Méthodes statistiques pour la modélisation des facteurs influençant la distribution et l'abondance de populations : Application aux rapaces diurnes nichant en France

Kévin Le Rest

THESE

Pour l'obtention du Grade de

DOCTEUR DE L'UNIVERSITE DE POITIERS
(UFR des sciences fondamentales et appliquées)
(Diplôme National - Arrêté du 7 août 2006)

Ecole Doctorale : Sciences pour l'Environnement Gay Lussac.

Secteur de Recherche : Biologie de l'environnement, des populations, écologie.

Présentée par :

Kévin Le Rest

****

Méthodes statistiques pour la modélisation des facteurs influençant la distribution et l'abondance de populations

Application aux rapaces diurnes nichant en France

****

Directeur de Thèse :

David Pinaud & Vincent Bretagnolle

****

Soutenue le 19 Décembre 2013

devant la Commission d'Examen

****

JURY

Rapporteur	**Liliane Bel**	Professeur, AgroParisTech, Paris
Rapporteur	**Frédéric Gosselin**	Ingénieur, IRSTEA, Nogent s/ Vernisson
Examinateur	**Stéphane Dray**	CR, UMR 5558, Lyon
Examinateur	**Didier Bouchon**	Professeur, UMR 7267, Poitiers
Directeur de thèse	**David Pinaud**	IR, UPR 1934, Chizé
Directeur de thèse	**Vincent Bretagnolle**	DR, UPR 1934, Chizé

RESUME

Face au déclin global de la biodiversité, de nombreux suivis de populations animales et végétales sont réalisés sur de grandes zones géographiques et durant une longue période afin de comprendre les facteurs déterminant la distribution, l'abondance et les tendances des populations. Ces suivis à larges échelles permettent de statuer quantitativement sur l'état des populations et de mettre en place des plans de gestion appropriés en accord avec les échelles biologiques. L'analyse statistique de ce type de données n'est cependant pas sans poser un certain nombre de problèmes. Classiquement, on utilise des modèles linéaires généralisés (GLM), formalisant les liens entre des variables supposées influentes (par exemple caractérisant l'environnement) et la variable d'intérêt (souvent la présence / absence de l'espèce ou des comptages). Il se pose alors un problème majeur qui concerne la manière de sélectionner ces variables influentes dans un contexte de données spatialisées. Cette thèse explore différentes solutions et propose une méthode facilement applicable, basée sur une validation croisée tenant compte des dépendances spatiales. La performance de la méthode est évaluée par des simulations et différents cas d'études dont des données de comptages présentant une variabilité plus forte qu'attendue (surdispersion). Un intérêt particulier est aussi porté aux méthodes de modélisation pour les données ayant un nombre de zéros plus important qu'attendu (inflation en zéro). La dernière partie de la thèse utilise ces enseignements méthodologiques pour modéliser la distribution, l'abondance et les tendances des rapaces diurnes en France.

ABSTRACT

In the context of global biodiversity loss, more and more surveys are done at a broad spatial extent and during a long time period, which is done in order to understand processes driving the distribution, the abundance and the trends of populations at the relevant biological scales. These studies allow then defining more precise conservation status for species and establish pertinent conservation measures. However, the statistical analysis of such datasets leads some concerns. Usually, generalized linear models (GLM) are used, trying to link the variable of interest (e.g. presence/absence or abundance) with some external variables suspected to influence it (e.g. climatic and habitat variables). The main unresolved concern is about the selection of these external variables from a spatial dataset. This thesis details several possibilities and proposes a widely usable method based on a cross-validation procedure accounting for spatial dependencies. The method is evaluated through simulations and applied on several case studies, including datasets with higher than expected variability (overdispersion). A focus is also done for methods accounting for an excess of zeros (zero-inflation). The last part of this manuscript applies these methodological developments for modelling the distribution, abundance and trend of raptors breeding in France.

REMERCIEMENTS

Je tiens à remercier tout particulièrement mes deux encadrants (David et Vincent) pour m'avoir offert l'opportunité de réaliser cette thèse dans un cadre fort agréable et d'avoir accompagné mon travail tout en me laissant libre dans mes choix. Je remercie également la LPO et les bénévoles qui ont récoltés les données, ainsi que la région Poitou-Charentes et le conseil général des Deux-Sèvres qui m'ont cofinancé.

Je remercie aussi tous les membres du laboratoire de Chizé, qui directement ou indirectement ont participés à la réalisation de ce travail. Merci en particulier aux cuistots (Christophe, Micheline et Arnaud) qui m'ont fait à mangé tous les midis pendant 3 ans ! Merci aussi aux membres de l'administration et à l'équipe de l'atelier pour leur aide quotidienne sur le bon fonctionnement du laboratoire. Merci beaucoup à l'équipe Agripop, à Isabelle, Luca et Nicolas pour les séminaires, à Bertrand pour l'aide en stats, à Sylvie pour les animations, et à tous les étudiants présents durant mes années passées à Chizé.

D'autre part je tiens aussi à remercier chaleureusement Joël et à Pascal pour leur aide tout au long de cette thèse mais aussi pour leur sincère sympathie. Merci aussi à Paul & à Boen pour avoir réussi à me supporter pendant tout ce temps dans le même bureau ; à Fabrice pour avoir supporté Bertrand chanter pendant si longtemps ; aux Chucks qui sont partis bien trop vite ; à Thibault et à Hervé pour les sorties 'naturalistes' du dimanche ; à Rocheteau pour les morilles et les coins de pêche ; aux DIVA pour leur placard ; aux deux Hélène pour leur gentillesse ; à Steeve pour ses mails incompréhensibles et à l'ensemble des personnes avec qui il m'a été possible d'échanger pendant ces trois années.

Enfin je ne peux pas finir ces remerciements sans penser à mes collocs, Mathieu, Jade, Aurélie, Damien et J.F. à quart-temps, puis à ma famille et à Manon. Un énorme merci à eux pour avoir été présents pendant ces années.

SOMMAIRE

AVANT PROPOS

La présence d'êtres vivants est sans conteste la plus curieuse des particularités de notre planète. Sans doute issue à l'origine d'une grosse soupe physico-chimique, la vie a donné lieu à une multitude d'entités utilisant des ressources pour croître et se multiplier. Mais la diminution de ces ressources ou la modification des conditions physico-chimiques environnantes peuvent l'affecter de manière substantielle. La vie est en effet très sensible aux perturbations de son environnement et plusieurs crises majeures ont déjà failli l'éradiquer. Ces grandes crises d'extinction semblent toujours dues à des modifications importantes de l'environnement. Les causes les plus probables bien que discutées, une importante glaciation, des éruptions volcaniques cataclysmiques ou l'impact d'un météore sur la Terre, ont toujours été d'ordre 'catastrophe naturelle'. Aujourd'hui de nombreux scientifiques s'accordent sur le fait que nous sommes aux prémices d'une sixième grande crise d'extinction. Cette fois, l'utilisation irrationelle des ressources par l'espèce humaine pourrait bien en être la cause principale. Depuis notre apparition sur Terre, nous avons en effet profondément modifié notre environnement et avons déjà provoqué de nombreuses catastrophes écologiques (marées noires, déforestation, pollution des cours d'eau). Un exemple bien connu est celui de l'île de Pâques où la disparition de plusieurs espèces endémiques est indéniablement due à l'espèce humaine. En effet même si ce n'est pas les pascuans qui ont épuisé leur île, comme cela a été soutenu par Jared Diamond en 2005 dans sa célèbre œuvre intitulée « Effondrement » ('*Collapse*'), ce sont tout de même des hommes qui, lors de la colonisation par les européens, l'ont détériorée de manière irréversible. Un autre exemple tout aussi frappant est l'éradication, en quelques dizaines d'années, de l'oiseau qui était sans doute l'un des plus abondants au monde, le Pigeon migrateur américain *Ectopistes migratorius*. Mais l'espèce humaine n'influe pas seulement de manière directe les êtres vivants, elle les influence aussi indirectement en modifiant leur environnement. En particulier, tous les changements globaux d'origine anthropique, tel que les changements d'usage des sols mais aussi les changements climatiques sont susceptibles de les affecter à long terme.

Quoi qu'il en soit, il en résulte un déclin rapide de la diversité du vivant dont il est nécessaire de se préoccuper rapidement. Heureusement, l'importance de la biodiversité, y compris pour l'espèce humaine, commence à préoccuper bon nombre de scientifiques, de politiques et d'économistes. En particulier, un nombre conséquent de moyens, dont des moyens financiers non négligeables[1], sont en train d'être mis en place pour enrayer ce déclin (parcs naturels, zones Natura 2000, mesures agro-environnementales, aires marines protégées, Trame verte et bleue, *et cetera*). Afin de mesurer la perte de biodiversité et de déterminer si ces moyens sont suffisants ou non pour la contrer, il est nécessaire de mettre en place des indicateurs de biodiversité. Sans ces indicateurs, statuer sur l'état réel de la biodiversité sur notre planète reste de l'ordre des présomptions, ce qui n'a que peu de poids lorsque des

[1] « Quand on parle pognon, à partir d'un certain chiffre, tout le monde écoute » Michel Audiard

décisions importantes doivent être prises à l'échelle d'un pays ou à l'international. Les indicateurs de biodiversité doivent être capables de donner des informations sur l'état des populations d'êtres vivants sur notre planète mais aussi de leurs tendances afin de détecter des signaux d'alarme ou au contraire des signaux positifs. Pour tenir compte à la fois de l'échelle à laquelle les organismes interagissent avec leur environnement et celle à laquelle sont prises les décisions concernant la protection de la biodiversité, il apparaît judicieux que le suivi des populations soit réalisé à large échelle géographique et ce sur une longue période. De plus en plus de programmes de suivi opèrent à cette échelle mais l'analyse des données qui en découlent n'est pas sans poser un certain nombre de difficultés. Ceci constitue précisément l'objet de cette thèse.

INTRODUCTION GENERALE

1) Pourquoi suivre les populations

Qu'est ce qu'une population ?

