existantes, celles appropriées pour modéliser les réponses individuelles des espèces ou groupes d'espèces (Table 2 p. 50, "Ind"), de celles qui permettent leur modélisation simultanée et de manière jointive (Table 2 p. 50, "Mult").

2.1.3.2 Les modèles paramétriques *versus* non-paramétriques

De manière générale et en accord avec l'évolution historique des méthodes, une importante distinction est faite entre les méthodes paramétriques et non paramétriques, et linéaires ou non linéaires. La Table 2 (p. 50) rend compte des méthodes existantes utilisables pour la modélisation des assemblages, dans chacune de ces catégories.

Les modèles statistiques paramétriques sont basés sur des théories et hypothèses. Ils supposent une quantité de connaissances *a priori* sur les paramètres (la distribution statistique des réponses). Leur objectif est d'obtenir des modèles facilement interprétables, excluant par exemple les prédicteurs corrélés.

Les premiers modèles développés sont basés sur la définition de la niche d'Hutchinson (1957), normale ou tout au moins uni-modale (un seul optimum). Les méthodes d'ordination et d'analyses factorielles, permettent de résumer les données multivariées sur des axes principaux et d'estimer les distances à la niche (distance de Malahanobis) pour obtenir des cartes de disponibilité des habitats, chaque pixel étant attribué à un indice de conformité ("suitability") à la niche.

L'analyse canonique des correspondances (CCA) ainsi que l'Analyse Discriminante Linéaire (LDA) ont été développées pour discriminer des espèces ou des groupes d'espèces dans l'espace environnemental. Bien que ces méthodes relèvent de l'ordination des données écologiques et de la classification respectivement, leurs formes algébriques sont équivalentes (Zhu et al., 2005). Le développement d'analyse discriminante de Mixture de Gaussiennes (MDA) permet de modéliser des niches multi-modales (Hastie and Tibshirani, 1993). Récemment, une méthode d'ordination d'espèces multiples a été développée pour donner davantage de poids aux espèces rares et pour s'abstenir de la distribution Gaussienne et densité-dépendante de la niche : l'"Outline Mean Index" (OMI) qui consiste à séparer les points de la niche en fonction de leurs position sur l'axe de marginalité, c'est-à-dire leur déviation par rapport à l'espace écologique de la zone. Il ne dépend donc ni de l'abondance ni de la moyenne de l'espèce (Doledec et al., 2000).

L"'Ecological Niche Factor Analysis" (ENFA) a été développé pour modéliser les présences d'espèces individuellement (Hirzel et al., 2002). Elle permet alors de mettre en évidence des axes factoriels principaux capables de maximiser la marginalisation (écart de la moyenne de la niche à la moyenne du site) et la spécialisation (ratio des écarts types de la niche sur la moyenne du site) de la niche dans l'espace environnemental de l'ensemble de la zone étudiée ; la niche étant définie telle une Gaussienne. L'utilisation de distances moyennes géométriques permet de s'abstenir de présupposés sur la distribution des points dans l'espace environnemental (Hirzel and Arlettaz, 2003).

Enfin, les modèles de régression linéaire généraux sont très largement utilisés pour modéliser les données de présence/absence et produire directement des probabilités d'occurrences. L'avantage de ces modèles réside dans le fait qu'ils peuvent être utilisés pour modéliser une grande diversité de distributions. Ils permettent de discriminer les présence/absence d'une espèce (distribution binomiale) comme les présences de plusieurs espèces (distribution multinomiale) (MacCullag and Nelder, 1989).

Par opposition, les modèles non-paramétriques ne requièrent pas l'estimation de paramètres tels que ceux nécessaire à la formalisation de la distribution de la réponse et des prédicteurs. Le point commun entre les modèles de distribution au niveau des communautés et les méthodes statistiques modernes non-paramétriques est qu'ils ont la particularité de pouvoir s'adapter à des données éparses ("sparce") ainsi qu'à des distributions complexes (Elith et al., 2006). De tels modèles peuvent alors aider à la modélisation des distributions des habitats qui peuvent être complexes lorsque l'on considère l'ensemble des espèces qu'ils contiennent (Olden et al., 2006; Leathwick et al., 2006; Elith et al., 2006).

L'analyse va identifier les paramètres en s'ajustant au mieux aux données et en ce sens elle est qualifiée de "data-mining" (Hastie et al., 2001). Elle permet donc de détecter et de décrire les patrons des données sans idées préconçues sur ce qu'ils devraient être, c'est-à-dire sans *a priori* quant à la forme de la relation entre les prédicteurs et la réponse (Hachacka et al., 2007). Ces analyses peuvent être caractérisées par le fait qu'elles produisent automatiquement des prédictions précises à partir des données, avec la capacité de "scanner" un large nombre de prédicteurs et d'identifier les plus importants. L'objectif est d'inclure toutes les variables qui peuvent maximiser les performances prédictives du modèle (être informatif). Ces méthodes peuvent être qualifiées de "Machine learning" lorsqu'elles correspondent à des algorithmes capables d'apprendre de leurs expériences pour améliorer leurs performances (Mjolsness and DeCoste, 2001). Elles ont été développées initialement pour de larges jeux de données avec un nombre important de prédicteurs et essentiellement dans le contexte de la classification. Ceci dit, il a été montré qu'elles pouvaient être performantes dans le cas de faibles tailles d'échantillons (Elith et al., 2006). Elles s'avèrent utiles si les connaissances *a priori* sur les patrons sont minimales et que les hypothèses ne sont pas clairement développées et constituent en quelque sorte un outil de prédilection pour les analyses exploratoires. Leur inconvénient majeur tient au fait que les modèles développés sont difficilement compréhensibles et les relations statistiques ajustées difficilement interprétables. Cependant, dans le domaine des distributions des espèces, quelques études se sont intéressées à améliorer cet aspect (Elith et al., 2005).

Ainsi l'on peut considérer que les modèles des enveloppes (basés sur les présences seulement)

qui définissent une région dans l'espace multivarié incluant les conditions favorables pour l'espèce par des coefficients de similarités (DOMAIN), sont les formes les plus simples de modèles non paramétriques (Barry and Elith, 2006). En effet, l'utilisation du point le plus proche de celui prédit pour estimer l'indice de similarité s'apparente à un lissage et à un procédé non paramétrique.

Les modèles de régressions peuvent être aussi non paramétriques. Les variables explicatives sont alors remplacées par des fonctions qui leurs sont appliquées. Ainsi les modèles généraux additifs (GAM) sont une modification des GLM par l'ajout de fonction spline (de lissage) (Hastie and Tibshirani, 1990). Aussi les modèles de régressions multiples adaptatives (MARS) transforment chaque variable explicative par une série de fonction de bases (Friedman, 1991). Contrairement aux fonctions spline des GAMs, les fonctions de bases sont plus facilement extrapolables sur des données indépendantes du jeu d'entraînement et estimables pour des réponses multiples. Elles permettent ainsi la modélisation de plusieurs espèces simultanément.

Dans le domaine de la classification, l'analyse discriminante peut être considérée comme équivalente en partie à une régression multiple linéaire réalisée sur les scores optimaux obtenus dans l'espace environnemental réduit (Hastie et al., 1994). Ainsi, l'incorporation de régressions non paramétriques a donné naissance à des analyses discriminantes dites flexibles (FDA) (Hastie et al., 1994). Une telle incorporation est également possible dans une analyse de correspondance (Zhu et al., 2005). Les arbres de classification (CART) permettent également de diviser l'espace des variables environnementales en régions présentant des réponses similaires par l'estimation de noeuds (Hastie and Tibshirani, 1993). Les "Support vector machines" (SVM, Vapnick 1995) ont été appliqués pour discriminer des présence/absence d'espèces de manière individuelles (Guo et al., 2005).

Des méthodes non basées sur la régression mais plutôt sur l'intelligence artificielle ont été développées avec succès comme par exemple les réseaux de neurones artificiels (Artficial Neural Network) pour modéliser les réponses jointes des espèces (Olden et al., 2006). Les algorithmes génétiques, basé sur un jeu de lois ont également été utilisés pour modéliser les distributions spatiales d'espèces individuelles (GARP, Stockwell and Peters, 1999).

Enfin, on notera que l'estimation de prédictions des habitats à partir du théorème de Bayes est réalisable et a été utilisée plus rarement pour classifier des communautés et établir leurs distributions (Ter Braak et al., 2003; Termansen et al., 2006) ou pour développer des modèles de régressions hiérarchiques (Gelfand et al., 2006). De telles méthodes peuvent alors être incorporées dans des modèles paramétriques ou non paramétriques.

De nombreuses études de comparaisons de méthodes ont mis en évidence des capacités de prédictions des modèles non-paramétriques tels que les GAMs, les réseaux artificiels de neurones et MARS, plus élevées que les modèles linéaires (GLM) (Moisen and Frescino, 2002). Aussi, les arbres de régression, MAXENT (maximum entropy) et MARS étaient plus performants que les autres techniques de régression ainsi que celles développées sur les présences seulement (Elith et al., 2006).

Malgré cette dichotomie, il est important de préciser que les deux approches, paramétrique et non paramétrique, sont connectées par des outils communs tels que les méthodes de ré-échantillonage (bootstrap) ou les techniques de régression que nous aborderons plus loin.

2.1.4 Construire un modèle prédictif

2.1.4.1 Les principales étapes

Comme nous l'avons brièvement décrit plus haut et tel que décrit dans Guisan and Zimmerman (2000), construire un modèle prédictif est un développement en plusieurs étapes (Figure 2.1).

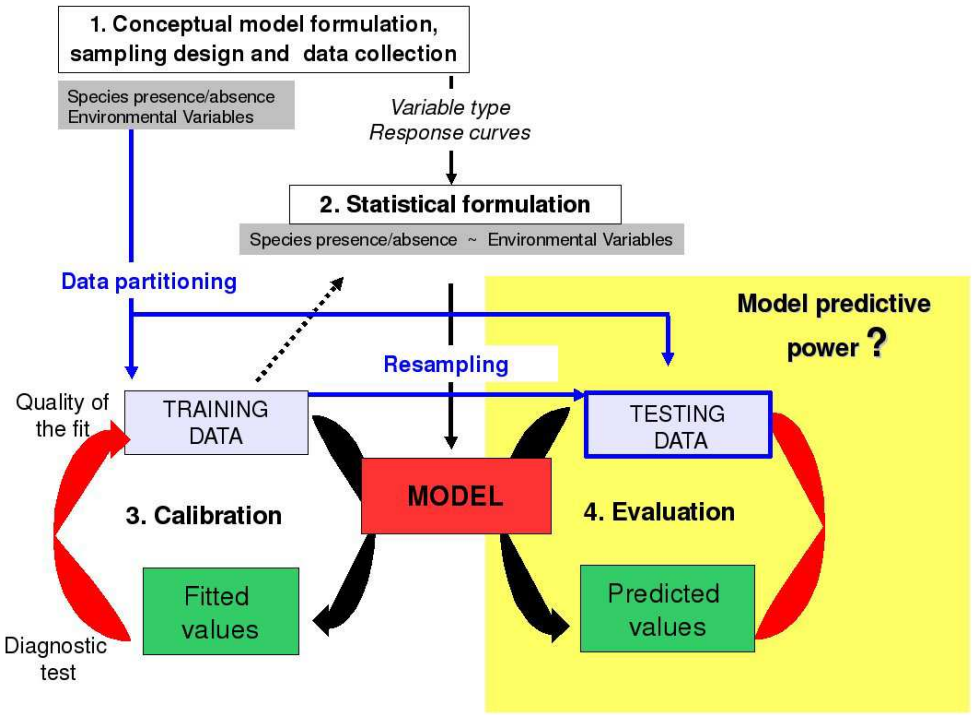

FIG. 2.1 – Schéma conceptuel résumant les différentes étapes de la construction d'un modèle prédictif, inspiré de Guisan and Zimmerman (2000)

1. La formulation conceptuelle du modèle

   Avant toute réalisation du modèle et idéalement la collecte des données, il apparaît important de clairement identifier la variable que l'on veut expliquer (diversité, abondance, présence/absence) et l'objectif de la modélisation. Puis, pour guider la stratégie d'échantillonnage, de réunir les connaissances sur les variables environnementales accessibles, celles qui manquent, et leur pertinence pour comprendre la distribution de la réponse mesurée à l'échelle considérée (l'occurrence des espèces). Il s'agit de définir d'abord l'échelle à laquelle on veut réaliser des prédictions et de fait celle à laquelle on va échantillonner.

2. La formulation statistique

   Cela consiste à choisir : i) la relation statistique établie entre la réponse et les prédicteurs (distribution de la réponse et structure du modèle) et de ce fait ii) la méthode de modélisation statistique la plus appropriée à cette dernière, aux concepts et à l'objectif de la modélisation.

   Á l'issue de ces questionnements, certains peuvent rester sans réponses. Cela peut concerner le choix de la méthode de modélisation. Il est alors prudent de considérer un jeu de

modèles candidats ou de techniques de modélisation, correspondant chacune à plusieurs hypothèses de travail (Johnson, 2004). Après les avoir développées leurs performances prédictives pourront être comparées. Ces incertitudes peuvent aussi concerner le choix des variables à considérer dans le modèle.

3. La calibration

   Elle consiste à mesurer l'ajustement des paramètres du modèle et des valeurs prédites aux données observées, c'est-à-dire la déviance du modèle dans le cas de la régression (R^2 ou D^2) ou d'autres critères adaptés à la méthode de modélisation (Guisan and Zimmerman, 2000).

   C'est aussi au cours de cette étape que la sélection des variables les plus explicatives ("model selection") doit être réalisée (Guisan and Zimmerman, 2000; Hastie et al., 2001).

   La sélection du modèle s'approchant du meilleur consiste à choisir le modèle le mieux supporté par les données, dont la calibration est maximale. La sélection de modèle, comme sa validation de manière générale, peut être réalisé en fonction de plusieurs mesures (Burnham and Anderson, 1998) :

   – des tests d'hypothèses effectués à la suite d'analyse de la variance des prédictions (Likelihood ratio test,stepwise selection)
   – des méthodes analytiques et l'optimisation de critères des modèles basés sur l'estimation de leur déviance et du maximum de vraisemblance ("likelihood") (AIC, BIC)
   – des méthodes dites *ad hoc* puisqu'elles consistent à ré-échantillonner le jeu de données original et à estimer la variance des erreurs de prédiction (Validation croisée, Bootstrap).

   L'utilisation des tests d'hypothèses, assumant l'existence d'un modèle nul et donc vide d'information est considéré aujourd'hui désuète par rapport à la comparaison d'un jeu de modèles candidats par les critères tels que l'AIC qui pénalise les performances du modèle par sa complexité (nombre de paramètres), en accord avec le principe de parcimonie (Burnham and Anderson, 1998; Johnson, 2004). En revanche l'utilisation des méthodes de ré-échantillonage dans le contexte de la sélection de modèles, bien que plus coûteuse en terme de temps de calcul (en fonction du nombre d'itérations), se présente comme une solution alternative à la théorie de l'information (Burnham and Anderson, 1998).

4. La prédiction et l'évaluation des prédictions

   Une fois le modèle validé sur les observations, les prédictions sur de nouvelles observations peuvent être réalisées. Cette étape permet d'estimer la capacité de généralisation du modèle en dehors de ces points d'entraînement. Idéalement, elle s'effectue sur un jeu de données indépendant (erreur externe). En pratique, on a souvent recours à l'évaluation interne qui consiste à ré-échantillonner des données à partir du jeu initial.

2.1.4.2 L'erreur de prédiction : une définition inter-disciplinaire

Dans mon travail, je me suis particulièrement intéressée à l'étude de la sélection des techniques de modélisation. Ceci a été réalisé en estimant **l'erreur de prédiction** interne et externe des modèles développés d'après la définition qui en est donnée par les statistiques modernes mais aussi par le champ disciplinaire des distributions spatiales des espèces.

L'estimation statistique de l'erreur

La séparation en 3 jeux de données : d'entraînement ("training"), de validation ("validating"), et de test ("testing"), est intrinsèque à l'effort de modélisation des statistiques modernes (Hastie et al., 2001). L'erreur de prédiction peut être estimée sur chacun de ces jeux de données.

Elle peut alors être décrite en terme de biais, de variance et de complexité. Les performances prédictives peuvent être exprimées par le taux d'erreur attendue ("error rate") du "fit" du modèle $\hat{f}(X)$ en un point X=x0, en utilisant la fonction de perte ("Loss function", L) au carré telle que (Hastie et al., 2001) :

$$Err(x_0) = L(Y - \hat{f}(X)) = E[(Y - \hat{f}(X))^2|X = x_0] = \sigma_\epsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0))$$

Où X est un vecteur de variables explicatives (prédicteurs) et Y la réponse.

Le **biais** ($Bias$) est alors défini comme l'espérance de la différence entre les probabilités observées et la moyenne des valeurs attendues estimées par le modèle, tel que : $Bias(\hat{f}(x_0)) = E[(f(X) - E[\hat{f}(X)|X = x_0])|X = x_0]$. La **variance** ($Var$) du modèle est définie comme l'espérance du carré de la déviation des valeurs prédites autour de leur moyenne, telle que : $Var(\hat{f}(x_0)) = E[(E[\hat{f}(X)|X = x_0] - \hat{f}(X))^2|X = x_0]$.

L'objectif de la sélection de modèle est de trouver celui qui minimise l'erreur de prédiction : le biais et la variance. Pour cela, on peut donc premièrement considérer **l'erreur d'entraînement** (ou l'erreur apparente) estimée sur les observations, telle que (Hastie et al., 2001) :

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

Or, celle-ci ne peut être utilisée pour l'estimation de l'erreur de prédiction car elle ne considère pas entièrement la complexité du modèle. En effet, on ne peut éviter de considérer la dépendance de l'erreur de prédiction du modèle à sa complexité. Le biais et la variance s'opposent dans leurs effets en fonction de cette complexité. Ainsi, Hastie et al. (2001) insistent sur la description de ces variations. Un modèle sera d'autant plus performant sur les observations qu'il sera complexe et capable de coller au mieux aux données. Cependant, il sera peu performant sur un jeu de données quelque peu différent des observations puisque ne prenant pas en compte la variabilité des prédicteurs. Aussi, l'erreur de prédiction aura une forte variance en fonction des valeurs des prédicteurs (instabilité). Un tel modèle sera alors difficilement généralisable et son erreur de prédiction sera importante. On parle alors d"'over-fitting" du modèle. L'inverse sera observé, "under-fitting", si l'écart entre les données et les prédictions, c.a.d le biais, est important. Dans ce cas, le modèle ne décrit pas correctement la réalité, et de ce fait il sera encore faiblement généralisable.

Les modèles non-paramétriques étant sujets à "l'over-fitting" sur le jeu d'entraînement, leur validation, usuellement réalisée sur le jeu d'entraînement par les statistiques classiques pour la sélection de modèle (AIC, likelihood ratio test), s'effectuera directement sur un jeu de données de validation, c'est-à-dire sur une portion du jeu de données initial non utilisé pour l'entraînement ou sur le jeu de données d'entraînement ré-échantillonné.

Les statistiques modernes ont développé des méthodes d'estimation de l'erreur test à partir de ré-échantillonage des observations qui peuvent être utilisées en l'absence de jeu de données

test, ce qui est souvent le cas quand la taille d'échantillonnage est faible. Nous nous sommes concentrés sur l'estimation de l'erreur de prédiction dans le cas d'une faible taille d'échantillon (cf Axe 1).

L'estimation de l'erreur dans l'espace écologique et géographique

L'erreur de prédiction des modèles de distribution des espèces ne peut être seulement décrite du point de vue des statistiques mais doit être également définie dans l'espace multivarié écologique et géographique dans lesquels se distribuent les données.

La difficulté dans l'interprétation des erreurs réside dans le fait qu'elles ne sont pas nécessairement corrélées dans l'espace géographique et écologique. Ainsi, certaines erreurs sont a-spatiales (globales) et ne dépendent que des relations statistiques entre les variables environnementales et la réponse. De telles erreurs peuvent être engendrées par la qualité des données ou la modélisation. Les erreurs liées aux données peuvent correspondre à (Barry and Elith, 2006; Elith et al., 2002) :

i) l'oubli de variables environnementales importantes, celles ci étant souvent indirectement corrélées aux processus écologiques ;

ii) la taille de l'échantillon concernant le nombre de points d'échantillonnage qui peut s'avérer insuffisante pour mettre en évidence les effets des variables environnementales de manière robuste. Aussi de faibles prévalences (fréquences des points échantillonnés où l'espèce est présente) sont souvent observées pour les espèces rares quelque soit le nombre de points échantillonnés.

iii) le biais dans l'échantillonnage qui peut limiter l'exploration de la variabilité (agrégation dans certains milieux, à proximité des routes...)

iv) les erreurs de mesures (erreur de positionnement GPS...)

Les erreurs liées à la modélisation concernent essentiellement la formulation statistique des relations et la capacité du modèle à s'adapter à la qualité des données. Du fait de la diversité des sources d'erreurs liées aux données on est souvent face à des distributions beaucoup plus complexes que les relations théoriques envisagées avec les variables environnementales à disposition (optimum pour l'espèce). Par exemple, l'oubli de certaines variables environnementales corrélées à d'autres incorporées dans le modèle peut entraîner une distribution non directement causée par les variables considérées. D'autre part, certaines espèces vont avoir des distributions très limitées sur quelques portions du gradient résultant et une discontinuité de la réponse peut être observée. Aussi, la taille de l'échantillon va contraindre la complexité du modèle. Pour les données "sparse", l'utilisation de modèles simples est recommandée mais sera davantage biaisée. Ainsi il sera question de choisir le modèle ou la technique de modélisation qui minimise au plus l'erreur globale.

D'autre part, certaines erreurs varient en fonction de la position des données dans l'espace géographique (erreurs locales). Elles peuvent se retrouver dans le jeu d'entraînement du fait par exemple de l'agrégation des points d'échantillonnage entraînant une auto-corrélation spatiale (biais des données). Les erreurs spatiales peuvent également être décelées à une plus large échelle lorsque les données sont extrapolées sur des données indépendantes. Dans ce cas, elles limitent l'extrapolabilité du modèle. Elles peuvent provenir de l'oubli de variables environnementales qui ont une importance localement, d'une variation spatiale de la niche pour l'espèce, ou encore de

facteurs biogéographiques. Il apparaît alors important de tester et si possible de réduire par son incorporation la structure spatiale des données d'entraînement et d'estimer les erreurs spatiales sur un jeu de données indépendant de celui d'entraînement.

2.2 Écologie comportementale du chien domestique

2.2.1 Outils moléculaires et comportement de défécation

L'analyse des distributions des fèces permet de mettre en évidence les micro-foyers où la contamination peut être potentiellement concentrée dans les environnements des hommes et des hôtes intermédiaires (cf paragraphe 1.2.2).

De nouveaux outils moléculaires, la PCR pour la détection d'ADN et le test ELISA pour la détection de coproantigènes ont été développés récemment pour *Em* (Dinkel et al., 1998; Deplazes et al., 1999; Eckert and Deplazes, 2004). Ces méthodes, contrairement aux autres méthodes de détection du parasite (autopsies et purges) sont non invasives, peuvent être utilisées pour estimer la présence du parasite dans les fèces à une fine résolution spatiale et sont applicables aux fèces d'espèces sauvages. La détection de coproantigènes est plus sensible, rapide et moins onéreuse que les analyses PCR davantage utilisées pour confirmer les résultats de tests ELISA (Deplazes et al., 2003). En revanche elle peut être moins sensible que ces dernières dans les cas de faibles charges parasitaires (Dinkel et al., 1998).

Ces méthodes ont été appliquées avec succès en Europe (Raoul et al., 2001b, 2003; Stieger et al., 2002; Guislain, 2006; Robardet et al., 2008) par exemple pour estimer les taux d'infection dans les fèces de renards et étudier les distributions spatiales de la contamination. En Chine des analyses PCR ont été réalisées pour mettre en évidence la contamination dans quelques fèces de chiens récoltées aléatoirement (Wang et al., 2009) ou récupérées de chiens autopsiés (Zhang et al., 2006).

2.2.2 Analyse du mouvement et comportement de prédation

2.2.2.1 Utilisation de l'espace par les chiens domestiques

Les données de télémétrie permettent d'obtenir un semis de points par individu reflétant la distribution spatiale de son activité. Comme nous l'avons vu plus haut dans le cas des populations, l'utilisation de l'espace, loin d'être homogène et aléatoire, reflète plutôt d'une sélection de l'habitat par l'animal. Elle est alors mesurée dans le cas des distributions d'individus par le pourcentage de temps passé dans différentes zones. Une fonction de densité de probabilités, appelée la Distribution d'Utilisation de l'animal peut être estimée à partir des localisations de l'animal (Worton, 1989). Ainsi, l'aire dans laquelle l'individu passe la majorité de son temps et réalise ses activités de routine est définie comme son domaine vital (Okubo and Levin, 2001). On le distinguera alors de son territoire qui renvoie à la zone défendue de ce même domaine.

D'autre part, la considération de la dimension temporelle des distributions des positions rend possible l'analyse du mouvement des individus. Il s'agit alors de considérer les distributions comme des processus dynamiques. Ici encore, le mouvement des animaux est rarement aléatoire (telle une marche aléatoire ou "random walk") mais plutôt corrélé ou orienté et peut être modélisé (Turchin, 1998; Okubo and Levin, 2001). Les trajectoires sont alors définies comme une

succession de mouvements (distance parcourue entre deux instants) caractérisés par plusieurs paramètres tels que la vitesse ou l'angle entre deux positions. Ainsi, les vitesses diminuent lorsque l'individu atteint une zone d'intérêt, par exemple dans les zone de recherche actives et potentiellement de prédation (Kareiva and Odell, 1987). Les comportements à l'intérieur de cette zone pourront être identifiés par des mouvements rapides et de larges angles (Zollner and Lima, 1999). Les vitesses seront aussi plus grandes lors de trajectoires excursives (de recherche extensive) que dans les zones de repos. L'analyse des trajectoires peut donc permettre l'identification de comportements ou activités au sein des domaines vitaux tels que la prédation, le repos ou la défense de territoire.

2.2.2.2 Distribution des hôtes définitifs et transmission

Définir les aires d'activité de l'hôte définitif d'*Em* revient à définir les aires à l'intérieur desquelles il contamine l'environnement et se contamine par la prédation d'hôtes intermédiaires, c'est-à-dire l'aire de transmission où la rencontre a lieu. Tel que pour le renard urbain, la proximité des chiens aux environnements humains est déterminante pour évaluer leur rôle dans la contamination de l'environnement (Eckert et al., 2000).

L'étude de l'utilisation des habitats définis comme l'environnement des hommes (les habitations et villages) et des hôtes intermédiaires (habitats optimaux de ces hôtes) pourra alors aider à la quantification de la rencontre entre les hôtes ("taux de transmission"). D'autre part, l'analyse des trajectoires pourrait aider à caractériser les comportements à risque pour la contamination des chiens tel que celui de prédation des hôtes intermédiaires. De plus, l'étude de la variabilité des paramètres d'utilisation de l'espace et des comportements spatiaux entre individus donne l'opportunité de tester l'influence des traits d'histoire de vie des individus sur les comportements "à risque" de transmission d'*Em*.

Les comportements spatiaux des hôtes définitifs d'*Em* ont déjà été estimés chez le renard en zone urbaine en Angleterre (Doncaster et al, 1991), en Suisse (Deplazes et al., 2004), en France (Robardet et al., 2008) ou en Allemagne (Thoma, 2005). L'utilisation de l'espace a permis la distinction aux alentours de la ville de Zurich entre les renards urbains et ruraux, la taille des domaines vitaux étant plus élevée pour les renards ruraux (Deplazes et al., 2004). Les renards urbains ayant de petits domaines vitaux sont potentiellement une source importante d'infection pour l'homme puisqu'ils peuvent acquérir le parasite en consommant des hôtes intermédiaires dans les alentours de la ville puis contaminer par leurs fèces les zones habitées. Ainsi, dans l'agglomération de Nancy, des aller-retours entre les zones urbaines et péri-urbaines ainsi qu'une forte exploitation des sites contenant des populations de rongeurs ont été observés et représentent un réel risque de contamination des populations humaines (Robardet et al., 2008).

Les comportements spatiaux du chien domestique varient principalement en fonction de leur degré de domestication et de leurs dépendance aux sociétés humaines. Une classification est couramment utilisée dans la littérature et distingue les chiens de propriétaires ("owned dogs"), les chiens errants n'appartenant à personne mais utilisant l'environnement des hommes ("stray dogs"), et les chiens sauvages ("feral dogs") indépendants des populations humaines. Cependant, de telles catégories ne sont pas si évidentes en nature et le processus de "féralisation", c.a.d

du passage d'une catégorie à l'autre, peut être considéré comme un processus comportemental ontogénique (Boitani et al., 1995). Un chien peut en effet devenir errant après abandon par ses maîtres ou devenir sauvage par son acceptation dans une nouvelle meute. Ainsi, tel qu'il a été observé chez les renards urbains, les chiens errants des zones urbanisées ont des tailles de domaines vitaux plus faibles (de 2-11 à 61 ha) que les chiens sauvages dont les domaines s'apparentent à ceux des autres canidés sauvages (entre 400 et 5800 ha). Ceci tient essentiellement à la disponibilité des ressources, la taille réduite des groupes et des interactions sociales limitées (Boitani et al., 1995). Les données concernant uniquement les chiens de propriétaires sont à notre connaissance inexistantes. Or, dans le contexte de la transmission d'*Em*, cette catégorie de chiens demeure celle à la plus forte probabilité de contact avec l'homme.

2.3 Plan méthodologique du mémoire

2.3.1 Modélisation des distributions spatiales de micro-mammifères

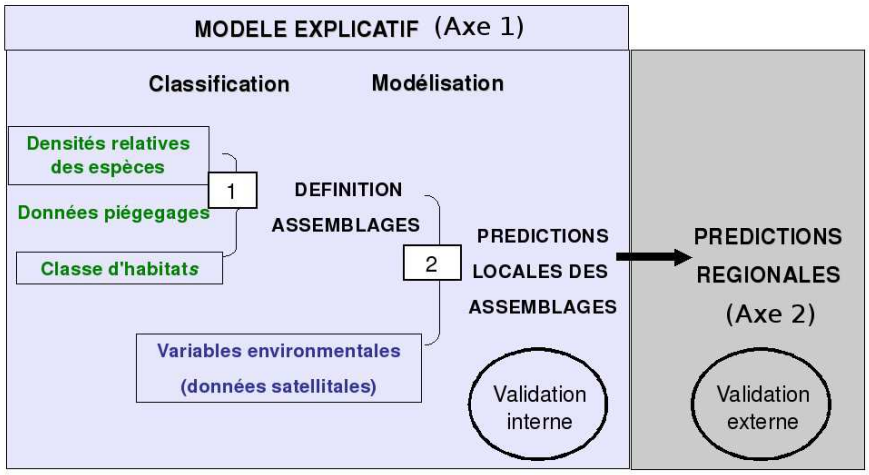

FIG. 2.2 – Schéma conceptuel des étapes de la modélisation (Axes 1 et 2)

Nous proposons un canevas pour la construction de modèles statistiques capables d'expliquer et de prédire au mieux les relations entre les assemblages d'espèces de micro-mammifères et leur environnement dans une région située dans l'Ouest du Sichuan en Chine.

La construction de tels modèles s'est déroulée en deux grandes étapes qui déclinent cette partie de l'étude en deux grands axes (Figure 2.2) :

- Axe 1 : la construction de modèles explicatifs localement
- Axe 2 : les prédictions régionales et leurs évaluations

Á chacune de ces étapes, la confrontation d'une multitude d'outils de modélisation à la nature de nos données de piégeage (présence/absence), la stratégie d'échantillonnage (basée sur avis d'expert) et la taille (faible) de notre jeu de données, a soulevé plusieurs questions méthodologiques auxquelles nous avons tenté de répondre.

2.3.1.1 Axe 1 : Construction d'un modèle explicatif

Cette construction s'est déroulée en deux étapes : 1) la définition des assemblages et 2) leurs modélisation dans l'espace multivarié.

Une méthodologie initialement développée sur des données de piégeage du Ningxia (Raoul et al., 2008) a été appliquée dans deux sites d'études du Sichuan pour définir les assemblages. Les assemblages ont ensuite été modélisés en fonction de variables environnementales issues de données satellitales (Stratégie "Classifier puis modéliser"). Ce faisant, à différentes étapes méthodologiques de la modélisation, de l'entraînement aux prédictions, nous nous sommes confrontés à différentes méthodes que nous avons comparées puis sélectionnées, concernant :

- la méthode de validation, pour l'estimation de l'erreur de prédiction
- la technique de modélisation ("response shape") des assemblages

En complément de ces recherches et au regard des limites de la stratégie de modélisation des assemblages "Classifier puis modéliser", nous avons également testé les capacités prédictives de la modélisation des réponses multiples des espèces ("Classifier et prédire simultanément").

2.3.1.2 Axe 2 : Vers un modèle prédictif des assemblages de micro-mammifères

Nous nous sommes attachés à prédire les distributions régionales des assemblages, c'est à dire au delà des aires d'entraînement des modèles et/ou sur l'étendue incluant nos deux sites d'étude. Ce faisant, nous avons évalué les capacités prédictives régionales de nos modèles, identifié quelques sources d'erreurs de prédiction et proposé des solutions pour les réduire. Cela a été réalisé i) de manière empirique sur les données de terrain des deux sites d'études utilisées précédemment, ii) par une expérience de simulation complémentaire.

(i) Les prédictions régionales des modèles développés localement et régionalement ont été évaluées et comparées. Ce faisant, nous avons exploré les effets de sources d'erreurs d'extrapolation liées à notre jeu de données (la prévalence des assemblages) et à la modélisation (le type de réponse modélisée).

(ii) L'expérience de simulation nous a permis d'explorer l'effet de la taille des échantillons des observations pour chaque assemblage sur les prédictions réalisées sur un jeu de données indépendant. Ce faisant nous avons pu quantifier et comparer les sensibilité aux tailles d'échantillons des différentes techniques de modélisation utilisées dans l'axe 1.

2.3.2 Axe 3 : Le rôle du chien domestique (*Canis lupus familiaris*) dans la transmission d'*Em*

Les comportements "à risque" du chien domestique ont été étudiés au sein des habitats des hommes et des hôtes intermédiaires dans des sites d'études situés sur le plateau tibétain (Sichuan).

L' analyse des distributions spatiales des fèces a été effectuée pour mettre en évidence les zones dans les villages à risque de transmission potentiellement élevé.

Aussi, l'estimation des prévalences fécales par des outils moléculaires récemment développés, nous a permis de quantifier le rôle du chien dans la contamination effective de l'environnement humain, ce relativement par rapport à celui du renard.

Enfin, à partir de trajectoires nocturnes de chiens équipés de colliers GPS nous avons cherché à i) estimer l'utilisation des habitats des hommes et des micro-mammifères et, ii) mettre en évidence l'influence des paramètres individuels des chiens sur leurs comportements spatiaux.

Tab. 2.1 – Classification des principaux algorithmes (liste non exhaustive) couramment utilisés pour modéliser les distributions des habitats pour des espèces ou assemblages.

Para/Non para	Algorithme	Méthode	Nature réponse	Ind/Mult	Dist	Strat	Objectif
Para	ENFA/MADIFA	ORDI	Qual (P)	Ind	Gaussian	1	Expl
	CCA/RDA	ORDI/REG	Quant/Qual	Mult	Gaussian	1/3	Expl/Pred
	OMI	ORDI	Quant	Mult	NA	3	Expl
	GLM	REG	Quant/Qual (PA)	ind/mult	Famille exponentielle	1/3	Expl/Pred
	LDA	ORDI	Qual (P PA)	ind/mult	Gaussian	1	Expl/Pred
	MDA	CLASSI	Qual (P PA)	ind/mult	Multiple Gaussian	1	Expl/Pred
	VGLM	REG	Qual (P PA)	ind/mult	Famille exponentielle ou pas	1/3	Expl/Pred
Non para	BIOCLIM, MAXENT	ENV	Qual (P)	Ind	NA	3	Expl/Pred
	Regression tree	CLASSI	Qual (P PA)	ind/mult	NA	1/3	Expl/Pred
	GAM	REG	Qual (PA)	Ind	NA	3	Pred
	Genetic algorithm (GARP)	INTEL	Qual(PA)	Ind	NA	3	Pred
	VGAM	REG	Qual(PA)	Mult	NA	3	Pred
	FDA	CLASSI	Qual (PA,P)	ind/mult	NA	1/3	Pred
	MARS	MULTIREG	Qual (PA,P)	ind/mult	NA	1/3	Pred
	Neural network	INTEL	Pres/abs	Ind-Mult	NA	3	Pred
Autre	Bayesian classification	BAYES	Pres/abs	Ind-Mult	NA	1/3	Pred

Les techniques de modélisation sont classifiées en fonction de :
- leur caractère paramétrique (Para) ou non paramétrique (Non para),
- la famille de méthodes à laquelle elles appartiennent : ordination ("ORDI") ; régressions ("REG") ; classification ("CLASSI") ;intelligence artificielle ("INTEL") ; ou bayésienne ("BAYES").
- la nature des réponses des espèces : qualitative ("Qual") et de présence seulement ("P") ou de présence/absence ("PA"), ou quantitative("Quant"),
- le caractère individuel ("Ind", pour chaque espèce indépendamment) ou multiple ("Mult", pour toutes les espèces simultanément) de la réponse,
- la distribution de la réponse (Dist), "NA" signifie que la distribution est libre (non paramétrique),
- la stratégie de modélisation des assemblages à laquelle elle répond (stratégie 1 ou 3),
- l'objectif de la modélisation : explicatif ("expl") ou prédictif ("pred")

Travaux de recherche

Chapitre 3

Axe 1 : Construction d'un modèle explicatif - Modéliser la distribution spatiale des assemblages de micro-mammifères

3.1 Introduction

3.1.1 Contexte environnemental et épidémiologique

Dans les travaux présentés ci-après nous nous sommes intéressé aux distributions spatiales des espèces de micro-mammifères dans 2 aires d'études situées dans l'Ouest du Sichuan, dans les comtés de Maerkang et de Rangtang (Chine). Dans ces aires des contreforts du plateau tibétain, les prévalences humaines sont faibles (0.88 % sur 571 patients et 1.48 % sur 675 patients respectivement) en comparaison à celles observées dans le comté de Shiqu sur le plateau (entre 4 % sur 475 patients à Niga et 9.4 % sur 631 patients à Yiniu) (Li et al., Submitted). Ces aires d'études constituent des zones de vallées montagneuses (situées entre 3000 et 4500 m) davantage forestières que le plateau et offrant une large diversité paysagère : des zones agricoles (prairies de fauche, cultures en terrasse), des forêts (bouleaux, conifères, rhododendrons, chênes), des stades intermédiaires (buissonneux et arbustifs) le long de gradients de déforestation/aforestation et de déprise agricole (figure 3.1 et figure 3.2).

FIG. 3.1 – Paysage de Rangtang

FIG. 3.2 – Paysage de Maerkang

3.1.2 Rappel des objectifs et des questions

Au regard de la qualité de notre jeu de données (faible taille d'échantillon, large diversité d'habitats, espèces rares), nous avons cherché à modéliser les distributions spatiales des groupes d' habitats partageant les mêmes espèces de micro-mammifères, c'est-à-dire des habitats d'assemblages d'espèces. Ce travail s'est déroulé en deux grandes étapes : la classification des réponses des espèces puis la modélisation des assemblages (Stratégie 2, Ferrier and Guisan (2006)). Cette recherche fait l'objet d'un article publié (Vaniscotte et al., 2009) et présenté dans l'Axe 1a.

Premièrement, face à la grande diversité, dans nos sites d'études, d'habitats échantillonnés, la grande richesse en espèces dont la majorité étaient rares, et la difficulté d'interprétation de telles diversités, nous nous sommes intéressé à définir des assemblages de manière quantitative par une modélisation statistique et multivariée des densités relatives des espèces.

Dans un deuxième temps, nous avons modélisé les distributions spatiales des habitats des assemblages définis. La définition des habitats effectuée sur avis d'experts est qualitative et de fait limitée dans une optique prédictive puisqu'elle réduit l'environnement des espèces à une variable qualitative. Nous nous sommes alors attachés à définir de manière quantitative les habitats pour ces groupes d'espèces dans l'espace des variables environnementales issues de données satellitales. Ce faisant nous avons exploré principalement deux aspects méthodologiques :

i) l'estimation de l'erreur de prédiction, en choisissant parmi les méthodes statistiques non-paramétriques la plus appropriée à notre cas d'étude,

ii) la sélection de la technique de modélisation la plus appropriée dans un contexte de discrimination des assemblages.

Enfin, nous avons testé, en complément de cette première analyse, la capacité des variables environnementales à expliquer directement, c'est à dire sans passer par la définition des assemblages, les densités relatives jointes des espèces. Cela constitue alors une première étape dans l'évaluation de la stratégie de modélisation et classification simultanées des assemblages. Cette analyse complémentaire constitue l'Axe 1b.

3.2 Axe 1a : Modélisation et discrimination spatiale des assemblages de micro-mammifères dans l'Ouest du Sichuan

Ci-dessous des précisions méthodologiques complémentaires au contenu de l'article sont apportées concernant i) la définition des assemblages, ii) les méthodes de modélisation et iii) l'estimation de l'erreur de prédiction.

3.2.1 Précisions méthodologiques

3.2.1.1 Étape 1 - Classifier : définition des assemblages

Dans notre cas d'étude, la classification des habitats, bien que qualitative et réalisée sur avis d'experts, constitue l'unique description de terrain des habitats (variable proximale) dans la zone étudiée. De surcroît, en l'absence d'autres données, elle a guidé la stratégie d'échantillonnage, réalisée de manière stratifiée dans chacune de ces classes.

La méthode de classification des points d'échantillonnage utilisée ici est basée sur la modélisation des densités relatives des espèces (succès de capture) en fonction des classes d'habitats définies *a priori*. Elle a pour objectif de réduire la redondance dans les informations paysagères apportées par les classes d'habitats pour expliquer les densités relatives des espèces.

Un Modèle Linéaire Généralisé (GLM) ayant une fonction de lien multinomiale est utilisé. En effet, pour chaque piège i (unité statistique), la probabilité de capturer une espèce suit une loi multinomiale puisque une des K espèces présentes dans l'environnement du piège peut être capturée, ce de manière exclusive. Les covariables considérées sont alors m classes d'habitats mais peuvent également être des variables liées au protocole d'échantillonnage et qui peuvent influencer le succès de capture (type de piège, la nuit de capture, ...).

Ainsi, pour chaque espèce j (de 1 à K) et pour chaque piège i, on obtient une équation de régression logistique (figure 3.3, 1). Les probabilités de capture des espèces sont obtenues par la fonction de lien du GLM multinomial (figure 3.3, 2). La réponse d'une espèce est conditionnée non seulement par les covariables mais également par les réponses des autres espèces. Les réponses des espèces sont donc relatives les unes par rapport aux autres et leurs probabilités somment à l'unité. Enfin, les paramètres des équations de régression sont optimisés en maximisant la vraisemblance multinomiale (figure 3.3, 3)).

La définition des assemblages consiste à regrouper les habitats partageant les espèces. Cela est effectué en regroupant deux à deux les données de captures des classes d'habitats, donnant alors m^2 nouveaux modèles et un nombre correspondant d'itérations de l'estimation des paramètres. Á chaque itération ou combinaison de classes, les paramètres du modèle multinomial sont ré-estimés en optimisant la vraisemblance et un nouveau Critère d'Information d'Akaike est calculé ($AIC_{nouveau}$) et est comparé au critère du modèle original (AIC_{ori}) par le calcul de leur différence ($\Delta AIC = AIC_{nouveau} - AIC_{ori}$). Un ΔAIC négatif indique alors que le regroupement des classes a apporté de l'information, c'est-à-dire que les différences observées dans les réponses des espèces entre les deux classes d'habitats ne correspondent pas à une réponse fonctionnelle du point de vue des assemblages de micro-mammifères. Le regroupement est alors conservé et considéré comme une nouvelle classe d'habitats (une super-classe) et comme variable explicative du modèle. La procédure de regroupement des classes est ré-itérée jusqu'à

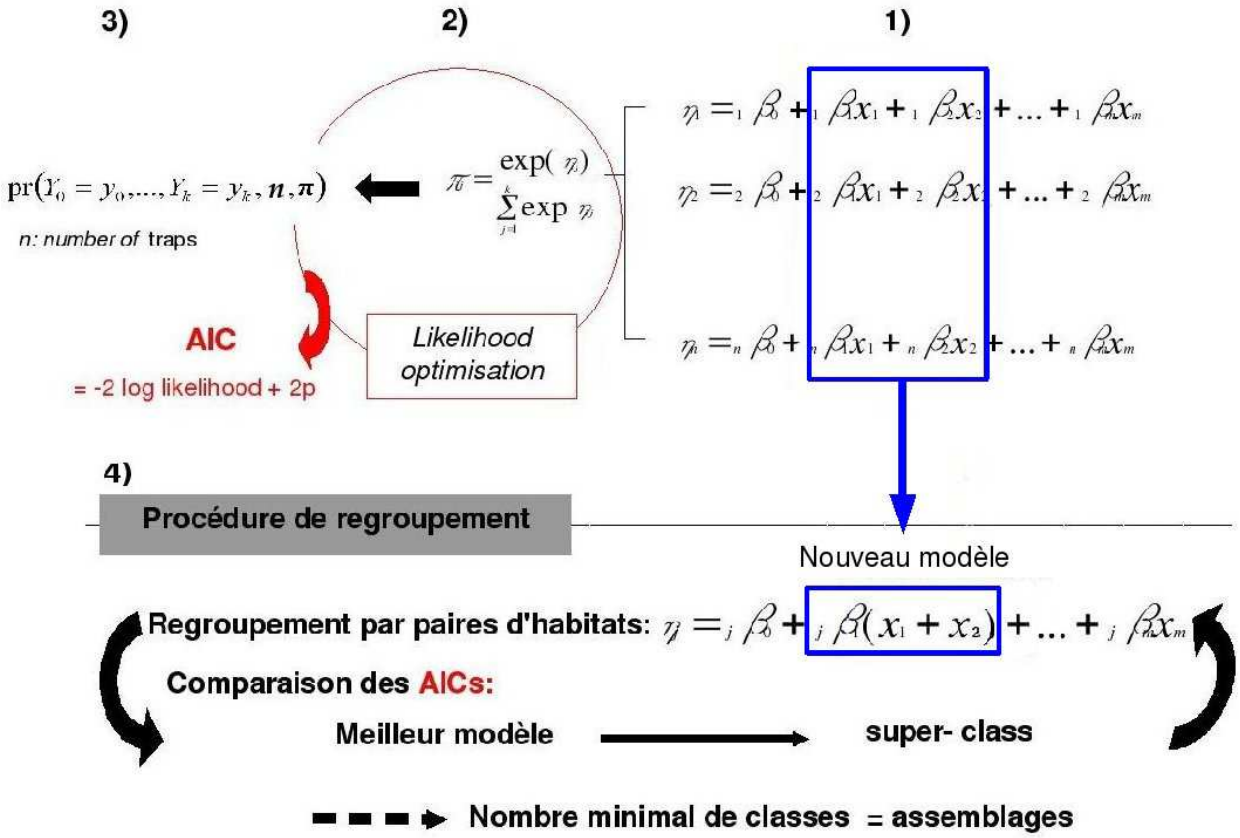

FIG. 3.3 – Modélisation multinomiale des densités relatives des espèces (de 1 à 3) et procédure de regroupement des classes d'habitats (4). Cf explication dans le texte.

l'obtention du modèle le plus parcimonieux (AICc minimal). Une nouvelle classification des habitats est ainsi obtenue et chaque classe composite correspond à l'habitat d'un assemblage ainsi qu'à une distribution unique de probabilités d'occurrence des espèces. Cette procédure a été implémentée dans le logiciel R (R, 2005). Elle s'est avérée utile et concordante avec les avis d'experts pour définir des assemblages dans la province du Ningxia le long de gradients de déforestation et de reforestation (Raoul et al., 2008).

3.2.1.2 Étape 2 - Modéliser les distributions des assemblages : un contexte de classification

La variable expliquée par nos modèles est la présence/absence d'assemblages de micro-mammifères en certains points géoréférencés. Ces dernières étant mutuellement exclusives, notre analyse se situe dans un contexte de classification des réponses des assemblages. La réponse est donc qualitative et attribuée à l'un des groupes préalablement identifié que l'on cherche à discriminer (classification supervisée). L'espace écologique est défini ici par certaines variables environnementales issues de données satellitales. Nous modéliserons donc une représentation partielle de l'habitat des assemblages.

Pour discriminer les assemblages de micro-mammifères, nous avons exploré les performances discriminatives de quatre techniques de modélisation statistiques couramment utilisées : les analy-

ses discriminantes, linéaires (Linear Discriminant Analysis, LDA) ou de mixture de Gaussiennes (Mixture Discriminant Analysis, MDA) (Hastie et al., 2001), les Modèles linéaires généralisés et Multinomiaux, Multinomial Model, MM) (MacCullag and Nelder, 1989) et les Régressions Multiples et Adaptives par fonctions de lissage (Multiple Adaptive Regression Spline, MARS) (Hastie et al., 2001).

Les analyses discriminantes et les régressions estimeront les probabilités d'appartenir à chaque classe. Classifier consistera alors à assigner à l'observation la classe pour laquelle la probabilité est maximale (méthode "softmax"). De telles méthodes de classification seront alors appelées à définir des "frontières de décision" ("decision boundaries") pour délimiter des classes dans l'espace multivarié des variables environnementales (Hastie and Tibshirani, 1993; Hastie et al., 1994). Leurs comparaisons développées dans l'article permettront de mettre en évidence la nature d'une telle "frontière de décision" correspondant à la distribution des assemblages et qui sera :

- linéaire si modélisée par LDA et MM (Figure 3.4, a et c), les classes étant représentées par des distributions Gaussiennes et multinomiales respectivement,
- non-linéaire si modélisée par MDA et MARS (Figure 3.4, b et d), les classes étant représentées par des distributions multi-Gaussiennes et libres respectivement.

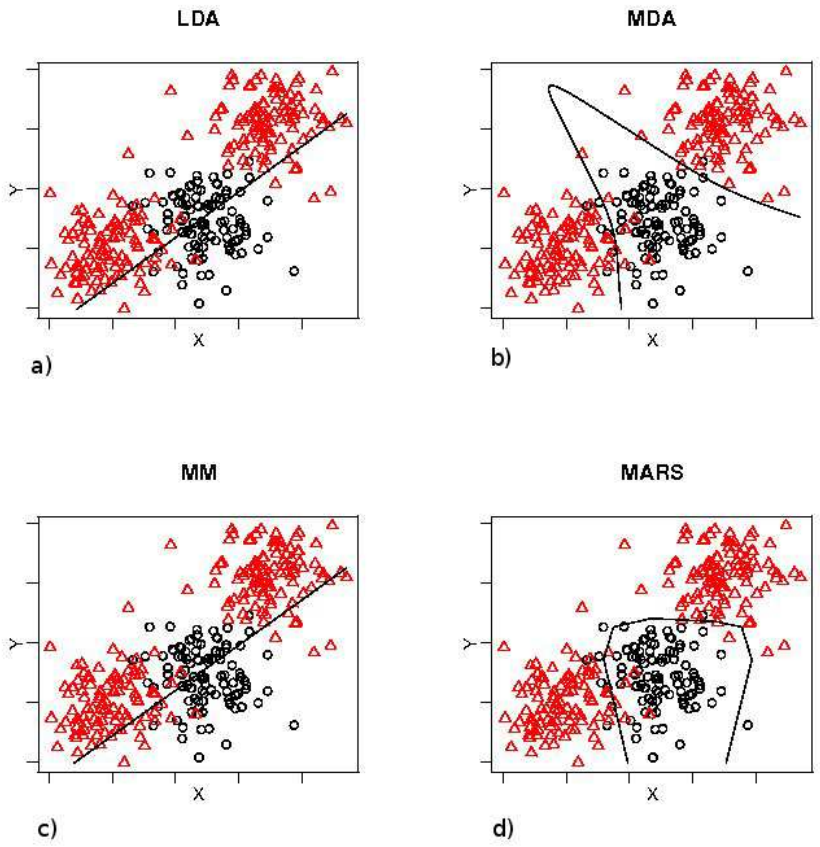

FIG. 3.4 – Exemple de "frontières de décision" obtenues pour différentes méthodes de modélisation (cf texte pour les abréviations), pour classifier l'appartenance d'une population à deux groupes dont les individus sont représentés par des cercles et des triangles.

3.2.1.3 Estimer l'erreur de prédiction : la validation croisée et le bootstrap 632+

L'erreur de prédictions a été estimée sur un jeu de données de validation. Étant donné la faible taille de nos échantillons, nous avons eu recours à une méthode de ré-échantillonnage des données.

L'estimation de l'erreur (telle que définie dans l'introduction) peut être qualifiée d'optimiste (sous-estimation) si les mêmes données sont utilisées pour entraîner et tester les prédictions du modèle.

Ainsi, la méthode de la validation croisée est couramment utilisée pour estimer l'erreur de prédiction corrigée de son optimisme. Elle consiste à extraire une donnée de l'échantillon, à entraîner le modèle sur les données restantes puis à estimer les erreurs de prédiction du modèle sur la donnée écartée. Ceci est répété pour chaque observation jusqu'à ce que leur ensemble (n observations) ait été utilisé. L'erreur de prédiction du modèle est alors estimée comme la moyenne des erreurs de prédiction observée sur toutes les observations, telle que :

$$\widehat{Err}_{CV} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}^{-i}(x_i))$$

L'erreur de prédiction (L) mesure alors l'écart entre les observations, (y_i) et les valeurs prédites par le modèle, $\hat{f}(x_i)$. Dans le contexte de la classification, L est souvent considérée comme égale à 1 si la donnée est correctement prédite et à 0 sinon.

Cette estimation de l'erreur est cependant sujette à une forte variance puisque les jeux d'entraînement diffèrent peu et que les erreurs sont estimées à chaque itération sur une seule observation. On peut alors étendre la méthodologie à un groupe de K données qui est exclu du jeu d'entraînement du modèle (n-K), mais qui est utilisé pour estimer les erreurs de prédiction. Une telle partition permet alors, en augmentant le nombre d'observations utilisées pour estimer les erreurs de prédiction, de réduire la variance de l'estimation. Cependant un désavantage de cette méthode réside dans le fait que l'erreur de prédiction n'est estimée qu'une seule fois sur chaque observation.

Le bootstrap se présente comme une méthode alternative à la validation croisée pour estimer l'erreur de prédiction. Il consiste à tirer B échantillons de manière aléatoire et avec remise à partir de la loi empirique sur l'échantillon de départ. Chaque échantillon est constitué d'observations (x_i appartenant à X) se répétant chacune un certain nombre de fois. Le modèle est estimé sur chaque échantillon. L'erreur de prédiction de chaque B modèle correspond à la moyenne des erreurs de prédiction estimés en chaque observation x_i. Si le nombre de ré-échantillonnage B est suffisamment grand l'espérance de l'erreur peut être alors approximée par la moyenne de l'erreur estimée sur les B échantillons. L'erreur de prédiction du modèle est alors la moyenne de ces erreurs sur B échantillons telle que :

$$\widehat{Err}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{b}(x_i))$$

Ainsi, contrairement à la validation croisée, l'erreur de prédiction en chaque observation est estimée en moyenne B fois. Cette procédure faisant intervenir une moyenne sur B échantillons permet une réduction considérable de la variance de l'estimation de l'erreur de prédiction. On

peut alors dire que le bootstrap lisse la variabilité de la prédiction en chaque point (Efron and Tibshirani, 1995).

Á la manière de la validation croisée, le bootstrap peut être utilisé pour corriger l'optimisme de l'estimation de l'erreur. Pour chaque observation x_i les prédictions sont réalisées à partir des échantillons qui ne contiennent pas cette observation (C^{-i}). L'erreur, pour chaque observation, est considérée comme la moyenne des erreurs estimées des C^{-i} échantillons. L'erreur de prédiction du modèle est alors considérée comme la moyenne des erreurs de ces observations, c'est le "leave-one-out" bootstrap tel que :

$$\widehat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\epsilon C^{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

Or, en l'état, le problème de l'optimisme n'est qu'en partie résolu. Il est estimé que 63.2 % des observations (distinctes) sont contenues dans chaque ré-échantillonage alors que 36.8 % d'entre elles ne sont pas considérées (Hastie et al., 2001). Ainsi l'estimation de l'erreur par bootstrap correspondrait à une partition en 2 groupes du jeu de données (telle une cross-validation) et peut être biaisée pour un si faible nombre de partitions. De surcroît, pour chaque ré-échantilonnage, l'erreur estimée sur toutes les observations est biaisée puisque certaines observations n'ont jamais été entraînées (36.8 %). Ainsi, la formulation du bootstrap 632 a été développée pour réduire ce biais (Efron and Tibshirani, 1997). $\widehat{Err}^{(1)}$ est estimé seulement pour 63.8 % des observations. Le taux d'erreur estimé sur le jeu de données non bootstrappé (erreur apparente $\overline{err}$; cf Introduction) sera alors attribué à l'autre fraction des données (63.2 %).

On obtient alors l'estimateur dit "632" de l'erreur :

$$\widehat{Err}^{(632)} = 0.368 \times \overline{err} + 0.632 \times \widehat{Err}^{(1)}$$

Avec,

$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

Enfin, une dernière correction concerne la réduction de l'optimisme de l'erreur apparente. Le taux d"'overfitting" est calculé tel que :

$$\hat{R} = \frac{\widehat{Err}^{(1)} - \overline{err}}{\hat{\gamma} - \overline{err}}$$

où $\hat{\gamma}$ corresponds au taux de"non-information" ou à l'erreur engendrée par le modèle nul (sous indépendance des prédicteurs et des observations). Il n'y a pas d'overfitting si $\widehat{Err}^{(1)} = \overline{err}$. En revanche R=1 si l'overfitting équivaut à $\hat{\gamma} - \overline{err}$.

Il vient alors corriger l'estimation $\widehat{Err}^{(632)}$ par l'estimateur apellé "632+" de l'erreur :

$$\widehat{Err}^{(.632+)} = (1 - \hat{w}) \times \overline{err} + \hat{w} \times \widehat{Err}^{(1)}$$

avec

$$\hat{w} = \frac{0.632}{1 - 0.368\hat{R}}$$

Le poids w est compris entre 0.632 si l'overfitting est nul (R=0) et 1 si l'overfitting est maximal (R=1).

Le bootstrap 632+ a été développé dans le contexte de la classification (Efron and Tibshirani, 1997) puis a été appliqué avec succès pour estimer l'erreur de prédictions de modèles visant à cartographier les distributions des espèces tels que l'algorithme "Random Forest" ou "CART" (arbre de classification et de régressions) (Merler et al., 1996; Furlanello et al., 2003), ou MARS (Leathwick et al., 2006). Cependant malgré les preuves analytiques (telles qu'exposées plus haut) ou empiriques (Merler et al., 1996; Steyerberg et al., 2001; Molinaro et al., 2005) des capacités de cette méthode à réduire le biais et la variance des erreurs de prédiction, ses applications restent encore assez rares dans notre domaine de recherche.

3.2.2 Article : des assemblages de micro-mammifères dans l'ouest du Sichuan (Chine)

Ecological Modelling 220 (2009) 1218–1231

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Modelling and spatial discrimination of small mammal assemblages: An example from western Sichuan (China)

Amélie Vaniscotte[a,*], David R.J. Pleydell[a,f], Francis Raoul[a], Jean Pierre Quéré[b], Qiu Jiamin[d], Qian Wang[d], Li Tiaoying[d], Nadine Bernard[a], Michael Coeurdassier[a], Pierre Delattre[b], Kenichi Takahashi[c], Jean-Christophe Weidmann[e], Patrick Giraudoux[a]

[a] Department of Chrono-environment, UMR UFC/CNRS 6249, USC INRA, Université de Franche-comté, Place Leclerc, 25030 Besançon Cedex, France
[b] Campus International de Baillarguet, INRA, CBGP-UMR 1062, CS 30016 Montferrier sur Lez, 34988 Saint Gély du Fesc Cedex, France
[c] Hokkaido Institute of Public Health, Kita 19, Nishi 12, 060-0819 Sapporo, Japan
[d] Institute of Parasitic Diseases, Sichuan Center for Disease Control and Prevention, Chengdu 610041, Sichuan, China
[e] 3 rue de Buffard, 25440 Liesle, France
[f] UMR BGPI, CIRAD TA A-54/K, Campus International de Baillarguet, 34398 Montpellier Cedex 5 France

ARTICLE INFO

Article history:
Received 28 April 2008
Received in revised form 13 February 2009
Accepted 18 February 2009

Keywords:
Small mammal assemblages
Habitat distribution modelling
Mixture discriminant analysis
Multiple adaptive regression spline
Environmental gradients

ABSTRACT

We investigate the relationship between landscape heterogeneity and the spatial distribution of small mammals in two areas of Western Sichuan, China. Given a large diversity of species trapped within a large number of habitats, we first classified small mammal assemblages and then modelled the habitat of each in the space of quantitative environmental descriptors. Our original two step "classify then model" procedure is appropriate for the frequently encountered study scenario: trapping data collected in remote areas with sampling guided by expert field knowledge.

In the classification step, we defined assemblages by grouping sites of similar species composition and relative densities using an expert-class-merging procedure which reduced redundancy in the habitat factor used within a multinomial logistic regression predicting species trapping probabilities. Assemblages were thus defined as mixtures of small mammal frequency distributions in discrete groups of sampled sites. In the modelling step, assemblages' habitats and environments of the two sampled areas were discriminated in the space of remotely sensed environmental descriptors. First, we compared the discrimination of assemblage/study areas by linear and non-linear forms of discriminant analysis (linear discriminant analysis *versus* mixture discriminant analysis) and of multiple regression (generalized linear models *versus* multiple adaptive regression splines). The "best" predictive modelling technique was then used to quantify the contribution of each environmental variable in discriminations of assemblages and areas.

Mixtures of Gaussians provided a more efficient model of assemblage coverage in environmental space than a single Gaussian cluster model. However, non-linearity in assemblage response to environmental gradients was consistently predicted with lower deviance and misclassification error by multiple adaptive regression splines. The two study areas were mainly discriminated along vegetation indices. However, although the normalized difference vegetation index (NDVI) could discriminate forested from non-forested habitats, its power to discriminate assemblages in Maerkang, where a greater diversity of forest habitat was observed, was seen to be limited, and in this case NDVI was outperformed by the enhanced vegetation index (EVI). Our analyses highlight previously unobserved differences between the environments and small mammal communities of two fringe areas of the Tibetan plateau and suggests that a biogeographical approach is required to elucidate ecological processes in small mammal communities and to reduce extrapolation uncertainty in distribution mapping.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

1.1. Modelling distributions of assemblages

Modelling spatial distributions of species is a developed and promising research field which can both explain and predict the effects of environmental descriptors on species presence/absence

* Corresponding author. Tel.: +33 3 81 66 57 14; fax: +33 3 81 66 57 97.
E-mail address: amelie.vaniscotte@univ-fcomte.fr (A. Vaniscotte).

0304-3800/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2009.02.019

in space. This relies on defining species' habitats (Guisan and Zimmerman, 2000) following Hutchinson's concept of ecological niche (1957). Recent reviews are found in Pulliam (2000) and Hirzel et al. (2002).

In situations where a large number of species co-occur in a large number of sampled sites, habitat definition for individual species can be complicated by species interactions. Defining habitat for rare or "shy" species can be impossible when presence is difficult to detect. By contrast, the full extent of habitats for dominant species can be elusive when the species is found in a large proportion of sampled sites. In these situations community level modelling constitutes a useful tool that provides a synthesis of such data sets by reducing their complexity to a much smaller set of higher-level entities. One such higher level entity can be that of an assemblage, i.e. a group of taxonomically related species which share the same habitat (Ferrier and Guisan, 2006). The habitat of an assemblage can therefore be defined in reference to a group of sites in which similar species composition and densities are observed.

Community-level spatial modelling involves three main steps (Ferrier and Guisan, 2006): (i) classification of species and/or sites into groups, (ii) statistical formulation to assess relationships between groups and environmental descriptors, and (iii) predictive mapping of groups. Classification (i) is often realized through ordination methods without incorporating environmental information, e.g. the TWINSPAN algorithm (Hill, 1979; Legendre and Legendre, 1998). Grouping species into discrete assemblages involves making assumptions on species distributions (Olden, 2003): first, grouped species are treated in a similar way regarding their responses to environmental descriptors; then, species are assumed to belong to discrete mutually exclusive groups. In regards to Hutchinson's concepts of ecological niche and associated species' niche specificity and marginality, such assumptions seem at odds with species distributions in nature (Doledec et al., 2000; Hirzel et al., 2002). Moreover, clustering procedure can be biased by the subjective choice of similarity criterion and decision threshold (Anderson and Clements, 2000). In the modelling step (ii) groups are typically modelled against environmental variables, individually *via* GLM or GAM, or simultaneously using polychotomous regression or discriminant analysis (Lehmann et al., 2002; Gibson et al., 2004; Ferrier and Guisan, 2006). Modelling techniques have been deeply investigated in a "model then classify" approach by which species responses are grouped into assemblages after prediction of their occurrences (Lehmann et al., 2002; Gibson et al., 2004; Ferrier and Guisan, 2006). Joint modelling of multi-species responses has been recently developed to account for species interactions. This can help elucidate those variables that have strong effects on the whole community but might not be identified relevant in species level analyses (Leathwick et al., 2006). This has been successfully achieved through non-linear models such as Multiple Adaptive Regression Splines (MARS) (Friedman, 1991; Moisen and Frescino, 2002; Leathwick et al., 2006) and artificial neural networks (Olden, 2003; Olden et al., 2006). However, these methods consider the model response to be presence of the individual species in each sampled site, defined using a threshold probability (usually 0.5) on observation frequencies, which can be viewed as a constraint for rarely observed species or sparse data sets.

1.2. Modelling landscape disturbance effects on small mammal distributions

Effects of landscape disturbances on small mammal communities have been investigated by defining assemblages using several classification methods on the basis of species trapping data. For example, Butet et al. (2006) used ordination methods and species/sites proximity in co-inertia axes while Giraudoux et al. (1998) defined assemblages by subjective criteria. Clustering algorithms have also been used to classify habitats into groups according to species composition dissimilarity (Krasnov et al., 1996).

On the basis of an *a priori* expert defined and qualitative habitat nomenclature, Raoul et al. (2008) produced an objective and reproducible classification of assemblages. The set of sampled habitat classes was reduced into a smaller set using information theory and assuming a multinomial model for the small mammal trapping data. Each *a posteriori* habitat class plus associated estimates of species trapping frequencies defined an assemblage. This approach presents several advantages over currently popular classification methods in the "Classify then model" approach: first, classification of assemblages was performed *via* a reclassification of sampled habitat classes *a priori* identified in the field and thus incorporates information gained from an expert oriented sampling strategy (Pearce et al., 2001); secondly, by considering each assemblage as a localised picture of species composition and relative densities, classification was performed at the habitat level instead of the species level, thus the crude assumptions of species responses mentioned above were avoided; finally, in the reclassification step, models were compared and selected using Akaike Information Criterion (AIC) which can be viewed as an objective method (Burnham and Anderson, 1998). However, this classification method limits the definition of assemblages' habitats to qualitative and *a priori* habitat classes limited to the sampling design. Consequently it suffers from predictive power limitations, i.e. assemblage definitions cannot be spatially extrapolated beyond the sampled sites.

1.3. Context, hypothesis and objectives

In China, the spatial distribution of small mammal species has been shown to be modified by landscape disturbances such as overgrazing and fencing practices on the Tibetan plateau (Wang et al., 2004; Raoul et al., 2006), deforestation in Gansu (Giraudoux et al., 1998) and afforestation in Ningxia (Raoul et al., 2008). We aimed to investigate the relationship between landscape heterogeneity and the spatial distribution of small mammal assemblages in two forested areas located in the fringes of the Tibetan plateau (Sichuan, China). There, forest management leads to landscape heterogeneity and likely drives changes in the spatial distributions of small mammals.

Our main methodological challenge was to develop a predictive model for trapping data sets of diverse taxon, realized in remote areas and thus constrained by a limited sampling effort, a large number of rare species trapped with low frequency and an expert oriented sampling design. There is currently a need to adapt habitat modelling methodologies in order to fit and extract the maximum information possible from such data sets which is common in conservation studies. Here, we developed an original two step "Classify then model" procedure to address these issues.

In the classification step, we applied the Raoul et al. (2008) expert-class-merging procedure to summarise our trapping data by defining assemblages. Then, our major technical contribution was to incorporate such assemblage definitions into a predictive modelling framework. In the modelling step, the predictive limitations of the initial expert-class-merging assemblage definitions was overcame by extending the definition of the habitat associated with each assemblage using a set of quantitative variables extracted from remote sensors. By doing so, we addressed some currently debated methodological issues in the field of species distribution modelling. The modelling step aimed to answer two questions:

(i) which modelling technique and associated response-factor relationships predict assemblage occurrence with lowest prediction error?

(ii) what are the contributions of each environmental variable in assemblage discrimination and prediction?

Among the theoretical hypothesis supporting current species modelling methodological frameworks, the shape of response curves has often been neglected and there is a need to consider it at the interface between statistical methods and ecological realism (Austin, 2002, 2007; Guisan et al., 2006). The richness of existing statistical models offers the opportunity to compare different ecological theories underlying observed patterns of species distribution in environmental space. While previous studies have shown non-linear models provide better fits and predictions of multiple species responses than linear models (Doledec et al., 2000; Olden, 2003; Leathwick et al., 2006), we tested if this observed pattern could hold when species responses are considered as a whole, i.e. when assemblages and not species are used as the response variable. Several methods exist to discriminate known groups by continuous variables. We compared the discrimination ability of two widely used classifiers, multiple logistic regression and discriminant analysis, and tested several forms of assemblage response curves: linear *versus* non-linear and single Gaussian *versus* Gaussian mixture, respectively.

The selection of a relevant set of predictors for building predictive models also remains an active current issue which is complicated by the nature of environmental predictors (direct/indirect) and their interactions (Austin, 2007). Here, instead of selecting a single "best" model, we investigated the overall contribution of each environmental variable within a set of models.

Because trapping data were collected in two distinct areas, the same questions were answered at the regional spatial extent to investigate discrimination of the two study areas in the environmental space.

2. Materials and methods

2.1. Small mammal species data sets

2.1.1. Study areas

Two study sites were investigated in western Sichuan province (central China) in the vicinities of Rangtang and Maerkang cities (approximatively 100 km apart) (Fig. 1), in June 2004 and July/September 2005, respectively. The sampling area in Maerkang ranged from 2950 to 4100 m altitude. The 2005 mean annual temperature and yearly average rainfall were 8.9° C and 811.5 mm, respectively (data source: Maerkang Center for Disease Control). Forested areas were either coniferous or birch and oak. Afforestation measures were in place and a large number of young plantation forests were found. Rhododendron forests and pastures were observed at higher elevations. Villages were often surrounded by ploughed fields and were situated close to rivers with abandoned terraced fields at higher elevation. In Rangtang, elevation of sampled area ranged from 3350 to 3900 m and in 2005 yearly average temperature and yearly average rainfall were 5.4° C and 854.3 mm, respectively. Landscape was mainly composed of grassland and shrub. Forest was less abundant than in Maerkang and was more frequently coniferous than broad-leaf.

2.1.2. Sampling protocol

Sampling was undertaken in *a priori* defined habitats identified in the field, i.e. habitats classified on the basis of apparent dissimilarities in vegetation structure and dominant genus composition. In Maerkang and Rangtang, 18 and 12 habitats were sampled, respectively (Table 2). Four habitats were found to be similar in the two locations: "Forest Rhododendron/coniferous", "Forest coniferous willow bushes understory", "Stream bushes" and "Slope bushes" (Table 2).

Extensive standard trapping (Giraudoux et al., 1998; Raoul et al., 2006) was undertaken in each habitat. Each standardized trapline

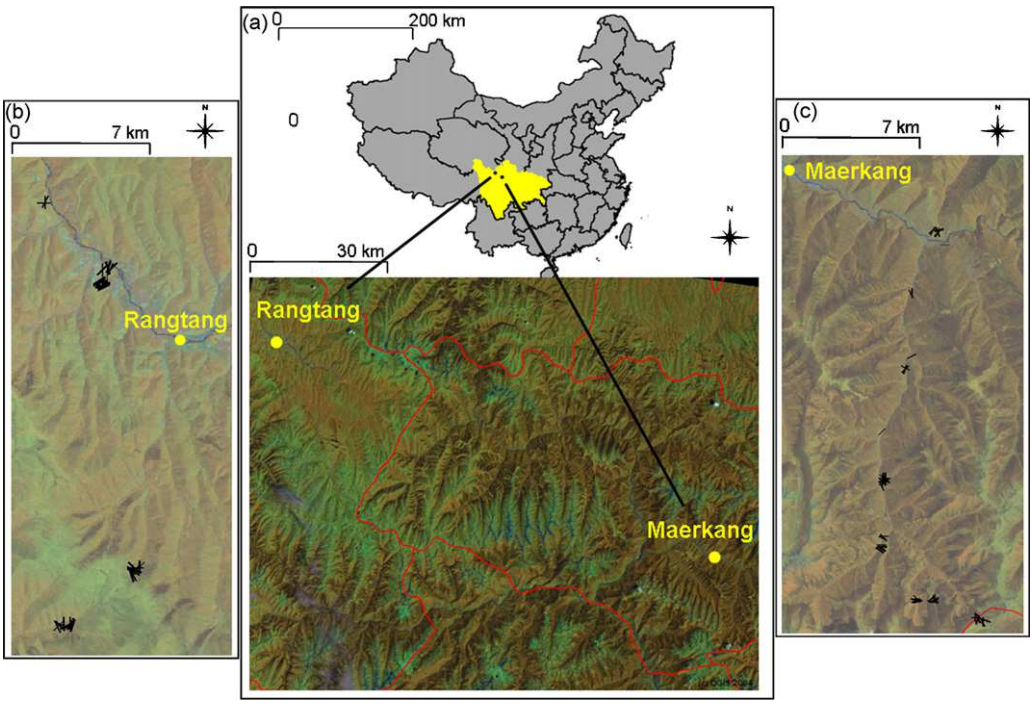

Fig. 1. (a) Location of Maerkang and Rangtang study areas in Sichuan province, China. Black and red lines delineate provinces and counties boundaries. Lines delineate province and county boundaries. (b) and (c) represent locations of traplines (black lines) around Rangtang and Maerkang cities, respectively. Locations are plotted on a false colour composite satellite image in which red, green and blue correspond to Landsat ETM bands 4, 5 and 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

consisted of 25 traps spaced 3 m apart. Two types of traps were used: small break back traps (SBBT), for animals not heavier than 100 g and big break back traps (BBBT) for larger individuals. Each trap was set for three nights (unless non-controlled factors dictated otherwise, e.g. trap theft), checked every morning and re-set/re-baited as necessary. We use the term *trap-night* to refer to a single trap set for one night. A total of 8095 trap-nights (4603 in Maerkang; 3492 in Rangtang) in 122 traplines (66 in Maerkang; 56 in Rangtang) were set; differential trapping effort by habitat is reported in Table 2.

Species were identified at the Centre de Biologie et Gestion des Populations (J.P. Quéré) using the following references: Corbet (1978); Fen and Zheng (1985); Gromov and Polyakov (1978); Gromov and Erbajeva (1995) and Smith and Xie (2008). *Apodemus penninsulae*, *Apodemus draco* and *Apodemus latronum* identifications were confirmed using cytochrome b sequencing. Nomenclature followed Wilson and Reeder (2005).

2.2. Assemblage definition

2.2.1. Statistical model

The response variable, the category (i.e. species or empty trap) observed in a given trap-night, was assumed to follow a multinomial distribution (Raoul et al., 2008). Relationships between habitat classes and species distribution were modelled using log linear multinomial regression with *a priori* selected habitat classes as explanatory variables (McCullag and Nelder, 1989). The three nights per trap and trap type were included as factors in the regression. Within each study site, a preliminary model comparison *via* AIC indicated no evidence for a night effect, so the factor was removed for further analyses.

2.2.2. Expert-class-merging procedure

Following Raoul et al. (2008), pairs of habitat classes (H_i and H_j) were "merged" by imposing an equality constraint on their regression coefficients (i.e. $\beta_{H_i} = \beta_{H_j}$). Iterative application of this constraint to all habitat class pairs lead to K new models which were compared via the difference in AICc (ΔAICc_{ij}) between each new model and the original (unconstrained) one. The model/merge providing the largest ΔAICc_{ij} identified the most redundant habitat class distinction and the two corresponding classes were fused into a single super-class. The process was iterated until no further evidence of class redundancy was observed, i.e. once $\Delta\text{AICc}_{ij}<2$(Burnham and Anderson, 1998). Finally, we computed the Simpson diversity index of each resulting assemblage. All statistical algorithms were implemented in R (R Development Core Team, 2008). Multinomial models were fitted using the `multinom` function of the `nnet` package.

2.3. Assemblage habitat modelling

2.3.1. Environmental descriptors

The environment at trapline locations was described using remotely sensed data layers corresponding to: spectral responses of Landsat Enhanced Thematic Mapper (ETM) bands; elevation; slope; and sun exposure. A (July 2005) multi-spectral ETM image from Landsat 7 was obtained from the U.S. Geological Survey (`landsat.usgs.gov`). ETM bands 3, 4 and 7, corresponding to red, near-infrared and middle-infrared, respectively, each with a 30 m resolution, were used for the analyses. Other ETM bands were omitted from further analysis due to strong correlations with the selected bands. Using the ETM image, the normalized difference vegetation index (NDVI) and Enhanced Vegetation Index (EVI) were computed (Huette et al., 1999). NDVI quantifies the difference between photosynthetic activity related absorption in the visible range and reflectance in the near-infrared which is related to electro-magnetic emission by plants. NDVI is thus correlated to vegetation biomass. However, the NDVI is known to suffer from saturation when vegetative biomass is high and to be sensitive to canopy background in open forested areas (Huette et al., 2002). Therefore, we computed EVI, defined as

$$EVI = \frac{G \times (NIR - RED)}{L + NIR + C_1 \times RED - C_2 \times BLUE}$$

where L is the canopy background and snow correction, C_1 and C_2 are coefficients of aerosol "resistance" terms and G the gain factor with $L = 1$, $C_1 = 6$, $C_2 = 7.7$ and $G = 2.5$.

A digital elevation model (DEM) was obtained from the SRTM program (http://srtm.usgs.gov/). Slope, aspect and elevation were derived from the DEM in GRASS GIS (http://grass.itc.it/). The sun index (SI), which provides a proxy for net incident solar radiation, was estimated as (Gibson et al., 2004):

$$SI = \cos(aspect) \times \tan(slope) \times 100.$$

In general traplines were not confined within the bounds of single pixels. In order to construct the data matrix the mean of each variable was estimated from 10 points regularly spaced along each trapline. This was performed using `spgrass6`, the R/GRASS interface available at http://r-spatial.sourceforge.net/xtra/xtra.RHnw.html.

2.3.2. Discriminant models

We explored if the set of environmental descriptors could: explain small mammal assemblage distributions locally, i.e. within each study area; and discriminate between the two study areas using data from trapline locations, a regional level analysis. Essentially this entailed discrimination of previously identified assemblages/traplines in the environmental space and thus presents a supervised classification problem. Model performances of two currently popular methods were compared: discriminant analysis (DA) and logistic multiple regression (LMR). Both techniques were tested: in linear form, linear discriminant analysis (LDA) (Fisher, 1936) and multinomial generalized linear model (GLM); and in non-linear form, mixture discriminant analysis (MDA) and Multiple Adaptive Regression Splines (MARS).

2.3.2.1. MARS. Multiple adaptive regression splines (Friedman, 1991) enable modeling multiple categorical responses with weak assumptions on the shape of the response. The regression modeling takes place in an enlarged space defined by the set of basis functions, i.e non-linear transformations of covariates, and their products, enabling identification of complex non-linear responses (Hastie et al., 2001, 1994). In the context of species distribution modelling, MARS incorporated within GLMs has been shown to outperform other non-parametric and non-linear modelling techniques on real (Leathwick et al., 2006) and simulated (Moisen and Frescino, 2002) data sets. Here we test performance of MARS incorporated into a multinomial GLM.

2.3.2.2. MDA. Non-linear decision boundaries in classification problems can also be modelled indirectly, i.e. without specifying and parameterising its form in the model. Mixture Discriminant Analysis is such an approach, generalising the linear decision boundaries of LDA by incorporating mixtures of multivariate Gaussians to discriminate a single class (Hastie and Tibshirani, 1993). MDA divides a given class into multiple subclasses each following a multivariate normal distribution with unique mean vector and a common covariance matrix. As far as we know, this method has been rarely tested in the context of niche spatial modelling except by Ter Braak et al. (2003). Here, maximum likelihood-based param-

eter estimation was achieved using the expectation maximisation (EM) algorithm (cf. Appendix A). Starting values for the EM algorithm were obtained by the Learning Vector Quantisation clustering algorithm (Hastie and Tibshirani, 1993).

All computations were performed in R (R Development Core Team, 2008). LDA was computed using the `lda` function of the "MASS" package. MARS and MDA were computed using the `mars` and `mda` functions of the `mda` package.

2.3.3. Evaluation and comparison of modelling techniques' performances

In order to compare modelling techniques on the basis of their statistical properties, we only compared models of similar structure in their explanatory components, i.e. including all available explanatory variables (Segurado and Araujo, 2004; Potts and Elith, 2006).

2.3.3.1. Evaluation criteria. Following Pearce and Ferrier (2000), models were evaluated on their ability to (i) provide reliable predictions (as soft classifiers) and (ii) discriminate between occupied and unoccupied sites (as hard classifiers). Prediction reliability was quantified using residual deviance (RD) (Pearce and Ferrier, 2000; Leathwick et al., 2006) equivalent to minus twice the log likelihood of the fitted probabilities, i.e. for N observations

$$RD = -2 \times \sum_{i=1}^{N} \log Pr(\hat{y}_i = y_i).$$

Discriminative performance (ii) is composed of omission (true positive fraction) and commission (false positive fraction) error (Pearce and Ferrier, 2000; Anderson et al., 2003). Here, predicted assemblages were mutually exclusive, i.e. the commission error of one assemblage corresponded to the omission error of another. Consequently, discriminative performance was assessed by omission error only using the true presence misclassification error rate (Segurado and Araujo, 2004).

2.3.3.2. Model validation. Predictive ability was assessed using the above mentioned criteria on independently re-sampled testing data sets using the bootstrap 632+ (Efron and Tibshirani, 1995) which provides the least biased and variant model re-sampling evaluation method (Hastie et al., 2001). Bootstrap 632+ was developed to reduce optimistic error estimation that arises with the original bootstrap due to commonality between the observations of training and testing data sets. Originally used to estimate error rates of per-subject prediction rule (e.g. misclassification error rate) it has recently and successfully been applied to evaluate predictive models of species or disease distribution (Steyerberg et al., 2001; Leathwick et al., 2006; Potts and Elith, 2006). We computed the bootstrap 632+ in R (R Development Core Team, 2008), using the `bootpred` function of the `boostrap` package and the `errorest` function of the `ipred` package. For each criteria the mean and standard deviation across 1000 bootstrap samples was calculated.

2.3.4. Effects of environmental descriptors on assemblage distributions

To simplify the assessment of effect size we selected the classification methods that provided the lowest bootstrapped misclassification error rate and residual deviance.

2.3.4.1. Effect size. Multicollinearity between factors is known to inflate the variance of regression coefficients, alter model predictive performance and complicates identification of real effects of factors on data variability (Legendre and Legendre, 1998). Hierarchical partitioning (MacNally, 2000, 2002), widely used in explanatory modelling (Gibson et al., 2004; Greaves et al., 2006; Olivier et al., 2000), permits to partition the variance explained by a model in order to isolate the independent *versus* joint contributions of each variable. We used this method to estimate the independent effects of each variable on the log-likelihood. Significance of these independent effects was assessed using a permutation test. The independent effect was re-estimated using 100 random permutations of the covariates, the mean and standard deviation of the resulting distribution permitting a *z-test* of the original independent effect.

The effect size of each environmental descriptor on the probability of observing each assemblage were assessed. All possible additive combinations of variables were fitted. For each combination we calculated the mean predicted probability of each assemblage at sites where that assemblage was observed. The difference in predicted probabilities

$$\Delta\mu_{j10} = \mu_{j1} - \mu_{j0}$$

were computed, where μ_{j1} and μ_{j0} represent the mean predicted probabilities across models which included and excluded the j th variable, respectively. Means and 95% quantile intervals were estimated over 200 cross-validation iterations. Predictive effects were considered for further analysis if the lower 95% quantile was greater than zero.

2.3.4.2. Response shape along environmental gradients. The shape of assemblage responses were visualised by plotting predicted probabilities with respect to each relevant environmental variable (Elith et al., 2005). Predictions of assemblage occurrence probabilities were made across the range of the variable of interest whilst all other variables were fixed to their mean value among sites corresponding to the assemblage in question.

All computations were made in R (R Development Core Team, 2008). Independent effects were estimated using the `combos` and `partition` functions of the `hier.part` package. The *z.test* was adapted from the `rand.hp` function of the same package.

3. Results

3.1. Trapping success: small mammal data

A total of 173 small mammals were trapped in Rangtang (90) and Maerkang (83) including 10 rodent, 1 lagomorph and 4 insectivore species (Table 1). Four species were trapped in both areas:

Table 1
Sampling success for each species (number of trapped individuals) obtained in each study area. The order of each species is indicated: rodent (R), insectivore (I) or lagomorph (L).

	Order	Location	
		Maerkang	Rangtang
Empty traps		4516	3396
Apodemus draco (Apdr)	R	17	0
Apodemus latronum (Apla)	R	14	0
Apodemus peninsulae (Appe)	R	3	68
Eospalax fontanieri (Eofo)	R	1	0
Eozapus setchuanus (Eose)	R	1	1
Microtus irene (Miir)	R	1	1
Microtus limnophilus (Mili)	R	0	13
Micromys minutus (Mimi)	R	6	0
Niviventer confucianus (Nico)	R	19	0
Sicista concolor (Sico)	R	1	0
Chodsigoa hypsibia (Chhy)	I	10	0
Sorex cylindricauda (Socy)	I	1	0
Sorex thibetanus (Soth)	I	2	0
Uropsilus soricipes (Urso)	I	2	0
Ochotona cansus (Occa)	L	5	7

Table 2
Maerkang (M1–M4) and Rangtang (R1–R4) assemblages obtained from expert-class-merging procedure. The number of traplines (Nb lines), habitat composition, number of trap-nights per sampled habitat (Nb trap-nights), species trapped (dominant and other) and Simpson diversity index are reported for each assemblage.

Assemblage	Nb lines	Habitat composition	Nb Trap-nights	Dominant species	Others species	Simpson diversity
M1	20	Bushes north slope	75	*Niviventer confucianus*	*Chodsigoa hypsibia*	2.60
		Bushes former terrace	744		*Apodemus latronum*	
		Bushes birch culture	300			
		Grassland	447			
		Forest oak near Village	301			
		Set aside field	749			
		Stream near culture	300			
M2	7	Culture	1050	*Micromys minutus*	*Eozapus setchuanus*	2.08
					Eospalax fontanieri	
					Apodemus latronum	
M3	34	Forest regeneration and plantation	719	*Ochotona cansus*	*Apodemus peninsulae*	7.86
		Stream bushes and trees	325		*Apodemus draco*	
		Forest coniferous and willow bushes	600		*Apodemus latronum*	
		Forest oak	444		*Chodsigoa hypsibia*	
		Stream willow bushes	294		*Sorex thibetanus*	
		Stream bushes	444		*Sorex cylindricauda*	
		Forest rhododendron	746		*Uropsilus soricipes*	
		Forest rhododendron and coniferous	747		*Sicista concolor*	
					Microtus irene	
M4	5	Forest white birch wet	306	*Apodemus draco*	*Apodemus latronum*	1.97
		Forest red birch understorey	438		*Niviventer confucianus*	
					Uropsilus soricipes	
R1	23	Field bank	717	*Apodemus peninsulae*	*Microtus limnophilus*	1.07
		Slope grass and sparse bushes	300		*Eozapus setchuanus*	
		Field bank	369			
		Culture border	671			
R2	18	Bottom valley grassland and bushes	600	*Apodemus peninsulae*	*Microtus limnophilus*	1.25
		Bottom valley grassland and bushes (Namuda)	577			
		Village garden	599			
		Forest coniferous and willow bushes	898			
R3	2	Fenced grassland	48	*Microtus limnophilus*	*Apodemus peninsulae*	1.25
R4	14	Forest rhododendron and coniferous	175	*Ochotona cansus*	*Microtus irene*	2.48
		Slope bushes	746		*Apodemus peninsulae*	
		Stream bushes	1041		*Microtus limnophilus*	

Ochotona cansus, *Eozapus setchuanus*, *Apodemus peninsulae*, and *Microtus irene*.

3.2. Assemblage definition

3.2.1. Maerkang

Four assemblages were identified (M1–M4) (Table 2, Fig. C.1). Assemblage M1, which included seven habitats subjected to human influences, was dominated by the rat *Niviventer confucianus*. M2, the Culture habitat, was not merged with other classes and was dominated by *Micromys minutus* which was specific to that habitat. M3 included forest and bush was the richest ($n = 11$) and most diversified (7.86) assemblage. Trapping probabilities were an order of magnitude lower in M3 than in other assemblages and nothing was trapped in Forest Oak and Stream bushes. M4 grouped white and red birch forests and had the lowest diversity. Trapping frequencies of *Apodemus draco* in M4 were greater than for any other species in any other assemblage.

3.2.2. Rangtang

Four assemblages were identified (R1–R4) (Table 2, Fig. C.2). R1, which included habitats in close vicinity to culture, was dominated by *Apodemus peninsulae* and provided the lowest diversity among all assemblages. In R2, which included valley bottom bushes, Forest coniferous/willow bushes and Village garden, *Apodemus peninsulae* was dominant but trapped with lower probability than in R1. R3 corresponded to the Fenced grassland class which was not merged with other classes. Only two species were found: *Microtus limnophilus* which was trapped at highest probability in R3, and *Apodemus peninsulae*. R4 grouped Forest rhododendron/coniferous with Slope and Stream bushes. Diversity was highest in this assemblage and *Ochotona cansus* was specific to it.

Details of AICc and Δ AICc evolution according to the number of the expert-class-merging procedure iterations, for Maerkang and Rangtang study areas, are available in Fig. B.1.

3.3. Assemblage habitat modelling

Modelling techniques were applied to discriminate all assemblages previously defined except assemblage R3 since it included only 2 traplines (Table 2).

3.3.1. Modelling techniques performances (Table 3)

3.3.1.1. Discriminant analysis. Concerning Maerkang assemblages and study sites discrimination, bootstrapped RD and error rate of MDA were approximately twice the RD and error estimated on training data. MDA was a more discriminant and reliable predictor than LDA in terms of lower bootstrapped error rates (Δ Error $(Maerkang)_{LDA-MDA} = 0.012$; $\Delta Error(Region)_{LDA-MDA} = 0.038$) and RD ($\Delta RD(Maerkang)_{LDA-MDA} = 5.831$; $\Delta RD(Region)_{LDA-MDA} = 9.672$). In Rangtang, training data-based estimates indicated MDA gave a more deviant fit than LDA although again bootstrapped RD ($\Delta RD(Rangtang)_{LDA-MDA} = 1.99$) and error ($\Delta Error(Rangtang)_{LDA-MDA} = 0.043$) were lower than for LDA.

Mixture discriminant analysis was also a more discriminatory and reliable model than the multinomial GLM in the Maerkang ($\Delta Error(Maerkang)_{MN-MDA} = 0.055$; $\Delta RD(Maerkang)_{MN-MDA} = 2.089$) and regional level ($\Delta Error(Region)_{MN-MDA} = 0.028$; $\Delta RD(Region)_{MN-MDA} = 10.781$) analysis. However, in Rangtang, MN outperformed MDA ($\Delta Error(Rangtang)_{MDA-MN} = 0.012$; $\Delta RD(Rangtang)_{MDA-MN} = 20.704$).

Finally, in each of the three analysis, MARS perfectly fitted the training data ($RD_{MARS,train} = 0.00$; $Error_{MARS,train} = 0.00$). Despite this overfitting, MARS was the most reliable and discriminatory

Table 3
Residual deviance (RD) and misclassification error rates (Error) obtained for each modelling procedure (linear discriminant analysis (LDA), mixture discriminant analysis (MDA), multinomial logistic regression (MN) and multiple adaptive regression spline (MARS)), for each study case, on training (train) and testing (boot) data sets. For the test data set, means (boot mean) and standard deviances (boot sd), estimated over 1000 bootstrap 632+ iterations, are indicated.

	LDA		MDA		MN		MARS	
	RD	Error	RD	Error	RD	Error	RD	Error
Maerkang								
Train	86.645	0.212	55.251	0.121	70.886	0.197	0.001	0.000
Boot mean	111.322	0.270	105.491	0.258	107.589	0.313	53.585	0.221
Boot sd	0.092	0.001	0.109	0.001	0.084	0.001	0.049	0.000
Rangtang								
Train	46.815	0.182	71.922	0.073	13.060	0.055	0.000	0.000
Boot mean	67.082	0.308	65.095	0.265	44.391	0.253	41.144	0.232
Boot sd	0.092	0.001	0.069	0.001	0.129	0.001	0.046	0.001
Study sites								
Train	51.693	0.106	17.506	0.033	41.230	0.081	0.000	0.000
Boot mean	56.292	0.111	46.620	0.073	57.401	0.101	16.927	0.029
Boot sd	0.001	0.000	0.012	0.000	0.027	0.000	16.927	0.029

model, providing bootstrap predictor errors 0.037, 0.021 and 0.044 lower than the next most discriminant models, outlined above, of Maerkang assemblages (MDA), Rangtang assemblages (MN) and study sites (MDA), respectively.

3.3.2. *Effects of environmental descriptors on assemblage distributions*

3.3.2.1. Maerkang. Significant independent contributions to model goodness-of-fit were observed for elevation and ETM band 7 (Table 4). Among all variables, elevation provided the largest independent contribution (I%= 39.7). It increased predicted probabilities for all assemblages bar M2 on training and bootstrapped data sets (Table 5). A threshold was observed at about 3600 m altitude above which M1 was seldom observed, the probability of M2 and M4 dropped rapidly and M3 became predominant (Fig. 2). Enhanced Thematic Mapper band 7 consistently discriminated all assemblages. The probability of M2 (cultures) reached a distinct peak at intermediate values (40) where the probability of M4 approached zero.

Despite the non-significance of their independent contributions, vegetation indices influenced mean predicted probabilities (Tables 4 and 5). On average, inclusion of NDVI increased predicted probabilities of M1 by 0.087. A change was observed at NDVI = 0.45 with M1 being more probable than M2 or M4 below and less probably than other assemblages above this point, respectively (Fig. 2).

Table 4
Percentage of the independent contribution over all independent effects (I %) and total contributions (Total) in model goodness of fit estimated for each variable and for each study case. The statistical significance (sig.) of the Z.score based on upper 0.95 confidence limit is indicated by (*).

	ETM7	Elevation	Slope	SI	NDVI	EVI
Maerkang						
I%	20.912	39.196	11.461	2.208	11.068	15.156
Z.score	2.13	9.36	0.17	−0.15	−0.03	0.52
sig	*	*	ns	ns	ns	ns
Total	23.096	39.691	15.099	3.601	11.091	16.627
Rangtang						
I%	16.410	23.318	21.160	2.263	29.058	7.790
Z. score	1.9	5.88	4.13	−1.1	2.92	−0.27
sig	*	*	*	ns	*	ns
Total	17.251	18.888	9.694	2.369	26.061	5.812
Region						
I %	26.842	37.132	8.529	0.316	21.024	6.156
Z. score	12.68	16.4	3.79	0.19	7.35	2.74
sig	*	*	*	ns	*	*
Total	31.888	58.427	16.210	−0.549	39.418	13.850

Inclusion of EVI significantly increased the probability to predict M2 and M4 whose bootstrapped means increased by 0.104 and 0.093, respectively. The range of EVI corresponding to M2 occurrence being lower than that of M4 (birch forests). Finally, slope and SI had no predictive power.