Une population peut se définir comme un ensemble d'individus d'une même espèce occupant une zone géographique commune et se reproduisant entre eux (voir Millstein 2010). Elle se caractérise par plusieurs paramètres tels que son effectif, son taux de croissance (augmentation, stabilité, diminution de l'effectif), sa densité (le nombre d'individus par unité de surface) et sa répartition (aire occupée par la population). Ces paramètres sont des éléments clés pour déterminer son état et en particulier sa viabilité (Shaffer 1981). Par exemple, une population constituée de peu d'individus a besoin d'une densité minimum pour se maintenir car il existe alors une relation positive entre son taux de croissance et sa densité (Allee 1931, 1938 ; Allee *et al.* 1949 ; Stephens *et al.* 1999 ; Courchamp *et al.* 1999, 2008 ; Stephens & Sutherland 1999). Pour sa conservation, il sera donc primordial de veiller à ce que sa densité ne descende pas en dessous d'un certain seuil. Les paramètres démographiques, qui sont la survie, la dispersion et la reproduction, vont quant à eux aider à comprendre les mécanismes qui régissent les différents paramètres de la population (voir Simberloff 1986 ; Robinson *et al.* 2004).

Suivre une population

Suivre une population consiste à déterminer l'évolution de l'ensemble ou d'une partie de ses paramètres au cours du temps (voir par exemple Baillie 1990). Les suivis de populations jouent donc un rôle crucial pour détecter à temps d'éventuels signaux d'alarme chez les espèces menacées et de prendre des décisions opportunes quant à leur conservation (Baillie 1990, 1991 ; Robinson *et al.* 2004 ; Greenwood *et al.* 2008). Mais ces suivis ne sont pas seulement intéressant pour les espèces menacées. Les espèces communes ont aussi un rôle primordial dans le fonctionnement des écosystèmes. Ainsi, un faible déclin de ces espèces peut résulter en une perte importante de biomasse, impactant de manière significative l'ensemble de l'écosystème (Gaston & Fuller 2008). Le fait de suivre les espèces communes permet en particulier de déterminer pourquoi elles sont abondantes aujourd'hui, ce qui pourra donner des informations essentielles pour leur conservation future. Même le suivi d'espèces en augmentation peut donner des informations précieuses. Par exemple, les espèces invasives peuvent entrer en compétition avec des espèces indigènes et donc expliquer les causes de leur déclin (Mooney & Cleland 2001 ; Gurevitch & Padilla 2004).

Implications pour le calcul des indicateurs de biodiversité

Le suivi de populations permet de statuer sur l'état de santé des espèces, de leur attribuer un statut de conservation (IUCN, www.iucnredlist.org), et sont donc souvent à la base du calcul des indicateurs globaux de biodiversité (voir Buckland *et al.* 2005 ; Pereira &

Cooper 2006 ; Butchart *et al.* 2010 ; Jones *et al.* 2011). L'ensemble des espèces mériterait d'être suivi, mais cela est bien sûr impossible à réaliser. On privilégiera donc certains groupes d'espèces, dites indicatrices, c'est-à-dire capables de donner des informations sur un plus grand nombre d'espèces (voir Kremen 1992 ; Caro & O'Doherty 1999 ; Carroll *et al.* 2001 ; Carignan & Villard 2002 ; Sergio *et al.* 2005). Pour le calcul d'indicateurs de biodiversité, il reste néanmoins important qu'un grand nombre de groupes taxonomiques différents soit étudié, ceci afin de rendre compte de l'évolution de la biodiversité à tous les niveaux (Andelman & Fagan 2000 ; Carignan & Villard 2002). Les organismes vivants étant très différents les uns des autres, il est nécessaire de mettre en place des types de suivi différents.

2) Comment suivre les populations

Méthode usuelle

Une des méthodes les plus utilisées pour suivre une population est la capture-marquage-recapture (CMR, Leslie & Chitty 1951 ; Leslie 1952 ; Leslie *et al.* 1953). Elle consiste à capturer des individus d'une population, leur attribuer un identifiant unique (une marque), les relâcher et les recapturer à nouveau au cours de plusieurs sessions de capture. La CMR permet d'estimer la taille de la population ainsi que ses paramètres démographiques (voir Lebreton *et al.* 1992 ; Conn *et al.* 2006) et est donc un outil avantageux pour suivre une population. Mais, pour que les résultats soient valides, il est nécessaire de poser plusieurs hypothèses qui sont rarement vérifiées en pratique. En particulier, il est nécessaire que la population soit close, c'est-à-dire qu'il n'y ait pas d'échanges avec l'extérieur de la zone étudiée, sinon les estimations peuvent être fortement biaisées (Kendall 1999). Des méthodes ont été développées pour utiliser la CMR sur des populations dites 'ouvertes' mais les résultats restent alors très sensibles à d'autres hypothèses comme le fait que la probabilité de capture et le taux de survie doivent être identiques entre individus (Jolly 1965 ; Seber 1965 mais voir aussi Pledger *et al.* 2010). Du fait de ces hypothèses sous-jacentes et de son coût non-négligeable dans sa mise en œuvre (capture, manipulation), la CMR est plutôt à réserver pour le suivi d'une zone de petite taille et où l'espèce d'intérêt est fidèle à la zone étudiée.

Le problème d'échelle

Les espèces (et c'est particulièrement le cas des espèces communes) occupent le plus souvent de larges aires géographiques, c'est-à-dire de l'ordre de plusieurs centaines voir milliers de kilomètres carrés. Elles font face à une grande variabilité des conditions rencontrées et donc à une grande variabilité dans les paramètres de la population. Elles fonctionnent alors plutôt sous forme d'une métapopulation, c'est-à-dire où plusieurs populations sont interconnectées (Levins 1969 ; Hanski & Gilpin 1991 ; Hanski *et al.* 1996). Afin de tirer des conclusions globales, et non pas seulement restreintes à un sous ensemble de la population, il est donc nécessaire d'adapter les suivis (Dickinson *et al.* 2010 ; Jones 2011). Ils doivent en particulier être réalisés à une échelle géographique qui soit large par rapport aux

capacités de dispersion de l'espèce étudiée, sinon les conclusions risquent d'être limitées pour sa conservation sur l'ensemble de son aire de répartition (Baillie *et al.* 2000). Par exemple, il arrive qu'une espèce soit présente dans une zone défavorable où son taux de croissance est inférieur à 1 (appelé puits, voir Pulliam 1988) ; conserver cette zone n'a alors que peu d'intérêt car la présence de l'espèce ne résulte en fait que de la colonisation par des individus provenant d'autres zones à proximité qui sont plus favorable, c'est-à-dire où le taux croissance est supérieur (appelé source).

Les décisions importantes prises pour la conservation de la biodiversité concernent elles aussi de larges zones géographiques, répondant ainsi aux menaces globales qui pèse sur notre planète (Convention on Biological Diversity 2010 ; Perrings *et al.* 2010, 2011). Dans le but de concilier l'échelle de fonctionnement des populations et l'échelle à laquelle les décisions sont prises, il semble inévitable d'effectuer les suivis sur de larges échelles géographiques (Jones 2011).

Les suivis à large échelle

Suivre une population à large échelle, comme par exemple à l'échelle d'un pays ou d'un continent, impose des contraintes non négligeables et nécessite potentiellement des moyens logistiques et /ou financiers importants (voir Dickinson *et al.* 2010 ; Jones 2011). Ainsi, dans la grande majorité des cas, il sera impossible de suivre l'ensemble de la zone et il faudra donc échantillonner seulement un sous-ensemble représentatif de celle-ci. Aussi plutôt que de suivre les individus sur la zone étudiée (comme c'est le cas pour la CMR), il sera préférable de suivre des sites et d'y quantifier les individus présents (voir Jhala *et al.* 2011 ; Jones 2011). On s'intéresse alors à la variation de densité d'individus dans l'espace et dans le temps, ce qui permettra dans un deuxième temps de donner des informations sur l'état de santé de la population comme son effectif, sa répartition et sa tendance. Dans ce cas, les paramètres démographiques sont la plupart du temps ignorés car difficiles à acquérir malgré qu'ils soient fondamentaux pour une compréhension approfondie des processus affectant la population (voir par exemple Gregory *et al.* 2004). Néanmoins, cette compréhension n'est pas toujours essentielle dans un premier temps, et ce particulièrement si l'on cherche simplement à détecter des signaux montrant que l'espèce est en déclin ou en augmentation (voir par exemple Fuller *et al.* 1995 ; IUCN, www.iucnredlist.org).

Il existe un grand nombre de techniques pour détecter les individus d'une espèce (voir Yates 1949 ; Seber 1982, 1986, 1992 ; Schwarz & Seber 1999 ; Royle & Nichols 2003 pour un aperçu). Elles doivent être suffisamment précises pour pouvoir donner des informations pertinentes sur la population (White 2001 ; Anderson 2001 ; Legg & Naggy 2006). La technique la plus élémentaire est le relevé de la présence/absence de l'espèce sur des surfaces de petite taille (carré, rectangle ou cercle), reflétant par exemple la taille d'un territoire chez les espèces territoriales. Cependant, bien que la présence/absence permet de dresser une carte de la distribution de l'espèce sur la zone étudiée, elle ne permet que rarement de déterminer une taille de population ou une tendance car l'abondance n'est pas toujours reliée au taux de

présence (Nielsen *et al.* 2005). Des méthodes permettent alors d'estimer l'abondance à partir de la présence/absence, par exemple en utilisant la relation entre l'abondance et le nombre de sites occupés (He & Gaston 2000 ; Holt *et al.* 2002), voir même leur agencement spatial (Conlisk *et al.* 2009). Mais ces méthodes utilisent des hypothèses fortes qui limitent leur validité (Conlisk *et al.* 2007 ; He & Gaston 2007). Ce type de relevé est donc plutôt à réserver lorsque la présence de plusieurs individus sur les unités de surface échantillonnées est anecdotique, soit parce que l'espèce est rare, soit parce que l'unité d'échantillonnage est petite (Joseph *et al.* 2006). Un relevé plus contraignant, mais aussi plus informatif, consiste à quantifier tous les individus présents sur les sites suivis, c'est-à-dire déterminer l'abondance. Avec une telle mesure, il sera cette fois possible d'estimer à la fois la répartition des individus, leur nombre et même leur tendance si le relevé est fait sur une longue période.