3.3.2.2. Rangtang. Independent effects of NDVI, elevation, ETM band 7 and slope were all significant (Table 4). NDVI provided the highest independent contribution (I%=29.058) and significantly improved mean predicted probabilities for all assemblages, this improvement being greatest for R4 (Forest rhododendron/coniferous and Slope bushes) (0.30) which was not observed when NDVI exceeded 0.25 (Table 5, Fig. 3). By contrast, above this threshold, the probability of R2, and to a lesser extent R1, dropped considerably.

Elevation increased predicted probabilities for all assemblages, the size of the effect being greatest for R1 (fields and bushes) which was predominant at lower altitudes and did not occur above 3650 m. Above this level, the probability of R4 reached a plateau and the probability of R2 also increased. Inclusion of ETM band 7 significantly increased prediction accuracy for R1 and R4. Band 7 values greater than 70 corresponded to predominant and subordinant predicted probabilities of R2 and R4, respectively, the latter being optimally distributed in the range 50–65. Inclusion of slope significantly increased predictive probabilities at R2, R1 and R4 sites (ordered by decreasing effect size). R2 was the predominant assemblage in flat areas and R1 became predominant when slopes became steeper than 10°. Finally, SI and EVI had no predictive effects.

3.3.2.3. Between study sites. All variables had significant and independent contributions in model goodness-of-fit except SI which negatively contributed in the model likelihood (Table 4). Rangtang was associated with ETM band 7 below 50, steep slopes, EVI above 0, NDVI below 0.4 and elevations in the range 3600–3900 m. Maerkang was associated with ETM band 7 below 50, NDVI above 0.4 and elevations below and above 3500 m and 3900 m, respectively (Fig. 4).

4. Discussion

4.1. *Assemblages as mixtures of species distributions*

The ability of the expert-class-merging procedure to cluster observations from a large number of habitats into a small number of super-classes highlights that apparently different habitats

Table 5
Contributions of each variable to change in the mean predicted probabilities of each assemblage at sites where it was observed. Estimates based on training data (mean train), cross-validated means (Mean CV P1-P0) and confidence intervals (qt0.05 CV P1-P0) are shown. ∗ indicated variable effects with lower 95% quantile greater than zero, i.e significant effetcs.

	Maerkang				Rangtang		
	M1	M2	M3	M4	R1	R2	R4
ETM 7							
Mean train	0.158	0.254	0.096	0.255	0.101	0.113	0.098
Mean CV P1-P0	0.107	0.143	0.049	0.212	0.030	0.003	0.015
qt0.05 CV P1-P0	0.079	0.068	0.021	0.133	0.008	−0.017	0.000
qt0.95 CV P1-P0	0.140∗	0.234∗	0.073∗	0.293∗	0.053∗	0.018	0.029∗
Elevation							
Mean train	0.363	0.133	0.278	0.433	0.161	0.128	0.123
Mean CV P1-P0	0.250	0.060	0.329	0.283	0.207	0.073	0.127
qt0.05 CV P1-P0	0.212	−0.007	0.290	0.160	0.178	0.038	0.078
qt0.95 CV P1-P0	0.278∗	0.121∗	0.357∗	0.364∗	0.237∗	0.102∗	0.163∗
Slope							
Mean train	0.076	0.092	0.038	0.197	0.150	0.228	0.071
Mean CV P1-P0	0.018	0.028	−0.021	0.036	0.190	0.307	0.052
qt0.05 CV P1-P0	−0.012	−0.050	−0.044	−0.046	0.159	0.271	0.028
qt0.95 CV P1-P0	0.049	0.075∗	0.003	0.096	0.222∗	0.342∗	0.079∗
SI							
Mean train	0.005	−0.003	0.003	0.012	−0.002	0.007	0.008
Mean CV P1-P0	−0.016	−0.042	−0.009	−0.030	−0.009	−0.006	−0.011
qt0.05 CV P1-P0	−0.032	−0.195	−0.019	−0.058	−0.011	−0.009	−0.017
qt0.95 CV P1-P0	−0.002	0.017	−0.001	−0.001	−0.007	−0.004	−0.008
NDVI							
Mean train	0.087	0.106	0.045	0.140	0.084	0.195	0.300
Mean CV P1-P0	0.034	0.011	−0.003	0.043	0.041	0.084	0.218
qt0.05 CV P1-P0	0.008	−0.030	−0.019	−0.006	0.014	0.057	0.186
qt0.95 CV P1-P0	0.063∗	0.066∗	0.014	0.092∗	0.069∗	0.111∗	0.245∗
EVI							
Mean Train	0.104	0.237	0.055	0.180	0.043	0.042	0.058
Mean CV P1-P0	0.029	0.104	0.000	0.093	−0.014	−0.014	−0.010
qt0.05 CV P1-P0	−0.008	0.037	−0.019	0.035	−0.024	−0.021	−0.016
qt0.95 CV P1-P0	0.063∗	0.179∗	0.022	0.169∗	−0.005	−0.006	−0.004

in fact provide approximately equivalent habitat quality for many small mammal species. However, the influence of a species in the assemblage definition is directly related to trapping frequency, the definition being dominated by those species trapped in largest number (e.g. *Apodemus peninsulae*), whilst rare or shy species trapped in low numbers (e.g. *Eozapus setchuanus*) play a relatively small role in the assemblage definition. It is well known that spatial distributions of small mammals can be driven by population dynamics (Giraudoux et al., 1997, 2007). The assemblages defined in the present study constituted a spatio-temporal snapshot of a process of complex interactions between multi-species metapopulations (Guisan and Thuiller, 2005). The current data set, with its lack of a temporal component, thus limits analyses to the description of "potential habitats" for the identified small mammal

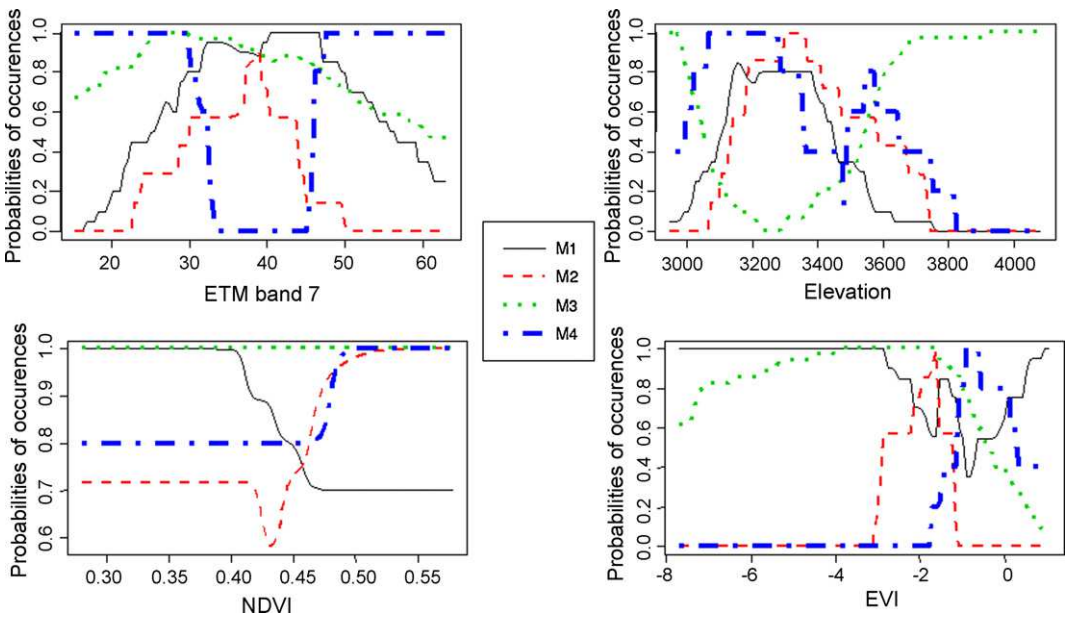

Fig. 2. Responses (predicted probabilities) of Maerkang assemblages along those continuous gradients for which predictive power for at least one assemblage was identified (Table 5).

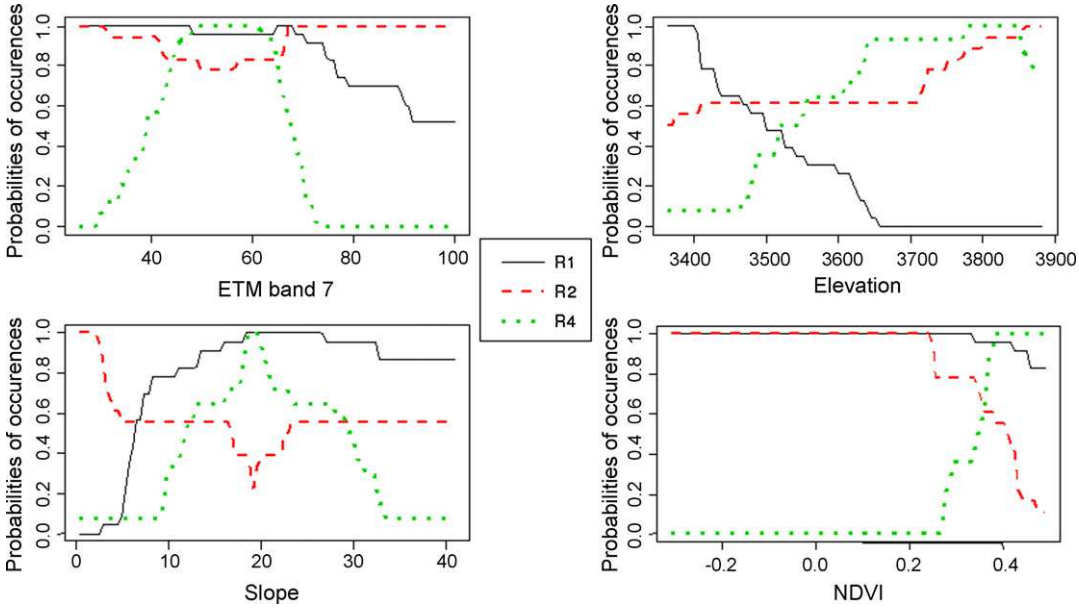

Fig. 3. Responses (predicted probabilities) of Rangtang assemblages along those continuous gradients for which predictive power for at least one assemblage was identified (Table 5).

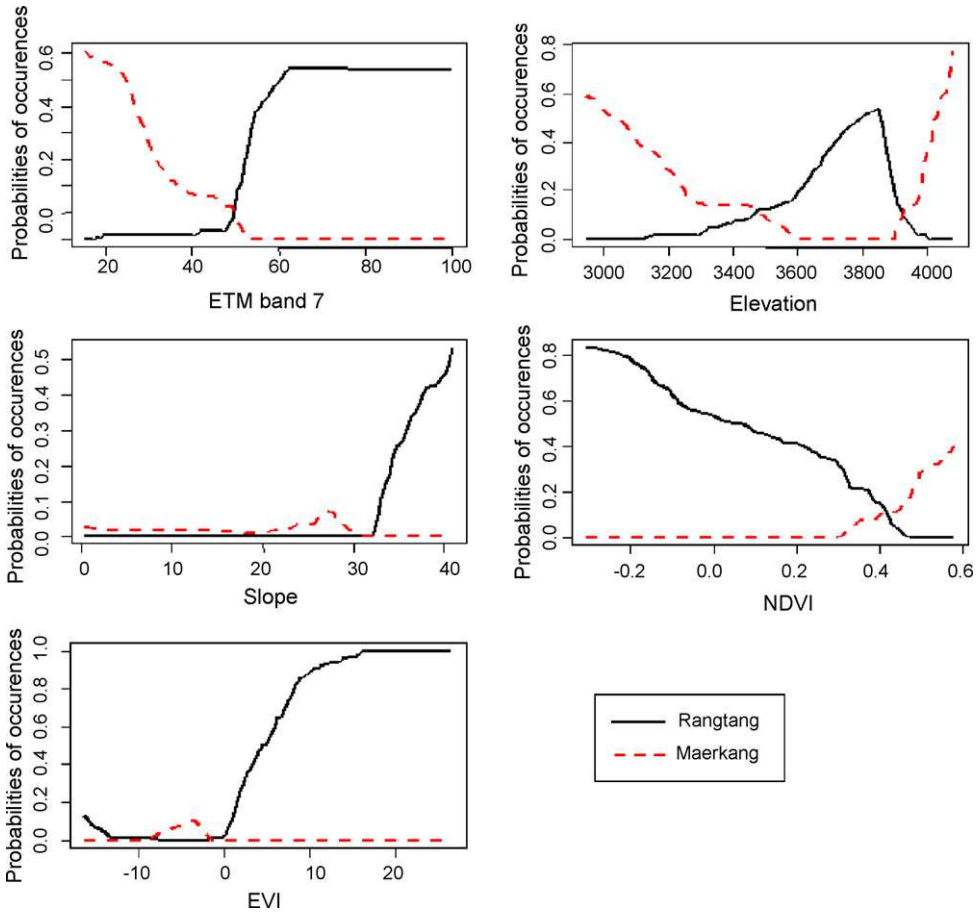

Fig. 4. Responses (predicted probabilities) of the two study areas along those continuous gradients for which predictive power for at least one assemblage was identified.

assemblages (Guisan and Thuiller, 2005; Araujo and Guisan, 2006; Guisan et al., 2006).

4.2. Assemblage habitat distribution in environmental space

4.2.1. Linear or non-linear discriminant analysis for small sample size problem?

Linear discriminant analysis is said to perform well when sample size is small because of the simple boundaries it provides between classes (Hastie et al., 1997). However, our analyses showed prediction reliability and class discrimination were consistently lower for MDA than for LDA. MDA's discrimination of assemblages in several sub-classes, permitting identification of non-linear boundaries, clearly helped capture important extra within class variability. This result re-enforces the need to use non-linear models in community modelling even when sample size is small (Munoz and Felicisimo, 2004).

MDA allows multi-modal representation of an assemblage in environmental space. By contrast, MN assumes assemblage responses can be represented as simple linear combinations of variables rescaled to the 0–1 range via a link function. Despite a larger number of parameters, MDA was more discriminant and less deviant than MN for discrimination of Maerkang assemblages and site environments. This result highlights the importance of modelling sub-class clusters in those case studies. By contrast, Rangtang assemblages were better discriminated by MN than by MDA suggesting a more homogeneous distribution of those assemblages in environmental space. MN was also a more performant predictor of Rangtang data than LDA despite the two models sharing the same logit regression form (Hastie et al., 2001). Regardless of linearity/non-linearity assumptions, discriminant analysis was not the most appropriate method for modelling the distribution of Rangtang assemblages for which sample sizes were smallest.

4.2.2. Parametric or non-parametric non-linear modelling of assemblage distributions?

Multiple adaptive regression splines advantageously extends the logistic response of multinomial GLMs to include multi-modal variation of the linear predictor with respect to a given variable. MARS has previously been shown to be a more appropriate mapping technique for *Fagus* species than logistic multiple regression (Munoz and Felicisimo, 2004) although more generally there has been difficulty to demonstrate better performance of MARS over linear models on real data sets (Moisen and Frescino, 2002). Here, MARS consistently provided more discriminatory and reliable predictions than LDA, MDA and GLM in all three of our cases studies. The non-linear structures in class distributions were more accurately captured by a model which did not rely on parametric assumptions regarding assemblage distribution in environmental space. These promising prediction results of MARS were obtained despite strong evidence of overfitting (i.e. selection of an excessive number of basis functions) the training data. This observation is encouraging since the superior predictive performance of MARS could clearly be improved *via* the increased parsimony obtainable with model selection procedures such as step-wise pruning (Friedman, 1991).

Assuming each *a priori* habitat was representable as a single cluster in environmental space, a mixture representation of an assemblage (i.e. super-class), including the diversity of species responses, and thus the requirement of non-linear decision boundaries for assemblage discrimination, becomes natural. However, processes underlying modelling of assemblages might have been complexified here in part from the fact that most of the variables in question were, at best, indirectly related to small mammal dynamic processes. Guisan and Zimmerman (2000) suggested that the use of direct environmental variables provides a physiologically mechanistic character to a model that is not apparent when indirectly related variables are used. We further suggest that the use of indirect variables in fact complicates response curves. Multi-modal responses in the space of indirect variables might correspond to relatively homogeneous responses among variables directly related to processes.

4.3. Explanatory and predictive power of environmental descriptors

4.3.1. Within versus between study area discrimination

The observed between site differences in small mammal fauna, environmental conditions and pertinence of environmental variables provided strong evidence that two distinct biogeographical zones had been identified. The degree of abruptness/smoothness in the transition between these zones was not identifiable given the scale of the sampling design and clearly further work is required before the process of small mammal communities on the fringes of the Tibetan plateau are fully elucidated.

The discrimination between Maerkang and Rangtang at high and low NDVI values, respectively suggests greater productivity in terms of vegetative biomass in Maerkang. This corresponds with our *a priori* land cover classification since forested habitats were more widespread and diversified in Maerkang than in Rangtang. By contrast, within study sites, NDVI and EVI were useful discriminators of different assemblages. In Maerkang, NDVI significantly improved predictions of M1 (bushes, grassland an oak forests near villages), while EVI helped discriminate birch forests (M4) and culture (M2). In Rangtang, NDVI discriminated forest (R4), valley-bottom vegetation (R2) and non-forested bush (R1), whereas EVI displayed no discriminative ability. The fact that NDVI was less useful in Maerkang, where 9 in 18 habitats were forested, than in Rangtang, where only 2 in 12 habitats were forested habitats, probably reflects the saturation problem, i.e. NDVI's limited ability to discriminate among forest types (Huette et al., 1999).

Variables such as elevation, ETM band 7 and slope had descriptive and predictive power but their indirect relation to small mammal resources renders interpretation of their effects difficult (regardless of the mixture problem outlined above). Moreover, these variables are subjected to the law of relative site constancy (Guisan and Zimmerman, 2000; Randin et al., 2006) which complicates the comparison of assemblage distributions between the two study areas along these gradients.

The discriminatory power of environmental variables differed according to the study area (Maerkang *versus* Rangtang) and to the spatial extent of the training area (local *versus* regional analysis). In a predictive mapping purposes, one should be aware of such variation in order to select a set of environmental variables appropriate for the required extent, resolution and location of the predictive map (Guisan and Thuiller, 2005).

4.3.2. Improving habitat definition

Numerous improvements to the current habitat descriptions can be envisaged. Firstly, numerous species have been shown to respond to landscape level effects whereby densities respond to composition or structure of landscape in a surrounding neighbourhood. Examples include *Tetrao urogallus* (Graf et al., 2005), *Echinococcus multilocularis* (Giraudoux et al., 2003), *Arvicola terrestris* (Fichet-Calvet et al., 2000; Giraudoux et al., 2007; Morilhat et al., 2007) and *Microtus arvalis* (Delattre et al., 1999, 2006; Duhamel et al., 2000), the latter two being small mammal species of the sub-family *Arvicolinae*. Here effects of composition and structure in buffers surrounding traplines were not considered, largely due to identifiability issues associated with the required increased level of data mining and the small sample size.

Secondly, trapline locations were often observed to lie within mixed pixels where numerous habitat types contribute to the observed spectral response. Improved predictive performance might therefore be achievable by addressing the mixed pixel problem for which fuzzy set approaches (Foody, 2000) and super-resolution techniques (Tatem et al., 2002) exist.

Thirdly, descriptors of structure and composition of the first vegetative strata, being more directly related to small mammal resources would be pertinent (Catling and Coops, 1999; Pearce et al., 2001; Gibson et al., 2004). This move to direct variables would be of particular relevance in a predictive modelling setting since indirect variables limit model transferability across large areas (Guisan and Thuiller, 2005; Randin et al., 2006). Understory modelling with very high resolution satellite remotely sensed data has recently been applied in sparse forests (Jianxi et al., 2007). Further, helicopter-borne laser scanner data has been used to provide high resolution DEMs from which understory structural and textural characteristics are easily extracted (Hirata et al., 2003). Further research is needed before remotely sensed indices of understory structure can be incorporated into small mammal species distribution models.

4.3.3. Toward building process-based models

As outlined above, statistical models can help to resolve some theoretical hypotheses on species/environment relationships (e.g. shape of the response curves, selection of influent environmental factors) and thus provide basis for the development of process-based models. Inversely, process-based models can serve statistical modelling and an iterative procedure incorporating both approaches had been advocated (Austin, 2007). Here, prediction precision might be increased once a deeper understanding of the ecological processes driving small mammal distributions on the fringes of the Tibetan plateau is achieved. It would be helpful to investigate landscape composition effects on small mammal communities in time and space. Ultimately, such relationships could be incorporated into a Structural Equation Modelling framework (Guisan et al., 2006; Austin, 2007).

4.4. Conclusion and perspectives

Our results showed mixtures of Gaussians provided better descriptions of the environmental space occupied be small mammal assemblages than single cluster model. However, the Gaussian mixture model was in turn outperformed by MARS with its flexible ability to detect non-linearity. ETM band 7, vegetation indices, elevation and slope all helped discriminate assemblages. Their predictive effects varied according to the location and spatial scale of the training area. However, predictive performance might be improved given further investigation of methodological issues and a deeper understanding of underlying ecological processes, i.e. given: a deeper knowledge of the responses of multiple-species to environmental variation; a closer match between spectrally derived variables and small-mammal resources; or, improve data set quality (sampling size). In our study, MARS was prone to over-fitting, thus its transferability and outperformance over other techniques should be assessed cautiously prior to extrapolation beyond the sampling area (Randin et al., 2006). Moreover, because it is a classifier (i.e. assumes that pre-defined groups exist), used here to model assemblage responses jointly, prediction of new (i.e. untrained) assemblages in space and time was not possible. Two distinct biogeographical areas regarding small mammal assemblages and environmental conditions might have been identified here, but the current sampling design does not permit elucidation of the transition between the two areas. Mapping small mammal assemblages in the area between Rangtang and Maerkang would require a sampling design spanning the ranges of each species within both the geographical and environmental space separating the two areas (Murphy, 2007).

Acknowledgments

This work was supported by grants number RFATW-00-002 and RO1 TW001565 from the Fogarty International Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Fogarty International Center or the National Institutes of Health. We address many thanks to our Chinese colleagues working for the Sichuan Center for Diseases Control of Chengdu for their logistic support. We express many thanks to Kurt Galbreath, Department of Ecology and Evolutionary Biology, Cornell University, Ithaca (New York), for his help in identification of *Ochotona* species. We address a special thanks to Stéphane Chrétien of the Department of Mathematics, UMR CNRS 6623, for his constructive critiques and corrections of the statistical component of this study. We also address many thanks to Alexandre Hirzel for his interest and attention in a pre-submission review of this paper.

Appendix A. MDA algorithm

The algorithm computes the mixture density for each class j divided in R_j subclasses each defined by their own mean $\mu_{(j,r)}$ and a common covariance matrix Σ, such as

$$m_{j(x)} = P(X = x|G = j) = |(2\pi)^d \Sigma|^{-1/2} \sum_{r=1}^{R_j} \pi_{jr} e^{-D_{(x\mu_{jr})}/2}$$

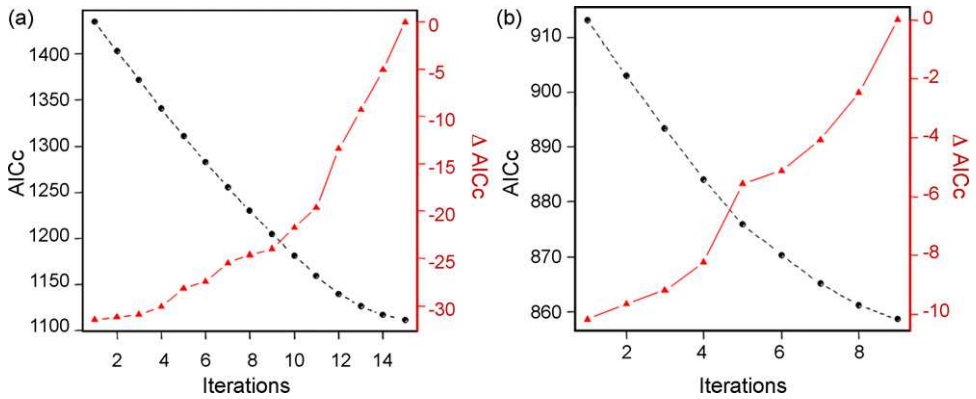

Fig. B.1. Redundancy reduction of the expert-class-merging procedure for (a) Maerkang and (b) Rangtang data sets illustrated by decreasing AICc and increasing Δ AICc w.r.t. to the number of iterations of the procedure.

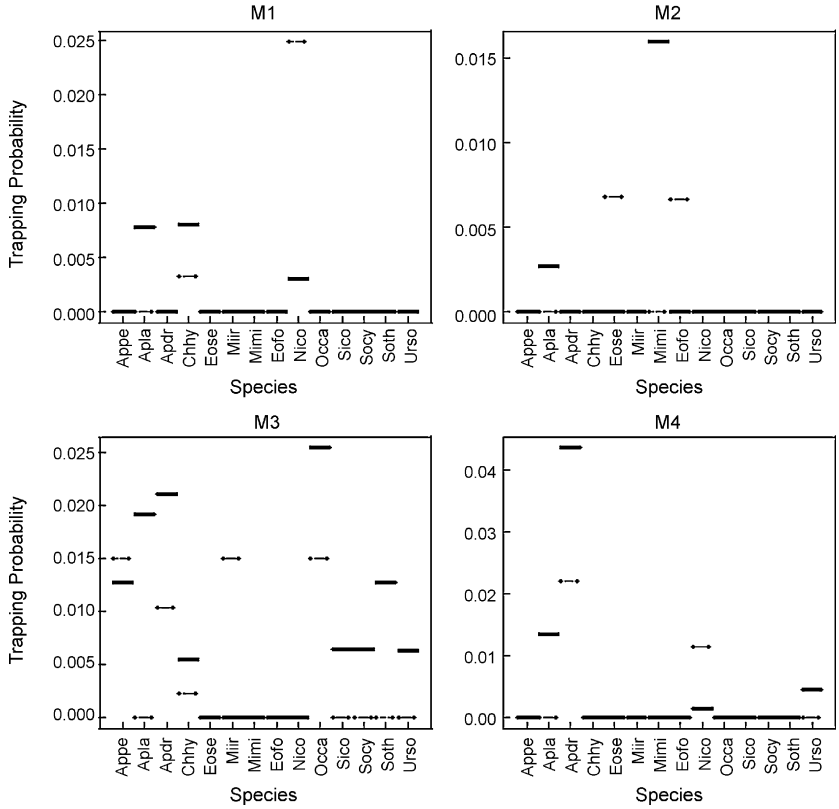

Fig. C.2. Species trapping joint probabilities predicted by the multinomial model for each assemblage (M1 to M4), in Maerkang. Dashes and full lines correspond to small and big traps, respectively. Correspondence between species names and their abbreviations is available in Table 2.

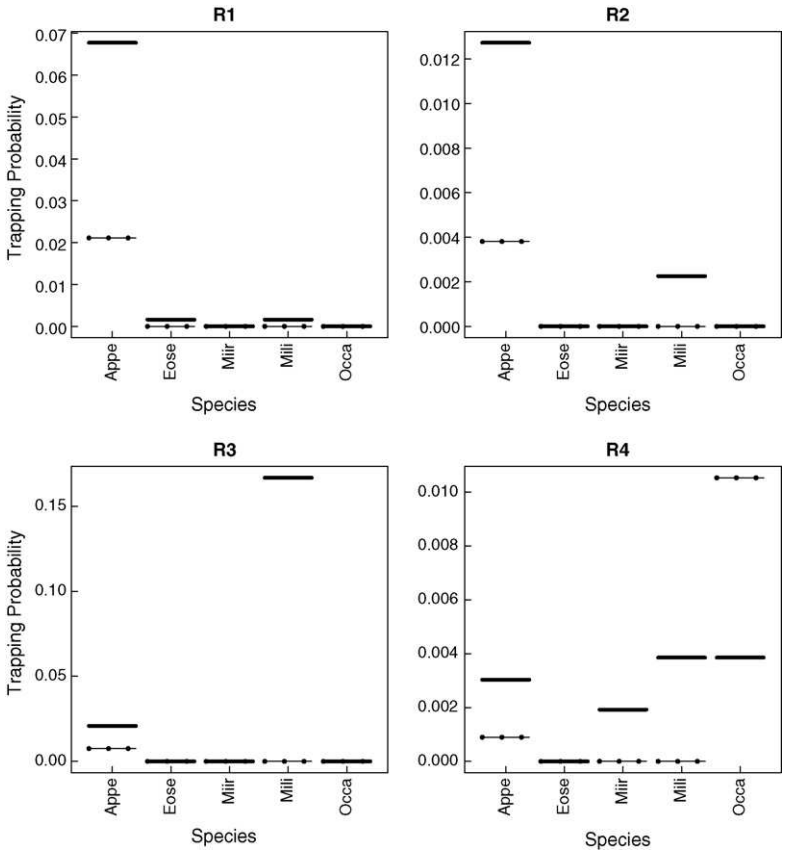

Fig. C.3. Species trapping joint probabilities predicted by the multinomial model for each assemblage (R1 to R4), Rangtang. Dashes and full lines correspond to small and big traps, respectively. Correspondence between species names and their abbreviations is available in Table 2.

where $D(x, y)$ is the Mahalanobis distance between x and the y and the mixing probability π_{jr} are unknown model parameters. Then, the conditional log-likelihood for the data l^{mix} is computed and maximised by the EM (expectation maximisation) algorithm with

$$l^{mix}(\mu_{rj}, \Sigma, \pi_{jr}) = \sum_{i=1}^{N} \log m_{gi}(x_i)$$

Expectation maximisation (EM) algorithm is an iterative two-step procedure: first, it estimates the mixing probabilities for each subclass and class and then, it conditionally optimises the mean and covariance for each subclasses.

Appendix B. Expert-class-merging procedure

See Fig. B.1.

Appendix C. Species predicted probability distributions in Maerkang and Rangtang assemblages

See Figs. C.1 and C.2.

References

Anderson, M., Clements, A., 2000. Resolving environmental disputes: a statistical method for choosing among competing cluster models. Ecol. Appl. 10 (5), 1341–1355.

Anderson, M., Lew, D., Peterson, A., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

Araujo, M., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. J. Biogeogr. 33, 1677–1688.

Austin, M., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecol. Model. 157 (18), 101–118.

Austin, M., 2007. Species distribution models and ecological theory: a critical assessment and some possible new approaches. Ecol. Model. 200, 1–19.

Burnham, K., Anderson, D., 1998. Model Selection and Multimodel Inference, 2nd ed. Springer-Verlag, New York.

Butet, A., Paillat, G., Delettre, Y., 2006. Factors driving small rodents assemblages from field boundaries in agricultural landscapes of western France. Landsc. Ecol. 21, 449–461.

Catling, P.C., Coops, N.C., 1999. Prediction of the distribution and abundance of small mammals in the eucalypt forests of south-eastern Australia from airborne videography. Wildl. Res. 641, 650.

Corbet, G., 1978. The mammals of the Palearctic Region: a taxonomic review. British Museum (Natural History). Cornell University Press, Londres.

Delattre, P., Clarac, R., Melis, J., Pleydell, D., Giraudoux, P., 2006. How moles contribute to colonization success of water voles in grassland: implications for control. J. Appl. Ecol. 43 (2), 353–359.

Delattre, P., De Sousa, B., Fichet-Calvet, E., Quéré, J., Giraudoux, P., 1999. Vole outbreaks in a landscape context: evidence from a six year study of microtus arvalis. Landsc. Ecol. 14 (4), 401–412.

Doledec, S., Chessel, D., Gimaret-carpentier, C., 2000. Niche separation in community analysis: a new method. Ecology 81, 2914–2927.

Duhamel, R., Quéré, J., Delattre, P., Giraudoux, P., 2000. Landscape effects on the population dynamics of the fossorial form of the water vole (Arvicola terrestris sherman). Landsc. Ecol. 15 (2), 89–98.

Efron, B., Tibshirani, R., 1995. Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Technical report (tr-477), Dept. of Statistics, Stanford University.

Elith, J., Ferrier, S., Huettmann, F., Leathwick, J., 2005. The evaluation strip: a new robust method for plotting predicted responses from species distribution models. Ecol. Model. 186, 280–289.

Fen, Z., Zheng, C., 1985. Studies on the pikas (genus Ochotona) of China. Taxonomic and distribution. Acta Theriol. Sin. 5, 269–289.

Ferrier, S., Guisan, A., 2006. Spatial modelling of biodiversity at the community level. J. Appl. Ecol. 43 (3), 393–404.

Fichet-Calvet, E., Pradier, B., Quéré, J., Giraudoux, P., Delattre, P., 2000. Landscape composition and vole outbreaks: evidence from an eight year study of Arvicola terrestris. Ecography 23 (6), 659–668.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eug. 7, 179–188.

Foody, G., 2000. Estimation of sub-pixel land cover composition in the presence of untrained classes. Comput. Geosci. 26, 469–478.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Statist. 19, 1–67.

Gibson, L., Wilson, B., Aberton, J., 2004. Landscape characteristics associated with species richness and occurrence of small native mammals inhabiting a coastal heathland: a spatial modelling approach. Biol. Conserv. 120, 75–89.

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

Giraudoux, P., Delattre, P., Habert, M., Quéré, J., Deblay, S., Defaut, R., Duhamel, R., Moissenet, M., Salvi, D., Truchetet, D., 1997. Population dynamics of fossorial water vole (Arvicola terrestris scherman): a land use and landscape perspective. Agric. Ecosyst. Environ. 66, 47–60.

Giraudoux, P., Pleydell, D., Raoul, F., Vaniscotte, A., Ito, A., Craig, P.S., 2007. Echinococcus multilocularis: why are multidisciplinary and multiscale approaches essential in infectious disease ecology? Trop. Med. Health 35 (4), 293–299.

Giraudoux, P., Quéré, J., Delattre, P., Bao, G., Wang, X., Shi, D., Vuitton, D., Craig, P., 1998. Distribution of small mammals along a deforestation gradient in Southern Gansu, central China. Acta Theriol. 43 (4), 349–362.

Graf, R., Bollmann, K., Suter, W., Bugmann, H., 2005. The importance of spatial scale in habitat models: Capercaillie in the swiss alps. Landsc. Ecol. 20 (6), 703–717.

Greaves, R., Sanderson, R., Rushton, S., 2006. Predicting species occurrence using information-theoretic appraoches and significance testing: an example of dormouse distribution in Cumbria, UK. Biol. Conserv. 130, 239–250.

Gromov, I.M., Erbajeva, M.A., 1995. The Mammals of Russia. Russian Academy of Sciences, St Petersburg.

Gromov, I.M., Polyakov, I.Y., 1978. Voles (*Microtinae*). Fauna of the USSR, Mammals. vol. 3. Brill E.J., Publishing Company.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J.C., R, A., T, H., 2006. Making better biogeographical predictions of species' distributions. J. Appl. Ecol. 43 (3), 386–392.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009.

Guisan, A., Zimmerman, N., 2000. Predictive habitat distribution models in ecology. Ecol. Model. 135, 147–186.

Hastie, T., Tibshirani, R., 1993. Discriminant analysis by gaussian mixtures. J. R. Stat. Soc. Ser. B (Methodological) 58 (1), 155–176.

Hastie, T., Tibshirani, R., Buja, A., 1994. Flexible discriminant analysis by optimal scoring. J. Am. Stat. Assoc. 89 (428), 1255–1270.

Hastie, T., Tibshirani, R., Buja, A., 1997. Flexible discriminant and mixture models. In: Kay, J., Titterington, D. (Eds.), Neural Networks and Statistics. Oxford University Press.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning. Springer-Verlag.

Hill, M.O., 1979. DECORANA - A FORTRAN Program for Detrended Correspondence Analysis and Reciprocal Averaging. Ecology and Systematics. Cornell University, Ithaca, NY.

Hirata, Y., Sato, K., Sakai, A., Kuramoto, S., Akiyama, Y., 2003. The extraction of canopy-understory vegetation-topography structure using helicopter-borne lidar measurement between a plantation and a broad-leaved forest. In: Geoscience and Remote Sensing Symposium, 2003 IGARSS apos;03 Proceedings. 2003, IEEE International 5, pp. 3222–3224.

Hirzel, A., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology 83 (7), 2027–2036.

Huette, A., Didan, K., Miura, T., Rodriguez, E., Gao, X., Ferreira, L., 2002. Overview of the radiometric and biophysical performance of the modis vegetation indices. Remote Sens. Environ. 83, 195–213.

Huette, A., Justice, C., van Leeuwen, W., 1999. Modis vegetation index (mod 13) algorithm theoretical basis document. version 3.

Jianxi, H., Feng, M., Wenbo, X., 2007. Retrieval of vegetation understory information fusing hyperion and panchromatic quickbird data in the method of neural network. In: Geoscience and Remote Sensing Symposium, 2007, IGARSS 2007. IEEE International, vol. (23–28), pp. 4315–4318.

Krasnov, B., Shenbrot, G., Khokhlova, I., Ivanitskaya, E., 1996. Spatial patterns of rodent communities in the Ramon erosion cirque, Negev Highlands, Israel. J. Arid Environ. 32 (3), 319–327.

Leathwick, J., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptative regression splines for statistical modelling of species distributions. Ecol. Model. 199, 188–196.

Legendre, P., Legendre, L., 1998. Numerical Ecology., 2nd ed. Elsevier Science B.V., Amsterdam.

Lehmann, A., Overton, J.M., Leathwick, J., 2002. GRASP: generalized regression analysis and spatial prediction. Ecol. Model. 157, 189–207.

MacNally, R., 2000. Regression and model building in conservation biology, biogeography and ecology: the distinction between and reconciliation of "predictive" and "explanatory" models. Biodivers. Conserv. 9, 655–671.

MacNally, R., 2002. Multiple regression and inference in conservation biology and ecology: further comments on identifying important predictor variables. Biodivers. Conserv. 11, 1397–1401.

McCullag, P., Nelder, J., 1989. Generalized Linear Models, 2nd ed. Chapman and Hall, London.

Moisen, G.G., Frescino, T.S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecol. Model. 157 (2–3), 209–225.

Morilhat, C., Bernard, N., Bournais, C., Meyer, C., Lamboley, C., Giraudoux, P., 2007. Response of Arvicola terrestris scherman populations to agricultural practices and Talpa europaea abundance in Eastern France. Agric. Ecosyst. Environ. 122, 392–398.

Munoz, J., Felicisimo, A.M., 2004. Comparison of statistical methods commonly used in predictive modelling. J. Veg. Sci. 15 (2), 285–292.

Murphy, H.T., 2007. Accounting for regional niche variation in habitat suitability models. Oikos 116 (12), 99–110.
Olden, J., 2003. Species-specific approach to modelling biological communities and its potential for conservation. Conserv. Biol. 17, 854–863.
Olden, J., Joy, M., Death, R., 2006. Rediscovering the species in community-wide predictive modeling. Ecol. Appl. 16 (4), 1449–1460.
Olivier, I., Mac Nally, R., York, A., 2000. Identifying performance indicators of the effects of forest management on ground-active arthropod biodiversity using hierarchical partitioning and partial canonical correspondence analysis. For. Ecol. Manag. 139, 21–40.
Pearce, J., Cherry, K., Drielsma, M., Ferrier, S., Whish, G., 2001. Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. J. Appl. Ecol. 38, 412–424.
Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133, 225–245.
Potts, J.M., Elith, J., 2006. Comparing species abundance models. Ecol. Model. 199 (2), 153–163.
Pulliam, H., 2000. On the relationship between niche and distribution. Ecol. Lett. 3, 349–361.
R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3–900051-07–0. http://www.R-project.org.
Randin, C.F., Dirnbock, T., Dullinger, S., Zimmermann, N.E., Zappa, M., Guisan, A., 2006. Are niche-based species distribution models transferable in space? J. Biogeogr. 33 (15), 1689–1703.
Raoul, F., Pleydell, D., Quéré, J., Vaniscotte, A., Rieffel, D., Takahashi, K., Bernard, N., Wang, J., Dobigny, T., Galbreath, K., Giraudoux, P., 2008. Small mammals assemblage response to deforestation and afforestation in Central China: a multinomial based modelling approach. Mammalia 72, 320–332.
Raoul, F., Quéré, J., Rieffel, D., Bernard, N., Takahashi, K., Scheifler, R., Ito, A., Wang, Q., Qiu, J., Yang, W., Craig, P., Giraudoux, P., 2006. Distribution of small mammals in a pastoral landscape of the Tibetan plateau (Western Sichuan China) and relationship with grazing practices. Mammalia 70 (3–4), 214–225.
Segurado, P., Araujo, M.B., 2004. An evaluation of methods for modeling species distributions. J. Biogeogr. 31 (10), 1555–1568.
Smith, A., Xie, Y., 2008. A Guide to the Mammals of China. Princeton University Press, Princeton.
Steyerberg, E., Harell, J.F., Borsboom, G., Eijkemans, M., Vergouwe, Y., Habbema, J., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J. Clin. Epidemiol. 54 (8), 774–781.
Tatem, A., Lewis, H., Atkinson, P., Nixon, M., 2002. Super-resolution land cover pattern prediction using a hopfield neural network. Remote Sens. Environ. 79, 1–14.
Ter Braak, C., Hoijtink, H., Akkermans, W., Verdonschot, P., 2003. Bayesian model-based cluster analysis for predicting macrofaunal communities. Ecol. Model. 160 (3), 235–248.
Wang, Q., Vuitton, D., Qiu, J., Giraudoux, P., Xiao, Y., Schantz, P., Raoul, F., Li, T., Wen, Y., Craig, P., 2004. Partial fencing as a possible risk factor for human alveolar echinococcosis in pastoral herdsmen communities of Sichuan, China. Acta Trop. 90, 285–293.
Wilson, D.E., Reeder, D.M., 2005. Mammals Species of the World: A Taxonomic and Geographic Reference. Smithsonian Institution Press, Washington, Londres.

3.3 Axe 1b : Classifier et prédire simultanément

3.3.1 Objectifs et méthodologie

La modélisation des assemblages est limitée puisqu'elle restreint les prédictions à des entités discrètes, traduisant des réponses identiques pour l'ensemble des espèces assemblées (cf Introduction 2.1.2).

Nous avons donc testé le pouvoir explicatif des variables environnementales quantitatives pour modéliser directement les densités relatives des espèces piégées (stratégie "Classifier et modéliser simultanément"). Le pouvoir explicatif de telles variables a été comparé à celui des classes d'habitats et des assemblages issus de leur regroupement. Cela revient à tester l'utilité de la définition des classes d'habitats par les experts, et de la définition d'assemblages qui en découle, pour expliquer la variabilité des densités des espèces entre les lignes de pièges.

Les densités relatives des espèces ont été modélisées par des modèles multinomiaux, linéaires généralisés (GLM) et non linéaires (MARS) incorporant trois types possibles de variables explicatives : i) les classes d'habitats définies *a priori* sur le terrain, ii) les assemblages (groupes d'habitats) et iii) les variables environnementales quantitatives utilisées dans Vaniscotte et al. (2009). Les performances prédictives de ces modèles ont été comparées au regard de leurs AICc. Conformément aux résultats précédents, les modèles incluaient un effet du type de piège. Concernant le modèle incluant les variables environnementales quantitatives, le modèle le plus parcimonieux (AICc minimal), parmi toutes les combinaisons possible de variables, a été sélectionné.

3.3.2 Résultats et discussion

Tab. 3.1 – Comparaison des modèles (GLMs) incorporant les variables environnementales quantitatives, les classes d'habitat (Habitat) et les assemblages (Assemblages) ainsi que le modèle nul (1). Les AICc, les poids d'Akaike (w_i) ainsi que le nombre de paramètres estimés (K) sont ordonnés en fonction de leurs écarts aux modèles nuls (Δ AICc). Les AICc obtenus pour les modèles MARS sont indiqués entre parenthèses.

		AICc	Δ AICc	w_i	K
Rangtang	Assemblages	908.91 (935.8)	0.00	1.00	5
	Altitude + Pente + NDVI + EVI	938.40 (974.5)	29.49	0.00	6
	Habitats	968.37 (984.2)	59.46	0.00	13
	1	1034.86 (935.8)	125.95	0.00	1
Maerkang	Assemblages	1176.94 (1274.3)	0.00	1.00	5
	Altitude + NDVI	1191.63 (1273.7)	14.68	0.00	4
	1	1274.16 (1274.3)	97.22	0.00	1
	Habitats	1508.13 (1334.2)	331.18	0.00	19

Exceptés pour les modèle nuls de Rangtang et de Maerkang ainsi que pour le modèle incluant les classes d'habitats pour Maerkang, les AICc des modèles MARS étaient supérieurs à ceux des

modèles GLMs. Aussi, les meilleurs modèles ont été obtenus dans chaque site d'étude par les GLMs. Par conséquent, nous avons basé la comparaison des modèles sur les modèles linéaires uniquement. Les variables environnementales retenues après sélection sont indiquées dans la Table 3.1.

Les assemblages (groupes d'habitats) étaient les variables les plus explicatives des densités relatives des espèces piégées à Maerkang et à Rangtang (Table 3.1). Aussi, les variables environnementales se sont avérées plus explicatives des densités relatives des espèces que les classes d'habitats définies *a priori*. Enfin, les classes d'habitats à Maerkang n'apportaient pas d'information supplémentaire que le modèle nul.

Dans notre cas d'étude où une grande diversité d'espèces ont été capturées dans une large diversité de classes d'habitats, la définition des assemblages, au préalable de la modélisation de leurs habitats dans l'espace multivarié, apparaît donc nécessaire. En effet, la définition de groupes d'habitats apporte une information supplémentaire pour expliquer les densités relatives des espèces que l'utilisation directe des variables environnementales quantitatives ne fournit pas. En revanche, le fait que les variables environnementales quantitatives soient davantage explicatives que les classes d'habitats peut être induit par la redondance de l'information contenue dans ces classes. Étant donné que les variables environnementales sont, par ailleurs, plus informatives que les modèles nuls, il serait envisageable de classifier les réponses des espèces directement dans l'espace multivarié qu'elles définissent. Des outils d'ordination (Analyse Canonique des Correspondances) ou de classification non supervisée (arbres de classification) (Legendre and Legendre, 1998) pourraient être utilisés pour produire une classification des espèces sur la base de variables environnementales quantitatives. Une telle classification pourrait être comparée à celle incorporant les habitats définis sur avis d'experts.

3.4 Principaux résultats du chapitre

- **Les distributions des assemblages**
  L' Analyse Discriminante de Mixtures (MDA) était plus performante que l'Analyse Discriminante Linéaire (LDA) pour modéliser les assemblages. La régression non linéaire et non paramétrique (MARS) était la technique la plus performante pour modéliser les distributions des assemblages le long de gradients environnementaux.

- **Le rôle des variables environnementales**
  Les environnements des deux sites d'étude sont discriminables avec une forte précision par des variables topographiques (altitude, pente), la bande ETM 7 et des indices de végétation (NDVI et EVI). Maerkang présente, par exemple, une biomasse végétale plus élevée que Rangtang. Des facteurs environnementaux peuvent donc être associés à des facteurs biogéographiques pour expliquer la diversité régionale des assemblages.
  Les variables environnementales issues d'images satellitales peuvent être utilisées pour discriminer les assemblages observés dans chaque site d'étude avec une erreur de classification minimale de 22%. De telles variables avaient des importances différentes en fonction du site d'étude. L'indice NDVI avait plus de poids dans la discrimination à Rangtang qu'à Maerkang alors que l'indice EVI était plus utile à Maerkang. De tels indices ont particulièrement permis de discriminer les habitats forestiers.

- **La stratégie de modélisation**
  Les assemblages pouvant être décrits comme des mélanges de distributions d'espèces, leurs discriminations se sont avérées complexes. Cependant, l'incorporation des classes d'habitats regroupées (des assemblages) apporte une information non négligeable pour expliquer les densités relatives des espèces en comparaison aux variables satellitales utilisées directement (stratégie "Classifier et modéliser simultanément").

Chapitre 4

Axe 2 : Vers un modèle prédictif : prédictions régionales des assemblages

4.1 Rappel des objectifs

La procédure de modélisation a jusqu'ici permis d'expliquer les présence/absence des assemblages dans chaque site d'étude par le choix de la technique de modélisation la plus appropriée ainsi que par la mise en évidence du rôle des variables environnementales satellitales. La construction d'un modèle prédictif sur une étendue régionale, c'est-à-dire au delà de l'aire d'entraînement du modèle, nécessite une évaluation externe des prédictions, de manière à évaluer les capacités des modèles à prédire sur des sites non échantillonnés et dans des contextes spatiaux et environnementaux différents.

Les études qui suivent constituent une exploration des capacités prédictives et régionales des modèles développée dans l'Axe 1. Le chapitre s'articule en deux parties dont les objectifs généraux sont d'évaluer les prédictions de manière externe, d'identifier les sources d'erreurs de prédiction et de proposer des solutions pour les réduire.

Nous avons principalement et initialement évalué les capacités prédictives et régionales des modèles et des données utilisés dans l'Axe 1. Au regard de notre stratégie d'échantillonnage dans la région incluant nos deux sites d'étude, deux échelles de modélisation sont réalisables pour prédire les distributions régionales des assemblages. Les modèles peuvent être entraînés i) localement, sur chaque jeu de données (Axe 1), puis extrapolés au delà du site d'étude et ii) régionalement, sur l'ensemble des jeux de données.

D'une part, connaître les possibilités et les limites de transférabilité des modèles entraînés localement, c'est-à-dire leur capacité à prédire sur des aires non échantillonnées est une étape importante dans l'évaluation des prédictions. De plus, prédire correctement sur des périodes temporelles ou spatiales indépendantes du jeu d'entraînement participe à des enjeux actuels en biologie de la conservation (impact du changement climatique, conservation de la diversité dans des zones peu étudiées ou sensibles). D'autre part, dans la perspective d'une cartographie régionale des assemblages, un modèle entraîné sur les deux sites d'étude est susceptible de prédire plus précisément les distributions des assemblages que les modèles locaux transférés puisque qu'il est entraîné sur la variabilité régionale des distributions.

Nous avons évalué et comparé les apports de chacune de ces échelles de modélisation pour prédire les distributions régionales des assemblages en répondant aux questions suivantes :

- *L'incorporation de la variabilité régionale des distributions des assemblages dans les modèles (modèle régional) peut-elle aider à discriminer les assemblages dans chaque site d'étude ?*

- *Les prédictions des modèles entraînés localement peuvent-elles être transférées entre les deux sites ?*

- *Les modèles entraînés régionalement sont-ils, comme on le suppose, plus appropriés que les modèles locaux pour prédire régionalement ?*

En parallèle, nous avons estimé les effets de deux principales sources d'erreurs de prédiction :
- la stratégie de modélisation, en comparant les performances d'une modélisation individuelle ou multiple des assemblages,
- la prévalence des assemblages, définie comme la proportion d'observations où l'assemblage est présent sur le nombre total d'observations, en établissant leurs corrélations avec les erreurs de prédiction.

Ces investigations font l'objet de l'article, prêt à soumission, présenté en section 2.2.

Dans un deuxième temps, une expérience de simulation a été réalisée afin d'explorer les effets d'une source principale d'erreurs de prédiction inhérente à nos données : la taille de l'échantillon (le nombre d'observations) pour chaque assemblage. Nous avons quantifié les variations des erreurs de prédiction induites par le nombre d'observations disponibles pour chaque assemblage. Les capacités prédictives et leur sensibilité aux tailles des échantillons ont été estimées et comparées entre les différentes techniques de modélisation (LDA, MDA, GLM et MARS) et entre les différents assemblages.

Cette expérience fait l'objet d'une publication en préparation, rapportée en section 2.3.

4.2 Article - Des données de piégeage aux prédictions régionales des habitats d'assemblages de micro-mammifères dans la province du Sichuan, Chine

From field trapping data to regional predictive mapping of small mammals assemblage in Sichuan province, China

Amelie Vaniscotte[*], Francis Raoul[*], David R. J. Pleydell[†] and Patrick Giraudoux[*]

Abstract

Aim: Evaluate the ability of different model strategies to predict and map the spatial distribution of small mammal assemblages in a region including two study areas separated by about 100 km and that differed regarding species diversity and assemblage definition.

Location: Western Sichuan (China)

Methods: We fitted Multiple Adaptive Regression Spline models on trapline data sets of each study area (locally), and of both study areas (regionally). Predictive performances of locally and regionally trained models were first evaluated and compared at sampling points (internally). Then, the abilities of locally trained models to predict on a regional extent, i.e on a distant area, were evaluated (external evaluation) and compared to the performances of the regionally trained model. Effects of assemblage prevalences as well as of the type of modelling process (multiple *versus* individual assemblage classification) on prediction accuracy were assessed.

Results: Locally trained models could not be extrapolated between study areas to explain and predict the regional assemblage distributions. Assemblage prevalences were correlated with their envelope breadths and constrained the amount of information on real assemblage distribution required to accurately discriminate them. Instead, a regional model, trained on multiple assemblage responses of both study areas was confirmed to be a more appropriate technique to build a regional classification map of species diversity. Moreover it could provide additional information to improve the discrimination of assemblages locally when sample sizes are limited.

Main conclusions: Regarding the modelling errors found, we could not resolve in our study case context, the well known conflict existing between the effects of biogeographical and environmental factors on assemblage distribution at a regional scale. Locally trained models should be extrapolated in homogeneous environmental and biogeographical area. Then, in the context of limited size of presence/absence regional data set of multiple species, a classification map of species diversity needs further evaluation along the whole regional environmental gradient.

[*]Department of Chrono-environment, UMR UFC/CNRS 6249, Université de Franche-comté, Place Leclerc, 25030 Besançon cedex.

[‡]UMR BGPI, CIRAD TA A-54/K, Campus International de Baillarguet, 34398 Montpellier Cedex 5 France

Corresponding author: Amélie Vaniscotte. Contact email address: amelie.vaniscotte@univ-fcomte.fr. Telephone number: (0033) 03 81 66 57 14. Fax number: (0033) 03 81 66 57 97

1 Introduction

Beyond their explanatory role, models of species habitat distributions are increasingly used for building predictive maps. They provide predictions that constitute a precious tool in multiple applied fields of ecology as listed in Manel et al. (2001). Moreover, the availability of remote sensing data allows to map predictions over large areas or across different time periods (Strauss and Biedermann, 2007; Pearman et al., 2008), output of crucial and growing interest in regional conservation planning and under global climate change context (Ferrier et al., 002b). Classifying species into groups and then predicting their distributions constitutes a suitable method when numerous rare species are observed (Ferrier and Guisan, 2006). It can help summarizing the complexity in species distribution and can be used as a surrogate for mapping biodiversity distribution patterns (Ferrier et al., 2004). Modelling habitats of species assemblages can be realized on multiple (joint) *versus* individual assemblage responses (Ferrier and Guisan, 2006). A classification map of multiple assemblages supposes that all fitted assemblages will be predicted, i.e that all assemblages potentially occurring within the prediction area are *a priori* known. Fit MARS (Multiple Adaptive Regression Splines) model on multiple responses and incorporate their interactions has been, for example, proved useful when low sample size for categorical responses is observed (Elith and Leathwick, 2007). By contrast, predicting individual groups presences/absences probabilities can be used in areas where species diversity is unknown.

Building a predictive map is challenging since it has to deal with many sources of error and uncertainties in predictions which could arise from the response and covariate data sets as well as from the modelling strategy (Elith et al., 2002; Barry and Elith, 2006). Field surveys provide accurate presence and/or absence species distribution data to build predictive maps. However, such data set, because they are collected in the field, are often limited to some particular locations and time periods. The sampling size (or sampling pressure) is known to limit the statistical power of the model both in its resolution, i.e the ability to model responses at points (Stockwell and Peterson, 2002), and spatial extent, i.e by limiting the trained variability in the parameter space. Beside the sampling size effect, species prevalences, i.e the proportion of presence data points over the total number of data points (presence and absence) relies on the ecological/biological characteristic of the study area. A rare species will be for example rarely recorded present whatever the sample size and some species could be trap-shy and hard to detect. This constitutes a problem of low probability event rather than data or modelling limitations (Jimenez-Valverde, 2006). Species prevalence is known to influence model performances *via* the amount of information necessary for modeling the habitat and to be correlated to predicted probabilities (McPherson et al., 2004; Olden et al., 2002). When multiple species are sampled at similar sampling points, the sampling size remains constant for each response category while prevalence could differ.

While calibrated on training data set limited to the sampling design, models are then used to predict on locations where species distribution is ignored. Therefore, beyond the model calibration on the training data set, model predictions need to be evaluated on independent area (model generalization) chosen so as to maximize the data representativeness of the model applications (Hastie et al., 2001; Vaughan and Ormerod, 2005). Resampling methods or data set partitioning are currently used for this purpose but can only provide reliable predictions in areas environmentally similar to the training areas (Vaughan and Ormerod, 2005). Model evaluation on new areas or time periods has been proven necessary in recent applications (Araujo et al., 2005; Strauss and Biedermann, 2007; Murphy and Lovett-Doust, 2007; Randin et al., 2006). Model transferability thus refers to its ability to provide good predictions of species distributions at unsampled areas and represents a new challenge (Peterson et al., 2007). In such a context, models should be evaluated for their abilities to predict accurately presence but also absence data (Fielding and Bell, 1997). Model transferability is known to suffer from several potential obstacles (Randin et al., 2006). First complication lies on the fact that species niche could vary in space following the law of relative

habitat constancy (Guisan and Zimmerman, 2000). Omission of substantial covariates related to local ecological processes could be another source of predictive errors. Moreover, the available environmental descriptors are often distal, e. g. satellite remote sensing data, and only provide surrogate variables for others more proximally related to "realized niches". Finally, biogeographical factors and history of species distributions arising at larger scale than the landscape (regional scale) could determine species geographical range. While they remain easily identifiable at the global (continent) or local (habitat) scales, effects of biogeographical *versus* environmental factors can be hardly discriminated at an intermediate scale (Krasnov et al., 1996).

We studied the distribution of small mammal assemblages in two study areas in western Sichuan, China, spaced by about a hundred kilometers. In such remote areas, small mammal assemblages have been defined on the basis of trapping field surveys and a regional shift in species composition has been found such that assemblages differed between the two study areas (Vaniscotte et al., 2009). The number of traplines set as well as the spatial extent of the sampling areas were limited by the time and heavy logistics required in such remote field area. Moreover, satellite data, i.e indirect predictors, remained the only covariates available to explain small mammal assemblage distributions. We showed that (Vaniscotte et al., 2009):

i) vegetation indices (NDVI and EVI), elevation and slope had a regional effect likely to explain the regional shift in assemblage distributions observed between the two study areas and had differential effects for explaining assemblage distributions in each study area.

ii) Multiple Adaptive Regression Spline (MARS) when run on multiple assemblage responses, was more performant to discriminate assemblages than Discriminant Analysis or Generalized Linear Models did.

Going a step further we aimed to explore the abilities of MARS to predict assemblage distribution regionally, i.e. within the extent including the two study areas. In our environmental and sampling context, predictive models can be trained at two spatial scales: locally, on the trapline data set of each study area and regionally, on the ensemble of both study area trapline data sets. We investigated the local and regional predictive abilities and limitations of those two spatial scales of modelling.

First, we asked: did and how assemblage occurrence probabilities predicted by the locally and regionally trained models differed between each others? This remains to test whether local environmental conditions were enough explanatory to model assemblage locally or if we could gain predictive power by a regionally fitted model. Predictions of the locally and regionally trained models were evaluated and compared at sampling points (internal model evaluation).

Then, a model trained on the whole regional variability is *a priori* supposed to predict more accurately assemblage regional distributions than locally transferred models do. We secondly aimed to verify such hypothesis by answering 2 questions:

i) Can locally trained models correctly predict the absences of assemblages on a study site distant of more than one hundred kilometers? Locally trained models were evaluated externally.

ii) Does a regionally trained model better capture the regional variability than transferred locally trained models do?

In parallel, we explored two additional questions relative to the potential sources of prediction errors arising in the context of limited sample size:

i) Which statistical formulation, i.e multiple *versus* individual assemblage response, provides better predictive performance on local and regional data sets?

ii) What are the effects of assemblage prevalences on assemblage presence/absence predictions from locally and regionally trained models?

2 Material and methods

2.1 Study areas and small mammal data sets

The two study areas were located in western Sichuan province, China, in the vicinities of Rangtang and Maerkang cities, and were separated by a distance of 125 kilometers (Figure 1). Climatic and landscape composition are provided in Table 1.

The small mammal data sets consisted in mutually exclusive assemblage presence/absence at 66 and 57 trapline locations, in Maerkang and Rangtang areas respectively. Four assemblages were defined independently in each study area: assemblages M1 to M4 in Maerkang and assemblages R1 to R4 in Rangtang. This was realized by a redundancy reduction procedure in the habitat covariate, on the basis of field trapping data set (Vaniscotte et al., 2009) (Table 2). Among the 15 species trapped, 4 of them, i.e *Apodemus peninsulae*, *Ochotona cansus*, *Eozapus setchuanus* and *Microtus irene* were trapped in both study area, i.e had regional geographical range.

2.2 Environmental factors

2.2.1 Environmental factor data set

Environmental factors selected were Enhanced Thematic Mapper (ETM) band 7, Elevation, slope, Enhanced Vegetation Index (EVI) and Normalized Differential Vegetation Index (NDVI) (Vaniscotte et al., 2009). Data were extracted from Landsat 7 image (July 2005 from U.S. Geological Survey) and a digital elevation model (DEM, SRTM program). Vegetation indices were calculated with ETM band 3, 4 and 1 (Huette et al., 1999). For each sampled trapline (a line of 25 traps spaced every 3 meters) we extracted its mean covariate response estimated from 10 points regularly spaced along it. Spearman correlation test indicated some significant correlation between factors ($p < 0.05$) but lower than 0.45 for all variable combinations. Therefore, all variables were considered for further modelling (Elith et al., 2006).

2.2.2 Discrimination of study area feature spaces

Ordination was used for a preliminary assessment as to whether Maerkang and Rangtang areas were environmentally distinct. The Mixture Discriminant Analysis (MDA) was the method of choice since it was previously shown to outperform Linear Discriminant Analysis (LDA) (Vaniscotte et al., 2009). One thousand randomly selected pixels in the image grid of each study area plus the sampling points of the trapline locations were discriminated. It provided an overview of the maximum discrimination possible between study areas and their sampling points in the environmental space and we could pointed out a potential regional overlapping. Then, we estimated for each assemblage its mean envelope breadth as the mean of Euclidean distance between each trapline and its envelope centroid (Kadmon et al., 2003). The effect of sampling prevalence on assemblage habitat distribution in this environmental space was assessed by estimating correlation between assemblage habitat breadths and sampling prevalences using Spearman correlation test.

2.3 Statistical modelling

Predictive models were build using training areas covering two spatial extents: we use “local model” to refer to models trained on data collected within just one of the two study areas and “regional model” to refer to models trained on data pooled from both study areas (Figure 2).

Multiple Adaptative Regression Splines (MARS) can be used to model categorical responses in relation to covariates which are transformed by pairwise basis functions (Hastie et al., 2001). MARS was chosen because i) it is a non-linear and non-parametric modelling technique, allowing for complex responses and ii) it is well suited to model rare responses in multiple response

data set (Leathwick et al., 2006; Elith and Leathwick, 2007). We fitted MARS to assemblage presence/absence data under two predictive mapping scenario:

- on multiple responses to predict multinomial probabilities of assemblage occurrence: basis functions were selected by minimizing the Residual Squared Errors among all responses and were then incorporated as covariates in a Multinomial GLM (Hastie et al., 2001; Leathwick et al., 2006).
- on individual assemblage responses to predict assemblage-wise occurrence probabilities: basis functions were selected by minimizing the Residual Squared Errors for each response independently and were then incorporated as covariates of Binomial GLMs.

All covariates were included in the model fittings and a backward pruning procedure was performed: basis functions that contributed negligibly to reduce the Residual Squared Errors, as estimated by a generalized cross validation procedure, were eliminated (Friedman, 1991; Hastie et al., 1994).

2.4 Internal model evaluation

Internal evaluation of predictions were assessed for local and regional models (Figure 2). Evaluation statistics were estimated on a re-sampled data set by the bootstrap 632 procedure (Efron and Tibshirani, 1995), a re-sampling method observed to be less prone to variability than cross-validation and more appropriate for low sample sizes (Hastie et al., 2001). Additional stratified (per assemblage) resampling was performed so as to avoid a sampling prevalence effect in the bootstrapped error estimate. This procedure was computed in R (R Development Core Team, 2008), using the `bootpred` function of the `bootstrap` package and the `errorest` function of the `ipred` package. Bootstrap 632 errors were estimated from 200 bootstrap iterations. Predictive performances were assessed in two predictive mapping contexts: overall assemblages and assemblage-wise classification of predicted probabilities.

2.4.1 Overall predictive performances

We estimated the misclassification error rate between all assemblages, where assemblage presence/absence were mutually exclusive. Also, the overall model Residual Deviance (RD) was estimated as a measure of prediction reliability, but since its calculation includes the number of observations, it could not be compared between local and regional models. Finally, confusion matrices between classified predictions and observations were drawn to identify regional prediction errors. Model overall predictive performances were estimated and compared for the two modelling techniques (multiple *versus* individual assemblages modelling).

2.4.2 Assemblage-wise predictive errors

The modelling technique providing the more accurate and reliable predictions was chosen for the following analysis. We focused the assemblage-wise model evaluation on the abilities for the model to discriminate assemblage presences from absences (model accuracy), performance of primary importance in a predictive mapping purpose.

Error measurement

For each assemblage, we defined its misclassification rate, i.e. the sum of all commission and omission errors. Presence of an assemblage, at a given sampling point, was considered above 0.5 predicted probability of presence. Comparison between local and regional model error rates was achieved by evaluating the regional model predictions on each study area data set rather than on the whole regional data set.

Assemblage prevalence effects

Correlation between assemblage prevalence and model accuracy were assessed over all assemblages by a Spearman correlation test. However, models can be highly accurate for extreme prevalences (very rare or abundant) while improvement of their prediction over chance predictions can remain slight (Vaughan and Ormerod, 2005). In order to account for this effect, we assessed whether model predictive errors differed from chance by simulating chance agreement for the error rate using randomization testing (Olden et al., 2002; Strauss and Biedermann, 2007). A null distribution of the error rate was generated by its estimation on randomly permuting observations (such as the sample size remains constant for each assemblage) (Steyerberg et al., 2001). This procedure was repeated as many times as the number of bootstrap iterations (200). The 95% bootstrap confidence interval for the chance distribution was estimated and errors lower than 0.05 quantile were considered to differ from chance.

2.5 External model evaluation

2.5.1 Predictions of local transferred models

Local model predictions were then evaluated on independent areas: predictions of locally trained Maerkang assemblages were realized on Rangtang trapline locations and *vice versa* (Figure 2). Model predictions were obtained for each assemblage individually. Since no assemblages were common between both study areas, tested points were all considered as absences. Consequently only the assessment of the commission error rate (specificity) could have been estimated and we only tested the transferability of absence predictions of the local models. Confusion matrices of transferred local predictions allowed us to identify potential confusions between predicted and observed Maerkang and Rangtang assemblages.

2.5.2 Comparison of local transferred and regional models predictive performances

Commission error rates of local transferred models were compared with those provided by the regional model estimated on the same sampling points and from the same fitting procedure (i.e. assemblage-wise). The significance of the differences in bootstrap error estimates between the regional and transferred local model predictions were assessed by computing the confidence intervals (95% and 99%) of the differences in their bootstrap error distributions, estimated on 200 bootstrap iterations. A confusion matrix of assemblage-wise classified predictions provided by the regional model was drawn. It allowed us to identify potential regional overlapping between Maerkang and Rangtang assemblages when predicted individually.

3 Results

3.1 Assemblage distributions in the regional and environmental space

Axes 1 and 2 of the Mixture Discriminant Analysis explained 98.5 % of the whole observed variability and discriminated the two study areas with a misclassification error rate of 0.08 (Figure 3). Study area ranges overlapped such that sampling points of assemblages R1, R2 but mainly R4 were found inside the range covered by assemblage M3.

Assemblage envelope breadth was positively correlated with assemblage prevalence (Spearman, rho=0.85, p = 0.01), such that M3 had the broadest range while assemblages M2, M4 and R3 the smallest ranges in the environmental space (figure 4). However, assemblage R1 had a larger prevalence but similar range than assemblages M1 and R2.

3.2 Internal model evaluation

3.2.1 Overall predictive performances

Comparison of modelling technique performances

The multiple responses model provided more accurate and less deviant bootstrapped predictions than the individual models fitted on Rangtang ($\Delta Error$=0.02; $\Delta Deviance$=20.45) and on the regional data set ($\Delta Error$=0.03; Δ $Deviance$=0.8) (Table 3). However, when fitted on Maerkang data set, the individual model provided slightly higher error rate and more deviant predictions than the multiple responses model ($\Delta Error$=0.02; $\Delta Deviance$=9.68). The confusion matrix (Table 4) indicated a between study area assemblage misclassification for the predictions provided by the regional model when trained on individual responses. However, such a regional misclassification concerned only one sampling point (R1 predicted for M3) when the regional model was fitted on multiple assemblage responses (Table 5).

Comparison of modelling level performances

Bootstrapped predictions provided by the Rangtang model were more accurate and less deviant than those provided by the Maerkang and regional models ($\Delta Error$=0.14 and $\Delta Error$=0.1, respectively, from the multiple responses models) (Table 3). The regional model provided misclassifications of Maerkang assemblages but not of Rangtang assemblages (Table 4 and 5).

3.2.2 Assemblage-wise predictive errors

Since the multiple responses model was more accurate and less deviant for 2 out of 3 data sets and was less prone to regional misclassification we used it for further further analysis.

No correlations were found between assemblage-wise predictive errors and their prevalences, either on local and regional model predictions (Spearman's ρ=0.43, p = 0.289 and ρ=0.57, p = 0.143, respectively).

Local model performances

In Maerkang, all assemblages were more accurately discriminated than by chance (Figure 5). The bootstrapped error rate was larger than 0.1 for M1 (Error=0.17) and M3 (Error=0.11) while it did not exceed 0.1 for M2 and M4. However, M1 and M3 predictions more departed from chance predictions than M2 and M4. In fact, the Euclidean distance between bootstrapped predictive errors and mean chance predictions was greater for M3 and M1 ($\Delta Error$=0.38 and $\Delta Error$=0.26) than M2 and M4 ($\Delta Error$=0.12 and $\Delta Error$=0.08).

Rangtang assemblages were discriminated better than by chance and with error rates lower than 0.1 by the Rangtang trained model, except R3 for which mean chance and its confidence intervals were close to zero (Figure 5). The maximally misclassified assemblage was R2 (Error=0.09). The difference between mean chance predictions and estimated error rate was higher for R1 ($\Delta Error$=0.43).

Comparison of modelling level performances

For Maerkang and Rangtang assemblages, excepted R3, error rates provided by the regional model all differed from chance predictions. The regional model provided equal or lower error rates than the local models did for all assemblages except for M3 which was predicted with a larger error rate of 0.05 by the regional model than by local models (Figure 5). In Rangtang, for all assemblages, the regional model was more accurate than the local models. This difference was maximum for R2 ($\Delta Error$=0.04).

3.3 External model evaluations

3.3.1 Predictions of local transferred models

Commission error rate significantly increased with assemblage prevalence (Spearman's ρ=0.85, p = 0.007).

Maerkang local model

When transferred on Rangtang training points, Maerkang model predicted absences of M1, M2 and M4 with low error rates (< 0.1) (Figure 6). However M3 was predicted to occur on Rangtang training points with an error rate of 0.52. Regarding bootstrapped confidence intervals, a significant difference in error rates was found between M3 and M2 (99 % CI=0.19-1) and M4 (95 % CI=0.02-1). The corresponding confusion matrix (Table 6) indicated that M3 was predicted at sites where the four Rangtang assemblages were observed and was most frequently mismatched with R4. Also, M1 was frequently predicted at sites occupied by R1.

Rangtang local model

Rangtang model predicted absences of R2 and R3 with low (0.01) and null error rates respectively (Figure 6). However, R1 was predicted across much of the Maerkang area (Error=0.62) and R4 was also badly discriminated (Error=0.29). A significant difference was found between R1 and R3 (95 % CI=0.17-0.80) as well as between R3 and R4 (95 % CI=0-0.50). Assemblage R1 was erroneously predicted at sites where all Maerkang assemblages were observed, the most frequent misclassification being at locations occupied by M1 (Table 7). Assemblage R4 was mainly predicted where M3 was observed, while R2 was predicted in 1 trapline location occupied by M3.

3.3.2 Comparison of local transferred and regional model predictive performances

Commission error rates estimated on the regional model predictions were not correlated with observed assemblage prevalences (Spearman's ρ=0.30, p = 0.464).

Error rates for assemblage M3 were significatively lower when predicted with the regional than by Maerkang local transferred model ($\Delta Error$=0.33; 99 % CI =0.27-1), while for the other assemblages, differences were not significant (Figure 6).

In Maerkang area, commission error rate were much lower or equal when predicted with the regional than by Rangtang local transferred model, such difference being significant for assemblage R1 ($\Delta Error$=0.62, 99 % CI = 0.01-0.80) and for R4 ($\Delta Error$=0.29, 95 % CI = 0.06-0.53) (Figure 6).

4 DISCUSSION

4.1 Local assemblage discrimination

The consideration of the regional assemblage variability and diversity in the model fitting provided a slight improvement in local discrimination of Rangtang assemblages and this performance gain was particularly large where the Rangtang model failed (assemblage R2). The fact that the size of the training sample used to train the regional model was twice as large as for the local models could explain this tendency. However, at Maerkang sites, the performance gain was not observed for assemblage M3. The relationships between the large diversity of habitat and species of this assemblage might rely on local environmental conditions that could have been neglected in the regional model fitting (Osborne and Surez-Seoane, 2002).

Observed assemblage prevalence was positively correlated to habitat range in the environmental space. This could be the result of an increasing of the sampled variability in habitat response with increasing number of observations per assemblage. Therefore, our knowledges on assemblage habitat real variability could have been biased by our sampling effort. Not taking into account the whole spatial turnover among habitats (β diversity) in assemblage composition could also alter the classification of community types (Ferrier, 2002). On the contrary, modelled envelope of R1 wasn't larger than those of two lower prevalence assemblages (M1 and R2). The habitat distribution variability was better captured for this assemblage and this could explain why it was the most accurately and less randomly predicted assemblage (Kadmon et al., 2003).

Considering improvement of the error rate over chance predictions allowed us to identify the bias in apparently correct predictions for low prevalence assemblages (Olden et al., 2006). We found that error rates were lower for extremely low prevalence assemblages (lower than 0.1) but model predictions were predicted with close or similar error than by chance. By contrast, the larger prevalence assemblages (M3 and R1) were the least randomly discriminated assemblages. Our knowledge and confidence in the discrimination of low prevalence assemblages was penalized by the lack of more observations. As emphasized above, the ability to discriminate assemblages also relies on their habitat ranges within the environmental space. Also, habitat variability should increase with landscape heterogeneity that assemblages incorporate. The required sampling size for an optimal discrimination should thus take into account landscape heterogeneity to estimate the optimal sampling effort per assemblage.

4.2 Regional predictions: local *versus* regional models?

4.2.1 Sources of prediction error in transferred predictions

The transferability of absence predictions for each assemblage was affected by observed assemblage prevalences. The most prevalent assemblages were extrapolated with large predictive error by transferred local models while low prevalence assemblages were not predicted on independent areas. Sampling prevalences could bias modelling during model fitting by over-predicting and under-predicting large and low prevalence responses respectively (Fielding and Bell, 1997; McPherson et al., 2004). Such clear prevalence effects prevented model transfer for sparsely sampled assemblages.

Even if it was the worst locally discriminated assemblage, R2 was rarely predicted in the Maerkang area by the local transferred Rangtang model. This result supports the hypothesis that regional variations of environmental factors could explain the regional shift in such assemblage distribution. However, MARS could not accurately discriminate absence of other trained assemblages in a different environmental context which emphasized the existence of modelling errors.

Assemblage R1, dominated by the murinae *Apodemus peninsulae*, provided the largest commission error in Maerkang, particularly at M1 sites (dominated by *Niviventer confucianus*) and the converse mismatch was true. Those two assemblages shared a close or common position in the environmental space while they did not share any species or habitat classes. They both inhabited non forested habitats linked with agricultural practices which could explain their bad discrimination. The heterogeneity of these agricultural landscape could not have been captured by the spectral environmental factors considered in our model. It is well known that the spatial distributions of small mammal species respond to landscape structure and composition at various spatial scales (Duhamel et al., 2000; Giraudoux et al., 1997; Butet et al., 2006; Morilhat et al., 2007). Landscape could directly influence habitat heterogeneity and complexity which are known to be strongly correlated to small mammals' assemblage diversity (Williams et al., 2002). A deeper analysis at the landscape level might be run to help discriminating such habitats. Land use history and management, information which was not available in the current study, could also explain such habitat discriminations (Dirnbck et al., 2004).

Large confusions between assemblage M3 predictions and observed Rangtang assemblages could be induced by the fact that the M3 habitat definition overlapped with signatures of some Rangtang traplines in the environmental space. Its dominant species *Ochotona cansus* was distributed in a large diversity of sampled habitats including coniferous/rhododendron forests which were sampled in both study areas. This suggests an absence of regional niche variation for those habitats between the two study areas such as they could be modelled by a regional model (Murphy and Lovett-Doust, January 2007). However, M3 and R4 did not constituted the same assemblage since other associated species had local distribution ranges. Habitats for such species could actually differ between study areas but could not have been discriminated spectrally and were confounded in the mean assemblage response.

Numerous errors arising from the modelling procedure, i.e prevalence effect, or arising from the data set, i.e missing landscape variables, affect inferences concerning the contribution of the regional environmental factors in the regional shift observed in small mammal assemblage regional distributions. Such errors also prevented us to extrapolate our local MARS predictions on an independent area just one hundred kilometers away within the same biogeographical area. Species and habitat diversities were organized in each area independently and could not be extrapolated to a mildly different environmental context. Among the numerous processes that determined the composition of local communities, those occurring at local scale, such as ecological interactions, landscape processes and local natural history sound of primary importance (Lomolino et al., 2005). Therefore, predictions of local models should be restricted in areas where the range of environmental factors are covered by the local data set (Murphy and Lovett-Doust, 2007).

4.2.2 The regional model predictive performances

While its local performance gain was contrasted between study areas, the regional model, not surprisingly, provided more reliable assemblage absence predictions than local transferred models did. The regional model was trained on the regional assemblage variability and consequently was able to predict it. Moreover, its assemblage predictions did not suffer a prevalence effect.

Predictions of the regional model were much less misclassified between the two study areas than those obtained from the transferred local models. In such a classification context of all assemblages, the species associated with *Ochotona cansus* and the variability observed in assemblage M3 and R4 habitat classes between Maerkang and Rangtang were sufficient informations to discriminate those assemblages sharing their dominant species. This confirms the recommendations of previous studies to use multiple response model fitting when low sample size for categorical responses is observed (Elith and Leathwick, 2007).

This statistical modelling could be viewed as an extension of traditional supervised image classification, that directly incorporate and predict multiple species response (Ferrier, 2002) and could be used in a regional assemblage classification context. A classification of multiple assemblage occurrences thus supposes that the species diversity potentially present in the region are *a priori* known. If not, false absence predictions could be realized at unsurveyed sites (Ferrier, 2002). Therefore, further sampling effort should be concentrated on the species diversity within the whole range of the environmental and geographical space separating the two study areas.

References

Araujo, M. B., Pearson, R. G., Thuiller, W., and Erhard, M. (2005). Validation of species-climate impact models under climate change. *Global Change Biology*, 11(9):1504–1513.

Barry, S. and Elith, J. (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology*, 43(11):413–423.

Butet, A., Paillat, G., and Y., D. (2006). Factors driving small rodents assemblages from field boundaries in agricultural landscapes of western france. *Landscape Ecology*, 21:449–461.

Dirnbck, T., Dullinger, S., and Chiarucci, A. (2004). Habitat distribution models, spatial autocorrelation, functional traits and dispersal capacity of alpine plant species. *Journal of Vegetation Science*, 15(1):77–84.

Duhamel, R., Quere, J.-P., Delattre, P., and P., G. (2000). Landscape effects on the population dynamics of the fossorial form of the water vole (arvicola terrestris sherman). *Landscape Ecology*, 15(10):89–98.

Efron, B. and Tibshirani, R. (1995). Cross-validation and the bootstrap: Estimating the error rate of a prediction rule. Technical report (tr-477), Dept. of Statistics, Stanford University.

Elith, J., Burgman, M. A., and Regan, H. M. (2002). Mapping epistemic uncertainties and vague concepts in predictions of species distribution. *Ecological Modelling*, 157(2-3):313–329.

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, T. A., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberon, J., Williams, S., Wisz, M. S., and Zimmermann, N. E. (2006). Novel methods improve prediction of species distributions from occurrence data. *Ecography*, 29(2):129–151.

Elith, J. and Leathwick, J. (2007). Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Diversity and Distributions*, 13(3):265–275.

Ferrier, S. (2002). Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? *Systematic Biology*, pages 331–363.

Ferrier, S., Drielsma, M., Manion, G., and Watson, G. (2002b). Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. ii. community-level modelling. *Biodiversity and Conservation*, 11:2309–2338.

Ferrier, S. and Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of applied ecology*, 43(3):393–404.

Ferrier, S., Powell, G. V. N., Richardson, K. S., Manion, G., Overton, J. M., Allnutt, T. F., Cameron, S. E., Mantle, k., Burgess, N. D., Faith, D. P., Lamoreux, J. F., Kier, G., Hijmans, R. J., Funk, V. A., Cassis, G. A., Fisher, B. L., Flemons, P., Lees, D., Lovett, J. C., and Van Rompaey, R. S. A. R. (2004). Mapping more of terrestrial biodiversity for global conservation assessment. *BioScience*, 54(12):1101.

Fielding, A. and Bell, J. (1997). A review methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, 24:38–49.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–67.

Giraudoux, P., Delattre, P., Habert, M., Qur, J., Deblay, S., Defaut, R., Duhamel, R., Moissenet, M., Salvi, D., and Truchetet, D. (1997). Population dynamics of fossorial water vole (arvicola terrestris scherman): a land use and landscape perspective. *Agriculture, Ecosystems and Environment*, 66:47–60.

Guisan, A. and Zimmerman, N. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, 135:147–186.

Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89(428):1255–1270.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The Elements of Statistical Learning*. Springer-Verlag.

Huette, A., Justice, C., and van Leeuwen, W. (1999). Modis vegetation index (mod 13) algorithm theoretical basis document. version 3.

Jimenez-Valverde, A. (2006). The ghost of unbalanced species distribution data in geographical model predictions. *Diversity and Distributions*, 12(4):521–524.

Kadmon, R., Farber, O., and Danin, A. (2003). A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, 13(3):853–867.

Krasnov, B., Shenbrot, G., Khokhlova, I., and Ivanitskaya, E. (1996). Spatial patterns of rodent communities in the ramon erosion cirque, negev highlands, israel. *Journal of Arid Environments 32(3): 319-327*, 32(3):319–327.

Leathwick, J., Elith, J., and Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptative regression splines for statistical modelling of species distributions. *Ecological Modelling*, 199:188–196.

Lomolino, M., Riddle, B., and Brown, J. (2005). *Biogeography*. Sinauer.

Manel, S., Williams, H. C., and Ormerod, S. J. (2001). Evaluating presence-absence models in ecology: the need to account for prevalence. *Journal of Applied Ecology*, 38(5):921–931.

McPherson, J., Jetz, W., and Rogers, D. J. (2004). The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, 41(5):811–823.

Morilhat, C., Bernard, N., Bournais, C., Meyer, C., Lamboley, C., and Giraudoux, P. (2007). Responses of arvicola terrestris scherman populations to agricultural practices, and to talpa europaea abundance in eastern france. *Agriculture, Ecosystems Environment*, 122(3):392 – 398.

Murphy, H. T. and Lovett-Doust, J. (2007). Accounting for regional niche variation in habitat suitability models. *Oikos*, 116:99–110(12).

Olden, J., Jackson, D., and Peres-Neto, P. (2002). Predictive models of fish species distributions: a note on proper validation and chance predictions. *Transactions of the American Fisheries Society*, 131:329–336.

Olden, J., Joy, M., and Death, R. (2006). Rediscovering the species in community-wide predictive modeling. *Ecological Applications*, 16(4):1449–1460.

Osborne, P. E. and Surez-Seoane, S. (2002). Should data be partitioned spatially before building large-scale distribution models? *Ecological Modelling*, 157(2-3):249 – 259.

Pearman, Peter, B., Randin, Christophe, F., Broennimann, Olivier, Vittoz, Pascal, Knaap, van Der, W. O., Engler, Robin, Lay, Le, G., Zimmermann, Niklaus, E., Guisan, and Antoine (2008). Prediction of plant species distributions across six millennia. *Ecology Letters*, 11(4):357–369.

Peterson, T., A., Papes, Monica, Eaton, and Muir (2007). Transferability and model evaluation in ecological niche modeling: a comparison of garp and maxent. *Ecography*, 30(4):550–560.

Randin, C., Dirnbock, T., Dullinger, S., Zimmermann, N. E., Zappa, M., and Guisan, A. (2006). Are niche-based species distribution models transferable in space? *Journal of Biogeography*, 33:1689–1703(15).

Steyerberg, E., Harell, J. F., Borsboom, G., Eijkemans, M., Vergouwe, Y., and Habbema, J. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54(8):774–781.

Stockwell, D. and Peterson, A. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, 148(1):1–13.

Strauss, B. and Biedermann, R. (2007). Evaluating temporal and spatial generality: How valid are species-habitat relationship models? *Ecological Modelling*, 204(1-2):104 – 114.

Vaniscotte, A., Pleydell, D. R., Raoul, F., Qur, J. P., Jiamin, Q., Wang, Q., Tiaoying, L., Bernard, N., Coeurdassier, M., Delattre, P., Takahashi, K., Weidmann, J.-C., and Giraudoux, P. (2009). Modelling and spatial discrimination of small mammal assemblages: An example from western sichuan (china). *Ecological Modelling*, 220(9-10):1218 – 1231.

Vaughan, I. and Ormerod, S. (2005). The continuing challenges of testing species distribution models. *Journal of Applied Ecology*, 42(4):720–730.

Williams, S. E., Marsh, H., and Winter, J. (2002). Spatial scale, species diversity, and habitat structure: small mammals in australian tropical rain forest. *Ecology*, 83(5):1317–1329.

5 Tables

Table 1: Climatic parameters (data source: Maerkang Center for Disease Control), number of sampled habitats (Forested, non forested and agricultural) and names of main tree species (in brackets), in Maerkang and Rangtang study areas.

	Maerkang	Rangtang
Mean annual temperature (C)	8.9	5.4
Mean annual rainfall (mm)	811.5	854.3
Elevation range (m)	2950-4100	3350-3900
Habitats sampled	18	11
Forested habitats	8 (birch, coniferous, rhododendron, oak)	2 (coniferous, rhododendron)
Non forested habitats	6	5
Agricultural habitats	4	4

Table 2: Small mammal assemblages defined in Maerkang and Rangtang (from Vaniscotte et al. (2009)). Number of merged habitat classes, species diversity (Simpson index), dominant species, first associated species, habitat classes, prevalences and corresponding sampling size (number of traplines) are provided for each assemblage. The total number of traplines is also reported for each area (N).

Maerkang	N=66			
Assemblages	M1	M2	M3	M4
Merged habitats	7	1	5	2
Species diversity	2.6	2.08	7.66	1.97
Dominant species	*Niviventer confucianus*	*Micromys minutus*	*Ochotona cansus*	*Apodemus draco*
Associated species	*Chodsigoa hypsibia*	*Eozapus setchuanus*	*Apodemus draco*	*Apodemus latronum*
Dominant habitat	Bushes, grassland, fields	Culture	Forests rhodo./coniferous	Forests birch
Prevalence	0.30	0.11	0.52	0.08
Sample size	20	7	34	5
Rangtang	N=57			
Assemblages	R1	R2	R3	R4
Merged habitats	4	4	1	3
Species diversity	1.07	1.25	1.25	2.45
Dominant species	*Apodemus peninsulae*	*Apodemus peninsulae*	*Microtus limnophilus*	*Ochotona cansus*
Associated species	*Microtus limnophilus*	*Microtus limnophilus*	*Apodemus peninsulae*	*Microtus limnophilus*
Dominant habitat	Field bank	Bushes and coniferous	Fenced grassland	Forest rhodo./coniferous
Prevalence	0.40	0.32	0.04	0.25
Sample size	23	18	2	14

Table 3: Misclassification error rates (Error) and Residual Deviance (RD) estimated for each modelling level (local *versus* regional) and modelling technique (multiple *versus* individual MARS).

		Maerkang		Rangtang		Regional	
		Error	RD	Error	RD	Error	RD
Multiple	Train.	0.18	43.38	0.09	18.11	0.07	37.55
	Boot. 632+	0.28	75.06	0.14	32.02	0.24	137.94
Individual	Train.	0.09	39.24	0.11	48.76	0.14	83.69
	Boot. 632+	0.26	65.38	0.16	52.47	0.27	138.74

Table 4: Confusion matrix of Maerkang and Rangtang assemblage classified predictions obtained from the regional model fitted on individual assemblage responses. Regarding its low sample size, M4 was never predicted to be present and was not considered in the confusion matrix.

	Observed							
Predicted	M1	M2	M3	M4	R1	R2	R3	R4
M1	18	1	5	0	0	0	0	0
M2	2	6	0	0	0	0	0	0
M3	0	0	28	0	1	0	0	5
R1	0	0	1	0	23	0	0	0
R2	0	0	0	0	0	17	0	0
R3	0	0	0	0	0	0	2	0
R4	0	0	0	0	0	0	0	14

Table 5: Confusion matrix of Maerkang and Rangtang assemblage classified predictions obtained from the regional model fitted on multiple assemblage responses.

	Observed							
Predicted	M1	M2	M3	M4	R1	R2	R3	R4
M1	17	1	3	0	0	0	0	0
M2	2	6	0	0	0	0	0	0
M3	1	0	30	0	0	0	0	0
M4	0	0	0	5	0	0	0	0
R1	0	0	1	0	23	0	0	0
R2	0	0	0	0	0	18	0	0
R3	0	0	0	0	0	0	2	0
R4	0	0	0	0	0	0	0	14

Table 6: Confusion matrix of Maerkang multiple responses model predictions at Rangtang observations.

	Observed			
Predicted	R1	R2	R3	R4
M1	10	2	0	0
M2	0	0	0	0
M3	5	7	1	8
M4	0	0	0	0

Table 7: Confusion matrix of Rangtang multiple responses model predictions at Maerkang observations.

	Observed			
Predicted	M1	M2	M3	M4
R1	20	7	11	3
R2	0	0	1	0
R3	0	0	0	0
R4	0	0	19	0

6 Figures

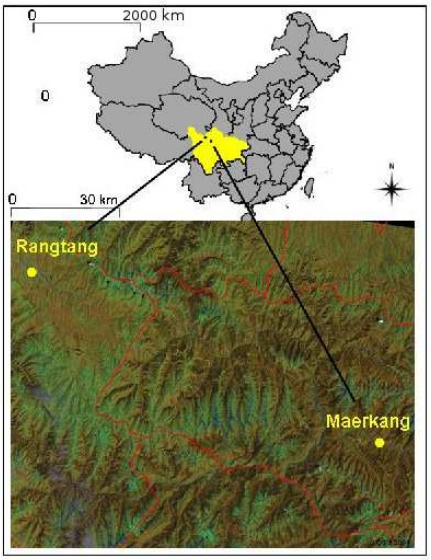

Figure 1: Geographical locations of Maerkang and Rangtang study areas in western Sichuan, China (datum: WGS84, projection: UTM 47N)

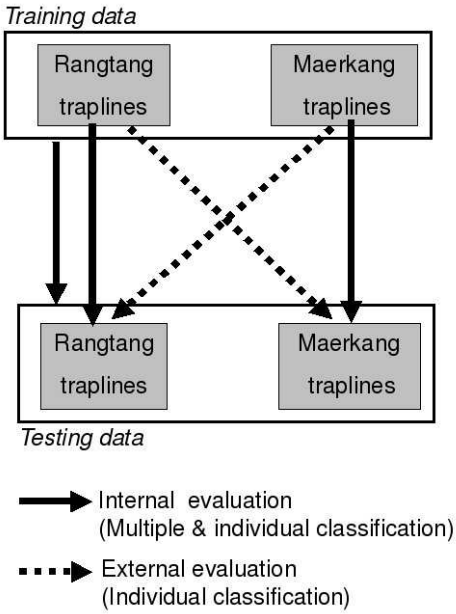

Figure 2: Internal and external model evaluations for the local (grey boxes) and regional (white boxes) models. For each type of evaluation, the modelling technique used (individual *versus* multiple assemblage responses) are provided.

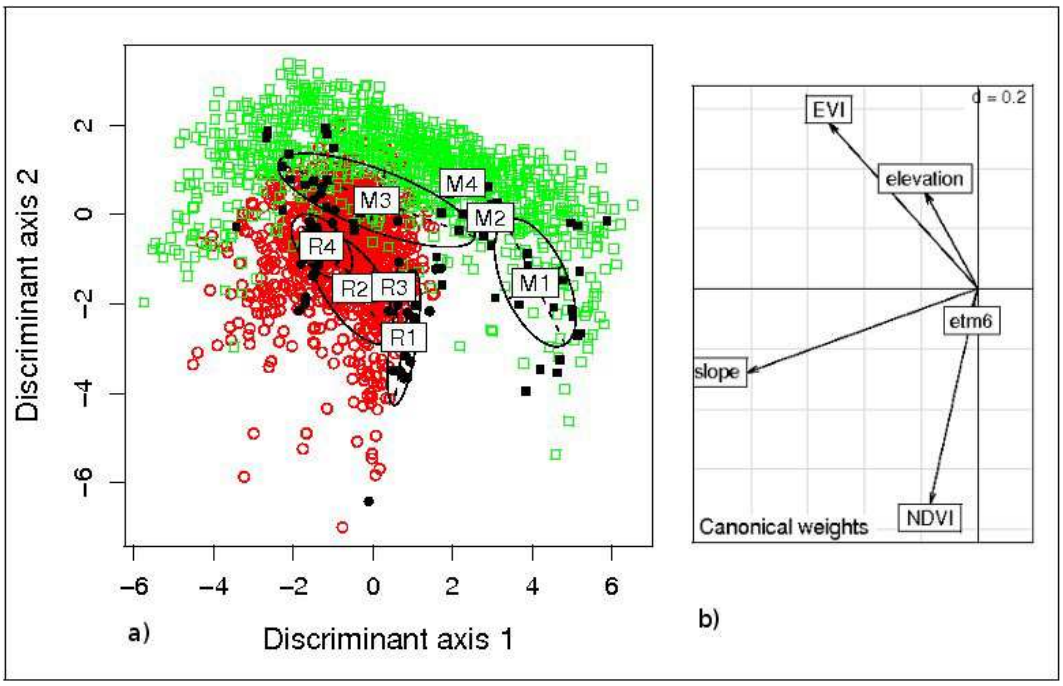

Figure 3: MDA space (2 first discriminant functions): a) Distribution of Maerkang (empty squares) and Rangtang (empty circles) random points as well as Maerkang (full squares) and Rangtang (full circles) traplines, 95 % of each assemblage inertia of Maerkang (M1, M2, M3 and M4) and Rangtang (R1, R2, R3, R4) were represented by ellipses labelled at their centroids; b) Canonical weights for each model covariate.