3) Tenir compte des contraintes liées au terrain

Délimiter une zone d'étude

Idéalement, la zone d'étude devrait couvrir l'ensemble de l'aire occupée par la population, voir même au delà si l'on veut détecter des changements dans son aire de distribution. Mais en pratique, cela est difficilement réalisable. En effet, il se pose alors des problèmes logistiques insolvables tels que l'impossibilité de faire des relevés dans des pays qui ne souhaitent pas participer au suivi. En dépit de la cohérence avec le fonctionnement de la population étudiée, les limites géographiques administratives, comme les frontières, sont souvent d'emblé un facteur limitant la taille de la zone d'étude. Nombre de suivis à large échelle se limitent donc à un pays (voir par exemple, *the breeding bird survey* - UK , www.bto.org/volunteer-surveys/bbs ; *the christmas bird count* - USA, birds.audubon.org/christmas-bird-count ; *the north american breeding bird survey* - USA, www.pwrc.usgs.gov/BBS ; l'observatoire rapaces - FR, observatoire-rapaces.lpo.fr ; le suivi temporel des oiseaux communs - FR, vigienature.mnhn.fr/page/le-suivi-temporel-des-oiseaux-communs-stoc ; *swiss ornithological institute monitoring* - CH, www.vogelwarte.ch/monitoring-en.html). L'avantage de faire des suivis à l'échelle d'un pays est sa cohérence avec l'échelle à laquelle les décisions pour la conservation seront prises (voir Jones 2011). Néanmoins, des projets existent pour étendre les suivis au-delà des frontières (*the biodiversity observation network*, www.earthobservations.org/geobon.shtml ; *the european biodiversity observation network*, Halada *et al.* 2009 ; Pereira *et al.* 2010) et il possible que dans les années à venir, des suivis internationaux, voir globaux, voient le jour.

La probabilité de détection

Un des problèmes majeurs des relevés de présence/absence et d'abondance est qu'ils nécessitent idéalement que tous les individus présents soient détectés sur les sites suivis[2] (voir Anderson 2001, 2003 ; Pearce & Ferrier 2001 ; Engeman 2003). Or, chez la plupart des

[2] au moins un individu dans le cas d'un relevé de présence/absence.

espèces, un certain nombre d'individus risquent de passer inaperçus, soit en raison de leur taille (microorganismes, insectes), soit en raison de leur comportement (évitement de l'homme, nocturne, sous-marine, sous-terraine). L'hypothèse d'exhaustivité n'est donc vraisemblablement jamais respectée, ce qui a d'ailleurs conduit à de fortes critiques et mises en garde vis-à-vis de ce genre de suivi (voir les échanges entre David R. Anderson et Richard M. Engeman, Anderson 2001, 2003 ; Engeman 2003). Si la probabilité de détection des individus est constante dans l'espace et dans le temps, la présence/absence ou l'abondance 'relative' à la détection sera tout de même une information importante puisqu'elle permettra de détecter les fluctuations de densité (du moins si la probabilité de détection n'est pas nulle). Mais en pratique, la probabilité de détection a très peu de chance d'être constante (Anderson 2001) et il est donc nécessaire de corriger la mesure. Pour ce faire une première stratégie consiste à effectuer un suivi intensif sur un sous ensemble des sites suivis. Ceci permettra de calibrer les autres données en utilisant le rapport entre l'abondance relative et l'abondance réelle en ces sites (Yates 1949 ; Cochran 1963 ; Kish 1965 ; Eberhardt & Simmons 1987). Une autre stratégie, plus aboutie, est d'adapter le type de relevé de manière à estimer en parallèle la probabilité de détection. Ainsi, si la détection est supposée variable entre observateurs, des protocoles de suivis à plusieurs observateurs sont préconisés (Cook & Jacobson 1979 ; Nichols *et al.* 2000 ; Alldredge *et al.* 2006). Si c'est plutôt la détection de l'espèce qui est problématique, on préférera la visite successive du même site (MacKenzie *et al.* 2002 ; Royle & Nichols 2003 ; Royle 2004). Pour ce dernier, il faut par contre faire l'hypothèse que les individus ne changent pas de lieu entre les visites, c'est-à-dire que la population soit close pendant toute la durée du suivi. Enfin, une autre méthode pour tenir compte de la probabilité de détection de l'espèce est l'échantillonnage par la distance (*distance sampling*, Burnham *et al.* 1980 ; Buckland *et al.* 1993, 2001, 2007). Il consiste à noter la distance entre l'observateur et les individus détectés, l'idée étant que la détection est de 100% à une distance nulle et décroît lorsque la distance d'observation augmente. Il est alors possible d'estimer une densité corrigée par la probabilité de détection. Cette mesure peut être réalisée sur des transects ou des points d'observations et nécessite aussi plusieurs hypothèses pour que les résultats soient valides, comme la bonne estimation des distances et la non-attraction-répulsion des individus vis-à-vis de l'observateur.

Compromis quantité/qualité des données, temps à allouer et financements

En pratique toutes les types de relevés ont des avantages et des inconvénients, qu'ils tiennent compte de la probabilité de détection ou non, et aucun n'est parfait (voir Seber 1986 ; Krebs 1999 ; Schwarz & Seber 1999 ; Buckland *et al.* 2000 ; Anderson 2001). Les relevés les plus fiables sont aussi ceux qui demandent le plus de temps pour être réalisés et ne sont donc pas toujours utilisables pour des suivis à large échelle (Jones 2011). Le choix de l'un d'entre eux résulte donc d'un compromis entre la quantité et la qualité des données, ce qui dépendra grandement des moyens humains et financiers disponibles. Pour les suivis à large échelle, il est nécessaire de mobiliser un nombre important d'observateurs, ce qui peut être réalisé à

moindre coût par la sollicitation d'un réseau de bénévoles (voir Dickinson *et al.* 2010 ; Levrel *et al.* 2010 ; Jiguet *et al.* 2011). Le nombre d'études scientifiques faisant appel à des bénévoles est littéralement en train d'exploser (voir Dickinson *et al.* 2010). Les bénévoles offrent des opportunités sans précédent pour suivre la biodiversité et permettent en particulier de suivre les populations à des échelles vraiment larges, ce qui est inenvisageable autrement (Devictor *et al* 2010 ; Dickinson *et al.* 2010). De plus, ces suivis sont à la fois un moyen d'investigation scientifique et de sensibilisation des citoyens (voir Cohn 2008). En contrepartie, les données peuvent être moins informatives car ce genre de suivi conduit à privilégier des relevés simples à réaliser et peu contraignants pour maximiser le taux de participation des bénévoles (Jones 2011). Il faut donc encore une fois faire un compromis, choisir un protocole suffisamment détaillé pour assurer la fiabilité des données, mais aussi suffisamment léger pour que les bénévoles veuillent bien participer au suivi (Dickinson *et al.* 2010 ; Jones 2011 ; Jiguet *et al.* 2011). Dans tous les cas, un nombre important d'observateurs génère aussi une plus forte hétérogénéité dans les données, par exemple en raison de capacités de détection différentes entre observateurs (Dickinson *et al.* 2010), et il sera nécessaire d'en tenir compte lors de l'analyse des données et de l'interprétation des résultats.

4) Modéliser les données

Pourquoi modéliser les données ?

Une fois le suivi réalisé et les données acquises, il reste à estimer les paramètres d'intérêt pour la population. Les sites échantillonnés vont permettre d'inférer ces paramètres mais l'utilisation brute des données permet rarement de les calculer de manière adéquate. En effet, même s'il est possible d'estimer un effectif simplement par une règle de 3 ([abondance totale relevée/surface total échantillonnée] * surface totale de la zone), rien ne permet de garantir la validité de cette estimation. Sa validité sera en particulier affectée dans le cas où les individus ne se répartissent pas de manière homogène, ce qui est très vraisemblablement toujours le cas dans la nature. L'objectif principal est de comprendre pourquoi et comment les individus se répartissent dans le temps et dans l'espace. Les données récoltées peuvent alors être confrontées à des données externes (variables explicatives) pour déterminer les facteurs influençant la distribution et l'abondance de la population étudiée. Par ailleurs, un autre objectif est souvent de prédire les paramètres d'intérêt sur l'ensemble de la zone étudiée. Par exemple, l'intérêt peut être de prédire des données d'abondance aux sites non-échantillonnés et ainsi d'obtenir une carte de distribution complète pour l'espèce étudiée. Les variables explicatives peuvent d'ailleurs elles aussi aider à la prédiction. Afin de répondre à ces objectifs, il est nécessaire de passer par une étape de modélisation des données, c'est-à-dire construire un modèle probabiliste adéquat[3] capable de rendre compte de l'information contenue dans les données ainsi que de sa précision.