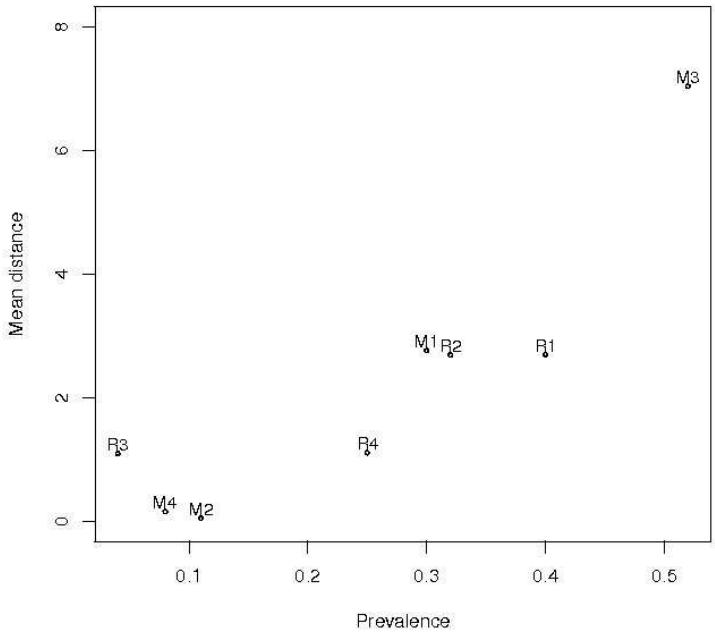

Figure 4: Mean sampling point distances from assemblage centroid (in MDA space), for each assemblage, according to its observed prevalence.

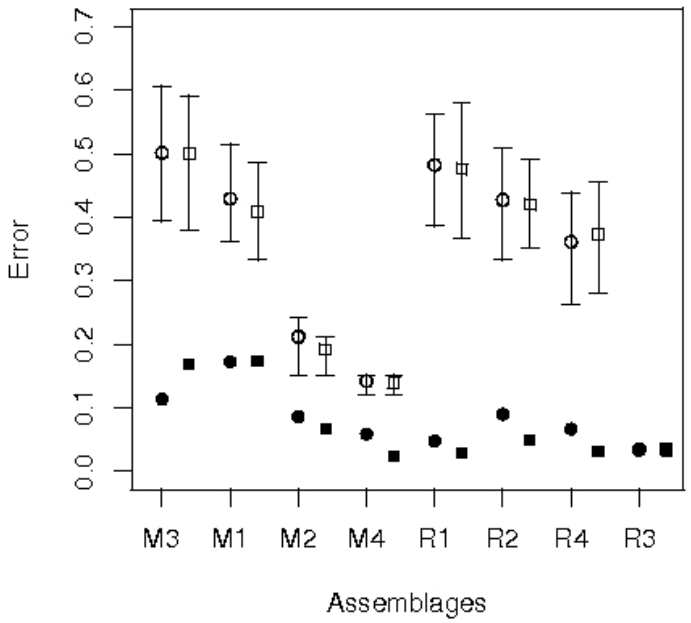

Figure 5: Misclassification error rate for each assemblage of Maerkang (M1, M2, M3 and M4) and Rangtang (R1, R2, R3 and R4) estimated from: i) the local (black circles) and regional (black squares) model predictions. Means and 95 % bootstrap confidence intervals of chance error are provided for the same models. Assemblages are ordered from the highest to the lowest prevalence. Model and chance error rate were null for assemblage R3.

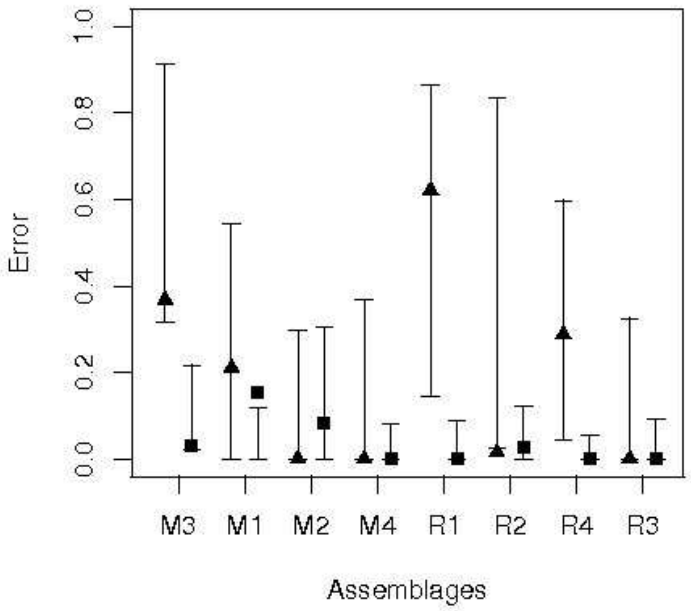

Figure 6: Misclassification error rate for each assemblage of Maerkang (M1, M2, M3 and M4) and Rangtang (R1, R2, R3 and R4) estimated from local transferred (diamonds) and regional (squares) multiple responses model predictions. Assemblages are ordered from the largest to the smallest prevalence. For the local model, errors of the individual response model (diamonds) are provided. The 90% confidence intervals estimated over 1000 bootstrap iterations are reported.

4.3 Article - Effet de la taille de l'échantillon sur la précision des prédictions

Sample size effects on predictive performances of multiple classes discriminant model

Amelie Vaniscotte, Francis Raoul and Patrick Giraudoux

Abstract

A simulation experiment has been realized in a real environmental context in order to investigate the effects of multiple class sample sizes on model predictions of their distributions. Quantification of sample size effects was assessed and compared between several modelling techniques and class envelope breadths. Satellital environmental variables expanded over a grid within an area of western Sichuan (China) have been classified in 4 groups by running an unsupervised Gaussian Mixture Model clustering algorithm. Class distributions obtained thus defined a group of points in which training and testing data sets could be randomly sampled.

Our results confirm the over-performance of MARS over Discriminant Analysis and Generalized Linear Models to accurately predict class distributions. However such abilities was valid for sample size larger than 32. The sample size per assemblage required to reach 90 % of the maximum accuracy was 88 for the MARS model. Also, even if not enough robust to drawn definitive conclusions, our results suggest that classification performances vary according to class envelope breadth in the environmental space. Those estimations were purely theoretical but provided an idea on the optimal sampling size required to maximise model performances. However, they cannot be directly applied in the context of assemblage distribution without considering assemblage prevalences and geographical ranges considered as constant in this experiment.

1 Introduction

Sampling size effects on prediction accuracy of habitat models are usually studied for 2 main purposes: i) to assess the optimal sampling size require to accurately predict species distributions and to guide sampling strategies and ii) to compare modelling technique predictive performances in order to select the most appropriate one given data quality.

The sample size is thus defined as the total number of sampling units or the total sampling pressure of species presence/absences data. It is recognised that limited sample size can be a source of errors in predictions of species habitats by limiting the amount of available data necessary to capture the variability of the habitat characteristics in a multivariate space (Barry and Elith, 2006).

When sampling species distribution, presence or absence can be recorded at each sampling points. Thus, effects of the sampling size needs to be defined according to species prevalence, i.e the proportion of sampling points representing its presence over the total number of sampling points. In fact, even if a large number of sites have been sampled, rare species generally occur in a few of them. The whole sampling pressure thus does not reflect the amount of information available to model such species distribution. It is therefore recommended to consider the species prevalences or the sizes of the presence-only data sets on model predictive performances (Barry and Elith, 2006). The minimum sample size required to accurately model species habitat distribution also relies on species niche breadth in the environmental space such as species with narrow niche required lower number of observations to be modelled than more generalist ones (Kadmon et al., 2003). Also, the geographical range of species distribution can influence species prevalences recorded in a given study area (McPherson et al., 2004).

Sampling size can induce errors arising in model building or in the evaluation of its predictions (McPherson et al., 2004). A limited sample size could be, for example, a source of uncertainty in parameter estimation or could increase the weights of outliers in habitat definition (Wisz et al., 2008). Such statistical

artefacts are known to vary according to the modelling techniques used (Wisz et al., 2008) that should be selected according to their ability to converge quickly (with the smallest sample size) to the greatest accuracy (Stockwell and Peterson, 2002). The size of the sampling can constrain model complexity, i.e the number of covariates and their interactions, thus limiting the ability to capture complex (not Gaussian and not linear) species/environment relationships. The use of non parametric and non linear modelling techniques is thus not recommended in case of small sample size data sets (Barry and Elith, 2006).

Predictive performances of Linear Regression and Discriminant Analysis are known to be sensitive to sampling size effects requiring a sample size comprised between 200 and 300 observations for accurate predictions (McPherson et al., 2004). GARP model (Genetic algorithm) has been shown to be more performant than Linear Regression to model small sample size data set requiring 10 observations to reach 90 % of the maximum accuracy (Stockwell and Peterson, 2002). Also, a comparative study of 12 algorithms found that flexible and non parametric models such as MARS were more prone to sample size effect than MAXENT or GARP (Wisz et al., 2008). However, a minimum number of 30 observations were required whatever the modelling technique used.

In a previous study, we modelled spatial distribution of small mammal assemblages in two study areas situated in Sichuan province (China), around Maerkang and Rangtang cities (Vaniscotte et al., 2009). We found that MARS was the more accurate modelling method to discriminate assemblage on re-sampling data sets compared to LDA, MDA and GLM. Also, we showed that differences in assemblage prevalences could partly explain their differential discrimination and prediction accuracy (Vaniscotte et al., in prep.). Therefore, we suppose that our previous modelling techniques comparison as well as model inferences were constrained by the little size of our training data sets (Vaniscotte et al., 2009). Here, we simulated classes (assemblages) distributions in the environmental context of one of our study area in order to experiment the effects of class sample sizes, i.e the number of class observations, on model prediction accuracy. We thus answer 2 main questions:

i) What is the maximum accuracy reachable and how model accuracy vary with sampling size, for all the modelling techniques used in our previous modelling framework?

ii) How variable is such sampling size sensitivity among classes which differed according to their niche breadth in the environmental space?

2 Methods

2.1 Simulating the training and testing data sets (artificial Gaussian Mixture distributions)

The environmental data consists in 4 environmental variables that provided the lowest misclassification error rate of small mammals assemblage distributions, i.e ETM7, elevation, NDVI and EVI such as defined in (Vaniscotte et al., 2009). Those variables expanded over a grid of 130065 pixels of 30 meters covering the Maerkang study area. Those data were clustered in 4 classes following Gaussian Mixture distributions, by running an unsupervised Gaussian Mixture Model clustering algorithm under the Mixmod software (Biernacki et al., 2006). The ICL criterion (Integrated Complete-data Likelihood) was used to select the best model among the 28 available in the software. This provided a simulated distribution of the 4 classes into such a feature space. Over the whole image grid, 84 %, 0.1 %, 13 % and 1.6 % pixels were thus labelled as classes 1 to 4 respectively. Regarding the small sample size available for the class 2, this former was not considered in further analysis.

In the context of classification of multiple groups, classes were mutually exclusive, i.e presence of one class corresponds with absences of the others, and the total number of observations (total number of pixels) were common for all classes. Thus, class sampling size was considered as the number of points where it was observed. One thousand and 100 data points were selected randomly within the image grid, in a stratified and balanced way for each class (a constant number per class), i.e within each response class sub-population, to define respectively the training and testing data sets. Hence, the prevalence remained constant among classes as it was set for species sampling size effects by (Stockwell and Peterson, 2002).

Generalized Linear Models, Linear Discriminant Analysis, Mixture Discriminant Analysis and Multiple Adaptive Regression Splines were fitted on a randomly sampled subset of the training data set at increasing size. Sampling size increased from 8 to 360 observations with 8 observations intervals. The upper limit was selected as the sample size above which variations in error rate did not exceed 0.001. The evaluation criteria for those models, the misclassification error rate, was then estimated on an independent testing data set. The training data set random sampling, model fitting and evaluation were repeated 1000 times for each training sample size such as an average error estimated over iterations was computed as a smooth estimate of its true value. This procedure was implement under R statistical software using the hand-make function (Team, 2004).

2.2 Model predictive errors

Minimum, mean and variance of the misclassification error rate estimated over the 1000 iterations were assessed for each modelling method to discriminate classes. Then, the minimum sampling size required for estimating 90 % percent of the maximum accuracy, i.e the lowest error rate reachable in the iterative modelling procedure, was determined Stockwell and Peterson (2002) and was compared along modelling techniques.

Model abilities to discriminate each class individually were then assessed. Class niche were defined here as data points distribution within the environmental space without any reference to niche ecological definition and was thus designed by the term "envelope". Class envelopes were represented by ellipsoids in the reduced environmental space defined by the 2 first discriminant functions of a MDA. Means and standard deviations of the Euclidean sampling points distances from class centroids were taken as envelope ranges and their standard variations in the discriminant space, respectively (Kadmon et al., 2003). We compared minimum reachable misclassification error rate as well as minimum sample size required to estimate 90 % percent of it between classes.

3 Results

3.1 Comparison of overall modelling technique performances

The minimum and mean error rate over 1000 iterations were ranked in increasing order as follows: MARS (min=0.03; mean=0.08), MDA (0.23 and 0.25), LDA (min=0.24 and mean=0.27) and MM (min=0.42 and mean=0.49) (Figure 1 a)). However, the variance of the testing error estimation over iterations was higher for MARS model (0.01 in mean over sampling sizes) than for the others modelling techniques (0 in mean over sampling sizes). However, the variance reached zero for sample size equal to 88 whatever the modelling technique (Figure 1 b)).

Test error rates decreased with increasing training sample sizes. Mean error rates were reached at training sample sizes varying from 8 to 88 observations according to the modelling technique (Figure 1 a)). Despite efforts made to reduce it (1000 iterations), some irregularities in such negative slopes were observed. Sudden decreases followed by increasing error rates were observed at sample size 24 for MDA and MM predictions and at sample size equal to 32 for MARS predictions. On average, the steepest slope (i.e the larger gap) between the minimum sample size estimate and the mean testing error rate was obtained for the MARS model. The MARS model provided lower error rate than MM, LDA and MDA for a sample size equal to or higher than 32.

Ninety % of the minimum error rate (Error rate = 0.089), was reached for a sample size equal to 88 for MARS model predictions (Figure 2). A similar minimum sample size was obtained for LDA. Finally, for MDA and MM, taking apart the discontinuities in error curves, minimum sample size required were 88 and 200 respectively.

3.2 MARS performances per assemblage

MDA discriminated the 3000 data points with a error rate of 0.25 and the two first discriminant functions explained 95.6 % of the whole data variability (Figure 3). Class 3 envelope was larger and more dispersed (mean=15.1; sd=11.8) than those of the other 2 classes (Figure 3). Class 1 envelope was slightly larger than of class 4 (mean=5.8 and mean=5.5 respectively) but was less dispersed (sd=6 and sd=12.3 respectively).

The maximum reachable accuracy was the lowest for class 1 (error=0.01) while it was the highest for the class 3 having the largest envelope (error=0.11) (Figure 4). Also, 90 % of the minimum reachable error rate required less observations for the class having the smaller envelope (72) than for class 3 and 4 (104).

4 Discussion and conclusion

Wisz et al. (2008) previously showed that MARS performances dropped drastically with decreasing sample size (occurrence records) of 10 species including birds, small vertebrate and plants, and were lower than GLM whatever the sampling size. In our study case, the steepest slope of the error rate distribution according to increasing sample size was obtained for MARS error rate. This suggests that MARS is highly sensitive to sample size effect of our simulated classes. However, in a context of classification of multiple class responses, we showed that MARS provided more accurate discrimination than did LDA, MDA and GLM above a sample size of 32 observations. Below such sample size, MARS performances were much lower and variable than those of the other modelling techniques. Those results emphasise that MARS is a more accurate predictive tool for discriminating classes than other parametric methods above 32 observations and that such predictive performances were not penalised by over-fitting problems, the major threat to this modelling technique performances (Leathwick et al., 2006). However, sample size required per class to achieve good predictive accuracy (lower than 0.1) for such non linear and non parametric class modelling (88 observations) was higher than 10 found for GARP modelling technique (Stockwell and Peterson, 2002).

In a classification context, the sample size required to reach % 90 of minimum accuracy by Discriminant Analysis was about 88 observations which was a lower size than those estimated by previous studies at about 150 and 300 occurrence data points to model and predict individual species distributions (McPherson et al., 2004; Cumming, 2000). Such results are not strictly comparable since techniques to estimate the optimal sampling size differ. However, the over-performance of multiple classes discrimination over individual species discrimination is understandable since multiple classes discrimination is fitted on a larger data set consisting of all class observations.

The sample size required to estimate 90% of the maximum accuracy estimated in this simulation study is clearly theoretical. In reality, it should vary according to assemblage prevalences and geographical ranges, parameters considered as constant in our study case. Sample size corresponds to a given amount of informations regarding niche breadth in the environmental space (range overlap) that decreases with sample size (Wisz et al., 2008). The amount of such variability encompassed by the observations should rely on assemblages and their niche breadths. Indeed we found that the minimum required sample size was lower for the class having the smallest niche breadth and a large standard deviation. Such results should however be considered cautiously since niche breadth measurement was based on a bell-shape niche and that we only compared 3 classes. In our study of small mammal assemblage distributions in 2 study sites in Sichuan (China), sample sizes varied from 5 to 34 for the 8 assemblages we defined (Vaniscotte et al., 2009). Regarding our simulation experiment we can expect a considerable improvement of MARS model prediction accuracy (estimated at 0.2 error rate on our real data set) by increasing sample size. Sample size per assemblage should thus be increased until capturing the whole niche variability (niche breadth) in the environmental space.

References

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. Journal of Applied Ecology 43, 413–423(11).

Biernacki, C., Celeux, G., Govaert, G., Langrognet, F., 2006. Model-based cluster and discriminant analysis with the mixmod software. Computational Statistics and Data Analysis 51 (2), 587–600.
URL `http://www-math.univ-fcomte.fr/mixmod/`

Cumming, G., 2000. Using between-model comparisons to fine-tune linear models of species ranges. Journal of Biogeography 27, 441–455.

Kadmon, R., Farber, O., Danin, A., 2003. A systematic analysis of factors affecting the performance of climatic envelope models. Ecological Applications 13 (3), 853–867.

Leathwick, J., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptative regression splines for statistical modelling of species distributions. Ecological Modelling 199, 188–196.

McPherson, J., Jetz, W., Rogers, D., 2004. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? Journal of Applied Ecology 41, 811–823(13).

Stockwell, D., Peterson, A., 2002. Effects of sample size on accuracy of species distribution models. Ecological Modelling 148 (1), 1–13.

Team, R. D. C., 2004. R: a language and environment for statistical computing. Fondation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0.
URL `http://www.R-project.org/`

Vaniscotte, A., Pleydell, D. R., Raoul, F., Quéré, J. P., Jiamin, Q., Wang, Q., Tiaoying, L., Bernard, N., Coeurdassier, M., Delattre, P., Takahashi, K., Weidmann, J.-C., Giraudoux, P., 2009. Modelling and spatial discrimination of small mammal assemblages: An example from western sichuan (china). Ecological Modelling 220 (9-10), 1218 – 1231.

Vaniscotte, A., Raoul, F., Pleydell, D., Giraudoux, P., in prep. From field trapping data to regional predictive mapping of small mammals assemblage habitats in sichuan province, china.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., Guisan, A., 2008. Effects of sample size on the performance of species distribution models. Diversity 38; Distributions 14, 763–773(11).

5 Figures

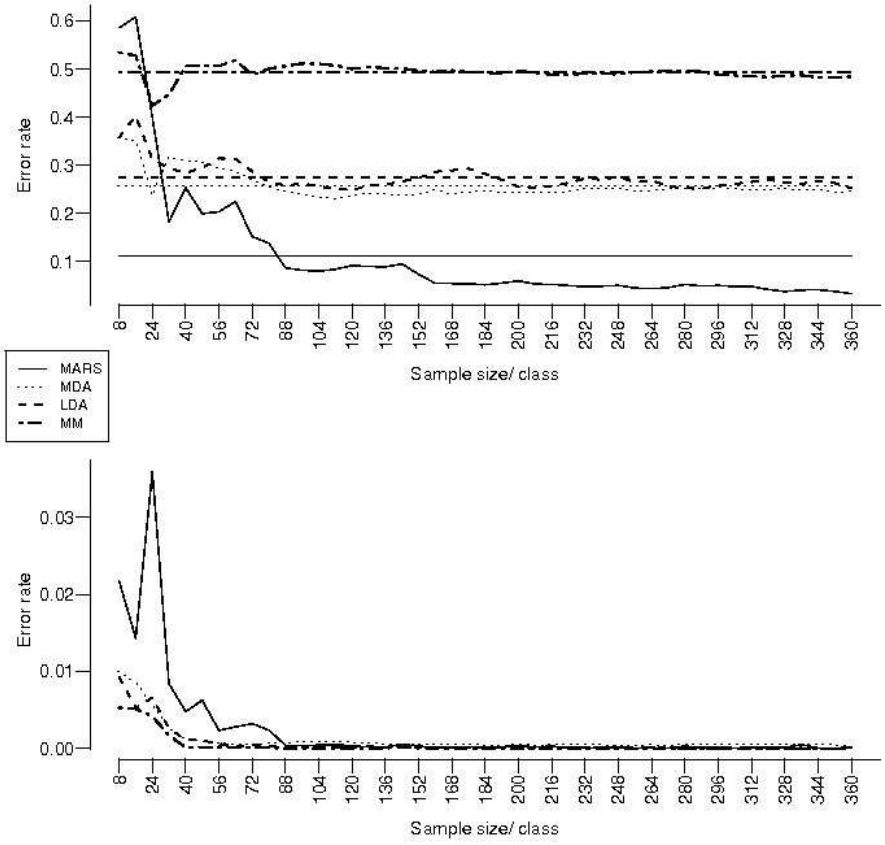

Figure 1: Test error rate estimated for each modelling method: MARS (black line), MDA (red line), LDA (green line) and MM (blue line) at increasing sample size from 8 to 360 observations.

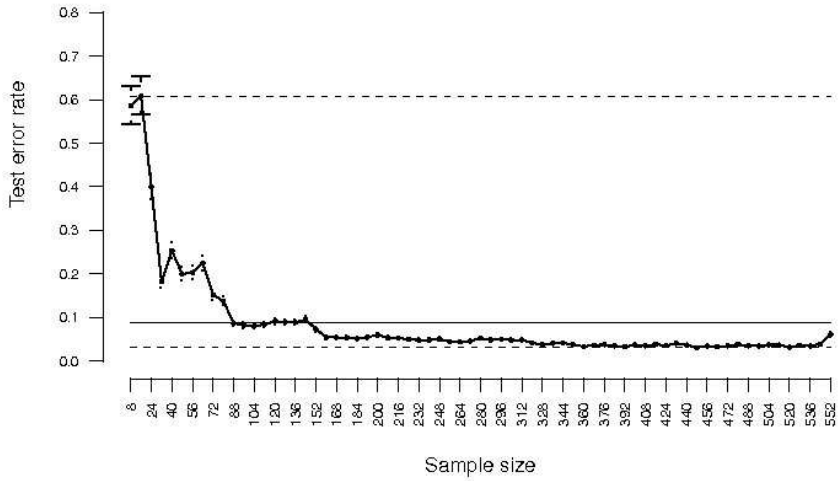

Figure 2: Error rates of MARS model predictions estimated at increasing sample size per class (full curve). Black line stands for 90 % of the maximum accuracy estimated for MARS predictions while dotted lines correspond to minimum and maximum error rates. Confidence intervals for mean error rate estimation are also provided at each sample size.

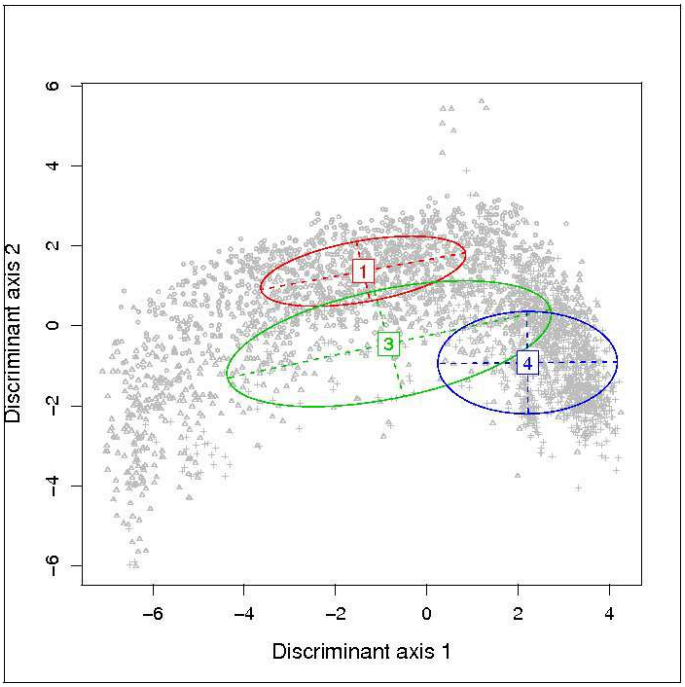

Figure 3: Distribution of class observations used to train and test the model (1000 per class) in the space defined by the 2 first discriminant functions. Envelope for each class is represented by an ellipsoid that incorporates 95% of class inertia.

4.4 Principaux résultats du chapitre

Évaluation externe des modèles locaux

Les modèles entraînés localement (MARS) étaient faiblement transférables : ils n'ont pas permis de prédire précisément les absences des assemblages sur des jeu de données indépendants et récolté sur un site distant d'une centaine de kilomètres du site d'entraînement du modèle.

Les faibles prévalences des assemblages sont susceptibles d'être une source d'erreur de prédiction. De plus, des confusions dans la définition de certains assemblages (une espèce généraliste à aire de répartition régionale), et dans la discrimination des habitats à l'échelle régionale, ont erroné nos prédictions.

L'expérience de simulation confirme les capacités prédictives supérieures de MARS sur les autres techniques au delà de 32 observations par assemblage. Elle souligne également l'importance de l'influence des tailles d'échantillons sur la qualité des prédictions et nous a permis d'estimer une taille d'échantillon optimale et théorique.

Construction d'un modèle régional

Dans notre contexte, la construction d'un modèle de classification régionale des assemblages, entraîné sur les données de piégeage obtenus sur les différents sites d'étude, est la méthode la plus appropriée (moindre erreur de prédiction) pour prédire.

Elle permet de cartographier les groupes d'espèces et d'habitats à condition de connaître le pool régional d'espèces susceptibles d'être présentes dans la région.

Chapitre 5

Axe 3 : Le rôle du chien domestique dans la transmission

5.1 Rappel des objectifs

Au regard des prévalences du parasite dans les populations humaines, l'intensité de transmission d'*Em* apparaît plus élevée dans les aires d'endémie situées sur le plateau tibétain que dans celles de ses contreforts (Maerkang, Rangtang). Les travaux antérieurs à ce travail ont mis en évidence que les facteurs écologiques de cette situation épidémiologique étaient principalement les fortes densités observées pour plusieurs espèces de micro-mammifères potentiellement hôtes pour *Em* favorisées par les pratiques agricoles et la disponibilité, en grande proportion, d'habitats optimaux dans ces paysages. D'autre part, les prévalences observées chez le chien domestique ainsi que les études des facteurs de risque apportent des preuves de l'implication du chien domestique dans la transmission à l'homme du parasite. Cela laisse supposer l'existence d'un cycle péri-domestique dans ces zones qui implique alors que i) le chien prédate des hôtes intermédiaires infestés et que ii) il contamine l'environnement des hommes. Cependant, peu d'études se sont penchées sur de tels comportements, si bien que l'intensité et les modalités de l'implication du chien dans le cycle sont à ce jour peu connues dans ces aires d'endémie.

Nous avons, dans ce chapitre, étudié les comportements à risque de chiens domestiques dans quatre villages du plateau tibétain situés dans le comté de Shiqu. Des outils d'écologie comportementale et de biologie moléculaire ont été utilisés pour analyser les comportements au niveau des populations.

Premièrement, l'analyse de la distribution des fèces de chiens et de leur contamination nous a permis de répondre à des questions relatives au rôle du chien dans la contamination de l'environnement :

Quelle est la part du chien dans la contamination de l'environnement des hommes en comparaison à celle de l'hôte définitif sauvage qu'est le renard ?

Comment se distribuent les fèces de canidés et leur contamination dans les environnements des hommes et des hôtes intermédiaires ?

Dans un deuxième temps, l'étude des comportements spatiaux, à partir de données de

télémétrie, nous ont permis d'explorer les distributions des zones où les activités à risque des populations de chiens ont lieu, et nous nous sommes demandés :

Quelle est l'intensité de fréquentation par les chiens des environnements des hommes et des micro-mammifères ?

Plus spécifiquement, nous avons tenté de localiser les zones où la rencontre entre les chiens et les micro-mammifères et par conséquent la contamination du chien peut se dérouler. Des analyses préliminaires de trajectoires individuelles ont été réalisées dans la perspective d'identifier des comportements "à risque" de recherche de proies. Enfin, l'influence de certains traits d'histoire de vie sur les variabilités inter-individuelles de ces comportements a été explorée.

Ces analyses ont été réalisées en collaboration avec les équipes de l'université de Stuttgart-Hohenheim (Thomas Romig) et de Salford (Jasmin Moss). Elles font l'objet d'un article de recherche prêt à soumission, présenté ci-après.

5.2 Article - Les fèces de canidés et le comportement des chien comme sources de contamination des villages tibétains

Dog behavior and contamination with *Echinococcus multilocularis eggs* in Tibetan villages

Amélie Vaniscotte[1], Francis Raoul[1], Marie-Lazarine Poulle[2], Thomas Romig[3], Anke Dinkel[3], Kenichi Takahashi[4], Marie-Héléne Guislain[2], Jasmin Moss[5], Tiaoying Li[6], Qian Wang[6], Jiamin Qiu[6], Patrick Giraudoux[1]

Department of Chrono-environment, UMR UFC/CNRS 6249, Universite de Franche-comte, 25030 Besanon cedex.

1. 2C2A-CERFE, rue de la Heronniere, 08240 Boult-aux-Bois, France.
2. Department of Parasitology, University of Hohenheim, 70599 Stuttgart.
3. Hokkaido Institute of Public Health, Kita 19, Nishi 12, 060-0819 Sapporo, Japan.
4. Cestode Zoonoses Research Group, Division of Biological Sciences, School of Environment and Life Sciences, University of Salford, The Crescent, Salford, M5 4WT, UK.
5. Institute of Parasitic Diseases, Sichuan Center for Disease Control and Prevention, Chengdu 610041, Sichuan, China.

Corresponding author: Amélie Vaniscotte, Department of Chrono-environment, UMR UFC/CNRS 6249, University of Franche-comte, 25030 Besançon cedex. E-mail address: amelie.vaniscotte@univ-fcomte.fr. Telephone number: (0033) 03 81 66 57 14. Fax number: (0033) 03 81 66 57 97

Abstract

On the Eastern Tibetan Plateau region of Sichuan province, China dogs are regarded as the main definitive hosts of *Echinococcus multilocularis* involved in a peri-domestic cycle through their predation on wild intermediate hosts (small mammals) and contamination of the human environment *via* deposition of infectious faeces. We analyzed and quantified dog behavioral paramaters important for the parasite transmission in four villages situated in Shiqu county by assessing the distribution of contaminated faeces and dogs' space utilization in human and small mammal environments. We found that in the four villages, dog faeces were the main source of contamination of the human environment (78 % of collected faeces was of dog origin, among which 23 % were positive for *E.multilocularis*) in comparison to fox faeces which were found less intensely. Contaminated dog faeces and host to host contact rate, were found to be heterogeneously distributed within 1500 m dog ranges around village houses, showing similar patterns in the four villages. Faeces and evidence of human exposure to the parasite were mainly distributed around dog owners' houses where dogs spent the majority of their time (0-200 m). Moreover contamination could be aggregated in some micro-foci within village areas where groups of dogs defecated preferentially. In parallel, small mammal densities decreased from the dog core areas to the periphery of the dog range to areas used occasionally (grassland surrounding villages). Male dogs were found to move farther than females. This study constitutes a first attempt to quantify the role of dogs in *Echinococcus multilocularis* peri-domestic cycles. Further incorporation of human behavioral parameters, individual factor of dog space variations and quantification of dog predation pressure are required for a more complete estimation of transmission risk to humans and contact rate between definitive and intermediate hosts.

Key words: transmission ecology, *Echinococcus*, domestic dogs, faecal contamination, spatial distribution, space utilization.

1 Introduction

Alveolar Echinococcosis (AE) is an often fatal zoonosis, distributed in arctic and temperate regions of the northern hemisphere. It is caused by the ingestion of the cestode *Echinococcus multilocularis* (*Em*) eggs which are disseminated into the environment by definitive hosts faeces. Definitive hosts are mainly canids and particularly red foxes (*Vulpes vulpes*) in the paleartic Euro-Asian region (Eckert and Deplazes, 2004). Canids become infected by ingestion of small mammal intermediate hosts. Humans can be considered as accidental intermediate hosts for this parasite. The incidence recorded for 100 000 inhabitants in endemic

areas of Europe can reach 0.74 per year (Vuiton et al., 2003). However, some foci of very high prevalences in humans were recorded in China, where a maximum of 14.3 % prevalence was observed in villages of the Eastern Tibetan Plateau (Sichuan province) (Li et al., 2005). The Tibetan plateau, situated between 3200 and 5000 m above sea level, is inhabited by a large diversity and abundance of host species and is regarded as a meta-stable reservoir for *Echinococcus multilocularis* (Giraudoux et al., 2006). Landscape is dominated by grassland and the main resource for people is provided by breeding yaks (*Bos grunniens*). The grazing pressure has been found to provide optimal habitats for the intermediate hosts of *Echinococcus multilocularis; Ochotona curzoniae* and *Microtus limnophilus* whose spatial distributions are clustered (colonies) (Raoul et al., 2006). Four wild canid species (*Vulpes vulpes*, *Vulpes ferrilata*, *Vulpes corsac* and *Canis lupus*) as well as the domestic dog (*Canis familiaris*) are known as potential definitive hosts. High prevalences of 59.1% and 57.1% have been recorded in Tibetan foxes and red foxes respectively (Qiu et al., 1995) while only 12 % and 14.8 % prevalence has been found in owned dogs following purgation studies in Shiqu county (Budke et al., 2005a; Wang et al., 2009). However, dogs are numerous in Tibetan villages where they are kept by humans for use as cattle guard dogs, and they are protected by the local religion (Buddhism). It is a rarity to see a house without dogs and stray dogs are frequently observed roaming inside and outside of villages.

In such context, dogs are currently assumed to be the definitive host in a peri-domestic cycle of *Echinococcus multilocularis* (parallel to the wildlife cycle) maintained by predator/prey interactions between dogs and small mammals. This assumption was confirmed by preliminary analysis of dog faeces collected on the Tibetan plateau which found small mammal remains in 28.8% of dog faeces (Wang et al., 2009). Furthermore, evidence for dog/small mammal interactions have been outlined by a positive correlation found between the proportion in the landscape of fenced pastures, optimal habitats for small mammals, and *Em* prevalences and worm burden in owned dogs (Wang et al., 2004; 2009). Finally, epidemiological studies have identified dog ownership, number of dogs owned and relationships with dogs (ie contact) to be highly correlated with AE prevalences in human communities living on the Tibetan plateau (Li et al., 2005; Budke et al., 2005a). Thus, it is hypothesised that dogs are responsible for the high human AE prevalence rates due to contamination of their environment with infected faeces. However, to our knowledge, no studies of the behavioral ecology of the dog population have been completed in this area.

Generally, studies on the role of behavioral ecology of dogs in parasite transmission cycles are rare, despite the fact that dogs are known to be hosts for 60 zoonosis (Macpherson et al., 2000). Behavioral ecology provides a means to study the role of dogs in transmission by identifying and quantifying relevant parameters such as their defecation and predation activities. Most studies of dog populations in an epidemiological context have concentrated their efforts on the population size as well as sex and age structure in order to develop control options (Anvik et al., 1974; WHO and WSPA, 1990). Only few studies have focused on the behavioral ecology of free-roaming dogs even when such animals are often considered as the main source of

contamination of the human environment due to their close proximity to humans (Craig et al., 2000; Macpherson, 2005; Thompson et al., 2009). The spatial distribution of the environmental contamination *via* the analysis of soil contamination or defecation behavior has been investigated in canid (Anvik et al., 1974; Habluetzel et al., 2003; Robardet et al., 2008) and felid species (Afonso et al., 2008). Modern non invasive methods have recently been developed to facilitate the assessment of definitive host infection status from fecal material via molecular analysis, e.g polymerase chain reaction (PCR) or immunological analysis, e.g. copro-antigen enzyme linked immunosorbant assays (ELISA) (Dinkel et al., 1998; Raoul et al., 2000). By contrast, the analysis of the movements of the definitive host have been very rarely incorporated into epidemiological studies (Robardet et al., 2008) despite its potential to explain infection patterns in domestic hosts. For example, the speed between successive relocations is likely to decrease when predator-prey interaction arise (Kareiva and Odell, 1987) and movement parameter can be used to identify foraging behavior.

On the Tibetan plateau, the planning of control programmes which prevent the transmission of *Echinococcus multilocularis* are currently based on two main hypotheses related to the dog's double role in transmission: their contamination of the human environment and their predation on the small mammals reservoir. In such a context, our general objective was to identity and quantify the role of domestic dogs in *Echinococcus multilocularis* peri-domestic transmission cycle in this area. We studied patterns of dogs' defecation and space utilization, and addressed the following questions:

i) Do dog or fox faeces constitute the main source of *Echinococcus multilocularis* contamination in the human environment?

ii) How are *Echinococcus multilocularis* contaminated canid faeces distributed *within* and at the *periphery* of the human environment (i.e. village areas)?

iii) What are the distributions (extent and frequency) of dog movement relocations within the human and wild intermediate host environment?

iv) How large is the inter individual variation of dog movement parameters and thus "risky" spatial behaviors?

v) Are there specific areas in the human environment where dogs could potentially be infected due to overlap of the dog ranges and small mammal colonies?

A new coprodiagnostic test (real-time PCR) combining parasite detection with discrimination of the canid host species from which the faceal sample derived provided us with a tool to quantify the role of dogs *versus* foxes in *Echinococcus multilocularis* environmental contamination. We then investigated the spatial distribution of canid faeces and those which were positive for *Echinococcus multilocularis* in relation to human and wild intermediate host population habitats (inside and outside of the village area). We also studied the distribution of individual dog nocturnal movements in relation to dog owners' houses. Dog individual traits (e.g. sex, relationships with owners) were analysed in order to explore inter-individual variability of

movement/spatial behaviour and the influence these traits had on dogs which would roam beyond the main dog population range. Finally, we analyzed as a preliminary step the spatial overlap between small mammal population distributions and dog space utilization.

2 Materials and methods

2.1 Study sites

Shiqu township (Sichuan province, China) is situated between 3900 and 4200 metres above sea level on the Tibetan Plateau. Four villages in the vicinity (20-85 km) of Shiqu township were investigated (Figure). The study commenced in May 2006 in Yiniu and Qiwu villages. In the following May (2007) these villages were revisited along with a further two sites: Mengsha and Jiefan. Villages were characterized by hundreds of scattered houses spread out along rivers and crossed by tracks. Within village, areas bare ground typified the earth in contrast to the surrounding landscape which was dominated by fenced or grazed grassland. Bushes were sparsely distributed on sloped grassland whilst wet grassland with micro-topography could be found near rivers.

The four villages were selected according to their epidemiological situation regarding alveolar echinococcosis prevalences recorded in humans (Li et al., 2005). Qiwu and Mengsha were identified as low prevalence areas (2 ± 0.9 % and 3 ± 5.8 %, respectively) while prevalences of 11 ± 13.3 % and 12 ± 8.8 % were observed in Jiefan and Yiniu. An epidemiological survey done twice in 2006 on owned attached dogs provided an approximative size of 31 to 64 owned dogs per village (Jasmin Moss, unpublished data) (Table 2). Dog population size was underestimated in Jiefan since only 70 % of the households have been sampled.

2.2 Canid faeces and environmental contamination

2.2.1 Faeces sampling and copro-assays

Canid faeces were sampled in one hectare quadrats located inside and outside of villages. A total of 33 quadrats have been sampled during 6 sampling surveys: 17 inside and 16 outside villages (Figure). Inside villages, quadrats were in close vicinity to houses, while outside villages quadrats were situated in grassland surrounding village cores. All the faeces found within each quadrat were counted and georeferenced systematically. Faecal densities were estimated as the number of counted faeces per quadrat.

Furthermore, in 27 out of the 33 quadrats we collected a minimum of 25 (in 2006) and 15 (in 2007) of the faeces samples for coprodiagnostic tests in order to detect DNA of *Echinococcus multilocularis* and canid host species. Two newly developed LightCycler-PCR assays were used:

1) a real-time multiplex nested PCR with hybridization probes and subsequent melting curve analysis for

the combined detection of *Echinococcus* species and host species DNA of *Vulpes vulpes*, *V. ferrilata*, *V. corsac* and *Canis lupus/familiaris* parallel in one reaction tube (Dinkel, unpublished data),

2) a real-time nested PCR with hybridization probes and subsequent melting curve analysis only for discrimination of the host species described above (Dinkel, unpublished data).

Additonally, we confirmed the detection of E. multilocularis by using a published nested PCR assay (Dinkel et al., 1998). DNA from fecal samples was extracted as described in Dinkel et al. 1998. To exclude false-negative results due to inhibition factors the isolated DNA from each fecal sample underwent an inhibition control.

Difference in *Echinococcus multilocularis* fecal prevalence between quadrats inside and outside of village areas was tested using a Chi-squared test.

2.2.2 Faeces distributions between and within quadrats

Random *versus* aggregated distributions of faeces density between quadrats were assessed in each sampling survey. The *ratio* of faeces variance density over its mean was used as an index of the aggregation. Null Generalized Linear Models assuming Gaussian, Poisson (random) or Negative Binomial (aggregated) faeces density distributions fitted on the whole data set were then compared. Response distribution providing the lowest AICc was used in further modelling.

We then tested the effect of the type of quadrat (inside *versus* outside village) on quadrat faeces density. Because houses were scattered and village boundaries hard to assess, we completed our analysis by investigating how faeces density varied according to the quadrat distance from houses. This provided us with additional information concerning the range of faeces distribution, and potential environmental contamination around houses. Generalized Linear models (GLM) incorporating the effect of the type of quadrat (inside *versus* outside village) and of the minimum distance from village houses (alone and associated) were evaluated and compared using AICc. Before running the models, we tested whether it would be benificial to add random effects on the year and village (sampling survey). The effects of those variables on the quadrat type and the minimum distance from houses were assessed by running GLMs and comparing their AICcs to the null model. Random effects were incorporated in models if those variables explained a part of the variability.

Random *versus* aggregated distributions of faeces within quadrats were then estimated by considering faeces as points in point spatial pattern analysis. The aggregation in each quadrat was estimated using the Clark and Evans index (R index) (Clark and Evans, 1954). This first order aggregation index was estimated as the *ratio* of the mean nearest neighbors distances over all points over the mean distance under random distribution assumption, corrected for edge effects in distance estimates. Indices were estimated using the function *clarkevans()* of the spatstat package under R statistical software (R Development Core Team, 2008).

We finally assessed if the proportion of quadrats with aggregative distribution differed inside and outside village areas using a Chi-squared test.

2.3 Dogs' nocturnal trajectories

2.3.1 Data collection

Two levels of dog freedom (amount of time they are tethered) were sampled: dogs attached all day and released at night (Owned Attached, OA dogs) and dogs always free (Owned Free, OF dogs). Dog owners supplied information regarding the sex of the dog and whether it usually visits summer pastures. The latter variable was named “pastoral activity” in the following analysis. Relocations were recorded by GPS collars (WildTrax, Bluesky Telemetry) that were set up on dogs during the day and removed the day after, one location being recorded every 10 minutes.

Since OA dogs were only free from thether at night, we focused on studying nocturnal trajectories for the whole dog population. The time period of relocations records was homogenized between dogs: the beginning of the night (8:00 pm) was set as the earliest hour at which owned attached dogs were known to be released, while the end of the night was considered as the latest hour at which the same dogs were known to be re-attached (8:00 am), providing trajectories of 12 hours for all dogs. However, despite efforts made to reduce it and regarding the lack of more accurate informations on the release and re-attachment hours, trajectories for OA dogs may include an unspecifiable time period when dogs were attached.

2.3.2 Distribution of individual dog nocturnal relocations

Relocation distribution were analyzed in relation to the distances from the dog to its release point (RP), i.e its owner’s house. The core area, classically used in the context of home range estimates (Seaman and Powell, 1990; Hodder et al., 1998), was used here as a mean to define areas around dog owners’ houses where dog relocations were aggregated, i.e where dogs spent the majority of their time. Minimum Convex Polygon (MCP) areas were estimated for different percentages of included relocations on the basis of their distances to their barycenter, i.e their release point. Friedman test followed by multiple comparisons were used to identify the percentage of relocations included in MCP estimation that lead to a significant increase in the area. The nocturnal core area for each dog was estimated as the MCP area that included the relocation below such threshold. We then estimated the maximum distance from the RP for dog relocations included in such nocturnal core areas. Then, dog trajectories were described by estimating the maximum distance travelled from the dog RP. Finally, activity was measured as the mean distance travelled per hour (Ciucci et al., 1997).

2.3.3 Characterization of excursive trajectories

Home range could be defined as “the area over which an animal or population normally travels in pursuit of

its routine activities", i.e including nesting sites, shelter, locations for resting, food-gathering and mating but excluding ventures sallies outside this area (Okubo and Levin, 2001). Excursive trajectories could be defined as animal movements falling outside this home range. Due to the limited period of sampled trajectories in this study (one night), dog home ranges could not be estimated. Instead, we defined their 'one night activity ranges', as their home range estimated on a delimited time period (Okubo and Levin, 2001). The strong dependence observed between individual dog relocations prevented us to estimate individual one night activity range. Therefore sub-sampling each dog trajectory results in the estimation of the populations' nocturnal activity range, which is less prone to relocation dependency bias and represents the main area used by the whole population (Boitani et al., 1995). This population nocturnal activity range included zones where individual interactions could take place.

The nocturnal activity range for the whole dog population was estimated for each sampling survey on the set of sub-sampled relocations within all dogs trajectories, i.e one location/hour/dog. Utilization Distribution functions were estimated by Kernel method (Worton). The smoothing parameter *h* was first estimated by the ad-hoc method available in the function kernelUD of the adehabitat package (R Development Core Team, 2008) providing an *href* value. However, this method is known to suffer from over-estimation of the area when multiple clustering is observed in the relocation data set, which was the case here since dog relocations were aggregated around their RP (Seaman and Powell, 1996). Also, in order to minimize the error of type II (inclusion of regions which are not part of the activity range), the optimal *h* was estimated for each sampling survey as a rescaling of *href* providing an area that did not connect clusters between which no moves (successive relocations) have been recorded. *href* was rescaled by factors varying between 0.9 and 0.1, every 0.1 interval. Areas provided by each rescaling were compared. The dog population nocturnal activity range was defined as the area in which dogs spent the majority of their time - we defined this on 90% of the relocations.

The delineation of the populations' nocturnal activity range allowed us to define excursive trajectories (dog relocations falling outside the dog population activity range). The number of such excursive relocations were assessed per dog and main patterns of excursive trajectories were identified. Then, excursive visiting paths were identified as at least 3 successive excursive relocations. Finally, as a potential surrogate for foraging behaviors, we explored the existence of aggregative relocation distribution for each dog by estimating the Clark and Evans index on excursive relocations.

2.3.4 Correlations of inter-individual movement parameters

Correlations between the maximum distance from the RP, the speed, the core area and the number of excursive relocations per dog were assessed and tested by Spearman correlation tests. Then effects of dogs' sex, freedom from tether and pastoral activity on normalized (log transformed) movement parameters were investigated among the dog population *via* Linear models. The effects on the number of excursive relocations

were assessed by a GLM assuming a Negative Binomial distribution for the response variable. For each movement parameter, we tested and compared the null model with all possible combinations of covariates, on the basis of their AICc. For each response, a preliminary test for the addition of random effects on the sampling survey and dog individuals was done. This was achieved by running the LM and GLM models against each of those variable and by comparing their AICc with the null model.