[3] '*… all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind…*' (Box & Draper 1987)

Les modèles de distribution corrélatifs en écologie

Classiquement, on cherche à relier les relevés de présence/absence ou d'abondance à des variables explicatives qui caractérisent les sites échantillonnés. Il existe de nombreuses méthodes pour déterminer ces liens (voir Segurado & Araújo 2004 ; Franklin 2009 ; Elith & Leathwick 2009). Les modèles de régression (et assimilés) sont les plus utilisés et les plus faciles à interpréter (Elith & Leathwick 2009). Ils consistent à expliquer la présence/absence ou l'abondance de l'espèce étudiée comme une somme pondérée d'autres variables (explicatives). Par modèle de régression[4], j'inclus ici toutes les formes de modélisation basées sur cette idée, que ce soit les modèles linéaires (LM, Adrien-Marie Legendre en 1805 et Carl Friedrich Gauss en 1809, voir Seal 1967), les modèles linéaires généralisés (GLM, Nelder & Wedderburn 1972), les modèles additifs généralisés (GAM, Hastie & Tibshirani 1986), ou leurs dérivées comme les modèles de régression avec effet aléatoire spatialement structuré ou non (LMM, Eisenhart 1947 ; Laird & Ware 1982 ; GLMM, Breslow & Clayton 1993 ; GAMM, Lin & Zhang 1999). Ces modèles sont dits 'corrélatifs', ce qui les distingue des modèles 'mécanistes' (voir Kearney & Porter 2009). Ils utilisent la corrélation entre l'abondance et les variables explicatives considérées pour quantifier les liens existants entre ces deux composantes. Il est important de noter que cette corrélation n'implique pas forcément une relation de cause à effet directe et peut être liée à des effets indirects (voir Sinclair *et al.* 2010). Il convient alors de rappeler qu'il existe deux types de modélisation : un premier type visant à expliquer les processus qui influent sur la distribution des espèces (les modèles explicatifs) et un second visant à maximiser le pouvoir prédictif du modèle de manière à l'utiliser ensuite pour prédire aux sites non-échantillonnés (les modèles prédictifs) (voir Guisan & Zimmermann 2000 ; Elith & Leathwick 2009). Un bon modèle prédictif n'est pas forcément un bon modèle explicatif et *vice versa*. Dans le premier cas, l'objectif est de maximiser le pouvoir prédictif du modèle au dépend de la compréhension des processus alors que dans le second cas, c'est la compréhension du système qui prime. En pratique, la distinction entre ces deux extrêmes n'est pas aussi claire et on cherche plutôt un modèle présentant les deux caractéristiques, c'est-à-dire un modèle explicatif ayant un fort pouvoir prédictif. Il est alors nécessaire de faire apparaître clairement les hypothèses écologiques testées, d'utiliser des variables explicatives adéquates mais aussi d'interpréter avec précaution ces modèles (Bahn & McGill 2007 ; Jiménez-Valverde *et al.* 2008).

L'estimation des paramètres

Le modèle de régression va permettre de quantifier l'influence des variables explicatives utilisées et par la suite d'estimer les paramètres de la population. Plusieurs méthodes existent pour y parvenir et le choix de l'une d'entre elles va principalement dépendre du type de modèle utilisé. Considérons le modèle linéaire classique pour n données tel que $y = X\beta + \varepsilon$ où

[4] C'est à Francis Galton que l'on doit le nom de régression suite à ses travaux sur l'hérédité de la taille des graines chez les plantes et sur l'hérédité de la taille chez l'être humain (voir Galton 1886). Il constatera en particulier que la taille des enfants régresse vers la moyenne lorsque les parents sont grands.

y est le vecteur de la variable réponse (n x 1), X est la matrice des j variables explicatives (n x j), β est le vecteur des j paramètres inconnus (j x 1) et ε est le vecteur (n x 1) des erreurs de moyenne 0 et de variance σ^2 tel que $\varepsilon \sim N(0, I\,\sigma^2)$. On cherche à quantifier l'influence de chacune des j variables explicatives utilisées, c'est-à-dire à connaître β, mais aussi à connaitre σ^2, la variance résiduelle (non prise en compte par les variables explicatives). Dans le cadre du LM, la manière classique d'estimer β est de minimiser la somme des erreurs au carré $|| y - X\beta ||^2$, ce qui conduit à l'estimateur bien connu de β : $b = (X'X)^{-1}X'y$ puis à celui de σ^2 : $s^2 = || y - Xb ||^2 / (n - j)$. Pour les données de présence/absence ou d'abondance, l'hypothèse de normalité n'est plus respectée et le LM est donc peu recommandé. L'utilisation d'un GLM (Nelder & Wedderburn 1972) permet de palier ce problème en conservant la forme de base du LM, ce qui en facilite l'interprétation. En effet, le GLM peut s'écrire $y \sim EF(g^{-1}(X\beta), I\Phi)$ où EF est une distribution de la famille exponentielle (la loi normale pour le LM), g est la fonction de lien (l'identité pour le LM) et Φ est le paramètre de dispersion (σ^2 pour le LM). Ainsi, on retrouve l'écriture du LM, $y \sim N(X\beta, I\sigma^2)$ comme cas particulier du GLM.

L'estimation des paramètres d'un GLM se fait en général par maximum de vraisemblance, la méthode des moindres carrés classique, basée sur l'hypothèse de normalité, étant alors peu adaptée. La vraisemblance est une mesure caractérisant la probabilité d'obtenir les données observées avec le modèle considéré. Son calcul dépend donc de la distribution (EF) considérée. Pour des données discrètes, c'est le produit des probabilités d'obtenir une donnée. Pour des données continues, on parle de densité de probabilité et non de probabilité. Dans le cas du LM, l'utilisation du maximum de vraisemblance conduit au même estimateur de β (b) que par moindres carrés. Par contre l'estimateur de σ^2 (s^2) est biaisé car il vaut $|| y - Xb ||^2 / n$ au lieu de $|| y - Xb ||^2 / (n - j)$. Ce biais n'a pas de conséquence importante lorsque le nombre de données (n) est suffisamment grand par rapport au nombre de paramètres du modèle (j). Mais cela montre néanmoins que le maximum de vraisemblance n'est pas adapté pour l'estimation des composantes de variance car il ne prend pas en compte la perte du nombre de degrés de libertés occasionnée par l'estimation des j effets fixes (Harville 1977). Ainsi pour les modèles incorporant plusieurs composantes de variance, comme les modèles à effets aléatoires (modèles mixtes ou hiérarchiques), le maximum de vraisemblance nécessite des modifications. On utilisera alors d'autres formes de vraisemblance qui peuvent estimer correctement les composantes de variance, comme la vraisemblance restreinte pour les LMM (*restricted maximum likelihood*, REML, Patterson & Thompson 1971).

Pour les GLMM, en plus du problème qui se pose pour le calcul des composantes de variance, la vraisemblance elle-même devient difficile voir impossible à calculer. Il devient donc nécessaire d'utiliser des approximations. Les plus classiques sont la quasi-vraisemblance pénalisée (*penalized quasi-likelihood*, PQL, Green 1987 ; Schall 1991 ; Breslow & Clayton 1993 ; Wolfinger & O'Connell 1993), la quasi-vraisemblance marginale (*marginal quasi-likelihood*, MQL, Goldstein 1991 ; Breslow & Clayton 1993), l'approximation Laplacienne (Tierney & Kadane 1986 ; Pinheiro & Bates 1995 ; Raudenbush et al 2000) ou encore l'approximation par quadrature (adaptative gaussienne – *adaptive gaussian quadrature*, AGQ,

Davidian & Gallant 1992 ; Pinheiro & Bates 1995 ; Pinheiro & Chao 2006). Les chaînes de Markov (*Monte Carlo Markov Chain*, MCMC, voir Gilks *et al.* 1996), développées dans le cadre bayésien (Gelman *et al.* 2004) permettent aussi d'estimer ce genre de modèle en procédant à de multiples échantillonnages et offrent des moyens efficaces pour résoudre des problèmes complexes (Breslow & Clayton 1993 ; Clayton 1996 ; Browne & Draper 2006). Elles nécessitent cependant des temps de calcul souvent très longs avant de donner des résultats stables (Breslow & Clayton 1993). Enfin citons l'approximation de Laplace imbriquée et intégrée (*integrated nested laplace approximations*, INLA, Rue *et al.* 2009) qui permet d'estimer ce genre de modèle en des temps records. Plusieurs auteurs ont mis à disposition une palette d'outils précompilés pour faciliter son utilisation par des utilisateurs lambda (www.r-**inla**.org). Les GLMM ne sont donc pas des outils faciles à manipuler et il faut retenir que les estimations peuvent être affectées de manière (très) importante par les outils utilisés (voir Bolker *et al.* 2009 pour une synthèse).

5) Respecter les postulats statistiques

L'indépendance des résidus

La validité des résultats obtenus lors de l'utilisation d'un modèle de régression dépend de la validité des postulats sous-jacents au modèle. Un postulat crucial mais souvent non respecté en pratique est l'indépendance des résidus $e = y - g^{-1}(Xb)$. S'il n'est pas respecté, les estimations peuvent être affectées, en particulier l'incertitude des estimations peut être biaisée (voir Cochrane & Orcutt 1949). Un cas concret de dépendance résiduelle se produit lorsque plusieurs mesures sont répétées sur les mêmes entités, par exemple les mêmes individus ou les mêmes unités d'échantillonnage. Cela peut être vu comme une forme de pseudo-réplication où le nombre de données ne reflète pas la quantité réelle d'information indépendante (voir Hurlbert 1984). Le nombre de degrés de libertés est alors souvent surévalué, ce qui conduit à des estimations faussement précises. Il faut alors tenir compte du fait que les mesures ont été réalisées sur les mêmes entités en incluant un facteur renseignant l'identité de ces entités. Ce facteur devra être traité en effet aléatoire lorsque les entités échantillonnées ne sont qu'un échantillon aléatoire de l'ensemble des entités qu'il est possible d'échantillonner dans la population, d'où le terme 'effet aléatoire' (Eisenhart 1947). Les entités effectivement échantillonnées ne sont alors qu'une partie restreinte d'un ensemble plus grand d'entités possibles dans la population. Les effets aléatoires mesurent la variabilité dans la population et doivent donc être pris en compte pour corriger l'estimation des effets fixes, et en particulier leur précision.

Un cas plus compliqué se produit lorsque les résidus ne sont pas structurés par entité, comme c'est le cas lorsque plusieurs mesures sont réalisées sur les mêmes individus ou sur les mêmes unités géographiques, mais sont structurées selon une composante continue. Les exemples les plus communs sont des résidus structurés en fonction de l'espace et/ou du temps, on parle alors d'autocorrélation spatiale et/ou temporelle résiduelle (Griffith 1987). Le

problème que pose l'autocorrélation d'une variable a depuis longtemps était mis en évidence (voir Hooker 1905) et a conduit à un développement faramineux de la méthodologie à son égard au cours du siècle dernier (Student <Gosset> 1914 ; Yule 1921 ; Bartlett 1935 ; Wold 1938 ; Von Neumann *et al.* 1941 ; Cochrane & Orcutt 1949 ; Moran 1950 ; Durbin & Watson 1950, 1951, 1971 ; Whittle 1953 ; Anderson 1954 ; Geary 1954 ; Box & Pearce 1970 ; Cliff & Ord 1972 ; Cressie & Hawkins 1980a, b ; Anselin 1988 ; Haining 1990 ; Getis 1990 ; Cressie 1993 ; Diggle *et al.* 1998 ; Cressie & Huang 1999 ; Griffith 2000 ; Diggle 2003 ; Cressie & Wikle 2011). La structure des résidus est vraisemblablement liée à la non prise en compte de certaines variables qui sont elles-mêmes structurées dans l'espace et/ ou dans le temps. Mais cela peut être aussi dû à des paramètres démographiques de la population tels que la reproduction et la dispersion des individus (Sokal & Oden 1978b ; Legendre & Fortin 1989 ; Legendre 1993 ; Koenig 1999 ; Lennon 2000 ; Legendre *et al.* 2002 ; Lichstein *et al.* 2002 ; Guisan & Thuiller 2005 ; Dormann 2007 ; Beale *et al.* 2010). On peut identifier deux cas d'autocorrélation des résidus ; le premier où les résidus sont plus ressemblants entre eux qu'ils ne le devraient (autocorrélation positive) et le cas où ils sont plus dissemblant (autocorrélation négative) (voir Sokal & Oden 1978a ; Griffith 1987, 2003, 2006a). Le premier cas conduirait à sous-estimer l'incertitude alors que le second à la surestimer (Griffith 1987 ; Legendre 1993). La présence d'autocorrélation négative résiduelle peut se manifester par des phénomènes de compétition, d'exclusion ou de territorialité. Elle est néanmoins rarement mise en évidence et ne sera donc pas considérée par la suite (mais voir Griffith 2006b pour plus de détails). Lorsque l'on cherche à déterminer la présence/absence ou l'abondance d'une espèce, l'utilisation de modèles de régressions classiques conduit fréquemment (voir toujours) à la présence d'autocorrélation spatiale et/ou temporelle résiduelle. En effet, il est peu probable que toutes les variables autocorrélées dans l'espace et/ou dans le temps soient prises en compte dans le modèle (Barry & Elith 2006 ; Beale *et al.* 2010). Par contre, en pratique, tous les jeux de données disponibles ne permettent pas toujours de le détecter. En particulier, un échantillonnage suffisamment important de la zone étudiée est nécessaire (voir Legendre *et al.* 2002 ; Miller *et al.* 2007). Par leur nature, les suivis à large échelle spatiale et/ou temporelle permettent souvent de mieux mettre en valeur ces structures ; en tenir compte est alors d'autant plus nécessaire.

Pour tenir compte de ce problème, il existe un nombre important de méthodes, des plus simples consistant à corriger le nombre de degrés de liberté pour aboutir à une inférence correcte (Clifford *et al.* 1989 ; Dutilleul *et al.* 1993), aux plus compliquées consistant à ajouter un terme spatial et/ou temporel explicite au modèle de régression (voir Anselin 1988 ; Cressie 1993 ; Diggle *et al.* 1998 ; Diggle 2003 ; Dormann *et al.* 2007 ; Beale *et al.* 2010 ; Cressie & Wikle 2011 ; Saas & Gosselin 2014). On fera en général l'hypothèse que le processus générant l'autocorrélation résiduelle est stationnaire (voir Myers 1989), c'est-à-dire qu'il a les mêmes caractéristiques quelque soit sa position dans l'espace et/ou dans le temps. Parmi les modèles spatialement explicites, on trouve tout d'abord les modèles dits autorégressifs où la valeur de la variable réponse au voisinage de l'observation (dans l'espace

et /ou dans le temps) est aussi utilisée comme variable explicative (Haining 1990 ; Lichstein *et al.* 2002). Comme beaucoup d'autres méthodes tenant compte du problème d'autocorrélation spatiale et/ou temporelle, ce genre de modèle a d'abord été développé pour l'analyse des séries temporelles (Yule 1921 ; Aitken 1935 ; Champernowne 1948 ; Cochrane & Orcutt 1949 ; Durbin 1960) et a ensuite été adapté au domaine spatial (Whittle 1954 ; Besag 1974). Ce genre de modèle n'est par contre pas adapté pour analyser des données de comptages, car ils ne permettent alors pas de modéliser de l'autocorrélation positive (Besag 1974 ; Cressie 1993 ; Griffith 2002). Une autre stratégie consiste à utiliser un effet aléatoire structuré de manière à ce que la valeur que prend l'effet aléatoire en un point soit dépendante des autres valeurs de cet effet aux points proches de l'espace et/ou du temps (GLMM, Laird & Ware 1982 ; Ware 1985 ; Liang & Zeger 1986 ; Cressie & Wikle 2011 ; Saas & Gosselin 2014). Ce terme permet donc de renseigner les liens entre les résidus selon une fonction de la distance et/ou du temps. Enfin, une autre méthode consiste à décomposer l'espace et/ou le temps en un ensemble de vecteurs orthogonaux (filtres) décrivant la relation entre les points à différentes échelles spatiales et/ou temporelle (voir Griffith 2002). Cette fois les données de comptages peuvent être analysées avec un GLM classique, ce qui en fait une méthode attractive pour l'analyse des données de présence/absence ou d'abondance (voir Borcard & Legendre 2002 ; Griffith 2003 ; Dray *et al.* 2006 ; Griffith & Peres-Neto 2006 ; Thayn & Simanis 2013).

La surdispersion

Quel que soit le type de suivi de population adopté, les données récoltées apparaissent sous une forme de données de comptage, qu'elles ne prennent que deux valeurs possibles 0/1 dans le cas de la présence/absence ou n'importe quel nombre entier positif ou nul dans le cas de l'abondance. Les comptages sont des données discrètes et nécessitent donc d'être traités comme tel (voir Bishop *et al.* 1975 ; O'Hara & Kotze 2010). La régression binomiale (par exemple probit ou logistic, Bliss 1934 ; Berkson 1944, 1953, 1955 ; Finney 1947 ; Nerlove & Press 1973 ; Nelder & Wedderburn 1972 ; Cox & Snell 1989) et la régression de Poisson (Haight 1967 ; Nelder & Wedderburn 1972 ; Frome *et al.* 1973 ; Griffith & Haining 2006) sont les outils usuels pour analyser, respectivement, les données de présence/absence et les données de comptages. Mais ces modèles ont une caractéristique particulière puisqu'ils n'ont qu'un paramètre à estimer : la moyenne. La variance est quant à elle considérée comme étant une fonction de la moyenne. Dans le cas d'une distribution binomiale, $Var(y) = n * p * (1-p)$ où n est le nombre de répétition de l'expérience (souvent n = 1 dans le cas de données de présence/absence) et où p est la probabilité de succès (probabilité de présence dans le cas de données de présence/absence) ; pour une distribution de Poisson, $Var(y) = E(y)$ (voir Hinde & Demétrio 1998). Ces relations entre variance et moyenne trouvent tout leur sens dans un contexte théorique où l'ensemble des processus aboutissant à la présence/absence ou l'abondance de l'espèce est bien modélisé et sont mesurés sans erreurs (Fisher 1941[5] ;

[5] 'The Poisson Series arises when equal samples are taken from perfectly homogeneous material' (Fisher 1941)

William 1982, 1996 ; Hinde & Demétrio 1998 ; Boes 2010). Mais, comme nous l'avons vu précédemment, il est illusoire de penser que toutes les variables pouvant affecter ces processus puissent être intégrées dans le modèle, soit parce qu'elles sont inconnues, soit parce qu'elles sont impossibles à mesurer sur le terrain (William 1982, 1996 ; Hinde & Demétrio 1998). Par exemple, un nombre important d'observateurs différents risque de conduire à une plus forte hétérogénéité dans les données. Si la variable caractérisant la capacité de détection des différents observateurs n'est pas renseignée dans modèle de régression, cela conduit alors à une mauvaise spécification de la variance, la moyenne restant théoriquement peu affectée (détection moyenne des observateurs). Si la variance est supérieure à celle suggérée par le modèle, on parle de surdispersion (voir Cox 1983 ; Hinde & Demétrio 1998). Si au contraire la variance est inférieure, on parle de sous-dispersion. En pratique, comme pour l'autocorrélation résiduelle négative, la sous-dispersion se produit plus rarement en écologie et ne sera donc pas détaillé par la suite (mais voir Famoye 1993 ; Faddy & Bosch 2001 ; Ridout & Besbeas 2004 pour des solutions méthodologiques et New *et al.* 2011 ; Guillera-Arroita *et al.* 2012 pour des exemples d'application).

Il existe de nombreuses solutions pour tenir compte de la surdispersion (voir Hinde & Demétrio 1998 pour une vue d'ensemble). La plus simple est sans doute de modifier la relation entre moyenne et variance sans pour autant spécifier une nouvelle forme de distribution, ce qui aboutit à une quasi-vraisemblance (Wedderburn 1974). Il s'agit donc d'une extension du modèle de régression et non d'un nouveau modèle. Cette méthode se base sur le fait qu'en l'absence de surdispersion, la déviance résiduelle (Déviance_{RES}) devrait être du même ordre que le nombre de degrés de liberté résiduels (ddl_{RES}) (Hinde & Demétrio 1998). Une correction intuitive de la relation entre moyenne et variance est donc de considérer que $\text{Var}(y)' = \Phi\text{Var}(y)$ où $\text{Var}(y)'$ est la nouvelle variance du modèle et Φ est le paramètre de dispersion tel que $\Phi = \text{Déviance}_{RES} / \text{ddl}_{RES}$. La relation attendue entre ces deux quantités est ainsi retrouvée en absence de surdipsersion, c'est-à-dire lorsque $\Phi = 1$. Ce type de traitement considère la surdispersion fixe mais d'autres formes plus complexes peuvent être utilisées (voir Hinde & Demétrio 1998). Une façon plus aboutie de tenir compte de la surdispersion est d'utiliser un autre modèle de régression que ceux classiquement utilisés, c'est-à-dire utiliser cette fois une nouvelle forme de distribution. Les plus utilisées sont la régression beta-binomiale (Ishii & Hayakawa 1960 ; Chatfield & Goodhardt 1970 ; Crowder 1978) et la régression binomiale négative (*negative binomial*, NB, Fisher 1941 ; Chatfield & Goodhardt 1970 ; Lawless 1987). La NB se décline sous de nombreuses formulations dont une est particulièrement attractive puisqu'elle est dans la famille exponentielle et fait donc partie des GLMs (voir Hinde & Demetrio 1998). Enfin, une autre méthode pour corriger la surdispersion consiste à ajouter un effet aléatoire à l'échelle de l'observation, renseignant le fait qu'il existe une corrélation entre les observations dans les données qui n'est pas prise en compte dans le modèle (Hinde & Demetrio 1998). Considérer une loi normale sur cet effet conduit au modèle logistique-normal (Pierce & Sands 1975 ; Willimas 1982 ; Hinde & Demetrio 1998) et Poisson-normal (Hinde 1982). Il s'agit alors de GLMMs qui nécessitent

soit des temps de calcul très long pour être estimés, par exemple en utilisant l'algorithme EM (Dempster *et al.* 1977) ou par MCMC (Gilks *et al.* 1996) soit l'utilisation d'approximations de la vraisemblance (voir la partie estimation des paramètres).