2.3.5 Spatial interactions between dogs and small mammal populations

In the four villages, radial transects- along which indices of activity (faeces and holes) of small mammals (mainly *Ochotona* and *Microtus spp.*) - were recorded within every interval of ten paces (Raoul et al., 2006). Transects started from the center of the village and ended in the surrounding grassland. We first estimated the minimum distance separating positive intervals for small mammal presence to all dog RP at each village. Differences in frequency distribution of positive intervals were assessed between the dog population activity areas and the excursive areas only. This was due to low or even null numbers of transect intervals being situated within individual dog one night core areas. A Generalized Linear Mixed Effects model assuming a logistic Binomial link function for the response, i.e the presence/absence of small mammals, including a random effect on the sampling survey, was used to investigate such differences. Model AIC was compared with the null model and the risks ratio was provided to identify the difference (Bailey and Alimadhi, 2007).

All statistical analyses were achieved using the R statistical software (version 5.4) (R Development Core Team, 2008). Movement parameter, kernel areas estimation and analysis of trajectories were estimated using the adehabitat package (Calenge, 2006).

3 Results

3.1 Canid faeces and environmental contamination

3.1.1 Faecal prevalences

A total of 980 faeces were counted and georeferenced in the four villages, among which 284 faeces were analyzed for host identification and *Echinococcus multilocularis* contamination (Table 1, details per quadrat provided in Appendix 1). PCR inhibition problems were observed for 74 (46.8%) of the 158 faeces collected in 2007 and 37 (23.8%) of the 126 faeces collected in 2006 (Table 1, details per quadrat provided in Appendix 1). These faeces could not be used for further analysis of fecal prevalences. Among the remaining 173 faeces, 142 were found to be dog faeces, 13 fox (*Vulpes ferrilata*) faeces and 25 could not be identified. The overall infection rate was 20.81%, with 22.53% and 15.38% respectively for dogs and foxes (Table 1). We failed to detect a significant difference in overall fecal prevalence distribution inside *versus* outside

villages (Chi-squared test, =0.55, df=1, p-value=0.5).

3.1.2 Faeces distribution between and within quadrats

Aggregation indices were greater than 1 in all sampling survey as well as for all confounded (Table 1). Clearly, faeces were over-dispersed in all villages and faecal density distributions were better captured on the whole data set by a model incorporating a Negative Binomial distribution (AICc=281.62 *versus* AICc=372.32 and AICc=2343.72 for Gaussian and Poisson distribution respectively). Consequently we used a Negative Binomial faecal distribution for further modelling.

We failed to detect effects of the year and the village on faeces densities (Δ_{AICc}=0.03, w_i=0.5 and Δ_{AICc}= 2.49, w_i=0.2, respectively). Those variables were thus not considered as random effects in the further models. We also failed to detect an effect of the type of quadrat on faeces densities (AICc=284.3, Δ_{AICc}=146.2, w_i=0). However, the model including the minimum RP distance provided the approximate best fit (AICc=138.1, w_i = 1). Faeces density reached a maximum (354 faeces/ha) in one quadrat situated at a minimum distance of 48 m from village houses while all quadrats situated farther than 200 m from village houses contained less than 14 faeces (Figure). Also, low faeces density was observed for 4 quadrats situated at less than 50 m from houses suggesting that the model particularly over-fitted faeces densities for those quadrats (Figure).

Finally, faeces were aggregated in 84% of the quadrats (Appendix 1). No significant difference was observed between the proportion of quadrats showing aggregation situated inside (88) and outside (81) villages (Chi-squared=0.01, df=1, p value=0.9).

3.2 Analysis of dog nocturnal trajectories

3.2.1 Sampling effort

Seventy-eight dogs were equiped for one night, 7 were equiped for 2 nights, 2 for 3 nights and 1 for 4 nights. A total number of 96 one night trajectories were thus recorded in the four villages (Table 2). Problems of GPS collar connections to satellites induced missing values that were excluded from the trajectories for further analysis. Thus, a mean of 19.5% (sd=22.9%) relocations were missed over all dog trajectories. The mean positioning error estimated over all dog relocations varied according to the sampling survey: 4.7 m (sd=2.9) and 10.5 m (sd=4.6) horizontally and 6.8 m (sd=5) and 15.20 m (sd=8.3) vertically.

3.2.2 Distribution of individual dog one night relocations

MCP areas differed among the percentages of relocations included in their estimation (Friedman test, Friedman chi-squared=837, df=9, p-value < 2.2e-16) (Figure 4). Most relocations were aggregated in areas smaller than the total activity area which relied on few exceptional relocations. When including 80% of the relocations, a median value of 10% and an outlier at a maximum of 67 % of the total activity area could be

captured among dogs. A very low proportion of relocations (the last 10%) provided 77% of the total area for half of the dogs and a maximum of 29% (except for three outliers). A significant difference was found between the inclusion of 80 and 100 % (p value=0.05). Therefore, one night core areas were estimated for each dog as the MCP area including 80% of its relocations.

Dogs were located at 10 to 1500 m from their release points but were generally under 250 m from this point, moving slowly (162 m/h on average) within a less than 1 ha core area for the large majority of them (Table 3). Considering the variability within the dog population, one night core areas as well as maximum RP distances were dispersed and aggregated around low values (standard deviations larger than the means). Fifty percent of the dogs had one night core areas smaller than 0.16 ha (Table 3) and stayed within a 36 m range around their RP. The same proportion of dogs travelled a maximum RP distance of 152 m. However, 5% of the dogs had one night core area larger than 1.19 ha, a maximum RP core distance at 115 m and travelled a maximum distance of 921 m.

3.2.3 Characterization of excursive trajectories

Optimal *h* (smoothing parameter) estimates for each sampling survey ranged from 35.7 (scaling factor=0.2) to 133.9 (scaling factor=0.5) (Table 4). The dogs' main activity areas ranged from 32.5 to 174.5 ha according to the village and sampling survey (Table 4). Thirty nine percent of the dogs had at least one relocation falling outside the population activity area (Table 4). The mean number of excursive relocations per dog varied from 2.7 to 5.1 depending on the sampling survey. The mean distance of such excursive relocations from the dog RP varied between 309 and 583 m according to the village, with a minimum distance of 58 m.

Among those trajectories, two patterns of excursive relocations were observed: visiting paths consisting of a minimum of 2 successive relocations (Figure 5, a) to d)), and “away and back” moves relying on isolated excursive relocation from the dog one night core areas (Figure 5, e)). Eleven dog trajectories were considered as “visitors” movements and five of them showed aggregative excursive relocations (mean(R)=0.85; sd(R)= 0.11) such as the trajectory illustrated on Figure 5, f). A maximum of 9 successive relocations, or 90 minutes excursive trajectories have been recorded for a male equiped in Qiwu 2007 and used to visit the summer pasture.

A maximum of 9 successive relocations (or 90 minutes excursive trajectories) have been recorded for a male dog equiped in Qiwu village, 2007 who was known to visit summer pastures.

3.2.4 Correlations of inter-individual movement parameters

As expected, all movement parameters were significatively correlated to each other (p values < 1.). Spearman correlation coefficients ranged from 0.53 between the number of excursions and the core area, to 0.84 between the maximum distance from RP and the speed. No effects of the individual, the year and the village have been identified using AICc comparisons. Therefore those variables were not incorporated as random

effects in the models. Dog sex slightly influenced dog maximum RP distances (LME, AICc=236.4, =2.4, =0.3), median area being larger for males (median=165) than for females (median=117). No effects of the other factors and on other movement parameters have been detected.

3.2.5 Spatial interactions of dog and small mammal populations

A total of 42 standard radial transects were done (Figure 6). Evidences for small mammal presence was found close to dog' owners' houses at a minimum distance varying between 32 and 122 meters according to the sampling survey (Table 5). Positive transect intervals (92%) have been identified inside dog individual one night core areas in Qiwu village (2007) (Table 5). The proportion of transect intervals with presence of small mammals was higher in the excursive area, ranging between 15.5 and 56.5%, than in the dog population activity area, ranging between 3 and 37% (GLME, AIC=5275, =138; Risks ratio=2.36 (1.86-2.93)).

4 Discussion

In the four villages situated in the area of highst endemicity for *Echinococcus multilocularis* of the Tibetan plateau, 7% of faeces analyzed were identified as Tibetan fox (*Vulpes ferrilata*) origin. This fox is generally known to inhabit grassland areas far from human influences (> 1000 m) (Gong and Hu, 2003; Wang et al., 2007) however our results support the additional utilization of village areas as a foraging area. Considering that 15% of fox faeces were infected, foxes should be as a source of human environmental contamination even if to a smaller extent than dogs. Indeed, 78% of faeces analyzed were identified as dog faeces and 23% of them were positive for *Echinococcus multilocularis*. Therefore our study provides additional data which suggests the dominant role of the dogs as the main definitive host responsible for the contamination of the human environment in this area.

All the dogs we collared shared a common pattern of their one night relocation distributions: they were aggregated around owners' houses. The area where the dogs spent the majority of their time at night ranged from 0.004 to 10 ha and the majority of the dogs (95%) stayed for 80% of their time at a maximum distance of 115 m from their owners' houses. That area might constitutes a small and easily defendable territory where aggressive behavior is mostly expressed (Macpherson et al., 2000; Boitani et al., 1995). In contrast, 40% of the dogs did excursions (i.e outside their population activity area) at distances varying from 58 to 1519 m from their release points (RP) and during a maximum time period of 90 minutes. Even if recorded for only 12 hours and at the individual level, the patterns of relocation distributions identified here are in agreement with those observed in feral or free-roaming dogs for which core areas (50% of their home range) corresponded to 5.71% of the total home range area, including dens, resting and retreat sites (Boitani et al., 1995). The large dispersion observed in the dog maximum RP distance, core area and core RP distance emphasized the existence of two types of dog' behavior: some fast and long distance movements *versus* a large proportion of

relatively slow movements concentrated around the owner's home. Furthermore, all the variables were correlated to each other suggesting that dogs which move farthest also have a tendency for (during the night they were radiotracked) larger core areas, speeds and excursive relocations. Also, the large dispersion observed in the number of excursive relocation and the diversity of trajectories ("away and back" and excursive paths) found among dogs also outlined a large heterogeneity in dog excursive behavior.

Because individual variability was not considered in our population-level sampling design, we could not generalize such results by categorizing dogs according to the value of their one-night movement parameter estimation. However our results agree with previous studies done at the individual level demonstrating inter-individual variability in free-roaming dogs' home ranges and discriminating sedentary (mean home range=2.6 ha) from wandering dogs (mean home range=927 ha) (Meek, 1999). Only the dog sex was found to partly explain this variability (males being more active than females) thus confirming previous studies (Boitani et al., 1995). The inter-individual variation in nocturnal behaviors might interact and hide the effects of the nature of dog/human relationships (pastoral activity and freedom level) on nocturnal spatial behavior. It is well known that dog-owner relationships influence social organization of dog and consequently their territory definition and spatial utilization- owned dogs having smaller home range size than owner-less dogs (Macpherson et al., 2000). Repeated trajectory records for each individual and stratified sampling of the population would help identify those effects.

Faecal densities were found to be higher in quadrats situated inside villages and were over-dispersed regarding their minimum distance from houses. Such aggregative distribution of faeces matched with the dog relocation distributions. The range of high faeces concentration (0 to 200 m) and the maximum density observed at 50 m fell within the maximum distance from the owner's house recorded within the core areas (345 m). Therefore, dogs mainly defecated within their core areas. In urban areas, concentration of dog faeces' have been observed in places close to houses, in parks and residential areas or in public playgrounds associated with a high rate of soil contamination by parasite, i.e *Toxocara*, eggs (Anvik et al., 1974; O'Lorcain, 1994). Concerning areas of the Tibetan Plateau, the close relationships and spatial proximity between dog and humans may cause a considerable increase of the *Echinococcus multilocularis* transmission risk. Given the average *Echinococcus multilocularis* fecal prevalence found (21%) and the dogs' spatial defecation behavior outlined above, we suggest that the village area, and particularly within about 200 meters around dog owners' houses, might be the place for a high contact rate between faeces and the human village population and cause high human exposition to the parasite. Consequently, the transmission risk is likely to decrease with the distance from the dog owner's house, if other risk behavior of people is kept constant (hygiene, contact with dogs) along this gradient. Since the villages we considered did not influence faeces spatial distribution patterns and prevalences, we could cautiously generalize such a general pattern to other villages of the study region. Furthermore, at a higher resolution, faeces were found to be aggregated within quadrats inside village area. Scent marking behavior and social interactions among dogs, arising within the

population area of activity, might explain such aggregation (Boitani and Ciucci, 1995). Faeces of different dogs are likely to be found within such micro-aggregation zones inside villages where *Echinococcus multilocularis* environment contamination might be particularly high.

In contrast to the village area, low densities of faeces were found further than 200 m from houses and outside village areas. This distance matched with the maximum dog core area RP distance and thus with the excursive areas. Therefore, dogs also defecated sporadically during their excursive trajectories at a maximum distance of 1519 m and therefore could contribute to the contamination of the wild reservoir (small mammals) mainly inhabiting grassland areas.

Infected small mammals can contain thousands of protoscoleces and dog infection pressure is high when predating on such preys (Budke et al., 2005b). The same authors ran an epidemiological model to estimate the contact rate between dogs and small mammals and found a value of 0.52 insult per year to explain actual dog prevalences in the same area (villages surrounding Shiqu township). This contact rate in the case of *Em* transmission is known to rely on prey densities and prevalences (Giraudoux et al., 2002) as well as on predator behavior (Hegglin et al., 2007). On the Tibetan plateau, relatively high prevalences have been recorded in *Microtus irene* (25%) and *Ochotona curzoniae* (up to 7.7%) (Qiu et al., 1999), species found in higher abundances and suceptible to undergo multi-annual fluctuations (Raoul et al., 2006). In the area studied here, our results suggest that spatial overlaps of dog movements with small mammal populations exist outside but also inside village area as a trade-off between small mammal population densities and dog space utilization. Indices for small mammal species presence were found with largest frequency outside dog activity areas, i.e in the areas where dogs would do only excursions. Additionally, aggregation was observed in some dogs excursive areas. Even if nocturnal excursions are done for the purpose of mating or to meet other dogs (authors personal obs.), they could also be motivated by the presence of small mammal colonies and lead to predation on this prey (authors personal obs.). Moreover, even in lower proportions than in the excursive area, small mammal occurrence was observed inside the dog population area, and even in dog core areas for one studied village. This suggests that contact between hosts could even occur inside this area and does not necessarily require excursions into the surrounding grassland. Also, a peri-domestic parasite cycle could be maintained in close vicinity of the dog owners' houses.

Further data are required to investigate the plasticity of dogs' predation behavior in relation to prey densities by dietary analysis. Since dog predation pressure is known to shift according to the variation in anthropogenic ressources (Macpherson et al., 2000), particular effort should be put on studying the effects of dog "domestication states" as defined by the feralization model developed in Boitani and Ciucci (1995), on predation behaviors. At the population level, our study constitutes a preliminary step in the quantification of the frequency of dog and small mammals interactions. Small mammal habitats have already been defined in this region of China (Raoul et al., 2006; Marston 2008) but further investigations are required to build an accurate predictive mapping of small mammal occurrence within the whole area surrounding villages.

Ultimately such contact rate could be of particular relevance as a parameter of epidemiological models (Budke et al., 2005b).

5 Acknowledgements

This work was supported by grants number RFATW-00-002 and RO1 TW001565 from the Fogarty International Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Fogarty International Center or the National Institutes of Health. We address many thanks to local people for their precious help in the fields and to the 2C2A-CERFE research team for their precious advices on space utilization analysis.

6 References

Afonso, E., Lemoine, M., Poulle, M.-L., Ravat, M.-C., Romand, S., Thulliez, P., Villena, I., Aubert, D., Rabilloud, M., Riche, B., Gilot-Fromont, E., 2008. Spatial distribution of soil contamination by *Toxoplasma gondii* in relation to cat defecation behaviour in an urban area. Inter. J. Parasitol. 38 (8-9), 1017-1023.

Anvik, J., Hague, A., Rahaman, A., 1974. A method of estimating urban dog populations and its application to the assessment of canine fecal polution and endoparasitism in Saskatchewan. Canine Veterinary Journal 15, 219-223.

Bailey, D., Alimadhi, F., 2007. logit.mixed: Mixed effects logistic. In: Zelig: Everyones Statistical Software.

Boitani, L., Ciucci, P., 1995. Comparative social ecology of feral dogs and wolves. Ethol. Ecol. Evol. 7, 4972.

Boitani, L., Francisci, F., Ciucci, P., Andreoli, G., 1995. Population biology and ecology of feral dogs in Central Italy. In: Serpell J., E. (Ed.), The domestic dog: its ecology, behaviour and evolution. Cambridge: Cambridge University Press.

Budke, C. M., Campos-Ponce, M., Qian, W., Torgerson, P. R., 2005a. A canine purgation study and risk factor analysis for echinococcosis in a high endemic region of the Tibetan plateau. Vet. Parasitol. 127 (1), 43-49.

Budke, C. M., Jiamin, Q., Craig, P., Torgerson, P., 2005b. Modeling the transmission of *Echinococcus granulosus* and *Echinococcus multilocularis* in dogs for a high endemic region of the Tibetan plateau. Inter. J. Parasitol. 35 (2), 163-170.

Calenge, C., 2006. The package "adehabitat" for the r software: A tool for the analysis of space and habitat use by animals. Ecol. Model. 197 (3-4), 516-519.

Ciucci, P., Boitani, L., Francisci, F., Andreoli, G., 1997. Home range, activity and movements of a wolf pack

in central Italy. The Zoological Society of London 243, 803-819.

Clark, P., Evans, F., 1954. Distance to nearest neighbour as a measure of spatial relationships inpopulations ecology. Ecology 35, 445-453.

Craig, P., Giraudoux, P., Shi, D., Bartholomot, B., Barnish, G., Delattre, P., Quere, J., Harraga, S., Bao, G., Wang, Y., Lu, F., Ito, A., Vuitton, D., 2000. An epidemiological and ecological study of human alveolar echinococcosis transmission in South Gansu, China. Acta Trop. 77, 167-177.

Dinkel, A., Von Nickisch-Rosenegk, M., Bilger, B., Merli, M., Lucius, R., Romig, T., 1998. Detection of *Echinococcus multilocularis* in the definitive host: Coprodiagnosis by PCR as an alternative to necropsy. J. Clin. Microbiol. 36 (7), 1871-1876.

Eckert, J., Deplazes, P., 2004. Biological, Epidemiological, and Clinical Aspects of Echinococcosis, a Zoonosis of Increasing Concern. Clin. Microbiol. Rev. 17 (1), 107-135.

Giraudoux, P., Delattre, P., Takahashi, K., Raoul, F., Quere, J., Craig, P., Vuitton, D., 2002. Transmission ecology of *Echinococcus multilocularis* in wildlife: what can be learned from comparative studies and multiscale approaches? In: Cestode zoonoses: Echinococcosis and Cysticercosis. Craig, P.S. And Pawlowski, Z., IOS Press.

Giraudoux, P., Pleydell, D., Raoul, F., Quere, J., Wang, Q., Yang, Y., Vuitton, D., Qiu, J., Yang, W., Craig, P., 2006. Transmission ecology of *Echinococcus multilocularis*: What are the ranges of parasite stability among various host communities in China? Parasitol. Inter. 55, 237-246.

Gong, M. H., Hu, J., 2003. The summer microhabitat selection of Tibetan fox in the northwest plateau of Sichuan. Acta Theriol. Sin. 23, 267-269.

Habluetzel, A., Traldi, G., Ruggieri, S., Attili, A., Scuppa, P., Marchetti, R., Menghini, G., Esposito, F., 2003. An estimation of *Toxocara canis* prevalence in dogs, environmental egg contamination and risk of human infection in the Marche region of Italy. Vet. Parasitol. 113, 243-252.

Hegglin, D., Bontadina, F., Contesse, P., Gloor, S., Deplazes, P., 2007. Plasticity of predation behaviour as a putative driving force for parasite life-cycle dynamics: the case of urban foxes and *Echinococcus multilocularis* tapeworm. Funct. Ecol. 21(9), 552-560.

Hodder, K., Kenward, R., Walls, S., Clarke, R., 1998. Estimating core ranges: a comparison of techniques using the Common buzzard (*Buteo buteo*). J. Raptor Res. 32 (2), 82-89.

Kareiva, P., Odell, G., 1987. Swarms of predators exhibit "preytaxis" if individual predators use area-restricted search. Am. Nat. 130 (2), 233.

Macpherson, C., 2005. Human behaviour and the epidemiology of parasitic zoonoses. Inter. J. Parasitol. 35, 1319-1331.

Macpherson, C., Meslin, F., Wandeler, A., 2000. Dogs, Zoonoses and Public Health. CABI Publishing, UK.

Marston, C., 2008. Spatial modelling of small mammal distributions in relation to parasite transmission in western china. Ph.D. thesis, School of Environment and Life Sciences University of Salford, Salford, UK.

Meek, P., 1999. The movement, roaming behaviour and home range of free-roaming domestic dogs, *Canis lupus familiaris*, in coastal New South Wales. Wildlife Res. 26, 847-855.

Okubo, A., Levin, S., 2001. Diffusion and ecological problems: modern perspectives. Springer verlag GMBH.

O'Lorcain, P., 1994. Prevalence of *Toxocara canis* ova in public playgrounds in the Dublin area of Ireland. J. Helminthol. 68, 237-241.

Qiu, J., Chen, X., Ren, M., Luo, C., Liu, D., Liu, X., 1995. Epidemiological study on alveolar hydatid disease in qinghai-xizang (tibetan) plateau. J Pract Parasit Dis 3, 106-109.

Qiu, J., Liu, F., Schantz, P., Ito, A., Carol, D., He, J., 1999. Epidemiological survey of hydatidosis in Tibetan areas of Western Sichuan Province. Archivos Internacionales de la Hidatidosis, 23-84.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

URL http://www.R-project.org

Raoul, F., Deplazes, P., Nonakac, N., Piarroux, R., Vuitton, D., Giraudoux, P., 2001. Assessment of the epidemiological status of *Echinococcus multilocularis* in foxes in France using elisa coprotests on fox faeces collected in the field. Inter. J. Parasitol. 31 (3-4), 1579-1588.

Raoul, F., Quere, J., Rieffel, D., Bernard, N., Takahashi, K., Scheifler, R., Ito, A., Wang, Q., Qiu, J., Yang, W., Craig, P., Giraudoux, P., 2006. Distribution of small mammals in a pastoral landscape of the Tibetan plateau (Western Sichuan, China) and relationship with grazing practices. Mammalia 70 (3-4), 214-225.

Robardet, E., Giraudoux, P., Caillot, C., Boue, F., Cliquet, F., Augot, D., Barrat, J., 2008. Infection of foxes by *Echinococcus multilocularis* in urban and suburban areas of Nancy, France: infuence of feeding habits and environment. Parasite 15, 77-85.

Seaman, D., Powell, R., 1990. Identifying patterns and intensity of home range use. Bears: Their Biology and Management. A Selection of Papers from the Eighth International Conference on Bear Research and Management, Victoria, British Columbia, Canada, February 1989 8, 243-249.

Seaman, D. E., Powell, R. A., 1996. An evaluation of the accuracy of kernel density estimators for home range analysis. Ecology 77 (7), 2075-2085.

Thompson, R. A., Kutz, S. J., Smith, A., 2009. Parasite zoonoses and wildlife: Emerging issues. Int. J. Environ. Res. Public Health 6 (2), 678-693.

Li, T., Jiamin, Q., Wen, Y., Craig, P., Xingwang, C., X., N., Ito, A., Giraudoux, P., Wulamu, M., Wen, Y., Schantz, P., 2005. Echinococcosis in Tibetan populations, Western Sichuan province, China. Emerg. Infect. Dis. 11 (12), 1866-1873.

Vuitton, D., Zhou, H., Bresson-hadni, S., Wang, Q., Piarroux, M., Raoul, F., Giraudoux, P., 2003. Epidemiology of alveolar echinococcosis with particular reference to China and Europe. Parasitology 127, 87-107.

Wang, Q., Raoul, F., Budke, C., Craig, P., Xiao, Y., Vuitton, D., Qiu, D., Pleydell, D., Giraudoux, P., in press. Grass height and transmission ecology of *Echinococcus multilocularis* in tibetan communities, china. Chin. J. Med..

Wang, Q., Vuitton, D., Qiu, J., Giraudoux, P., Xiao, Y., Schantz, P., Raoul, F., Li, T., Wen, Y., Craig, P., 2004. Fenced pasture: a possible risk factor for human alveolar echinococcosis in Tibetan pastoralist communities of Sichuan, China. Acta Trop. 90, 285-293.

Wang, Z., Wang, X., X., L., 2007. Selection of land cover by the Tibetan fox *Vulpes ferrilata* on the eastern Tibetan Plateau, western Sichuan Province, China. Acta Theriol. 52, 215-223.

WHO, WSPA, 1990. Guidelines for Dog Population Management. World Health Organisation, Geneva.

Worton, B. J., 1989. Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70 (1), 164-168.

Tables and Figures

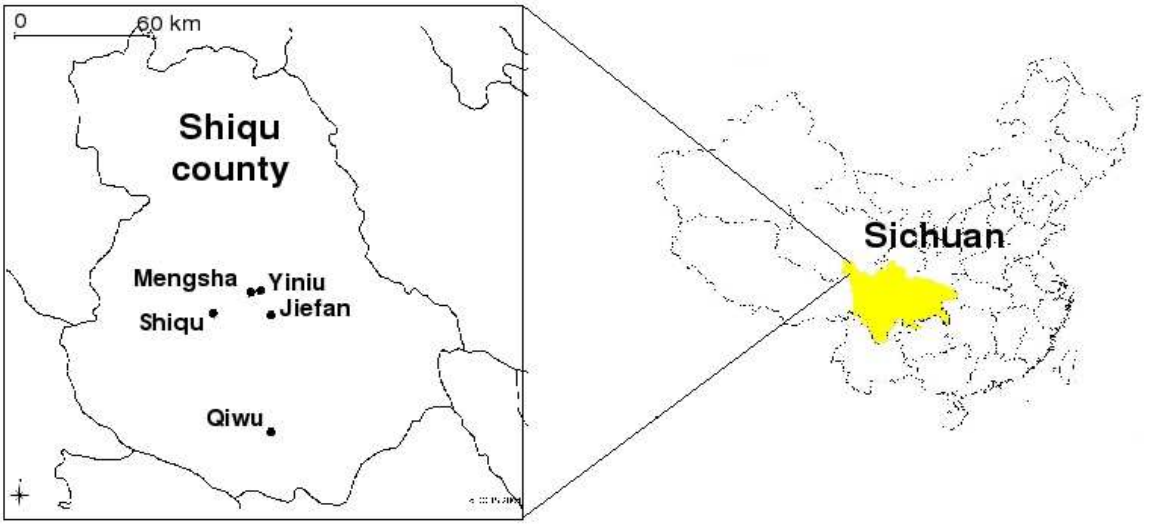

Figure 1: Localization of study sites in Serxu county, China.

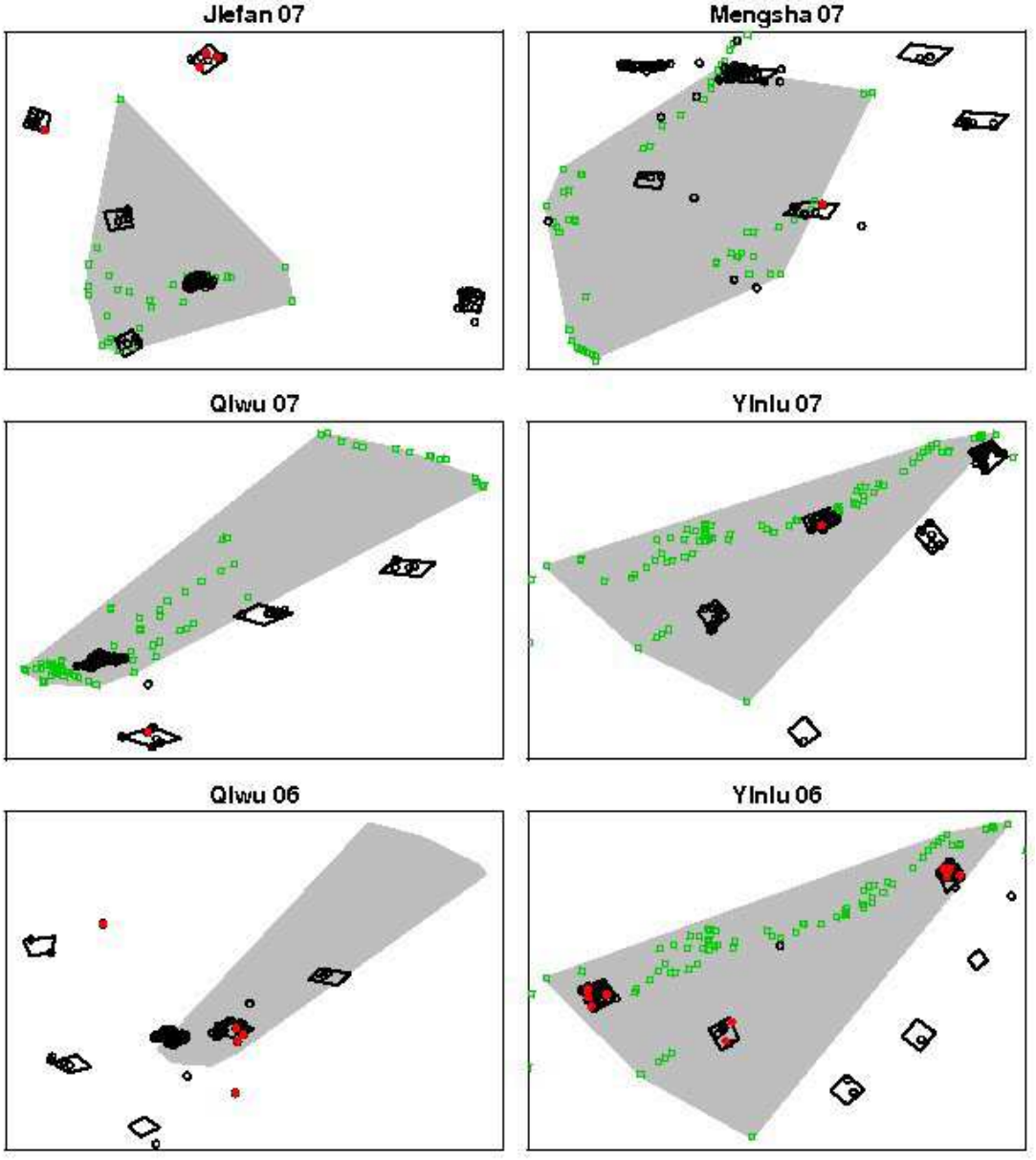

Figure 2: Faeces sampling protocol per sampling survey. Village areas (90 % MCP polygon) and village houses were represented by gray areas and small green squares respectively. Black squares, empty and full black circles represented quadrats, dog and fox faeces respectively.

5.3 Principaux résultats

Nos analyses spatiales et moléculaires de fecès apportent la preuve et permet de quantifier et de spatialiser le rôle majeur du chien dans la contamination de l'environnement des villages du plateau tibétains pour lesquels nous disposions de données éco-épidémiologiques.

En effet, d'une part les fèces de chiens constituent la principale source de contamination des villages étudiés en comparaison à celles de renards tibétains qui peuvent fréquenter et déféquer à l'intérieur des villages de manière sporadique.

D'autre part, les fèces de canidés et leur contamination potentielle sont distribuées de manière non aléatoire dans l'espace utilisé par les chiens : elles sont concentrées dans un périmètre d'environ 200 m autour des habitations et peuvent être agrégées dans des micro-foyers à l'intérieur des villages.

L'utilisation de l'espace par les chiens corroborent les comportements de défécation observés et apportent de nouvelles informations : la nuit, les chiens passent la majorité de leur temps dans un rayon d'environ 200 m autour des habitations humaines, mais certains peuvent également réaliser une à deux trajectoires excursives dans les prairies environnant les villages et à une distance maximale de 1500 m des habitations. De plus, les mâles sont plus sujets à s'éloigner des habitations que les femelles.

Une analyse spatiale préliminaire des interactions entre les chiens et les populations de micro-mammifères a permis une première quantification du taux de contact entre les hôtes. A l'inverse du comportement du chien, les densités de population de micro-mammifères augmentent depuis les aires coeur des chiens (0-50 m des maisons) jusqu'aux prairies pâturées, habitat optimal pour ces espèces. Ainsi, les zones d'interactions maximales entre les hôtes se situeraient donc dans les aires d'activités moyenne de la population de chiens domestiques.

Discussion générale

Chapitre 6

Modélisation des distributions spatiales des assemblages d'hôtes intermédiaires : apports, limites et amélioration du protocole

Dans la perspective de modéliser le risque de transmission d'*Em*, nous avons exploré le pouvoir explicatif et prédictif de modèles de distribution des assemblages de micro-mammifères développés sur des données de piégeage obtenues dans deux sites d'étude du Sichuan. La modélisation s'est déroulée en 3 étapes : définir des assemblages (Axe 1, étape 1), modéliser leurs habitats (Axe 1, étape 2) et prédire leurs distributions sur une étendue régionale (Axe 2) (Figure 6.1, p. 147). Chaque étape de la modélisation a apporté des éléments de connaissances écologiques ou éco-épidémiologiques concernant les distributions spatiales des populations. En parallèle, l'estimation des erreurs de prédiction a permis de mettre en évidence les limites pour atteindre notre objectif de notre jeu de données et des outils de modélisation statistique. Après un rappel des apports de notre travail aux connaissances méthodologiques et écologiques/épidémiologiques, nous discuterons des améliorations de notre protocole requises pour réduire "ce que les modèles n'expliquent pas" en accord avec "l'état d'esprit" de la construction de modèles prédictifs. Enfin, les applications de nos modèles à la modélisation du risque éco-épidémiologique seront discutées.

6.1 De la définition des assemblages

6.1.1 Apports méthodologiques

La définition d'assemblages de micro-mammifères s'est avérée utile pour analyser et interpréter des données de présence/absence d'espèces de micro-mammifères, caractérisées par une large diversité d'espèces dans une large diversité d'habitats échantillonnés et définis sur avis d'experts.

La réduction de la redondance (Figure 6.1, p. 147) par la procédure de regroupement des habitats était nécessaire pour expliquer au mieux les variabilités observées des densités relatives

des espèces entre les lignes de pièges. Une large diversité et redondance des classes d'habitats peuvent être problématiques lorsque l'échantillonnage est basé sur des cartes classifiées d'occupation du sol ou de types de végétation.

Bien que les habitats aient été définis sur avis d'experts, leurs regroupements ont été réalisés sur des critères objectifs et quantitatifs (optimisation de la vraisemblance des modèles, modélisation multinomiale). Un tel regroupement semble plus efficace que ceux plus couramment utilisés et réalisés sur avis d'experts (Ferrier and Guisan, 2006; Oliver, 2002) et ayant été montré inefficaces dans certains cas d'étude (Pearce et al., 2001). D'autre part, notre approche permet de partitionner la diversité sur la base des densités relatives et de la composition spécifique des lignes de pièges. Elle peut donc être utilisée pour estimer les distributions spatiales des propriétés macro-écologiques des communautés, résumées par des indices de diversité traditionnellement utilisées dans le domaine de l'écologie des communautés (Ferrier and Guisan, 2006; Ferrier et al., 2007). En parallèle, elle constitue en quelque sorte une classification et modélisation simultanées des densités relatives des espèces en fonction de variables qualitatives (les classes d'habitats). Elle permet donc d'informer également sur les compositions et densités relatives de chacune des espèces composant les assemblages ou les groupes d'habitats, ce qui peut être utilisable en biologie de la conservation par exemple.

On remarquera enfin que la définition des assemblages a été réalisée sur une période de temps limitée et ne prend pas en considération les effets des dynamiques des populations sur les distributions des espèces, ce qui pourrait entraîner des variations temporelles dans la définition des assemblages. Cependant, la considération de l'ensemble du "turn over" spatial de la zone d'étude peut permettre de compenser cette omission en intégrant, dans la définition des assemblages, différentes phases de la dynamique des populations.

6.1.2 Les assemblages comme groupes de transmission d'*Em*

L'identification, sans protocole spécifique d'échantillonnage, des espèces qui interviennent majoritairement dans la transmission du parasite est compliquée par la grande diversité des espèces potentiellement hôtes échantillonnées. La double information contenue dans les assemblages définis, diversité et densités relatives, peut s'avérer utile pour définir les groupes de transmission dans nos sites d'études, en référence au statuts infectieux des espèces identifiés par des études antérieures.

Comme nous l'avons décrit précisément en introduction, la définition d'assemblages d'espèces de micro-mammifères a été utilisée pour comprendre les distributions spatiales des groupes de transmission dans différents sites d'étude. Les paysages à risque de transmission sont alors caractérisés par la présence, en proportion importante, d'habitats optimaux pour une ou plusieurs espèces potentiellement hôtes et capables de montrer une dynamique cyclique de populations, qui peuvent alors atteindre des pics de fortes densités. Sur le plateau tibétain, il s'agit des prairies encloses et des zones buissonnantes dominées par *Microtus limnophilus* (susceptible à *Em*) et des prairies pâturées où abondent des espèces potentiellement hôtes du genre *Ochotona* (*O.curzoniae*, *O. cansus*) et *Cricetulus kamensis* (Raoul et al., 2006; Giraudoux et al., 2003). *Microtus irene*, également susceptible, avait comme habitat des zones rocheuses et des prairies rases à Stellaires. Dans le Gansu, le stade intermédiaire de la déforestation dominé par *Cricetulus longicaudatus*

et *Microtus limnophilus*, espèces potentiellement hôtes, est considéré comme l'habitat à plus haut risque de transmission (Giraudoux et al., 1998, 2003).

Dans nos deux sites d'études, *Microtus irene* a été piégé à de faibles densités (1 individu par site), dans les forêts mixtes de conifères et de rhododendrons. Ces groupes d'habitats forestiers étaient les plus diversifiés en espèces. Le genre *Ochotona* dominait ces mêmes assemblages en terme de densités relatives. Par opposition à la situation rencontrée sur le plateau tibétain, la diversité des habitats et, de fait, la richesse spécifique des assemblages ont tendance à être élevées dans les vallées des contreforts du plateau, prévenant alors les surabondances de quelques espèces (Giraudoux et al., 2003; Michel et al., 2007). Le cycle d'*Em* pourrait donc exister à très bas bruit à Maerkang et à Rangtang dans les habitats forestiers et serait alors qualifié de sylvatique. La grande hétérogénéité du paysage, l'absence de fortes densités d'hôtes intermédiaires dans des habitats assez éloignés des habitations humaines et des chiens domestiques, laissent supposer qu'un tel cycle fonctionne avec une moindre intensité que celui en place sur le plateau, comme cela a été observé dans d'autres provinces chinoises telles que le Ningxia ou le Gansu (Giraudoux et al., 2006).

Par opposition, les habitats situés près des habitations et sous l'influence agricole étaient moins diversifiés en espèces que les assemblages forestiers. Aussi, à Rangtang, les densités relatives étaient plus élevées pour les assemblages R1 et R3, correspondant aux habitats agricoles, dominés par *Apodemus peninsulae* et *Microtus limnophilus* respectivement, que dans les assemblages forestiers. *Microtus limnophilus*, piégé principalement dans les prairies encloses, était également présent à de moindres densités dans chacun des autres assemblages. Les fortes densités de cette espèce associées à sa présence dans tous les assemblages soulignent son caractère généraliste (large "niche") et indiquent la possibilité de dispersions dans une large diversité d'habitats non forestiers proches des habitations humaines, et par suite, un risque de fluctuations cycliques des densités de ces populations (Giraudoux et al., 2006). Cela suggère la possible existence dans la région de Rangtang de fortes densités ponctuelles d'hôtes intermédiaires pour le parasite pouvant alimenter un cycle péri-domestique ou sauvage.

6.2 De la modélisation des assemblages

6.2.1 Rôle des variables environnementales indirectes et distributions des assemblages

Nous avons trouvé que les variables environnementales issues d'images satellites pouvaient être utilisées pour discriminer des assemblages avec une précision relativement importante et supérieure à celle obtenue par le hasard. Les techniques non paramétriques de discrimination des assemblages se sont avérées les plus performantes pour prédire leurs distributions. Notre étude apporte donc des éléments supplémentaires en faveur de l'utilisation de telles méthodes, dans le cas de données éparses et concernant des assemblages d'espèces (Elith et al., 2006).

Parmi les variables topographiques, l'altitude, qui influence directement les distributions de la végétation et des habitats (étagement de la végétation), était la variable la plus influente. Ce résultat souligne l'intérêt de l'utilisation des gradients topographiques pour modéliser les distributions d'assemblages de micro-mammifères. Cet effet avait déjà été mis en évidence sur les

distributions des espèces individuellement. Ainsi, Gibson et al. (2004) ont montré que l'altitude et l'ensoleillement étaient négativement corrélés à la richesse spécifique de communautés de micro-mammifères du sud ouest de l'état de Victoria en Australie. Ces variables permettaient d'expliquer mieux que la complexité structurale de la végétation, les densités observées. De même, alors que les variables paysagères relatives à la fragmentation du paysage n'étaient pas informatives, la pente s'est révélée la seule variable explicative des densités de rongeur dans la péninsule du Panama où l'altitude était homogène entre les sites d'échantillonnage (Suzan et al., 2006).

Les indices de végétation peuvent être considérés comme des descripteurs plus directs des habitats et ont joué un rôle important pour expliquer les distributions des assemblages. Le NDVI était important à Rangtang pour discriminer les prairies de vallées mais aussi les forêts de conifères et de rhododendron. Á Maerkang, c'est davantage l'EVI qui aidait à la discrimination des prairies encloses et des forêts. Á notre connaissance, excepté dans le contexte de la transmission d'*Em* (Marston, 2008), peu d'études ont utilisé de telles variables pour modéliser les distributions de micro-mammifères. L'indice NDVI n'avait pas permis d'expliquer les distributions du campagnol roussâtre en Belgique (Linard et al., 2007). Nos résultats mettent en évidence l'utilité de telles variables indirectement corrélées à la ressource des espèces lorsque peu d'informations sur les habitats mesurées sur le terrain (proximales) sont disponibles (Pearce et al., 2001).

Cependant, les variables environnementales utilisées sont loin d'être exhaustives, comme l'indiquent les erreurs de prédictions. Actuellement, le réalisme de nos modèles est limité par l'utilisation de variables environnementales satellitales décrivant seulement indirectement l'écologie des populations. De plus, les patrons non linéaires et non Gaussiens observés correspondent à des mélanges dans l'espace des variables environnementales résultant de la superposition des distributions de chaque habitat. Ils ne nous informent donc pas sur la "forme" de la réponse des espèces le long des gradients environnementaux. L'amélioration des connaissances des processus écologiques à l'origine des distributions, puis leurs incorporations dans nos modèles, pourraient permettre de réduire les erreurs de prédiction liées aux choix de variables environnementales indirectes utilisées (Austin, 2002). Plusieurs paramètres influençant les distributions des populations de micro-mammifères pourraient être considérés et sont discutées ci-après.

6.2.2 Améliorations du protocole

6.2.2.1 Description des micro-habitats

Les variables satellitales ont permis de décrire l'environnement des lignes de pièges, c'est-à-dire les habitats pour les assemblages. De telles variables environnementales ont servi à décrire de manière indirecte ("surrogate variable") la structure et la composition de la végétation, ressources pour les espèces. Les habitats échantillonnés ont été définis initialement sur avis d'experts en fonction d'estimations sommaires des composition et structure des communautés végétales. Les classes regroupées apportaient plus d'informations que les variables quantitatives (altitude, pente, NDVI, EVI, indice d'ensoleillement) prises seules, pour expliquer la variabilité des densités relatives (Axe 1b). Cependant, cette information n'est pas quantitative et aucune relation entre les variables satellitales et les caractéristiques de structure et de composition des

habitats n'a été établie.

Or, les caractéristiques des micro-habitats et de la ressource pour les espèces de micro-mammifères (structure et composition de la végétation) ont été mises en évidence comme facteurs clefs pouvant expliquer les distributions des espèces. Ayant testé l'effet de plusieurs variables paysagères et topographiques, Catling et Coops (1999) soulignent l'importance primordiale de la strate arbustive pour modéliser les distributions de micro-mammifères. Dans les paysages agricoles français, les caractéristiques locales de l'habitat, telles que la largeur des haies par exemple, constituent le principal facteur de variation des densités (Michel et al., 2007). Cependant de telles données ne sont pas disponibles au delà des sites d'échantillonnage et ne peuvent donc pas être utilisées dans des modèles prédictifs.

L'établissement de corrélations entre les variables satellitales et de terrain pourrait aider à rendre compte des attributs des habitats considérés par les variables de substitution. Cela pourrait aider à sélectionner les variables les plus corrélées à la réalité de la structure et de la composition des habitats pour améliorer les capacités prédictives des modèles.

6.2.2.2 La dimension paysagère

Le paysage revêt de multiples acceptations de par son utilisation dans divers domaines, des sciences sociales à la géologie. De manière générale, il correspond à la perception de l'environnement pour les individus ou populations y habitant et qui en définissent alors sa fonction et sa dimension (échelle). Ainsi il peut être décrit comme "une portion d'espace perceptible par des individus" (Foltete, 2006). Dans notre contexte, le paysage est celui des populations de micro-mammifères et sera le lieu et le générateur de processus écologiques. Il peut alors être défini comme "un niveau d'organisation des systèmes écologiques, supérieur à l'écosystème" (Burel and Baudry, 1999).

Au delà des attributs environnementaux des micro-habitats, l'environnement des points d'échantillonnage peut être considéré à deux échelles pour expliquer les dynamiques et distributions des espèces de micro-mammifères (Bowman et al., 2000; Michel et al., 2007) : i) l 'échelle sectorielle, dans le voisinage proche des lignes de pièges ou "l'ensemble des systèmes biologiques reliés entre eux par des processus de dispersion" (Blondel, 1995) et, ii) paysagère, où les habitats constituent une mosaïque par leurs compositions et leurs structures.

La prise en compte des caractéristiques environnementales au voisinage du micro-habitat échantillonné s'est avérée utile dans la compréhension des distributions du genre *Ochotona* dans un rayon de 1500 m autour de transects (Marston, 2008). En France, Morilhat et al. (2008) ont montré que les variables paysagères liées aux surfaces et aux proportions des classes paysagères avaient des effets sur les densités du campagnol des champs dans un rayon variant de 400 à 800 m de la parcelle considérée.

Sur des étendues plus larges (régionale, n x 10 km), la composition du paysage influence la dynamique à plus large échelle (Giraudoux et al., 1997; Fichet-Calvet et al., 2000) et peut avoir des effets simultanés sur les distributions de plusieurs espèces de micro-mammifères. Dans l'est de la France, par exemple, les densités de rongeurs étaient stables dans les paysages à faible proportion de prairies permanentes et instables dans les paysages dominés par ce même type de prairies (Raoul et al., 2001a). Des méthodes d'ordination ont permis l'analyse des effets

de la composition du paysage sur l'ensemble des espèces composant les assemblages. Ainsi, un effet négatif de l'intensification de l'agriculture sur les fréquences d'occurrences des espèces rares et spécialistes et positif sur les espèces plus généralistes ont été montré dans l'ouest de la France (de la Péña et al., 2003). De même, l'hétérogénéité du paysage favorise la diversité des assemblages d'espèces de rongeurs dans les paysages d'agriculture intensive de l'ouest de la France (Butet et al., 2006).

Depuis la naissance de l'épidémiologie paysagère (Pavlowski, 1964), le paysage est considéré comme générateur du risque épidémiologique et peut être utilisé pour le prédire (Kitron, 1998; Ostfeld et al., 2005; Giraudoux et al., 2008). Dans notre contexte eco-épidémiologique, on peut s'attendre à ce que des perturbations paysagères, associées à des modifications des pratiques agricoles, influencent les densités des populations d'hôtes intermédiaires et donc du réservoir pour le parasite. En effet, en modifiant la composition du paysage, de telles perturbations peuvent augmenter la proportion d'habitats favorables, influencer les dynamiques des populations et augmenter ponctuellement les densités d'hôtes intermédiaires (Cf Introduction pour une synthèse).