L'inflation en zéro : un cas particulier de surdispersion

Lors d'un comptage, les zéros ont souvent un statut particulier qui peut prêter à confusion (Ridout *et al.* 1998). Par exemple, si l'on s'intéresse au nombre de cigarettes qu'une personne quelconque fume pendant une journée, le zéro peut provenir d'un fumeur qui n'a pas fumé de la journée ou d'un non-fumeur strict. Dans le premier cas, l'information apportée par le zéro va influer sur le nombre moyen de cigarettes fumées par jour chez les fumeurs alors que dans le second cas elle donne une information sur la proportion de non-fumeurs. Il existe donc deux sous-populations qui sont mélangées, les fumeurs et les non-fumeurs, dont une qui ne génère que des zéros (les non-fumeurs). Ne pas tenir compte de l'existence de ces deux sous-populations conduit à un cas particulier de surdispersion, l'inflation en zéro (voir Lambert 1992 ; Mullahy 1997 ; Ridout *et al.* 1998 ; Tu 2002 ; Preisser *et al.* 2012). Ce phénomène a particulièrement été mis en évidence dans le cas de la régression de Poisson et a conduit au développement de plusieurs outils pour en tenir compte. Une manière intuitive de traiter ce problème est de séparer les zéros et les comptages strictement positifs puis d'utiliser deux modèles de régression distincts pour chacun des deux types de données. Un premier modèle va être utilisé pour traiter les données de présence/absence (où toutes les données strictement positives sont remplacées par 1) et un second va être utilisé pour les données de comptages strictement positives. Ce type d'approche est appelé modélisation en deux parties (*hurdle model*, Mullahy 1986, *two-part models*, Heilbron 1994). Le modèle ajusté sur les données ayant des valeurs strictement positives doit utiliser une distribution tronquée en zéro pour tenir compte du fait que la valeur zéro ne peut pas se produire (voir Mullahy 1986 ; Welsh *et al.* 1996). Ce genre de modèle a beaucoup été utilisé en écologie pour modéliser les données d'abondance car elle permet de distinguer de manière directe les processus influençant sur la présence/absence et ceux influant sur l'abondance des organismes (Welsh *et al.* 1996 ; Martin *et al.* 2005).

Une autre méthode pour traiter le problème d'inflation en zéro, est de considérer un mélange de deux modèles au lieu de les modéliser séparément. Cela donne lieu aux modèles dits zéro-enflés (abrégé ZIM par la suite) dont la version la plus commune est le modèle zéro-enflé de Poisson (*Poisson with zero*, WZ, Mullahy 1986 ; Heilbron 1994 ; *zero-inflated Poisson*, ZIP, Lambert 1992). Cette fois le comptage observé est considéré comme étant un mélange entre un modèle de régression de Poisson ayant pour moyenne λ et un modèle de régression de Bernoulli (Binomial avec m = 1) donnant la probabilité d'absence en excès π (voir Lambert 1992 ; Greene 1994 ; Welsh *et al.* 1996 pour plus de détails). On peut bien sûr aussi modéliser la probabilité de défaut de présence p tel que $p = 1 - \pi$ mais la plupart des logiciels sont paramétrés pour modéliser la probabilité d'absence en excès, sans doute en raison du fait que Diane Lambert, qui a apporté une forte contribution au développement du

ZIP, les définit ainsi (Lambert 1992). De plus cette manière de présenter le ZIP permet de bien souligner le fait que la seconde partie du modèle s'intéresse aux zéros qui sont en excès par rapport à la distribution de Poisson. Si le nombre de zéros observé est en adéquation avec celui attendu par la loi de Poisson, cette seconde partie du modèle s'annule en donnant une probabilité d'excès d'absence qui est nulle. Il faut préciser que les outils corrigeant la surdispersion, tiennent déjà compte d'un nombre de zéros attendus plus important qu'en utilisant une loi de Poisson (Ridout *et al.* 1998). Mais ce genre de modèle, tel que la NB, ne corrige pas spécifiquement l'excès de zéros et n'est donc pas le mieux adapté en présence d'inflation en zéro. Une combinaison des modèles tenant compte de la surdispersion et de ceux tenant compte de l'inflation en zéro est facilement utilisable, par exemple en remplaçant la distribution de Poisson par une NB, ce qui donne la distribution binomiale négative zéro-enflée (*zero-inflated negative binomial*, ZINB, Greene 1994) dans le cas des modèles à mélange.

6) Sélectionner les variables explicatives

Le nombre de variables capables d'influencer les processus biologiques est très grand mais le nombre de données disponibles est quant à lui toujours limité. Il va donc falloir choisir les variables qui vont être utilisées dans le modèle et ce choix va influencer l'estimation des paramètres du modèle. En particulier, dans un modèle de régression, plus le nombre de variables explicatives utilisées est grand, plus l'estimation de la variance de leurs effets l'est aussi (Burnham & Anderson 2002). D'un autre côté, n'utiliser qu'un nombre limité de variables peut conduire à des résultats biaisés (Sessions & Stevans 2006). Il faut donc trouver un compromis entre le nombre de variables à utiliser et la qualité d'ajustement du modèle aux données. On parle souvent de compromis biais/variance ou du principe de parcimonie (voir la figure ci-dessous extraite de Burnham & Anderson 2002).

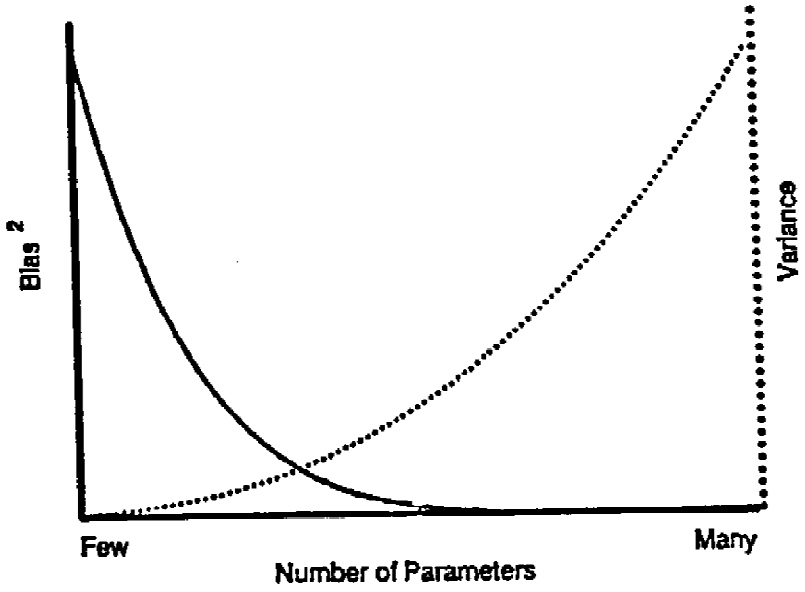


FIGURE 1.3. The *principle of parsimony*: the conceptual tradeoff between squared bias (solid line) and variance vs. the number of estimable parameters in the model (K). All model selection methods implicitly employ some notion of this tradeoff. The best approximating model need not occur exactly where the two curves intersect. Full truth or reality is not attainable with finite samples and usually lies well to the right of the region in which the best approximating model lies (the tradeoff region). Bias decreases and variance (uncertainty) increases as the number of parameters in a model increases.

Pré-sélection des variables

La première étape consiste à choisir des variables explicatives pertinentes, supposées influencer la présence/absence ou l'abondance de l'espèce étudiée. Il faut savoir que les individus d'une population ne se distribuent pas de manière aléatoire. Ils répondent à des contraintes qui peuvent leur être favorables ou défavorables (Hutchinson 1957 ; Guisan & Zimmermann 2000) et celles-ci se structurent à plusieurs échelles spatiotemporelles. Par exemple, chez les animaux, certaines contraintes vont déterminer la zone géographique que l'espèce peut occuper (aire de répartition), d'autres vont déterminer les domaines vitaux / territoires occupés par les individus, et encore d'autres vont déterminer les éléments occupés au sein de ces domaines vitaux / territoires (voir Johnson 1980). Il peut s'agir de contraintes abiotiques (non liées au vivant), comme la température ou la pluviométrie, qui déterminent la capacité d'une espèce à accomplir son cycle de vie (Pearson & Dawson 2003 ; Sekerciooglu *et al.* 2012). Mais il peut aussi s'agir de contraintes biotiques (liées au vivant) comme la présence d'une espèce proie ou la compétition intra- et interspécifique (Guisan & Thuiller 2005). Enfin, les paramètres démographiques vont aussi jouer un rôle majeur sur la distribution de la population (voir Pulliam 1988 ; Baillie *et al.* 2000).

Pour des raisons de facilité d'acquisition, les variables utilisées pour modéliser la distribution des individus sont la plupart du temps des variables décrivant le climat (Pearson & Dawson 2003 ; Sekerciooglu *et al.* 2012) et/ou l'habitat (Guisan & Zimmermann 2000). Néanmoins de plus en plus d'études soulignent la nécessité de tenir compte d'informations caractérisant mieux la population étudiée, comme ses paramètres démographiques, ses

contraintes physiologiques, les interactions biotiques intra- et interspécifiques (Austin 2002, 2007 ; Guisan & Thuiller 2005 ; McGill *et al.* 2007 ; Jiménez-Valverde *et al.* 2008 ; Kearney & Porter 2009 ; Buckley *et al.* 2010). Ces dernières sont néanmoins plus difficiles à mesurer et présentent un intérêt limité pour les modèles prédictifs. En effet, de telles données pourront difficilement être connues sur l'ensemble de la zone étudiée, ce qui limite leur utilisation pour prédire aux sites non-échantillonnés. Par exemple, la distribution d'un prédateur est sans doute très dépendante de la distribution de ses proies (voir la théorie sur distribution libre idéale, Fretwell & Lucas 1969 et celle sur la distribution despotique idéale, Fretwell 1972). Mais mesurer la variable 'quantité de proie disponible' sur l'ensemble de la zone d'étude nécessiterait d'autres suivis à part entière.

Le nombre de variables explicatives qu'il est possible de mesurer est aujourd'hui en pleine explosion. En particulier, les récents progrès en imagerie satellite permettent un accès libre à de nombreuses bases de données donnant des informations sur les habitats, comme sur température, l'ensoleillement, la pluviométrie, le type de couvert végétal ou encore des indices de production primaires (*Bioclim*, Hijman *et al.* 2005, www.worldclim.org ; *Corine Land Cover*, www.eea.europa.eu ; *MODIS vegetation indices*, Huete *et al.* 2002, lpdaac.usgs.gov). Ces informations sont particulièrement utiles pour les suivis à large échelle (Kerr & Ostrovsky 2003) car la mesure de ces données sur le terrain, en plus du suivi de l'espèce d'intérêt, demanderait un effort considérable (mais voir Sinclair *et al.* 2010 pour leurs limites).

Les critères de sélection

Pour sélectionner l'information la plus pertinente compte tenu des données disponibles, de nombreux critères statistiques ont été développés (voir Shao 1997 pour un aperçu). Ils sont pour la plupart basés sur le principe de parcimonie (ou Ockham's razor), qui veut que l'on ne doit pas complexifier les choses sans nécessité (voir Forster 2000). En écologie, le critère le plus utilisé est sans conteste l'*Akaïke information criterion* (AIC, Akaïke 1973, 1974) qui est en fait une approximation de la divergence de Kullback-Leibler (KLD, Kullback & Leibler 1951). Cette divergence caractérise la distance entre la vraie distribution des données (inconnue) et celle modélisée (voir Burnham & Anderson 2001 ; Richards 2008). Philosophiquement, l'AIC ne cherche pas à identifier un vrai modèle de dimension (ou complexité) finie et donc ne cherche pas à identifier le vrai ensemble de variables ayant générer les données. Il cherche plutôt un modèle de dimension raisonnable, parcimonieux, permettant de répondre à des questions à partir de l'ensemble de données disponibles (Stone 1979 ; Shibata 1981). Cette philosophie s'oppose à d'autres critères de sélection, qui au contraire supposent l'existence d'un vrai modèle de dimension finie et cherchent à l'identifier (voir Shao 1997). Ces derniers sont souvent dits consistants au sens où ils trouvent le vrai modèle lorsque $N \rightarrow +\infty$ si celui existe et est de dimension finie. Ils sont néanmoins inconsistants si le vrai modèle est de dimension infinie (Stone 1979). Le plus connu de ces critères est le *Bayesian information criterion* (BIC, Schwarz 1978), qui est

d'ailleurs souvent directement confronté à l'AIC alors qu'ils n'ont pas la même finalité. Les différences entre l'AIC et le BIC ont bien été résumées par Kenneth P. Burnham et David R. Anderson (Anderson & Burnham 1999 ; Burnham & Anderson 2004). En pratique, une des principales différences entre AIC et BIC est que le premier (l'AIC) considère que la dimension du vrai modèle peut être très grande et donc que la dimension du modèle à sélectionner devrait augmenter avec le nombre de données disponibles. Le second (le BIC) considère quant à lui qu'il existe un vrai modèle de dimension finie et donc que la dimension du modèle à sélectionner devrait être fixe quelque soit le nombre de données disponibles (Stone 1979 ; Anderson & Burnham 1999 ; Burnham & Anderson 2004). L'utilisation fréquente de l'AIC en écologie n'est pas sans lien avec ses propriétés. En effet, le nombre de variables capable d'influencer la présence/absence ou l'abondance d'une espèce est peut être bien infini (Burnham & Anderson 2002). L'AIC permet alors de choisir la complexité de modèle en accord avec le nombre de données disponibles.

L'AIC et le BIC se décomposent tout deux en une partie évaluation de l'ajustement aux données (par exemple la déviance : -2 * log-vraisemblance) et une partie « pénalité » proportionnelle à la complexité du modèle. Seule la pénalité change entre ces deux critères et vaut *2 * k* pour l'AIC alors qu'elle vaut *log(n) * k* pour le BIC où *k* est le nombre ne paramètres du modèle et *n* est le nombre de données disponibles (voir George 2000). De manière intuitive, on sent bien que l'AIC va avoir tendance à complexifier le modèle lorsque n augmente. La plupart des autres critères de sélection existants sont souvent soit équivalents à l'AIC comme le *Mallow's Cp* (Mallow 1973), le *leave-one-out* (LOO, Allen 1974 ; Stone 1974 ; Stone 1977), le *Sp* (Breiman & Freedman 1983), soit équivalents au BIC comme un cas particulier de *k-fold cross-validation* (k-fold CV, Geisser 1975 ; Shao 1997) ou un cas particulier du m*inimum description length* (MDL, Rissanen 1978 ; Hansen & Yu 1999, 2001). Il existe aussi des critères n'utilisant pas une pénalité fixe comme le *final prediction error* (FPE_{α}, Shibata 1984), la forme générale du MDL (Hansen & Yu 1999), la forme générale du k-fold CV (voir Shao 1993, 1997) ou le *generalized information criterion* ($GIC_{a(n)}$, Nishii 1984 ; Rao & Wu 1989). Ainsi ces derniers peuvent être vus comme des cas généraux de la plupart des autres critères de sélection. Par exemple l'AIC est un cas particulier du $GIC_{a(n)}$ si a(n) = 2 et le BIC en est un si a(n) = log(n) (voir Shao 1997).

Tenir compte de l'incertitude lors de la sélection des variables

Lors de la sélection des variables, il arrive fréquemment que plusieurs modèles candidats aient des performances similaires. Il est alors imprudent de dire que le modèle qui minimise le critère de sélection considéré est bien celui recherché. En effet, un autre modèle pourrait très bien être sélectionné en présence d'un autre échantillon de données (Burnham & Anderson 2002). Ce problème peut être relié au fait que le même échantillon est utilisé à la fois pour la sélection des variables et pour l'estimation (Miller 1984). Ne considérer que le meilleur modèle (celui minimisant le critère) pour l'inférence néglige l'incertitude présente lors de l'étape de sélection de variables, ce qui peut conduire à une inférence trop précise

(Miller 1984 ; Hoeting 1999) et donc à de fausses conclusions. Il est alors possible d'utiliser plusieurs modèles pour l'inférence et la prédiction, c'est le *model averaging* ou *multi-model inference*[6] (voir Hoeting 1999 ; Burnham & Anderson 2002 ; Richards 2008 ; Richards *et al.* 2011). Pour cela, il suffit d'attribuer des poids à chacun des modèles en compétition. Ainsi l'estimation des paramètres sera la moyenne pondérée des estimations des paramètres de chacun des modèles (voir Burnham & Anderson 2002 p.151-153). Notons néanmoins que l'avantage de ce type d'approche, par rapport au fait d'utiliser seulement le meilleur modèle, est controversé (voir Richards *et al.* 2011).

L'incertitude liée à l'étape de sélection des variables peut aussi être traitée en utilisant des méthodes de régularisation où l'inférence est basée sur une vraisemblance pénalisée par la valeur des paramètres β (voir Fu 1998). Dans ce cas, toutes les variables candidates sont utilisées au sein d'un même modèle, mais les coefficients sont réduits par la pénalité de sorte qu'ils prennent plutôt des valeurs proches de zéro (Ridge, Hoerl & Kennard 1970a, 1970b) voir exactement zéro (Bridge, Franck & Friedman 1993 ; Garotte, Breiman 1995 ; Lasso, Tibshirani 1996). Par simulation, Tibshirani a montré que la régression du type Ridge est la mieux adaptée en présence d'un grand nombre de variables ayant chacune un petit effet alors que celle du type Lasso est plus adaptée en présence d'un nombre modéré de variables ayant des effets plus ou moins forts. La sélection d'un sous-ensemble de variables (comme illustré dans la section précédente) reste la méthode la plus performante en présence d'un nombre faible de variables ayant des effets forts (Tibshirani 1996). Du fait de sa pénalité, le Lasso permet de combiner à la fois la sélection des variables et l'estimation des paramètres, ce qui en fait une méthode avantageuse. Une autre méthode de régularisation a récemment été proposée et combine les pénalités Ridge et Lasso : l'Elastic net[7] (Zou & Hastie 2005). Cette dernière est particulièrement intéressante lorsque le nombre de variables est supérieur au nombre de données et/ou lorsque les variables sont très colinéaires.

7) Objectifs de la thèse

Tenir compte de manière simultanée de ces problèmes

L'introduction qui précède montre à l'évidence que déterminer la distribution, l'abondance et les tendances d'une population n'est pas chose aisée, tant du point du vue de la récolte des données que de leur analyse. Il apparaît en particulier que l'étape de modélisation se heurte à de nombreux problèmes d'ordre statistique. Bien qu'il existe d'ores et déjà un large panel de méthodes disponibles pour résoudre ces problèmes, peu sont actuellement comprises par les écologues. Le **premier objectif** de cette thèse est d'explorer les possibilités offertes actuellement, afin de promouvoir leur utilisation auprès des écologues mais aussi de montrer

[6] Alan J. Miller avait déjà mentionné l'idée de considérer un nombre important d'ensemble de variables différents afin d'évaluer le biais lors de l'estimation du modèle final (voir Miller 1984).