Une augmentation des densités des populations d'*Ochotona cansus*, piégées dans des habitats des stades de reforestation dans nos sites d'étude, pourrait être engendrée par exemple par un changement de pratiques sylvicoles. Pour *Microtus limnophilus*, l'augmentation de surface de prairies clôturées telle qu'observée sur le plateau (Wang et al., 2004) pourrait engendrer une modification de la dynamique de population. La structure du paysage peut également influencer la transmission en influençant les dynamiques de population d'hôtes. Ceci a été vérifié dans le cas de la transmission d'un hantavirus par *Penomysens moniculatus* (Langlois et al., 2001) ou encore dans le cas de la transmission du virus de Puumala transmis par le campagnol roussâtre (Linard et al., 2007). Dans ces cas d'étude, il est supposé que les dispersions des populations d'hôtes et du pathogène sont accélérées par la fragmentation de l'habitat optimal ou par la proximité de ses tâches dans le paysage, qui maintiennent alors sa connectivité.

Dans notre travail, nous avons utilisé les valeurs numériques des bandes spectrales, ou de leurs combinaisons, pour décrire l'environnement des lignes de pièges. Nous n'avons pas utilisé de données d'images classifiées et paysagères puisqu'aucune classification commune aux deux sites n'était disponible. Ainsi, dans la mesure où il ne correspond pas à une catégorie d'occupation du sol ou à une unité fonctionnelle définie en amont par les écologues du paysage, l'environnement que nous avons défini n'a pas de sens paysager. Dans chaque site d'étude, l'intégration de variables paysagères pourrait aider à expliquer les distributions observées et à prédire le risque paysager. Sur une étendue plus large, la discrimination des assemblages entre les deux sites d'étude, par exemple ceux des paysages agricoles (R1 et M1), pourrait être améliorée en considérant la composition des paysages dans nos modèles. Des images classifiées ont été réalisées indépendamment à Maerkang et à Rangtang. Après avoir été validées sur le terrain, elles pourraient être utilisées pour tester l'incorporation des effets du paysage sur les prédictions des modèles localement, leur utilisation dans des modèles entraînés régionalement demeurant limitée.

D'autre part, la proximité des lignes de pièges a compliqué l'estimation des variables environnementales au voisinage de nos points d'échantillonnage. En effet, l'addition d'une telle

variabilité qui, de plus, est partagée par plusieurs lignes de pièges, a été considérée *a priori* comme source de confusion dans la discrimination des assemblages. Malgré tout, cela limite le réalisme écologique du modèle et des efforts doivent être effectués pour tester l'effet de l'intégration des effets de voisinage sur les prédictions.

6.2.3 La stratégie de modélisation : incorporer l'avis d'experts

Contrairement aux méthodes de classification classiquement utilisées à ce stade de la modélisation (algorithme TWINSPAM, clusterisation), la classification que nous avons réalisée incorpore l'information paysagère des classes d'habitats définies sur avis d'experts. Le fait que les classes regroupées expliquent davantage les densités relatives des espèces que les variables satellitales (Axe 1b) souligne l'importance de la définition de l'habitat sur avis d'experts (composition et structure de la végétation) pour expliquer la variabilité des densités, à condition que la réduction de la redondance soit effectuée (cf plus haut).

De manière générale, l'incorporation dans les modèles de distribution des habitats des avis d'experts s'avère indispensable lors d'études exploratoires et peut être utile lorsque les données sur les habitats sont limitées à des variables environnementales indirectes. Des outils d'incitation, tel que le logiciel "Elicidator" (James et al., 2010) ont été développés pour collecter les connaissances des experts sur les habitats et les distributions spatiales des densités en utilisant un support SIG. Ces informations considérées comme des priors dans une analyse bayésienne se sont avérées utiles pour modifier ou renforcer l'analyse des patrons des distributions spatiales de Pétrogale à queue touffue *Petrogale penicillata* ou d'oiseaux en Australie (Murray et al., 2009; Martin et al., 2005). De même, les distributions spatiales décrites par les experts peuvent être mises en relation avec des variables environnementales (topographiques, ensoleillement, distance à la forêt) pour établir des cartes d'habitats sur la zone étudiée (Yamada et al., 2003). Cependant, la variabilité des connaissances et le nombre d'experts interrogés semblent limiter la validité de telles cartes, qui peuvent alors être considérées davantage comme une base d'informations pour des études de terrain et des modélisations confirmatoires. D'autre part, le regroupement de classes de couverture du sol ou l'établissement d'indices de pertinence ("suitability") se sont avérés peu informatifs en comparaison des variables environnementales quantitatives, abiotiques (conditions climatiques, pédologiques, ...) ou biotiques (cartes de végétation), pour expliquer les distributions de 93 espèces de marsupiaux, de reptiles, d'oiseaux et de chiroptères en Australie (Pearce et al., 2001). Enfin, l'avis d'experts pour la sélection des variables explicatives dans les modèles de distribution d'oiseaux en Espagne s'est révélé peu informatif par rapport à une approche non supervisée de la sélection ainsi que pour l'extrapolation des prédictions sur des aires indépendantes (Seoane et al., 2005).

Dans notre étude, nous n'avons pas comparé les capacités prédictives de notre stratégie de modélisation des assemblages, incluant la variable habitat définie sur avis d'experts ("Classifier puis modéliser"), avec celle qui consiste à "Classifier et modéliser simultanément" et qui, s'abstenant de l'avis d'experts, reposerait sur un échantillonnage stratifié le long des gradients environnementaux. Or, étant données les nombreuses erreurs de prédiction des assemblages dans l'espace environnemental, dues à leurs natures mixtes et discrètes, il serait intéressant d'évaluer si la stratégie "Classifier et modéliser simultanément", malgré l'absence d'informations paysagères,

permet une modélisation des habitats plus précise et extrapolable.

6.3 De l'évaluation des prédictions

De manière générale, ce travail a permis d'explorer les limites des méthodes de modélisation des distributions des assemblages d'espèces de micro-mammifères. Les données satellitales disponibles sur de larges étendues ont permis de modéliser et de prédire les distributions des assemblages sur des étendues locales (site d'étude) mais aussi régionales. Nos travaux ont permis de discuter des possibilités prédictives des deux échelles d'entraînement des modèles sur les données ré-échantillonnées ou sur un jeu de données indépendant. Ainsi, plusieurs sources d'erreurs spatiales ou globales ont été identifiées et sont discutées ci-après.

6.3.1 La stratégie d'échantillonnage

La nature mixte des assemblages (mélange d'habitats) a compliqué leur modélisation ainsi que l'interprétation des effets des variables environnementales. En effet, l'effort de piégeage différait entre les habitats définis *a priori*, et les réponses des assemblages dans l'espace multivarié étaient alors définies essentiellement par les habitats les plus échantillonnés. Nous avons montré que les prévalences des assemblages étaient corrélées à la largeur de leurs niches. La modélisation des habitats des assemblages d'espèces requiert également de considérer une taille d'échantillon suffisante pour prendre en considération l'ensemble du "turn over" spatial dans la composition des assemblages (Ferrier et al., 2007).

L'échantillonnage devrait donc être réalisé de manière stratifiée, c'est-à-dire dans chaque combinaison de variables environnementales (classe d'habitat) (Guisan and Zimmerman, 2000; Hirzel and Guisan, 2002), qu'elle soit définie sur avis d'experts et/ou par une classification des variables environnementales réalisée au préalable. L'échantillonnage dans chaque strate devra être alors répété jusqu'à capturer l'essentiel de la variabilité pour chaque classe dans l'espace environnemental en s'assurant de sa possible discrimination. Idéalement, une classification de l'ensemble de la variabilité environnementale du site d'étude le long de gradients environnementaux connus pour influencer les distributions devrait être réalisée au préalable de la campagne de piégeage. Dans le cas d'une étude exploratoire, une classification non supervisée de l'image pourrait aider à la représentation d'une telle variabilité et être combinée à l'avis d'expert pour s'assurer de l'identification de chaque habitat échantillonné dans l'espace des variables environnementales utilisées ensuite pour modéliser les distributions. La taille de l'échantillon dépendra alors davantage de sa capacité à capturer cette variabilité plutôt que de l'aire de la strate utilisée dans certaines stratégies (Hirzel and Guisan, 2002). L'estimation de la variabilité environnementale échantillonnée pourra alors facilement être estimée au fil des sessions de piégeage par une simple cartographie de l'espace environnemental et des lignes de pièges. Un tel échantillonnage stratifié et régulier des habitats permettrait d'assurer la discrimination des assemblages, même constitués d'un seul habitat.

6.3.2 Les erreurs spatiales

6.3.2.1 La structure spatiale des données

Marston (2008) et Raoul et al. (2008) ont montré que l'auto-corrélation entre les intervalles des transects ou entre les lignes de pièges n'affectait pas les prédictions des modèles de distributions d'habitats, ni la définition des assemblages. De plus, les variables environnementales peuvent expliquer en partie la variabilité spatiale des résidus des modèles et par conséquent l'omission de telles variables dans les modèles peuvent également être source d'auto-corrélation entre les observations (Barry and Elith, 2006). Enfin, l'ajout d'une composante spatiale de manière explicite dans les algorithmes augmente la complexité des modèles, qui nécessitent alors une taille d'échantillon assez importante pour être validés. L'utilisation combinée de relations spatialement explicites et non linéaires s'avère donc délicate dans le cas d'une faible taille d'échantillon (Barry and Elith, 2006).

Au regard de notre faible taille d'échantillon, de la nature indirecte et non exhaustive des variables environnementales disponibles et de la complexité déjà existante des modèles multivariés utilisés (MDA, MARS), nous n'avons pas pris en compte l'agrégation des lignes de pièges par l'ajout d'une composante spatiale dans les algorithmes. Cependant, pour la construction d'un futur modèle et de cartes de prédiction robustes et dans le souci de respect des conditions des inférences statistiques, l'influence particulière de l'auto-corrélation des lignes de pièges sur les capacités prédictives des modèles de distribution des assemblages devrait être testée.

6.3.2.2 Le contexte environnemental et biogéographique

L'extrapolation des prédictions a été compliquée par les distributions géographiques et dans l'espace environnemental des deux sites d'étude.

Les signatures spectrales des 2 secteurs ne se recouvraient pas complètement dans l'espace des variables environnementales, la "niche" de Maerkang étant plus large que celle de Rangtang (cf Axe 2, 2). Ces différences de contexte environnemental peuvent expliquer pourquoi les modèles de Rangtang n'ont pas pu prédire dans les régions de l'espace environnemental sur lesquelles ils n'ont pas pu être entraînés. De telles erreurs peuvent être évitées en extrapolant les prédictions des modèles locaux uniquement dans des aires environnementalement similaires à l'aire d'entraînement du modèle, c'est-à-dire dans les régions du spectre où ils ont été entraînés (Murphy and Lovett-Doust, 2007). Pour ce faire, les représentations des variabilités des aires de prédictions et des sites d'échantillonnage peuvent être réalisées et comparées dans l'espace environnemental que définissent les principaux gradients environnementaux.

Les deux secteurs différaient également de par leurs contextes paysagers, où des processus écologiques particuliers, dits associatifs-interactifs tels que dispersion, extinction et recolonisation locale, peuvent se réaliser indépendamment dans chaque secteur pour expliquer les distributions observées (Blondel, 1995). Par exemple, la structure des taches de prairies clôturées, habitat optimal pour *Microtus limnophilus* peut influencer les dynamiques des populations de cette espèce. La considération du contexte paysager dans les modèles serait donc nécessaire pour l'extrapolation des prédictions dans des contextes différents.

Peuvent également intervenir des processus macro-évolutifs expliquant à l'échelle du continent la différenciation des espèces et la définition de zones dites biogéographiques (Blondel, 1995). Les diversités observées dans chaque secteur (locales) peuvent dépendre des distributions géographiques des espèces à de plus larges échelles (régionales) (Ricklefs, 2004). Ainsi, les domaines biogéographiques peuvent expliquer, en complément des facteurs écologiques, les structures locales des communautés de mammifères (Rodriguez et al., 2006) ou la variabilité des richesses des assemblages entre sites d'étude (Hortal et al., 2008).

Au regard des erreurs de modélisation énoncées plus haut, notre étude n'a pas permis de faire la part entre les effets liés aux processus paysagers et ceux liés aux processus macro-évolutifs, pour expliquer les différences de présence d'espèces entre les 2 zones d'étude et par suite les erreurs de transférabilités des modèles locaux. Une telle confusion a été également observée dans le cas du transfert de modèles locaux de distribution de 10 espèces d'oiseaux dans 2 zones situées en Espagne (Seoane et al., 2005). Les auteurs soulignent que l'intégration de ces facteurs, processus écologiques locaux et macro-évolutifs, est un problème irrésolu dans la modélisation des habitats. La connaissance à l'échelle continentale des aires de distribution de chacune des espèces nous permettrait d'établir le "pool" régional commun et disponible aux 2 secteurs, pouvant être soumis ensuite aux conditions locales de l'environnement. Une telle information rendrait possible, dans la limite de la faible résolution des données d'atlas, la délimitation de zones dites biogéographiquement homogènes, c'est-à-dire où les espèces pouvant y être observées sont connues et où les prédictions des modèles développés localement (les relations environnement/espèces) pourraient être validées.

6.3.3 La dynamique des populations

De manière générale, nos modèles n'incorporaient pas les dynamiques des populations. L'équilibre des réponses des espèces aux variables environnementales est en effet un fort présupposé des modèles de distribution des habitats potentiels des espèces (Guisan and Zimmerman, 2000). Or, bien que nous ne l'avons pas démontré dans cette étude, les dynamiques des populations peuvent entraîner des erreurs de prédiction et expliquer l'absence d'espèces ou d'assemblages là où leurs habitats sont prédits. L'influence des dynamiques de population sur les modèles de distributions spatiales devrait être particulièrement considérée dans le cas des micro-mammifères, dont les dynamiques sont largement influencées par la structure et composition du paysage (Giraudoux et al., 2007).

Le développement de modèles spatialement et temporellement explicites (dynamiques) est possible (Dullinger et al., 2004) mais requiert une connaissance précise de la biologie des espèces et apparaît difficilement réalisable dans le cas de dynamiques d'assemblages. En revanche, il est plus facile d'incorporer les facteurs de variations des dynamiques, telles que des séries temporelles des variables environnementales (Marston, 2008), comme variables explicatives dans les modèles. Comme nous l'avons vu plus haut, le ROMPA (Ratio of Optimal to Marginal Patch Area) peut être utilisé comme un indicateur de l'état dynamique d'une population. Ainsi, son incorporation dans les modèles statiques comme d'autres facteurs paysagers pourrait aider ici encore à intégrer les effets des dynamiques des populations sur les prédictions.

6.4 Perspectives pour la cartographie des assemblages d'hôtes intermédiaires

La modélisation des distributions des assemblages développée dans ce travail constitue un outil de quantification de la définition de leurs habitats optimaux. L'habitat est en effet défini dans un espace multivarié par une densité de probabilité. La probabilité d'occurrence sera d'autant plus forte que les caractéristiques environnementales dans la zone considérée se rapprocheront de celles où l'assemblage a été observé (dans le jeu d'entraînement), et donc de l'habitat optimal. Ainsi, on peut estimer de manière contiguë dans l'espace, la probabilité de présence des habitats, et si nécessaire, l'habitat optimal pourra être discrétisé par la définition d'un seuil de probabilité.

Les modèles statistiques constituent alors des outils de prédilection pour quantifier les ROMPA des espèces de manière contiguë et continue dans l'espace (Marston, 2008). En revanche, il est difficilement imaginable d'interpréter une carte de ROMPA pour des assemblages. En effet, la proportion de l'habitat optimal de l'assemblage dans le paysage n'aura pas la même influence sur les dynamiques de chacune des espèces le composant. Or, les assemblages sont considérés comme des entités discrètes supposant une réponse commune des espèces aux variations environnementales.

En contrepartie, les modèles utilisés ici permettent une cartographie des habitats des assemblages pouvant caractériser les paysages à risque définis alors par la présence d'assemblages déséquilibrés, où une espèce potentiellement hôte pour le parasite domine. Les assemblages offrant la possibilité d'espèces hôtes alternatives pourraient également être repérés. Á l'inverse, les assemblages riches en espèces et équilibrés (forestiers dans notre cas d'étude) indiqueraient des aires à faible potentiel de transmission. On pourrait alors se focaliser à prédire la distribution de certains assemblages "à risque" individuellement. Il est également envisageable de classifier les occurrences de tous les assemblages sur une zone, et ainsi de visualiser la diversité en espèces et en habitats, puis, à la manière d'une carte de classification d'occupation du sol, d'établir une cartographie régionale de la biodiversité (Ferrier et al., 2004).

Nos résultats suggèrent que les prédictions de modèles développés localement, dans un site d'étude, pourront être effectuées dans l'aire échantillonnée uniquement. Face aux erreurs de prédiction des modèles locaux, la cartographie des assemblages sur des étendues régionales nécessite d'entraîner le modèle sur plusieurs sites d'étude et de piégeages. Dans ce cas, les assemblages peuvent être classifiés, c'est-à-dire modélisés et prédits de manière multiple plutôt qu'individuellement. Cette préférence pour une classification multiple a déjà été soulignée dans le cadre d'une cartographie régionale de données d'atlas (Elith et al., 2007).

La diversité δ des espèces (entre sites) a, dans notre cas d'étude, compliqué les prédictions régionales des assemblages. Les assemblages ont été définis indépendamment sur les 2 secteurs puisqu'ils ne partageaient pas la majorité de leurs compositions en espèces et en habitats. Nous avons tenté de définir, au préalable, les assemblages à une échelle régionale c'est-à-dire sur l'ensemble des données de Maerkang et Rangtang mais cette approche s'est avérée également compliquée. Un seul assemblage était en effet commun aux deux sites qui incluait les forêts de conifères et de rhododendrons et les genres *Ochotona* et *Microtus* communs aux 2 sites mais il intégrait également des habitats et des espèces non partagées par les 2 secteurs.

L'interprétation des prédictions d'un tel assemblage, dans les zones non échantillonnées et où le pool régional d'espèces potentiellement présentes est inconnu, aurait alors été confuse. En revanche, l'incorporation dans les modèles des variations spatiales des paramètres, de manière explicite, pourrait permettre de remédier aux confusions entre des assemblages différents mais présentant des espèces communes (en l'occurrence R3 et M4). L'on perçoit bien ici les limites, dans l'extrapolation des prédictions, de la définition d'entités discrètes et non modulables au regard de leurs composition en espèces que sont les assemblages. Idéalement, la modélisation jointe des espèces à l'échelle régionale permettrait d'éviter de telles confusions.

Enfin, une cartographie prédictive des assemblages à l'échelle continentale pourrait être réalisée en utilisant des données d'atlas. Á une telle échelle, les zones biogéographiques pourraient être intégrées, comme variables explicatives, dans les modèles. Bien que d'une moindre résolution que les données de piégeage, de telles données sont accessibles pour la quasi totalité du pays et sont depuis peu disponibles en données spatialisées (Xie et al., 2004). De plus, de nombreuses méthodes ont été développées pour modéliser les distributions sur la base seulement de données de présence (Elith et al., 2006). Dans ce contexte, MARS s'est révélé être un outil intéressant pour réaliser des cartes de prédictions régionales de 226 espèces de plantes, d'oiseaux, de mammifères et de reptiles, dans 6 régions du monde différentes (Elith et al., 2007). Dans la perspective d'améliorer les prédictions locales, une combinaison des données d'atlas pourraient être combinées à celles de piégeage (classes d'habitats), mais cette méthode, n'a pas, à ce jour, montré de résultats satisfaisants (McPherson et al., 2006).

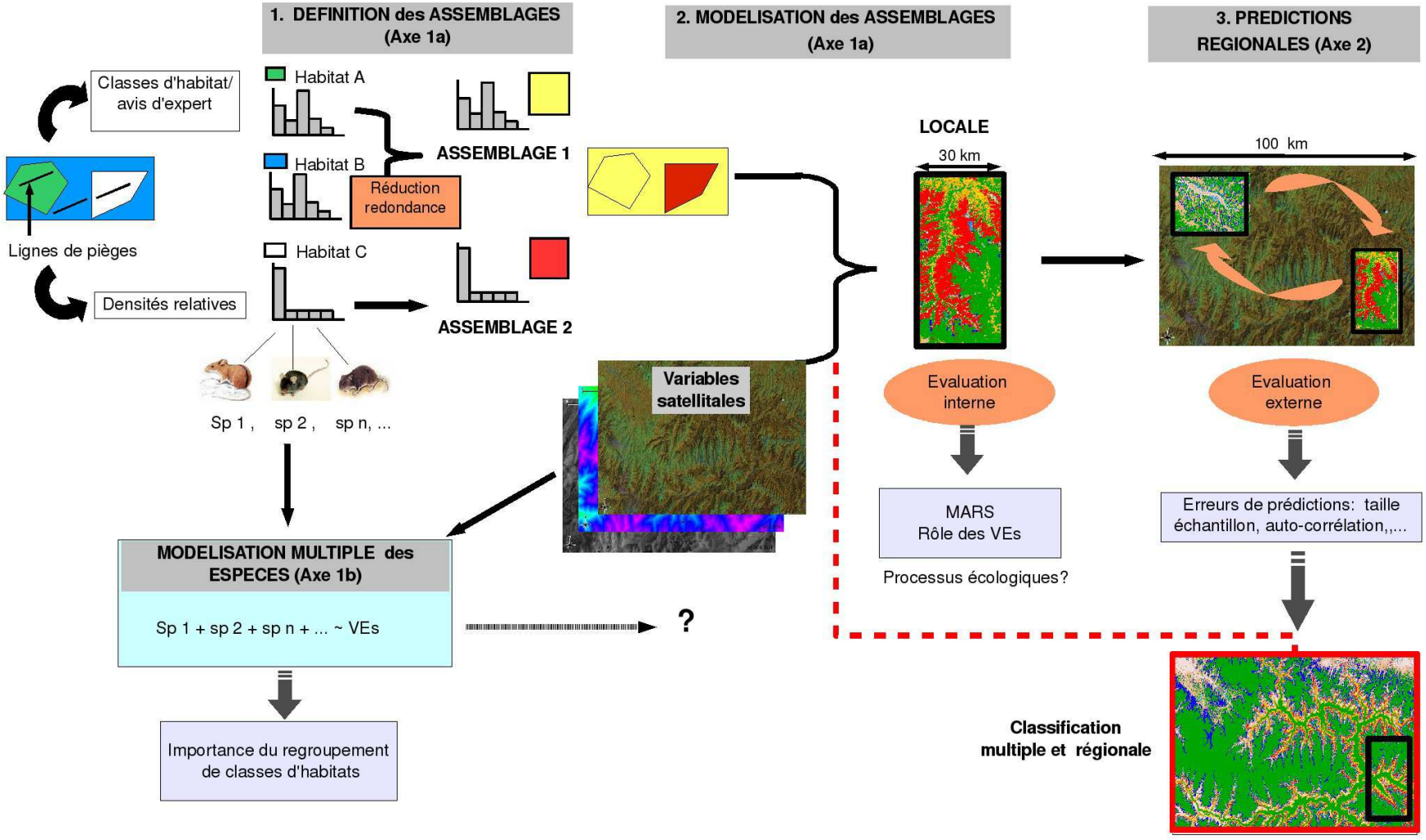

FIG. 6.1 – Schéma récapitulatif des principales étapes de la modélisation des distributions des assemblages de micro-mammifères, qui constituent les axes 1 et 2 de notre travail. Les principaux résultats sont indiqués dans les encadrés grisés et les points d'interrogation suggèrent les limites de l'approche.

Chapitre 7

Le rôle du chien dans la transmission

L'analyse des distributions de chiens domestiques ainsi que de leurs fèces (et contamination fécale) dans l'espace défini par l'environnement des hommes et des hôtes intermédiaires a permis une quantification partielle de la rencontre (du taux de contact) concernant i) les hôtes intermédiaires et les populations humaines avec la forme infectieuse du parasite (contamination environnementale) et ii) les hôtes définitifs avec les hôtes intermédiaires.

Dans l'environnement des villages, nous avons pu définir l'espace des chiens domestiques à l'intérieur duquel ils réalisent leurs comportements à risque (Figure 7.1).

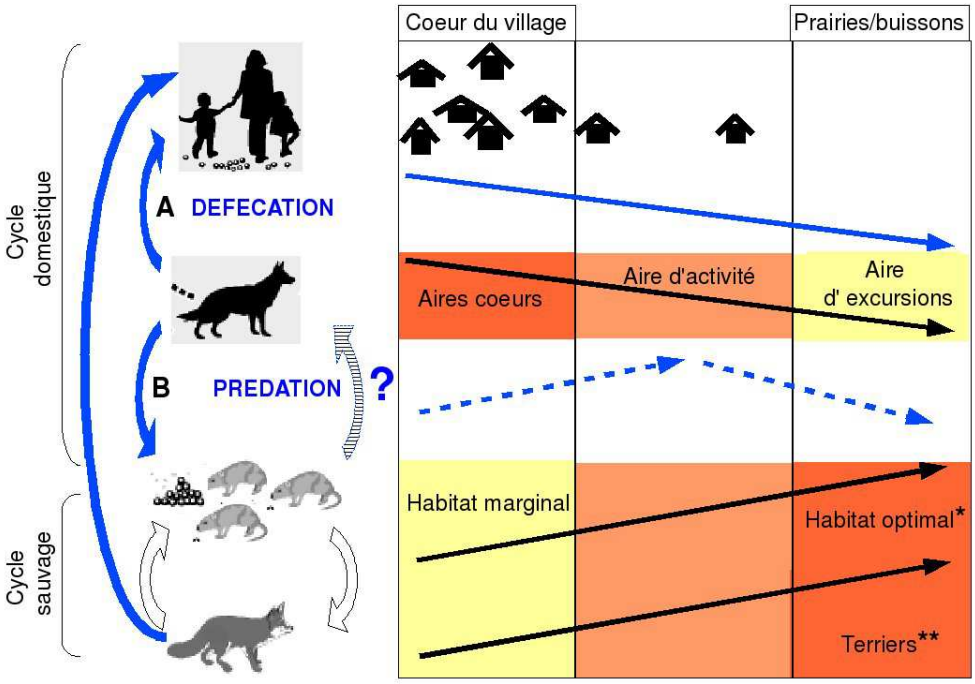

FIG. 7.1 – Représentation des fréquences d'utilisation de l'espace par les populations d'hôtes domestiques et sauvages et des comportements "à risque" du chien (représentés dans la partie gauche), en fonction de la distance aux habitations. Les comportements du chien impliqués dans la transmission sont mentionnés par des flèches bleues, pleines pour les comportements étudiés, et hachurées pour ceux uniquement discutés. "A" indique la contamination de l'environnement humain et "B" celle des habitats des hôtes intermédiaires.(*) (Wang et al., 2004; Raoul et al., 2006), (**) (Wang et al., 2007)

Dans ce cadre spatial, nous discuterons ci-après de la quantification des comportements à

risque déterminant le rôle du chien dans la contamination environnementale ainsi que de son mode de contamination.

7.1 Comment le chien contamine l'environnement ?

7.1.1 Contamination de l'environnement des hommes

7.1.1.1 Principaux résultats

L'analyse des contaminations fécales et la détermination par identification d'ADN de l'espèce hôte (chien *versus* renard) ont permis de rendre compte du rôle prédominant des chiens sur les renards dans la contamination de l'environnement humain. Ces résultats confirment l'existence d'un cycle péri-domestique où le chien intervient comme hôte définitif, établissant un lien entre le réservoir sauvage et les populations humaines (Figure 7.1). Cependant, l'existence de fèces de renards, même sporadique, à l'intérieur des villages, suggère la possible réalisation d'un cycle sauvage synanthropique dans ces villages du plateau tibétain tel qu'il a déjà été observé dans certaines agglomérations en Europe (Deplazes et al., 2004; Romig et al., 2006). Ainsi, l'existence d'un cycle péri-domestique et dans une moindre mesure celle d'un cycle sauvage synanthropique, à proximité des habitations humaines dans les villages tibétains, associée à de fortes densités de populations de chiens domestiques, apporte des éléments pour caractériser l'EA comme une zoonose émergente dans cette région endémique, telle que décrite dans certaines villes d'Europe (Eckert et al., 2000; Eckert and Deplazes, 2004). D'autre part, cette situation éco-épidémiologique est comparable à celle des villages côtiers de la mer de Bering sur l'île Saint Lawrence en Alaska. Dans ces zones d'endémie pour *Em*, l'existence d'un cycle péri-domestique à l'intérieur des villages, où les chiens sont supposés contaminer l'environnement des communautés d'esquimaux Yupik, a été mise en évidence, en parallèle du cycle sauvage se réalisant dans les zones de toundras (Rausch and Fay, 2002).

Par ailleurs, nous avons montré une agrégation des fèces et de leurs potentielles contaminations autour des habitations, devenant alors une source d'exposition au parasite pour les populations humaines (Figure 7.1). Cependant l'interprétation du risque de transmission aux hommes ne tient pas compte de l'aléa engendré par les comportements humains. Or, à l'échelle du village, les comportements des hommes peuvent modifier la probabilité de contact avec le parasite et donc la transmission. Souvent oubliés dans les programmes de contrôle des zoonoses (Macpherson, 2005), de tels paramètres devraient être considérés dans la modélisation du risque de transmission. Dans le cas d'*Em*, l'hygiène et la proximité aux chiens constituent des facteurs de risque de contact avec le parasite mais sont difficilement modélisables. Cependant, comme pour les autres hôtes, l'utilisation par l'homme de l'espace potentiellement contaminé n'est pas homogène. Il dépend alors des activités et des relations qu'entretiennent les populations avec leur environnement. Les comportements spatiaux et les fréquences d'utilisation de l'espace des populations humaines pourraient être analysés et mis en corrélation, localement, avec celles des populations de chiens. De plus, l'estimation de la variabilité de ces distributions en fonction de la zone géographique (par exemple entre pâtures d'hiver et d'été) pourrait aider à l'identification des zones géographiques particulièrement à haut risque de transmission (Schantz et al., 2003).

7.1.1.2 Applications au contrôle de la transmission

Des mesures de prévention concernant la contamination du sol par les chiens ont été réalisées pour contrôler la transmission de *Toxocara*, le plus répandu des parasites helminthes des chiens dans le monde. Le retrait des fèces par les propriétaires des chiens, ou l'interdiction d'accès des chiens aux aires de jeux publiques, ont été mis en place. Dans le cas d'*Em* en Chine, la caractérisation des zones à forte contamination dans l'environnement des hommes peut aider à la mise en place de telles mesures de contrôle. Cependant, l'accès à l'eau propre étant limitée et les conditions d'hygiène étant précaires, le retrait des fèces par les propriétaires pourrait devenir un facteur de risque de transmission.

7.1.2 Contamination des habitats des hôtes intermédiaires

Le fonctionnement du cycle péri-domestique est décrit comme dépendant du cycle sauvage, les hôtes intermédiaires étant supposés être contaminés essentiellement par les fèces de renards (Eckert et al., 2000; Deplazes et al., 2004; Romig et al., 2006). Au regard des faibles densités de fèces trouvées dans les habitats optimaux pour les micro-mammifères, nos résultats suggèrent que les chiens participent effectivement peu à la contamination du réservoir sauvage. Cependant, nos résultats ne permettent pas de confirmer le rôle prédominant du renard dans cette contamination en raison des faibles densités de fèces de cet hôte trouvées dans les prairies environnant les villages. Au regard des faibles densités de fèces rencontrées en général dans ces zones, un effort d'échantillonnage plus intense permettrait probablement de comprendre les implications relatives des chiens et des renards dans la contamination des habitats des hôtes intermédiaires.

D'autre part, nos résultats attestent de la présence de colonies de micro-mammifères (*Ochotona*, *Microtus* ou *Cricetulus*) à proximité des habitations. Aussi, *Cricetulus kamensis*, espèce susceptible pour *Em*, a été piégée dans le village de Tuan Jie à quelques dizaines de kilomètres de nos sites d'étude (Raoul et al., 2006). De plus, les renards sont supposés fréquenter de manière peu intensive l'espace du village du fait des fortes densités de chiens domestiques. Il est donc possible que le cycle péri-domestique se réalise presque indépendamment du cycle sauvage dans des micro-foyers à proximité des habitations. Dans ces aires restreintes, les probabilités de rencontre entre les micro-mammifères, les oeufs du parasite et les chiens et, par conséquent, la contamination canine, seraient favorisées.

7.2 Comment le chien se contamine ?

7.2.1 Éléments d'écologie comportementale du chien domestique

La contamination des chiens passe nécessairement par la prédation d'hôtes intermédiaires infectés. La prédation suppose la recherche de proies et leur consommation et le choix de réaliser ces actions correspond, pour les individus, à un compromis entre gain et perte d'énergie ("Optimal foraging") (Krebs, 1978).

En milieu urbain, le renard adopte un comportement de prédation opportuniste exerçant une pression de prédation plus faible sur les populations de micro-mammifères et profitant de la disponibilité de ressources anthropiques (Deplazes et al., 2004; Hegglin et al., 2007). Ainsi,

concernant la transmission d'*Em*, il a été montré que le cycle fonctionne moins intensément dans les zones de centres urbains où les domaines vitaux des renards sont plus petits et les proies moins abondantes, que dans les zones de périphérie (Deplazes et al., 2004). Mais il existe une zone intermédiaire entre la zone urbaine et sa périphérie, dans laquelle les prévalences des renards sont plus élevées qu'en ville et où la transmission du parasite pourrait être particulièrement intense du fait de densités de populations de prédateurs élevées (effet de la ville) et des habitats favorables pour les espèces proies d'arvicolinae qui y abondent (Deplazes et al., 2004; Fischer, 2003; Romig et al., 1999). Des comportements de prospection ont en effet été observés chez quelques renards urbains dans ces zones de périphérie (Robardet et al., 2008).

Chez le chien domestique, le caractère opportuniste de la prédation est susceptible d'être d'autant plus marqué que celui-ci ne dépend pas de la prédation pour sa survie. Contrairement au renard urbain pouvant prospecter dans les habitats optimaux des proies, on peut supposer que le chien n'investira que peu d'énergie dans la prédation. Sous cette hypothèse, on peut donc estimer la probabilité de rencontre entre les chiens et les hôtes intermédiaires par la superposition des espaces utilisés par ces deux populations.

Notre étude constitue une première tentative de quantification de la probabilité de contact entre ces hôtes. La présence de colonies de micro-mammifères et l'utilisation de l'espace par les chiens ont tendance à varier de manière opposée au regard de leurs distances aux habitations (Figure 7.1). En effet, les densités de micro-mammifères augmentent alors que la fréquence d'utilisation de l'espace par les chiens diminuent quand la distance aux habitations augmente. Á la manière des interactions renard/micro-mammifères, la probabilité de contact et le risque de transmission seraient donc plus importants dans des zones d'aire d'activité des chiens, excluant leurs aires coeur, où les densités de micro-mammifères sont intermédiaires. D'après nos données de télémétrie, cela correspondrait aux aires moyennes d'activités des chiens comprises entre 15 et 170 hectares autour des villages (cf Axe 3).

La proximité des chiens aux habitations humaines est plus forte que celle des renards si bien que les chiens peuvent être en contact avec les espèces de micro-mammifères occupant les habitations ou les jardins. On peut s'attendre à ce que la capture de proies soit facilitée dans de tels habitats de taille restreinte où le chien passe la plupart de son temps. Ainsi, l'hypothèse de la contamination des chiens dans les habitations a été supposée puisque des rongeurs "domestiques" ont été trouvés infesté sur le plateau tibétain (Qiu, com. pers.). Á l'heure actuelle les seules proies identifiées dans les fèces de chiens étaient du genre *Ochotona* et *Microtus*, dont les habitats optimaux se situent à l'extérieur des villages. Cependant, une grande majorité de macro-restes étant non identifiables, l'identification des items alimentaires et des proies du chien pourrait être améliorée par des outils moléculaires d'analyse du régime alimentaire (Valentini et al., 2009).

D'autre part, l'existence de comportements de prospection pourrait influencer l'estimation du taux de contact et devrait être pris en considération. Nous avons pu mettre en évidence l'existence de trajectoires excursives pour certains chiens, mais les comportements de prédation n'ont pas été étudiés. Cette recherche nécessiterait d'analyser les trajectoires elles-mêmes, c'est-à-dire les distributions des vitesses et des angles caractérisant les mouvements, et non pas uniquement les distributions spatiales des localisations. Aussi, les trajectoires excursives étant plutôt rares (1 à 2 par nuit) et pas systématiquement observées dans la population (16 % des chiens équipés), la détection de tels comportements nécessiterait de travailler sur les variations

individuelles des comportements spatiaux. De surcroît, des équations différentielles (telle que le modèle de la "marche aléatoire") peuvent être utilisées pour modéliser les trajectoires (Turchin, 1998). Récemment, des mixtures de modèles de "marches aléatoires" ont été utilisées pour décrire l'hétérogénéité des comportements au sein d'une même trajectoire et discriminer les comportements de repos et exploratoires dans des trajectoires d'élans (Morales et al., 2004). Dans notre cas d'étude, la classification des distributions de probabilités des paramètres du mouvement pourrait permettre une définition plus précise des aires coeur et des trajectoires excursives basée, dans nos travaux, uniquement sur la distance aux points de lâchers. Le développement d'un modèle ou d'une mixture de modèles capables de considérer la distribution non aléatoire des localisations, c'est-à-dire une agrégation autour des points de lâcher et des attractions vers des zones extérieures aux aires moyennes d'activités, est envisageable.

7.2.2 La modélisation spatiale de la rencontre

Le taux d'infection (β) par an a été estimé par simulation et s'avère d'une grande importance dans les modèles de transmission (Budke et al., 2005a; Torgerson, 2003). La probabilité de rencontre entre les hôtes intermédiaires et définitifs constitue un facteur pouvant expliquer en partie le taux d'infection de l'hôte définitif. Son estimation peut alors compléter l'information apportée par les facteurs épidémiologiques dans la transmission tels que les charges parasitaires et les prévalences des hôtes. Nos résultats de données de télémétrie soulignent l'apport de l'écologie comportementale pour une première approche de l'estimation du taux d'infection entre les chiens et les micro-mammifères tel qu'il avait été estimé entre les individus hôtes de la rage ou de la maladie de Carré (Kauhala and Holmala, 2006; Haydon, 2008).

Á une échelle sectorielle, notre étude a permis une première superposition spatiale des distributions des hôtes intermédiaires et définitifs. Cependant, l'estimation des distributions des fréquences d'utilisation de l'espace ont été dans notre étude discrétisées (non continues) et étaient non contiguës sur notre aire d'étude. En effet, les fréquences d'utilisation de l'espace par les chiens ont été classées en catégories et les présences des rongeurs n'étaient disponibles que le long de transects. Ainsi, dans la perspective de construction de cartes du risque de transmission dans ces villages, les probabilités de présence des populations d'hôtes définitifs et intermédiaires pourraient être superposées. Cela suppose quelques améliorations de notre protocole et de nos analyses. Concernant l'analyse de l'utilisation de l'espace par les chiens, une estimation correcte des fonctions d'utilisation de l'espace des populations pourra être obtenue en s'assurant que le nombre de chiens échantillonnés est suffisamment grand pour capturer l'essentiel du domaine vital de la population dans chaque village. Concernant les distributions spatiales des micro-mammifères, les méthodologies utilisées dans les axes 1 et 2 peuvent être appliquées pour prédire les distributions spatiales des espèces échantillonnées le long des transects ou dans des campagnes de piégeages. Marston (2008) a déjà modélisé les distributions d'*Ochotona* en fonction d'indices de végétations et de métriques paysagères dans le comté de Shiqu où nos suivis de chiens ont été réalisés. Cependant, ces cartes de prédictions ne sont pas validées si bien que l'erreur d'extrapolation entre les transects n'est pas estimée. Nos données de transects pourraient servir à valider ou à améliorer les prédictions établies des distributions dans chacun des villages où les chiens ont été suivis.

Enfin, les paramètres des distributions spatiales des hôtes intermédiaires (habitats optimaux) et définitifs (aires d'activités) peuvent venir alimenter le développement de modèles de transmission spatio-déterministes (Hansen et al., 2003; Milner-Gulland et al., 2004) pour modéliser la transmission d'*Em* dans ces zones du plateau tibétain.

7.2.3 Applications au contrôle des prévalences d'*Em*

Actuellement, le principal traitement des infections canines consiste à administrer des appâts contenant un vermifuge (anthelminthe), le Praziquantel. Réalisés avec succès en Allemagne (par voie aérienne) (Romig et al., 1999) ou au Japon (manuellement, par voie terrestre) (Craig et al., 1996), les traitements de masse se sont avérés longs, difficiles et coûteux. Actuellement en Chine, sur le plateau tibétain, dans le comté de Shiqu, des appâts aux Praziquantel sont administrés de manière systématique aux chiens domestiques. Parallèlement, un suivi épidémiologique des populations de chiens domestiques est en cours (Jasmin Moss, com. pers.). Il permet une surveillance des statuts infectieux dont la quantification peut être utilisée pour la gestion des programmes de contrôle (Macpherson and Craig, 2005).

Connaissant les sources de variations des paramètres écologiques du cycle, il est possible d'agir et de traiter dans les zones où les comportements de l'hôte définitif "à risque" pour les populations humaines se réalisent (défécation et prédation) (Eckert and Deplazes, 2004). Ainsi, à Zurich, des appâts au Praziquantel ont été administrés tous les deux mois à des renards situés dans six zones de 1 km^2 en périphérie urbaine régulièrement fréquentée par les renards et ayant accès aux hôtes intermédiaires (*Arvicola terrestris*). Ces zones, fréquentées régulièrement par les hommes et leurs chiens domestiques en tant qu'aires d'activités, sont à fort risque de transmission.

Concernant les chiens domestiques, partageant inévitablement les mêmes lieux de vie que leurs maîtres, il est envisageable de considérer leur comportement de prédation, et de ne traiter que les chiens ayant accès aux micro-mammifères infectés, ce toutes les 4 semaines (Eckert and Deplazes, 2004). Nos résultats indiquent que tous les chiens de propriétaires ont potentiellement accès aux micro-mammifères avec une légère tendance pour les mâles à prospecter dans les prairies environnantes. La détermination des variations inter-individuelles des comportements spatiaux "à risque" liés aux traits d'histoire de vie des chiens, mis en évidence par les études de facteurs de risque, pourrait aider à mettre en place des protocoles de contrôle sélectifs. Pour ce faire, les suivis des trajectoires devraient être réalisés sur un nombre plus important de chiens suivant un échantillonnage stratifié de chacune des catégories de chiens définies au préalable comme variables explicatives : la fonction du chien (chien de garde des troupeaux, chien de garde des habitations), son degré de domestication (de propriétaire, errants) et ses traits d'histoire de vie telle que la classe d'âge.

7.3 Degré de domestication et place du chien errant dans le cycle

La proximité des hommes aux animaux domestiques est à l'origine de nombreuses zoonoses (Macpherson, 2005). Une longue histoire lie les chiens et les hommes depuis les débuts de la domestication de cette espèce dans les temps préhistoriques, il y a quelques 12 à 15 milles

ans (5000 ans pour les chats) (Morey, 1994). Au cours de ce processus, l'homme a modifié les comportements des chiens. Ainsi, la taille du domaine vital mais aussi l'activité de prédation diminuent avec le degré de dépendance des chiens aux populations humaines.

Comme partout ailleurs dans le monde, la domestication du chien en Chine a engendré une population dépendante des activités et des comportements humains. Sur les plateaux, où le chien a un caractère sacré, il semblerait qu'il en existe au moins 2 catégories, les chiens de propriétaire et les chiens errants ou non nourris systématiquement par les hommes (l'existence de chiens sauvages n'étant pas prouvée à notre connaissance).

Notre étude a porté sur les comportements à risque des chiens domestiques puisque que la proximité qu'ils ont avec leurs maîtres constitue un facteur de risque et que la plupart des habitants en possèdent. Cependant, comme dans d'autres régions du monde dites "en développement", les populations de chiens n'y étant pas systématiquement contrôlées et la relation aux animaux de compagnie étant variable, les chiens errants y sont également très nombreux. Qui plus est, ces derniers peuvent être nourris dans les monastères. Alors qu'en Europe de l'ouest ce sont les chiens de propriétaires qui sont davantage à risque et contaminent les places publiques, dans les aires endémiques où ils sont nombreux, les chiens errants peuvent également tenir une place importante dans la transmission. Par exemple, en Australie (Brisbane), les chiens ferraux participent davantage aux infections humaines par la dissémination des oeufs de *Ankylostoma caninum* que les chiens domestiques (Macpherson, 2005).

Concernant la transmission d'*Em*, les chiens errants sont susceptibles d'exercer une plus forte pression de prédation sur les populations de micro-mammifères que les chiens domestiques. De surcroît, fréquentant les villages et étant nombreux, ils sont susceptibles de contaminer cet environnement. L'association de tels facteurs laisse supposer que ces chiens occupent une place dans le cycle péri-domestique d'*Em*. Sur les plateaux tibétains, le risque de transmission à l'homme pourrait donc être dupliqué par l'existence de 2 populations de chiens à risque, errants et domestiques. L'équipement par colliers GPS de chiens errants n'a pas été possible dans notre protocole expérimental et préliminaire. Bien que les distributions spatiales des fèces corroborent l'utilisation de l'espace des chiens domestiques équipés, nous ignorons à ce jour la contribution des chiens errants et non domestiqués dans la contamination environnementale.

Une méthode consisterait à marquer les fèces des chiens de propriétaires (étiquettes dans appâts) afin de les discriminer de celles des chiens errants. Cependant, l'identification des fèces de chiens errants supposerait alors de marquer toutes les fèces de chaque chien domestique, ce qui semble difficilement réalisable de manière optimale. En revanche, nos résultats suggèrent que les comportements spatiaux dans l'environnement des hommes et des micro-mammifères peuvent être utilisés pour définir les zones où les comportements potentiellement à risque (défécation, prédation) ont lieu ainsi que leur fréquences d'utilisation. Ainsi, la comparaison des distributions spatiales des individus entre les 2 catégories de chiens pourrait être réalisée en obtenant un échantillon de trajectoires de chiens errants comparable à celles de chiens de propriétaires obtenues lors de notre étude préliminaire.

Conclusion générale

L'étude de l'écologie spatiale des populations des hôtes pour *Em* a apporté quelques éléments de connaissance du fonctionnement de la transmission du parasite sur des zones d'endémies chinoises. D'une part, les assemblages d'espèces de micro-mammifères et les distributions relatives des espèces susceptibles au parasite ont été identifiées dans les paysages diversifiés des contreforts du plateau tibétain. D'autre part, le rôle primordial du chien dans la transmission du parasite à l'homme et l'existence d'un cyle péri-domestique en étroite relation avec le cycle sauvage ont été trouvés. Aussi, notre travail participe, dans une certaine mesure, au champs disciplinaire de l'éco-épidémiologie.

Les diverses notions (assemblages, habitats, comportements de défécation et de prédation,...) et les méthodes de collectes de données (techniques de piégeage, de collecte de fèces, de télémétrie) empruntées à notre discipline ont permis de considérer de multiples facteurs (environnementaux, individuels) et différents niveaux d'organisation biologiques pour expliquer les distributions et les comportements observés. De sucroît, ce travail, en accord avec l'orientation actuelle de la discipline, s'est orienté vers une exploration et appropriation d'outils de statistiques modernes (régressions non linéaires, bootstrap) et de spatialisation (SIG) qui ont été utiles pour l'analyse d'une telle complexité.

Cependant, au regard des erreurs de prédictions ou de définition obtenues, les modèles utilisés ne sont pas assez développés pour produire des prédictions robustes des distributions spatiales des populations de micro-mammifères ou de comportements de canidés. C'est pourquoi, nos résultats de modélisation ne peuvent être directement incorporés dans des modèles de transmission du risque éco-épidémiologique ni utilisés pour la gestion des populations et sont donc considérées comme préliminaires. En revanche, ils soulignent les limites des méthodes utilisées ainsi que des données dont nous disposions. Ils permettent de proposer de nouvelles améliorations des protocoles de collecte des données et de modélisation, de sorte à ce que l'aller-retour entre les différentes perceptions du problème, l'étude de terrain et la modélisation aboutisse à une meilleure compréhension des patrons de distribution observés et, par suite, à des prédictions plus précises et plus robustes. Le réalisme de nos modèles pourrait, par exemple, être amélioré par un protocole d'échantillonnage supervisé, alors que l'incorporation de l'autocorrélation spatiale permettrait de pallier les difficultés de terrain à collecter des données non corrélées. Une étroite interaction entre les suivis de terrain et le développement des modèles est préconisée pour parvenir au meilleur compromis entre la richesse des méthodes existantes en écologie et en statistiques et les limites logistiques et humaines de leurs applications.

Bibliographie

World Health Organization , 2006. The control of neglected zoonotic diseases: a route to poverty alleviation.
URL `http://www.who.int/zoonoses/Report_Sept06.pdf`

Anderson, M., Clements, A., 2000. Resolving environmental disputes: a statistical method for choosing among competing cluster models. Ecological Applications 10 (5), 1341–1355.

Araujo, M., Guisan, A., 2006. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33, 1677–1688.

Austin, M., 1999. A silent clash of paradigms : some inconsistencies in community ecology. Oikos 86 (1), 170–178.

Austin, M., 2002. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling 157 (18), 101–118.

Barry, S., Elith, J., 2006. Error and uncertainty in habitat models. Journal of Applied Ecology 43 (11), 413–423.

Begon, M., 2008. Effects of host diversity on disease dynamics. In : Ostfeld, R., Keesing, F., Eviner, V. (Eds.), Infectious Disease Ecology: Effects of Ecosystems on Disease and of Disease on Ecosystems. Princeton University Press, Princeton, NJ, pp. 12–29.

Blondel, J., 1995. Biogéographie. Approche écologique et évolutive. Masson.

Boitani, L., Francisci, F., Ciucci, P., Andreoli, G., 1995. Population biology and ecology of feral dogs in Central Italy. In : Serpell J., E. (Ed.), The domestic dog: its ecology, behaviour and evolution. Cambridge: Cambridge University Press., pp. 217–243.

Bowman, J., Forbes, G., Dilworth, T., 2000. The spatial scale of variability in small-mammal populations. Ecography 23, 328–334(7).

Budke, C. M., Campos-Ponce, M., Qian, W., Torgerson, P. R., 2005a. A canine purgation study and risk factor analysis for echinococcosis in a high endemic region of the Tibetan plateau. Veterinary Parasitology 127 (1), 43 – 49.

Budke, C. M., Jiamin, Q., Craig, P., Torgerson, P., 2005b. Modeling the transmission of *Echinococcus granulosus* and *Echinococcus multilocularis* in dogs for a high endemic region of the Tibetan plateau. International Journal for Parasitology 35 (2), 163 – 170.

Burel, F., Baudry, J., 1999. Écologie du paysage. Concepts, méthodes et applications. Paris, TEC & DOC.

Burnham, K., Anderson, D., 1998. Model selection and multimodel inference. 2nd ed. Springer-verlag, New York.

Butet, A., Paillat, G., Y., D., 2006. Factors driving small rodents assemblages from field boundaries in agricultural landscapes of western France. Landscape Ecology 21, 449–461.

Calenge, C., 2005. Des outils statistiques pour l'analyse des semis de points dans l'espace ecologique. These de doctorat, Université Claude Bernard, Lyon I.

Combes, C., 2001. Interactions durables. Ecologie et Evolution du parasitisme. Masson, Paris.

Craft, M. E., Hawthorne, P. L., Packer, C., Dobson, A. P., 2008. Dynamics of a multihost pathogen in a carnivore community. Journal of Animal Ecology 77 (6), 1257–1264.

Craig, P., Giraudoux, P., Shi, D., Bartholomot, B., Barnish, G., Delattre, P., Quéré, J., Harraga, S., Bao, G., Wang, Y., Lu, F., Ito, A., Vuitton, D., 2000. An epidemiological and ecological study of human alveolar echinococcosis transmission in south gansu, China. Acta Tropica 77, 167–177.

Craig, P., Liu, D., Shi, D., Macpherson, C., Barnish, G., Reynolds, D., Gottstein, B., Wang, Z., 1992. A large focus of alveolar echinococcosis in central China. The Lancet 340 (8823), 826–831.

Craig, P., The Echinococcosis Working Group in China, 2006. Epidemiology of human alveolar echinococcosis in China. Parasitology International 55, 221–225.

Danson, F. M., Armitage, R. P., Marston, C. G., 2008. Spatial and temporal modelling for parasite transmission studies and risk assessment. Parasite 15 (3), 463–8.

Danson, F. M., Graham, A. J., Pleydell, D. R. J., Campos-Ponce, M., Giraudoux, P., Craig, P. S., 2003. Multi-scale spatial analysis of human alveolar echinococcosis risk in China. Parasitology 127, 133–141.

Daszak, P., Cunningham, A., Hyatt, A., 2001. Anthropogenic environmental change and the emergence of infectious diseases in wildlife. Acta Tropica 78, 103–116.

Daszak, P., Cunningham, A. A., Hyatt, A. D., 2000. Emerging Infectious Diseases of Wildlife–Threats to Biodiversity and Human Health. Science 287 (5452), 443–449.

de la Péña, M., Butet, A., Delettre, Y., Paillat, G., Morant, P., Le Du, L., Burel, F., 2003. Response of the small mammal community to changes in western french agricultural landscapes. Landscape Ecology 18 (3), 265–278.

Delattre, P., Giraudoux, P., Quéré, J., 1990. Conséquences épidémiologiques de la réceptivité d'un nouvel hôte intermédiaire du taenia multiloculaire (*Echinococcus multilocularis*) et de la localisation spatiotemporelle des rongeurs infestés. C.R. Acad. Sci. Paris 310 (3), 339–344.

Deplazes, P., Alther, P., Tanner, I., Eckert, J., 1999. *Echinococcus multilocularis* coproantigen detection by enzyme-linked immunosorbent assay in fox, dog, and cat populations. The Journal of Parasitology 85 (1), 115–121.

Deplazes, P., Dinkel, A., Mathis, A., 2003. Molecular tools for studies on the transmission biology of echinococcus multilocularis. Parasitology 127 (Supplement), 53–61.

Deplazes, P., Hegglin, D., Gloor, S., Romig, T., 2004. Wilderness in the city: the urbanization of *Echinococcus multilocularis*. Trends in Parasitology 20 (2), 77–84.

Dinkel, A., Von Nickisch-Rosenegk, M., Bilger, B., Merli, M., Lucius, R., Romig, T., 1998. Detection of *Echinococcus multilocularis* in the definitive host: coprodiagnosis by PCR as an alternative to necropsy. Journal Of Clinical Microbiology 36 (7), 1871–1876.

Dobson, A., Foufopoulos, J., 2001. Emerging infectious pathogens of wildlife. Philosophical Transactions of the Royal Society of London B Biological Sciences. 29, 1001–1012.

Doledec, S., Chessel, D., Gimaret-carpentier, C., 2000. Niche separation in community analysis: a new method. Ecology 81, 2914–2927.

Dullinger, S., Dirnbock, T., Grabherr, G., 2004. Modelling climate change-driven treeline shifts: relative effects of temperature increase, dispersal and invasibility. Journal of Ecology 92, 241–252(12).

Dupuy, G., Giraudoux, P., Delattre, P., 2009. Numerical and dietary responses of a predator community in a temperate zone of europe. Ecography 31, 1–14.

Duscher, G., Pleydell, D., Prosl, H., Joachim, A., 2006. *Echinococcus multilocularis* in austrian foxes from 1991 until 2004. Journal of Veterinary Medicine, Series B 53, 138–144(7).

Eckert, J., Conraths, F., Tackmann, K., 2000. Echinococcosis: an emerging or re-emerging zoonosis ? International Journal for Parasitology 30, 1283–1294(12).

Eckert, J., Deplazes, P., 2004. Biological, Epidemiological, and Clinical Aspects of Echinococcosis, a Zoonosis of Increasing Concern. Clinical Microbiology Review 17 (1), 107–135.

Efron, B., Tibshirani, R., 1995. Cross-validation and the bootstrap: estimating the error rate of a prediction rule. Technical report, Dept. of Statistics, Stanford University.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the .632+ bootstrap method. Journal of the American Statistical Association 92 (438), 548–560.

Elith, J., Leathwick, J., 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity & Distributions 13 (3), 265–275.

Elith, J., Burgman, M. A., Regan, H. M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distribution. Ecological Modelling 157 (2-3), 313–329.

Elith, J., Ferrier, S., Huettmann, F., Leathwick, J., 2005. The evaluation strip: a new robust method for plotting predicted responses from species distribution models. Ecological Modelling 186, 280–289.

Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Peterson, T. A., Phillips, S. J., Richardson, K., Scachetti-Pereira, R., Schapire, R. E., Soberon, J., Williams, S., Wisz, M. S., Zimmermann, N. E., 2006. Novel methods improve prediction of species distributions from occurrence data. Ecography 29 (2), 129–151.

Elith, J., Leathwick, J. R. R., Hastie, T., 2008. A working guide to boosted regression trees. The Journal of animal ecology.

Ferrier, S., Drielsma, M., Manion, G., Watson, G., 2002b. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. ii. community-level modelling. Biodiversity and Conservation 11, 2309–2338.

Ferrier, S., Guisan, A., 2006. Spatial modelling of biodiversity at the community level. Journal of applied ecology 43 (3), 393–404.

Ferrier, S., Manion, G., Elith, J., Richardson, K., 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13 (13), 252–264.

Ferrier, S., Powell, G. V. N., Richardson, K. S., Manion, G., Overton, J. M., Allnutt, T. F., Cameron, S. E., Mantle, k., Burgess, N. D., Faith, D. P., Lamoreux, J. F., Kier, G., Hijmans, R. J., Funk, V. A., Cassis, G. A., Fisher, B. L., Flemons, P., Lees, D., Lovett, J. C., Van Rompaey, R. S. A. R., 2004. Mapping more of terrestrial biodiversity for global conservation assessment. BioScience 54 (12), 1101.

Fichet-Calvet, E., Pradier, B., Quéré, J., Giraudoux, P., Delattre, P., 2000. Landscape composition and vole outbreaks : evidence from an eight year study of *arvicola terrestris*. Ecography 23, 659–667.

Fischer, C., 2003. Relation in the presence of various parasites in the red fox (*Vulpes vulpes*) in Geneva. Swiss Medical Weekly 133 (S61).

Foltete, J., 2006. Paysage et mouvement : de l'écologie aux déplacements urbains : éléments pour une identification des paysages préférentiels. Rapport d'Habilitation à Diriger des Recherches. Université de Franche-Comté.

Friedman, J. H., 1991. Multivariate adaptive regression splines. Annals of Statistics 19, 1–67.

Furlanello, C., Neteler, M., Merler, S., Menegon, S., Fontanari, S., Donini, A., Rizzoli, A., Chemini, C., March 2003. Gis and the random forest predictor: integration in r for tick-borne disease risk assessment. In : Hornik, K., Leisch, F., Zeileis, A. (Eds.), Proceedings of the 3rd International Workshop on Distributed Statistical Computing, March 2003 ,Vienna, Austria.

Vol. DSC 2003.
URL http://www.ci.tuwien.ac.at/Conferences/DSC-2003

Gelfand, A., Silander, Jr., J. A., Wu, S., Latimer, A. M., Lewis, P. O., Rebelo, A. G., Holder, M., 2006. Explaining species distribution patterns through hierarchical modeling. Bayesian Analysis 1, 41–92.

Gibson, L., Wilson, B., Aberton, J., 2004. Landscape characteristics associated with species richness and occurence of small native mammals inhabiting a coastal heathland: a spatial modelling approach. Biological Conservation 120, 75–89.

Giraudoux, P., 1991. Utilisation de l'espace par les hôtes du ténia multiloculaire (*Echinococcus multilocularis*). Ph.D. thesis, Université de Bourgogne,Dijon.

Giraudoux, P., Craig, P., Delattre, P., Bao, G., Bartholomot, B., Harraga, S., Quéré, J., Raoul, F., Wang, Y., Shi, D., Vuitton, D., 2003. Interaction between landscape changes and host communities can regulate *Echinococcus multilocularis* transmission. Parasitology 127 (1), 121–131.

Giraudoux, P., Delattre, P., Habert, M., Quéré, J., Deblay, S., Defaut, R., Duhamel, R., Moissenet, M., Salvi, D., Truchetet, D., 1997. Population dynamics of fossorial water vole (*Arvicola terrestris scherman*): a land use and landscape perspective. Agriculture, Ecosystems and Environment 66, 47–60.

Giraudoux, P., Delattre, P., Takahashi, K., Raoul, F., Quéré, J., Craig, P., Vuitton, D., 2002. Transmission ecology of *Echinococcus multilocularis* in wildlife: what can be learned from comparative studies and multiscale approaches ? In : Cestode zoonoses: *Echinococcosis* and *Cysticercosis*. Craig, P.S. and Pawlowski, Z., IOS Press, pp. 251–266.

Giraudoux, P., Pleydell, D., Raoul, F., Quéré, J., Wang, Q., Yang, Y., Vuitton, D., Qiu, J., Yang, W., Craig, P., 2006. Transmission ecology of *Echinococcus multilocularis*: what are the ranges of parasite stability among various host communities in China ? Parasitology International 55, 237–246.

Giraudoux, P., Pleydell, D., Raoul, F., Vaniscotte, A., Ito, A., Craig, P., 2007. *Echinococcus multilocularis* : why are multidisciplinary and multiscale approaches essential in infectious disease ecology ? Tropical Medicine and Health 35, 293–299.

Giraudoux, P., Quéré, J., Delattre, P., Bao, G., Wang, X., Shi, D., Vuitton, D., Craig, P., 1998. Distribution of small mammals along a deforestation gradient in southern gansu, central China. Acta Theologica 43 (4), 349–362.

Giraudoux, P., Raoul, F., Pleydell, D., Craig, P., 2008. Multidisciplinary studies, systems approaches and eco-epidemiology : something old, something new. Parasite 15, 469–476.

Giraudoux, P., Vuitton, D., Bresson-Hadni, S., Craig, P., Bartholomot, B., Barnish, G., Laplante, J., Shi, D., Lenys, D., 1996. Mass screening and epidemiology of alveolar echinococcosis in France, western europe and gansu, central China: from epidemiology towards transmission

ecology. In : Uchino, J., Sato, N. (Eds.), Alveolar Echinococcosis, Strategy for Eradication of Alveolar Echinococcosis in the Liver. Fujishoin, Sapporo, pp. 197–211.

Glass, G., Cheek, J., Patz, J., Shields, T., Doyle, T., Thoroughman, D., Hunt, D., Enscore, R., Gage, K., Irland, C., Peters, C., Bryan, R., 2000. Using remotely sensed data to identify areas at risk for hantavirus pulmonary syndrome. Emerging Infectious Diseases 6 (3), 238–247.

Gottstein, B., Saucy, F., Deplazes, P., Reichen, J., Demierre, G., Busato, A., Zuercher, C., Pugin, P., 2001. Is high prevalence of *Echinococcus multilocularis* in wild and domestic animals associated with disease incidence in humans ? Emerging Infectious Disease 7 (3).

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8 (9), 993–1009.

Guisan, A., Zimmerman, N., 2000. Predictive habitat distribution models in ecology. Ecological modelling 135, 147–186.

Guislain, M., 2006. Etude à différentes fenêtres de perception, des facteurs impliqués dans la transmission d'*Echinococcus multilocularis*, parasite responsable d'une maladie émergente : l'échinococcose alvéolaire. Ph.D. thesis, Université de Franche-Comté.

Guo, Q., Kelly, M., Graham, C. H., 2005. Support vector machines for predicting distribution of sudden oak death in california. Ecological Modelling 182 (1), 75 – 90.

Hachacka, W.M.and Caruana, R., Fink, D., Munson, A., Riedewald, M., Sorokina, D., Kelling, S., 2007. Data-mining discovery of pattern and process in ecological systems. Journal of Wildlife Management 71 (7), 2427–2437.

Hall, L. S., Krausman, P., Morrison, M., 1997. The habitat concept and a plea for standard terminology. Wildlife Society Bulletin 25, 173–182.

Hansen, F., Tackmann, K., Jeltsch, F., Wissel, C., Thulke, H., 2003. Controlling *Echinococcus multilocularis*-ecological implications of field trials. Preventive Veterinary Medicine 60 (1), 91–105.

Hassan, R., Scholes, R., Ash, H., 2005. Human health : Ecosystem regulation of infectious diseases. In : Epstein, P., Githeko, A., Rabinovich, J., Weinstein, P. (Eds.), Ecosystems and human well-being. WHO, pp. 391–411.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, New York, 335 p.

Hastie, T., Tibshirani, R., 1993. Discriminant analysis by gaussian mixtures. Journal of the Royal Statistical Society. Series B (Methodological) 58 (1), 155–176.

Hastie, T., Tibshirani, R., Buja, A., 1994. Flexible discriminant analysis by optimal scoring. Journal of the American Statistical Association 89 (428), 1255–1270.

Hastie, T., Tibshirani, R., Friedman, J. H., August 2001. The Elements of Statistical Learning. Springer-Verlag.

Hay, S. I., 2000. An overview of remote sensing and geodesy for epidemiology and public health application. Advances in parasitology 47, 1–35.

Haydon, D. T., 2008. Cross-disciplinary demands of multihost pathogens. Journal of Animal Ecology 77, 1079–1081(3).

He, J., Qiu, J., Liu, F., Chen, X., Liu, D., Chen, W., 2000. Epidemiological survey on hydatidosis in tibetan region of western sichuan: ii. infection situation among domestic and wild animals. Chinese Journal of Zoology 5.

Hegglin, D., Bontadina, F., Contesse, P., Gloor, S., Deplazes, P., 2007. Plasticity of predation behaviour as a putative driving force for parasite life-cycle dynamics: the case of urban foxes and *Echinococcus multilocularis* tapeworm. Functional Ecology 21, 552–560(9).

Hirzel, A., Guisan, A., 2002. Which is the optimal sampling strategy for habitat suitability modelling. Ecological Modelling 157 (7), 331–341.

Hirzel, A., Hausser, J., Chessel, D., Perrin, N., 2002. Ecological niche factor analysis: how to compute habitat-suitability maps without absence data ? Ecology 83 (7), 2027–2036.

Hofer, S., Gloor, S., Muller, U., Mathis, A., Hegglin, D., Deplazes, P., 2000. High prevalence of *Echinococcus multilocularis* in urban red foxes (*Vulpes vulpes*) and voles (*arvicola terrestris*) in the city of Zurich, Switzerland. Parasitology 120, 135–142.

Hortal, J., Rodriguez, J., Nieto-Diaz, M., Lobo, J., 2008. Regional and environmental effects on the species richness of mammal assemblages. Journal of Biogeography 35, 1202–1214.

Hutchinson, G., 1957. Concluding remarks. Cold Spring Harbor Symp. Quantitative Biology, 415–427.

Ito, A., Urbani, C., Qiu, J., Vuitton, D., Dongchuan, Q., Heath, D., Craig, P., Zheng, F., Schantz, P., 2003. Control of echinococcosis and cysticercosis: a public health challenge to international cooperation in China. Acta Tropica 86, 3–17.

James, A., Choy, S. L., Mengersen, K., 2010. Elicitator: an expert elicitation tool for regression in ecology. Environmental Modelling & Software 25 (1), 129 – 145.

Johnson, J., 2004. Model selection in ecology and evolution. Trends in Ecology & Evolution 19 (2), 101–108.

Johnson, P. T. J., Hartson, R. B., Larson, D. J., Sutherland, D. R., 2008. Diversity and disease: community structure drives parasite transmission and host fitness. Ecology Letters 11, 1017–1026(10).

Kareiva, P., Odell, G., 1987. Swarms of predators exhibit "preytaxis" if individual predators use area-restricted search. The American Naturalist 130 (2), 233.

Kauhala, K., Holmala, K., 2006. Contact rate and risk of rabies spread between medium-size carnivores in southeast finland. Annales Zoologici Fennici 43, 348–357.

Kearney, M., 2006. Habitat, environment and niche: what are we modelling ? Oikos 115, 186–191.

Keesing, F., Holt, R. D., Ostfeld, R. S., April 2006. Effects of species diversity on disease risk. Ecology Letters 9 (4), 485–498.

Kitron, U., 1998. Landscape ecology and epidemiology of vector-borne diseases: tools for spatial analysis. Journal of Medical Entomology 35 (11), 435–445.

Kitron, U., 2000. Risk maps: transmission and burden of vector-borne diseases. Parasitology Today 16 (8).

Krebs, J., 1978. Optimal foraging: decision rules for predators. In : Krebs, J., Davies, N. (Eds.), Behavioural ecology, an evolutionary approach. Blackwell Scientific Publisher.

Lai, C., Smith, A., 2003. Keystone status of plateau pikas (*Ochotona curzoniae*): effect of control on biodiversity of native birds. Biodivers Conservation 12, 1901–12.

Langlois, J., Fahrig, L., Merriam, G., Artsob, H., 2001. Landscape structure influences continental distribution of hantavirus in deer mice. Landscape Ecology 16.

Leathwick, J., Elith, J., Hastie, T., 2006. Comparative performance of generalized additive models and multivariate adaptative regression splines for statistical modelling of species distributions. Ecological Modelling 199, 188–196.

Legendre, P., Legendre, L., 1998. Numerical Ecology. 2nd ed. Elsevier Science B.V, Amsterdam.

Li, T., Chena, X., Zhenb, R., Qiu, J., Qiu, D., Xiao, N., Ito, A., Wang, H., Giraudoux, P., Sako, Y., Nakao, M., Craig, P., Submitted. Widespread co-endemicity of human cystic and alveolar echinococcosis on the eastern tibetan plateau, China. Acta Tropica.

Linard, C., Tersago, K., Leirs, H., Lambin, E., 2007. Environmental conditions and puumala virus transmission in Belgium. International Journal of Health Geographics 6 (1), 55.

Liu, F., 1993. Prevalence of *Echinococcus granulosus* in dogs in the Xinjiang Uygur Autonomous Region, prc. In : Andersen, F. L. (Ed.), Compendium on Cystic Echinococcosis with Special Reference to the Xinjiang Uygur Autonomous Region,The People Republic of CHina. Provo, Utah, Brigham Young University, pp. 168–176.

MacCullag, Nelder, 1989. Generalized Linear Models. 2nd ed. London: Chapman and Hall.

Macpherson, C., 2005. Human behaviour and the epidemiology of parasitic zoonoses. International Journal for Parasitology 35, 1319–1331.

Macpherson, C., Craig, P., 2005. Dogs and cestode zoonoses. In : Dogs, Zoonoses and Public Health. Vol. 35. CABI Publishing, UK, Ch. 7, pp. 177–211.

March, D., Susser, E., 2006. The eco- in eco-epidemiology. International Journal of Epidemiology 35 (6), 1379–1383.

Marston, C., 2008. Spatial modelling of small mammal distributions in relation to parasite transmission in western CHina. Thesis report, School of Environment and Life Sciences University of Salford, Salford, UK.

Martin, T. G., Kuhnert, P. M., Mengersen, K., Possingham, H. P., 2005. The power of expert opinion in ecological models using bayesian methods: impact of grazing on birds. Ecological Applications 15 (1), 266–280.

McPherson, J. M., Jetz, W., Rogers, D. J., 2006. Using coarse-grained occurrence data to predict species distributions at finer spatial resolutions–possibilities and limitations. Ecological Modelling 192 (3-4), 499 – 522.

Merler, S., Furlanello, C., Chemini, C., Nicolini, G., 1996. Classification tree methods for analysis of mesoscale distribution of ixodex ricinus (acari: ixodidae) in trentino, italian alps. Journal of Medical Entomology 33 (6), 888–893.
URL `http://mpa.itc.it/papers/merler1996classifcation.pdf`

Michel, N., Burel, F., Legendre, P., Butet, A., 2007. Role of habitat and landscape in structuring small mammal assemblages in hedgerow networks of contrasted farming landscapes in brittany, France. Landscape Ecology 22, 1241–1253.

Milner-Gulland, E., Torgerson, P., Shaikenov, B., Morgan, E., 2004. The tapeworm *Echinococcus multilocularis* in kazakhstan: transmission dynamics in a patchy environment. In : Akcakaya, H., Burgman, M., Kindvall, O., Wood, C., Sjogren-Gulve, P., Hatfield, J.S. ans McCarthy, M. (Eds.), Species conservation and management.Case studies. Oxford Univ Press, Oxford, pp. 179–189.

Miterpakova, M., Dubinsky, P., Reiterova, K., Machkova, N., Varady, M., Snabel, V., 2003. Spatial and temporal analysis of the *Echinococcus multilocularis* occurrence in the Slovak republic. Helminthologia 40 (4), 217–226.

Mjolsness, E., DeCoste, D., 2001. Machine Learning for Science: State of the Art and Future Prospects. Science 293 (5537), 2051–2055.

Moisen, G. G., Frescino, T. S., 2002. Comparing five modelling techniques for predicting forest characteristics. Ecological Modelling 157 (2-3), 209–225.

Molinaro, A. M., Simon, R., Pfeiffer, R. M., 2005. Prediction error estimation: a comparison of resampling methods. Bioinformatics 21 (15), 3301–3307.

Morales, J. M., Haydon, D. T., Frair, J., Holsinger, K. E., Fryxell, J. M., 2004. Extracting more out of relocation data: building movement models as mixtures of random walks. Ecology 85 (9), 2436–2445.

Morey, D., 1994. The early evolution of the domestic dog. Scientific American, 336–347.

Morgan, E., Milner-Gulland, E., Torgerson, P., Medley, G., 2004. Ruminating on complexity: macroparasites of wildlife and livestock. TRENDS in Ecology and Evolution 19 (4), 181–188.

Murphy, H. T., Lovett-Doust, J., 2007. Accounting for regional niche variation in habitat suitability models. Oikos 116, 99–110(12).

Murray, J., Goldizen, A., O'Leary, R., McAlpine, C., Possingham, H., S., L. C., 2009. How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise ? a case study using brush-tailed rock-wallabies *Petrogale penicillata*. Journal of Applied Ecology 46, 842–851.

Okubo, A., Levin, S., 2001. Diffusion and ecological problems: modern perspectives. Springer Verlag GMBH.

Olden, J., 2003. Species-specific approach to modelling biological communities and its potential for conservation. Conservation Biology 17, 854–863.

Olden, J., Joy, M., Death, R., 2006. Rediscovering the species in community-wide predictive modelling. Ecological Applications 16 (4), 1449–1460.

Oliver, I., 2002. An expert panel-based approach to the assessment of vegetation condition within the context of biodiversity conservation: stage 1: the identification of condition indicators. Ecological Indicators 2 (3), 223 – 237.

Ostfeld, R., Glass, G., Keesing, F., 2005. Spatial epidemiology: an emerging (or re-emerging) discipline. TRENDS in Ecology and Evolution 20 (6), 328–336.

Ostfeld, R. S., Keesing, F., 2000. Biodiversity and disease risk: the case of lyme disease. Conservation Biology 14, 722–728(7).

Panteleyev, P., 1998. The rodents of the Palaearctic fauna: composition and areas. Moscow Institute of Ecology and Evolution of Russian Academy of Science.

Patz, J. A., Daszak, P., Tabor, G. M., Aguirre, A., Pearl, M., Epstein, J., Wolfe, N., Kilpatrick, A., Foufopoulos, J., Molyneux, D., Bradley, D., members of the Working Group on Land Use Change, Emergence., D., 2004. Unhealthy landscapes: policy recommendations on land use change and infectious disease emergence. Environmental Health Perspectives 112, 1092–1098.

Pavlowski, E., 1964. Natural foci of transmission diseases in connexion with the landscape epidemiology of zooanthropozoonses. Nauka, Moscow-Leningrad, 213 p.

Pearce, J., Cherry, K., Drielsma, M., Ferrier, S., Whish, G., 2001. Incorporating expert opinion and fine-scale vegetation mapping into statistical models of faunal distribution. Journal of Applied Ecology 38, 412–424.

Peterson, A., 2006. Ecological niche modelling and spatial patterns of disease transmission. Emerging Infectious Diseases 12 (12), 1822–1826.

Pffeifer, D., Hugh-Jones, M., 2002. Geographical information systems as a tool in epidemiological assessment and wildlife disease management. Rev sci tech Off int Epiz 21 (1), 91–102.

Pleydell, D., Raoul, F., Vaniscotte, A., Craig, P., Giraudoux, P., 2006. Micromammals and Macroparasites. Springer Japan, Ch. 25, Towards understanding the impacts of environmental variation on Echinococcus multilocularis transmission, pp. 545–564.

Pleydell, D. R. J., Raoul, F., Tourneux, F., Danson, F. M., Graham, A. J., Craig, P. S., Giraudoux, P., 2004. Modelling the spatial distribution of *Echinococcus multilocularis* infection in foxes. Acta Tropica 91 (3), 253 – 265.

Pleydell, D. R. J., Yang, Y. R., Danson, F. M., Raoul, F., Craig, P. S., McManus, D. P., Vuitton, D. A., Wang, Q., Giraudoux, P., 09 2008. Landscape composition and spatial prediction of alveolar echinococcosis in southern ningxia, China. PLoS Neglected Tropical Diseases 2 (9), e287.

Pulliam, H., 2000. On the relationship betwen niche and dsitribution. Ecology Letters 3, 349–361.

Qiu, J., Chen, X., Ren, M., Luo, C., Liu, D., Liu, X., 1995. Epidemiological study on alveolar hydatid disease in Qinghai- xizang (tibetan) plateau. Journal of Practical Parasitic Diseases 3, 106–109.

Qiu, J., Liu, F., Schantz, P., Ito, A., Carol, D., He, J., 1999. Epidemiological survey of hydatidosis in Tibetan areas of Western Sichuan Province. Archivos Internacionales de la Hidatidosis, 23–84.

R, 2005. Development core team, R: a language and environment for statistical computing. R foundation for statistical computing, vienna, austria. isbn 3-900051-07-0. online.
URL `http ://www.R-project.org`

Raoul, F., Defaut, R., Michelat, D., Montadert, M., Pepin, D., Quéré, J., Tissot, B., Delattre, P., Giraudoux, P., 2001a. Landscape effects on the population dynamics of small mammal communities : a preliminary analysis of prey-resources variations. Revue d'Ecologie, La Terre et la Vie 56, 339–352.

Raoul, F., Deplazes, P., Nonakac, N., Piarroux, R., Vuitton, D., Giraudoux, P., 2001b. Assessment of the epidemiological status of echinococcus multilocularis in foxes in France using elisa coprotests on fox faeces collected in the field. International Journal of Parasitology 31 (3-4), 1579–1588.

Raoul, F., Michelat, D., Ordinaire, M., Décoté, Y., Aubert, M., Delattre, P., Deplazes, P., Giraudoux, P., 2003. *Echinococcus multilocularis*: secondary poisoning of fox population during a vole outbreak reduces environmental contamination in a high endemicity area. International Journal for Parasitology 33 (9), 945 – 954.

Raoul, F., Quéré, J., Pleydell, D., Vaniscotte, A., Giraudoux, P., 2008. Small mammals assemblage response to deforestation and afforestation in Central China: a multinomial based modelling approach. Mammalia 72 (4), 320–332.

Raoul, F., Quéré, J., Rieffel, D., Bernard, N., Takahashi, K., Scheifler, R., Ito, A., Wang, Q., Qiu, J Yang, W., Craig, P., Giraudoux, P., 2006. Distribution of small mammals in a pastoral landscape of the tibetan plateau (western sichuan, China) and relationship with grazing practices. Mammalia 3-4, 214–225.

Rausch, R. L., 1995. Life-cycle patterns and geographic distribution of echinococcus species. In : Thompson, R. C. A., Lymbery, A. J. (Eds.), Echinococcus and Hydatid disease. Wallingford, Cab International, pp. 89–119.

Rausch, R. L., Fay, F., 2002. Epidemiology of alveolar echinococcosis, with reference to St. Lawrence Island, bering sea. In : Craig, P., Pawlowski, Z. (Eds.), Cestode zoononoses: echinococcosis and cysticercosis, an emergent and global problem. IOS Press, Amsterdam. The Netherlands., pp. 309–325.

Ricklefs, R., 2004. A comprehensive framework for global patterns in biodiversity. Ecology Letters 7, 1–15(15).

Ricklefs, R. E., Miller, G. L., 1999. Ecology. WH Freeman, New-York.

Rioux, J., Decamps, H., Lanotte, J., Combes, C., Théron, A., Pointier, J., Seytor, S., Delattre, P., Bougerol, C., 1977. Ecologie de la schitosomose en guadeloupe. Analyse du système épidémiologique. Documents pour un essais de modélisation. Revue d'Epidémiologie et Santé Publique 25, 483–519.

Robardet, E., Giraudoux, P., Caillot, C., Boue, F., Cliquet, F., Augot, D., Barrat, J., 2008. Infection of foxes by *Echinococcus multilocularis* in urban and suburban areas of Nancy, France: influence of feeding habits and environment. Parasite 15, 77–85.

Roberts, M., Aubert, M., 1995. A model for the control of *Echinococcus multilocularis* in France. Veterinary Parasitology 56, 67–74(8).

Rodriguez, J., Hortal, J., Nieto, M., 2006. An evaluation of the influence of environment and biogeography on community structure: the case of holarctic mammals. Journal of Biogeography 33, 291–303(13).

Romig, T., 2002. Spread of *Echinococcus multilocularis* in europe ? In : Craig, P., Pawlowski, Z. (Eds.), Cestode zoonoses: echinococcosis and cysticercosis. Amsterdam, IOS Press, pp. 65–80.

Romig, T., Bilger, B., Merli, B., Dinkel, A., Lucius, R., Mackenstedt, U., 1999. Bekampfung von *Echinococcus multilocularis* in einem hochendemiegebiet suddeutschlands. in: neuere methoden und ergebnisse zur epidemiologie von parasitosen. Tagung der Fachgruppe Parasitologie und parasitaÂšre Krankheiten, 172–183.

Romig, T., Thoma, D., Weible, A.-K., 2006. *Echinococcus multilocularis* ? a zoonosis of anthropogenic environments ? Journal of Helminthology 80 (02), 207–212.

Schantz, P., 1998. A survey of hydatid disease (echinococcosis) in tibetan populations in China: preliminary results. In : Proceedings of The First National Hydatidology Conference of

Imaging and Therapeutics, and International Symposium on Transmission and Control of Echinococcus Infection. Urumqi,Xinjiang Medical University., p. 152.

Schantz, P. M., Chai, J., Craig, P. S., Eckert, J., Jenkins, D. J., MacPherson, C. N. L., Thakur, A., 1995. Epidemiology and control of hydatid disease. In : Thompson, R. C. A., Lymbery, A. J. (Eds.), Echinococcus and Hydatid Disease. CAB International, UK., pp. 233–331.

Schantz, P. M., Wand, H., Qiu, J., Liu, F. J., Saito, E., Emshoff, A., Ito, A., Roberts, J. M., Delker, C., 2003. Echinococcosis on the tibetan plateau: prevalence and risk factors for cystic and alveolar echinococcosis in tibetan populations in Qinghai province, China. Parasitology 127, S109–S120.

Seoane, J., Bustamante, J., Diaz-Delgado, R., 2005. Effect of expert opinion on the predictive ability of environmental models of bird distribution. Conservation Biology 19, 512–522(11).

Smith, A., Xie, Y., 2008. A Guide to the Mammals of CHina. Princeton University Press, Princeton, NJ.

Soberon, J., Peterson, A. T., 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics.

Sorre, M., 1933. Complexes pathogènes et geógraphie medicale. Annales de Geógraphie 42 (235), 1–18.

Staubach, C., Thulke, H., Tackmann, K., Hugh-Jones, M., Conraths, F., 2001. Geographic information system-aided analysis of factors associated with the spatial distribution of *Echinococcus multilocularis* infections of foxes. American Journal of Tropical Medicine and Hygiene 65 (6), 943–948.

Stehr-Green, J. K., Stehr-Green, P. A., Schantz, P. M., Wilson, J. F., Lanier, A., 1988. Risk factors for infection with *Echinococcus multilocularis* in alaska. American Journal of Tropical Medicine and Hygiene 38, 380–385.

Steyerberg, E., Harell, J. F., Borsboom, G., Eijkemans, M., Vergouwe, Y., Habbema, J., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology 54 (8), 774–781.

Stieger, C., Hegglin, D., Schwarzenbach, G., Mathis, A., Deplazes, P., 2002. Spatial and temporal aspects of urban transmission of *Echinococcus multilocularis*. Parasitology 124 (06), 631–640.

Suzan, G., Giermakowski, J. T., Marce, E., Suzan-azpiri, H., Armien, B., Yates, T., 2006. Modelling Hantavirus reservoir species dominance in high seroprevalence areas on the Azuero peninsula of Panama. American Journal of Tropical Medicine and Hygiene 74 (6), 1103–1110.

Takumi, K., Van Der Giessen, J., 2005. Transmission dynamics of *Echinococcus multilocularis* ; its reproduction number, persistence in an area of low rodent prevalence, and effectiveness of control. Parasitology 131 (01), 133–140.

Taylor, L., Latham, S., Woolhouse, E., 2001. Risk factors for human disease emergence. Philosophical Transaction of Royal Society 356, 983–989.

Ter Braak, C., Hoijtink, H., Akkermans, W., Verdonschot, P., 2003. Bayesian model-based cluster analysis for predicting macrofaunal communities. Ecological Modelling 160 (14), 235–248.

Termansen, M., McClean, C., Preston, C., 2006. The use of genetic algorithms and bayesian classification to model species distributions. Ecological modelling 192, 410–424.

Thoma, D. A., 2005. Untersuchungen zum urbanen ubertragungszyklus von *Echinococcus multilocularis*. Ph.D. thesis, University of Stuttgart-Hohenheim.

Tiaoying, L., Jiamin, Q., Wen, Y., Craig, P., Xingwang, C., X., N., Ito, A., Giraudoux, P., Wulamu, M., Wen, Y., Schantz, P., 2005. Echinococcosis in Tibetan populations, Western Sichuan province, China. Emerging Infectious Diseases 11 (12), 1866–1873.

Tolle, F., 2005. Paysage et risque sanitaire. le cas de l'échinococcose alvéolaire, approche multiscalaire. Ph.D. thesis, Université de Franche-Comté.

Tolle, F., Pleydell, D., Combes, B., Cliquet, F., Piarroux, M., Giraudoux, P., Tourneux, F., 2005. Identification of environmental risk factors for the presence of *Echinococcus multilocularis*. In : 14th European Colloquium on Theoretical and Quantitative Geography. Tomar, Portugal. Actes parus sur CD-Rom.

Torgerson, P., 2003. The use of mathematical models to simulate control options for echinococcosis. Acta Tropica 85, 211–221(11).

Turchin, P., 1998. Quantitative analysis of movements: Measuring and Modeling Population Redistribution in Animals and Plants. Sunderland, MA: Sinauer Associates.

Valentini, A., Miquel, C., Nawaz, M., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J., Taberlet, P., 2009. New perspectives in diet analysis based on dna barcoding and parallel pyrosequencing: the trnl approach. Molecular Ecology Resources 9 (1), 51–60.

Vaniscotte, A., Pleydell, D. R., Raoul, F., Quéré, J. P., Jiamin, Q., Wang, Q., Tiaoying, L., Bernard, N., Coeurdassier, M., Delattre, P., Takahashi, K., Weidmann, J.-C., Giraudoux, P., 2009. Modelling and spatial discrimination of small mammal assemblages: an example from western sichuan (China). Ecological Modelling 220 (9-10), 1218 – 1231.

Veit, P., Bilger, B., Schad, V., Sch ?fer, J., Frank, W., Lucius, R., 1995. Influence of environmental factors on the infectivity of *Echinococcus multilocularis* eggs. Parasitology 110 (01), 79–86.

Viel, J., Giraudoux, P., Abrial, V., Bresson-Hadni, S., 1999. Water vole (*Arvicola terrestris scherman*) density as risk factor for human alveolar echinococcosis. American Journal of Tropical Medicine and Hygiene 61 (4), 559–565.

Vuitton, D., Zhou, H., Bresson-hadni, S., Wang, Q., Piarroux, M., Raoul, F., Giraudoux, P., 2003. Epidemiology of alveolar echinococcosis with particular reference to China and Europe. Parasitology 127, 87–107.

Wang, H., Ma, S., Cao, D. P., Zhao, H. L., WU, H. Y., 1999. General review of echinococcosis in Qinghai. In : Urumqi, I. f. E. D. R., Chinese.), C. I. (Eds.), In Proceedings of the Workshop on Echinococcosis Control Strategy in CHina. pp. 7–11.

Wang, Q., Qiu, J., Schantz, P., He, J., Ito, A., Liu, F., 2001. Investigation of risk factors for development of human hydatidosis among households raising livestock in tibetan areas of western sichuan province. Chinese journal of parasitology & parasitic diseases 19 (2), 93–6.

Wang, Q., Raoul, F., Budke, C., Craig, P., Xiao, Y., Vuitton, D., Qiu, D., Pleydell, D., Giraudoux, P., 2009. Grass height and transmission ecology of *Echinococcus multilocularis* in tibetan communities, China. Chinese Journal of Medicine.

Wang, Q., Vuitton, D., Qiu, J., Giraudoux, P., Xiao, Y., Schantz, P., Raoul, F., Li, T., Wen, Y., Craig, P., 2004. Partial fencing as a possible risk factor for human alveolar echinococcosis in pastoral herdsmen communities of sichuan, China. Acta Tropica 90, 285–293.

Wang, Z., Wang, X., X., L., 2007. Selection of land cover by the Tibetan fox *Vulpes ferrilata* on the eastern Tibetan Plateau, western Sichuan Province, China. Acta Theriologica 52, 215–223.

Woolhouse, M., 2002. Population biology of emerging and re-emerging pathogens. Trends in Microbiology 10, 3–7(5).

Worton, B. J., 1989. Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70 (1), 164–168.

Xie, Y., MacKinnon, J., Li, D., 2004. Study on biogeographical divisions of China. Biodiversity and Conservation 13.

Yamada, K., Elith, J., McCarthy, M., Zerger, A., 2003. Eliciting and integrating expert knowledge for wildlife habitat modelling. Ecological Modelling 165 (2-3), 251 – 264.

Yang, Y., Sun, T., Li, Z., Li, X., Zhao, R., Cheng, L., 2005. Echinococcosis, ningxia, China. Emerging Infectious Disease 11, 1314.

Zhang, Y., Bart, J., Giraudoux, P., Craig, P., Vuitton, D., Wen, H., 2006. Morphological and molecular characteristics of *Echinococcus multilocularis* and *Echinococcus granulosus* mixed infection in a dog from xinjiang, China. Veterinary Parasitology 139, 244–248.

Zhang, Y., Jin, S., Quan, G., Li, S., Ye, Z., Wang, F., 1997. Distribution of mammalian species in CHina. Beijing CHina Forestry Publishing House.

Zhou, H., 2001. Epidemiologie et ecologie de la transmission d'echinococcus multilocularis au xinjang, republique populaire de chine: etude preliminaire. Ph.D. thesis, Besancon, France.

Zhou, H., Chai, S., Craig, P., Delattre, P., Quéré, J., Raoul, F., 2000. Epidemiology of alveolar echinococcosis in xinjiang uygur autonomous region, China: a preliminary analysis. Annals of Tropical Medicine and Parasitology 94, 715.

Zhu, M., Hastie, T., Walther, G., 2005. Constrained ordination analysis with flexible response functions. Ecological modelling 192 (4), 524–536.

Zollner, P. A., Lima, S. L., 1999. Search strategies for landscape-level interpatch movements. Ecology 80 (3), 1019–1030.

Résumé

Cette thèse porte sur l'écologie de la transmission d'*Echinococcus multilocularis*, un parasite (cestode) dont le stade larvaire, transmis des faeces de canidés (hôte dénitif) aux micro-mammifères (hôtes intermédiaires), peut engendrer accidentellement une zoonose mortelle chez l'homme : l'échinococcose alvéolaire. En Chine, où sont observées les plus fortes prévalences humaines au monde, des recherches multidisciplinaires ont mises en évidence des perturbations anthropiques des écosystèmes génératrices de risque de transmission du parasite. Nous nous sommes intéressé à quantifier, spatialiser voire prédire certains des facteurs écologiques qui influencent la transmission du parasite dans différents sites endémiques de la province du Sichuan (Chine), à savoir : i) les distributions spatiales des assemblages de micro-mammifères et ii) les comportements des chiens domestiques intervenant dans la transmission.
Suite à une définition quantitative et objective des assemblages de micro-mammifères à partir de données de piégeages collectées dans 2 sites d'études, nous avons modélisé les distributions spatiales de leurs habitats en fonction de facteurs environnementaux issus de données satellitales et en utilisant différentes techniques de modélisation. Les régressions non linéaires multiples (MARS) discriminaient le plus précisément les assemblages le long des gradients d'altitude, de pente, de la bande ETM 7 et d'indices de végétation (NDVI et EVI). Cependant, les modèles développés localement n'étaient pas transférables sur un jeu de données distant d'une centaine de km. Un modèle de classification entraîné sur les données des deux sites d'étude apparaît être une méthode plus appropriée pour prédire les distributions régionales des assemblages. Les comportements de défécation et l'utilisation de l'espace des chiens domestiques ont été étudiés dans 4 villages du plateau tibétain. L'estimation des prévalences fécales par PCR montre le rôle prédominant du chien sur le renard dans la contamination de l'environnement des hommes. Les faeces et la contamination étaient agrégées autour des habitations humaines (entre 0 à 200 m) où le risque de transmission serait élevé. L'analyse des trajectoires nocturnes montre que les chiens passent la majorité de leur temps autour des habitations et peuvent réaliser des excursions à l'extérieur des aires d'activité des populations de chiens des villages, correspondant aux habitats où les indices de présence de micro-mammifères sont les plus fréquents.
Notre analyse par des outils statistiques et d'écologie spatiale a permis l'estimation et la modélisation de paramètres écologiques utiles pour comprendre et prévenir le risque de transmission et incorporables dans des modèles épidémiologiques. Les erreurs de prédictions ainsi que les limites de nos conclusions encouragent la recherche d'une meilleure adéquation entre la collecte de données éco-épidémiologique et la diversité des outils d'analyse disponibles.

Mots clefs : écologie de la transmission, assemblages de micro-mammifères, modélisation des habitats, chien domestique, éco-éthologie.

Abstract

This thesis concerns the transmission ecology of the parasite cestode *Echinococcus multilocularis* which larval stage, transmitted from canid faeces to small mammals, can accidentally cause a fatal zoonosis in humans : the alveolar echinococcosis. In China, where the highest prevalences in the world have been observed, interdisciplinary researches have outlined that some anthropogenic disturbances of ecosystems can increase the transmission risk of the parasite. We attempted to quantify, spatialize and when possible predict some of the ecological factors influencing parasite transmission in the Sichuan province (China) : i) the spatial distributions of small mammal assemblages and ii) the domestic dog behaviors involved in the transmission.
Small mammal assemblages were defined from trapping data sets collected in two study sites situated on the Tibetan plateau spurs. The spatial distributions of their habitats were modelled as a function of some environmental factors extracted from satellite data and using several modelling techniques. Non linear multiple regressions (MARS) best discriminated assemblages along elevation, slope, ETM band 7 and vegetation indices (NDVI and EVI) gradients. However, predictions of the locally trained models were not transferable on a data set distant from one hundred km. Classification model trained on the whole regional data set was a more appropriate method to predict assemblage distribution within a regional extent. The defecation behaviors and space utilizations of domestic dogs were investigated in 4 villages of the Tibetan plateau. The estimation of fecal prevalences by PCR emphasized the dominant role of dogs in human environment contamination in comparison to foxes. Faeces and their contaminations were aggregated around habitations (from 0 to 200m) where the transmission risk might be particularly high. Moreover, analysis of nocturnal dog trajectories shown that dogs spend the majority of their time around their owners' houses and that they can travelled excursive paths outside the mean activity area of the village dog populations where small mammal presence indices were the most frequent.
Our analysis done with statistical and spatial ecology tools allows to estimate and model ecological parameters useful to understand and prevent the transmission risk and that are incorporable in epidemiological models. Prediction errors and limitations of our conclusions call for the research of a better adequacy between eco-epidemiological data collection and the diversity of tools available to analyze them.

Keywords:transmission ecology, small mammal assemblages, habitat distribution modelling, domestic dogs, eco-ethology.