[7] "*It is like a stretchable fishing net that retains 'all the big fish'.*" (Zou & Hastie 2005).

leurs limites. Par ailleurs, les différents problèmes statistiques rencontrés, que se soit la non-indépendance des résidus, la présence de surdispersion ou encore l'inflation en zéro, sont traités séparément. Pourtant ils sont la plupart du temps rencontrés conjointement lors de l'analyse des données et des liens d'ordre biologique et/ou statistique peuvent exister entre eux. Ainsi la présence d'autocorrélation dans les résidus génère de la surdispersion (Griffith 2006a ; Griffith & Haining 2006 ; Haining *et al.* 2009) mais corriger la surdispersion ne permet pas pour autant de corriger l'autocorrélation des résidus. Le **deuxième objectif** de cette thèse est de présenter des travaux combinant les différentes méthodes existantes au sein d'une même approche afin de tenir compte simultanément des différents problèmes rencontrés. Enfin, certaines lacunes méthodologiques méritent encore d'être comblées. C'est particulièrement le cas pour l'étape de sélection des variables explicatives en présence d'autocorrélation et de surdispersion. Le **troisième objectif** de thèse est de présenter une solution pour la sélection des variables explicatives dans un tel cas. Tout au long de ce manuscrit, un intérêt particulier sera porté à un jeu de données traitant de l'abondance des rapaces en France. Celui-ci a été choisi dans le but de rendre plus concrets les aspects méthodologiques discutés mais aussi pour participer à l'amélioration des connaissances sur ces espèces à fort intérêt patrimonial en France.

Plan de thèse

Le premier chapitre de cette thèse va se focaliser sur le problème d'autocorrélation spatiale et montrer les conséquences que cela peut avoir sur les estimations d'abondance. Il souligne par ailleurs l'importance de l'étape de sélection de variable dans le processus de modélisation et en particulier l'utilisation d'un critère de sélection qui prend en compte la présence d'autocorrélation spatiale résiduelle.

Le second chapitre va quant à lui s'intéresser plus spécifiquement à ce dernier point. Il propose en particulier la généralisation et l'évaluation de la méthode de sélection de variables utilisée dans le premier chapitre.

Le troisième chapitre s'intéresse à la généralisation de cette méthode de sélection de variables dans le cadre de données de comptage où le modèle de régression de Poisson présente de la surdispersion. La question sous-jacente est alors de savoir si des modifications sont nécessaires pour utiliser la méthode dans de tels cas de figure et si oui, quelles sont-elles ?

Un autre aspect important, ignoré dans les trois premiers chapitres, concerne la présence d'inflation en zéro et la manière de la traiter. C'est l'objet du quatrième chapitre qui propose d'aborder cette question par simulation en s'intéressant plus particulièrement aux performances des modèles à mélange. Cette fois l'intérêt concerne autant la sélection des variables que l'estimation des paramètres. Par souci de simplicité, le cas des résidus autocorrélés ne sera pas abordé dans ce quatrième chapitre. Son objectif est de donner des pistes sur quand et comment utiliser ces modèles zéro-enflés en écologie.

Enfin, le dernier chapitre propose l'application de ces méthodes à un jeu de données sur l'abondance des rapaces en France. L'idée est d'intégrer ces méthodes dans une approche plus générale de modélisation ne concernant pas seulement l'étape de sélection de variable ou le choix de modèle à utiliser mais allant jusqu'à décrire de manière précise la distribution, l'effectif et les tendances des populations de rapaces en France.

CHAPITRE 1 : INTRODUCTION AUX PROBLEMES LIES A L'AUTOCORRELATION SPATIALE ET IMPLICATIONS POUR LA CONSERVATION

Ce premier chapitre de thèse a pour but d'introduire le problème d'autocorrélation spatiale résiduelle et de montrer l'importance d'en tenir compte pour tirer des conclusions valides. L'article présenté ci-après a fait l'objet d'une publication dans le journal international *Ecological Informatics*.

Accounting for spatial autocorrelation from model selection to statistical inference: Application to a national survey of a diurnal raptor

Authors: Le Rest Kévin*, Pinaud David and Bretagnolle Vincent
Centre d'Etude Biologique de Chizé (CEBC), CNRS UPR 1934, 79360 Beauvoir-Sur-Niort, France
Email addressees: lerest.k@gmail.com, pinaud@cebc.cnrs.fr and breta@cebc.cnrs.fr
*Corresponding author:
Email: lerest.k@gmail.com
Tel: +33 (0)5 49 09 35 13
Fax: +33 (0)5 49 09 65 26

Ecological Informatics 14 (2013) p. 17–24

Overview

Abstract

Planning actions for species conservation involves working at both an ecologically meaningful spatial scale and a scale suitable for implementing management or conservation plans. Animal populations and conservation policies often operate across wide areas. Large-extent spatial datasets are thus often used, but their analyses rarely deal with problems inherent to spatial datasets such as residual spatial autocorrelation, which can bias or even reverse results. Here we propose a procedure for analysing a large-scale count dataset integrating residual spatial autocorrelation in a Generalized Linear Model framework by combining and extending previously published methods. The first step concerns the selection of the environmental variables by a modified cross-validation procedure allowing for residual spatial autocorrelation. Then the second step consists in evaluating the spatial effect of the model using a spatial filtering approach based on the variogram parameters. We apply this method to the Black Kite (*Milvus migrans*) to estimate the distribution and population size of this species in France. We found some divergence in estimated population size between spatial and non spatial models, as well as in the distribution map. We also found that the uncertainty of the model was underestimated by the residual spatial autocorrelation. Our analysis confirms previous results, that residual spatial autocorrelation should be always accounted for, especially in conservation where false results may lead to poor management decisions.

Keywords: GLM; Population size; Residual spatial autocorrelation; Spatial cross-validation; Spatial filtering; Species distribution.

Abbreviations

The following abbreviations are used in this paper:
AIC: **A**kaïke **I**nformation **C**riterion
GLM: **G**eneralized **L**inear **M**odel
PCA: **P**rincipal **C**omponent **A**nalysis
RMSEP: **R**oot **M**ean **S**quared **E**rror of **P**rediction
RSA: **R**esidual **S**patial **A**utocorrelation

Introduction

Animal populations and conservation policies often operate across wide areas. Large-extent spatial datasets (Scheiner *et al.* 2000) can therefore be extremely valuable to determine population parameters for conservation purposes, e.g. the geographical distribution of species, its population size or trends. However, the statistical analyses used often ignore issues that may bias conclusions. In particular, they rarely deal with inference problems inherent from spatial datasets such as residual spatial autocorrelation (hereafter RSA), which may actually reverse observed patterns (Kühn 2007).

Spatial autocorrelation arises when the measure of a variable of interest in multiple sample units are not independent of each other (Griffith 1987), which often occurs in ecological data. Such spatial patterns are usually explained by environmental features (e.g. climatic variables or habitat structure) that are themselves spatially structured. Therefore, including all environmental variables that are spatially structured may be sufficient to remove RSA of a regression model (Diniz-Filho *et al.* 2003). However, it is often impossible to measure all spatially structured variables: for instance, variables accounting for social behaviour or for the availability of food resources, are very difficult to measure and often missing in the dataset. In such cases, the inclusion of all available variables does not fully remove RSA and thus the important assumption of independence of residuals is violated (see Dormann *et al.* 2007). It is well known that this problem mostly affects the uncertainty of statistical models (Legendre 1993; Legendre *et al.* 2002), i.e. the confidence interval around the regression coefficients, which is commonly measured by the standard error. A positive RSA, i.e. closer locations having more similar residuals values than others, tends to underestimate the true standard errors of parameters, which leads to an over-precise estimation of the regression coefficients. In turn this can lead to an erroneously low p-value, wrong R^2 and wrong likelihood (Legendre 1993; Legendre *et al.* 2002; Lennon, 2000; Hoeting *et al.* 2006).

RSA raises two mains concerns. The first relates to model selection, since classical criterion such as Akaïke Information Criterion (hereafter AIC) are biased in presence of RSA (see Cassemiro *et al.* 2007 ; Diniz-Filho *et al.* 2008 ; Hoeting *et al.* 2006). The most common strategy to overcome this problem involves correcting first the RSA by considering a spatially explicit model and then, using a classical criterion such as AIC. However, accounting for RSA for all biologically pertinent candidate models can be extremely time consuming, especially if the number of candidate models is high (see Craig *et al.* 2007). As a consequence, AIC is often used without accounting for RSA (see for example Kühn *et al.* 2009). Kissling & Carl (2008) proposed several strategies to choose the spatial structure that should be added to the model in order to correct for RSA, but they did not provide solutions for the selection of variables. The second concern relates to the model estimation since model parameters are not estimated correctly (Dormann 2007; Kühn 2007; Keitt *et al.* 2002). To overcome this problem, some tools were made available for Generalized Linear Models

(hereafter GLMs) (see Dormann *et al.* 2007; Carl & Kühn 2010). Among these, the spatial filtering techniques are recognized as one of the most efficient, both practically and theoretically (Dormann *et al.* 2007; Diniz-Filho *et al.* 2009). Spatial filtering consists in using a weighted distance matrix to address the issue of RSA, by adding several spatial filters (eigenvectors) to a GLM (see Diniz-Filho & Beni. 2005; Dray *et al.* 2006; Getis & Griffith 2002; Griffith 2000). However, there is evidence that the choice of the weight matrix highly influences the set of spatial filters and thus the model (Patuelli *et al.* 2006). In addition, although there are several possibilities for defining the weight matrix (see Getis & Aldstadt 2004; Tiefelsdorf *et al.* 1999), it remains mainly based on basic functions of the distance (binary, linear, quadratic) which may not always satisfy the complexity of the residual spatial structure underlined in the ecological processes.

In this paper, our aim is to provide a guideline for analysing spatial datasets integrating RSA within a GLM framework, by extending different methods within the same framework. As a first step, we deal with model selection, by using a cross-validation approach. In order to overcome the problem of RSA in the selection step, we use a threshold distance between the training and the validation sets to ensure that they are fully independent. The second step consists in accounting for the RSA of the selected model. We use a spatial filtering technique, where the weighted matrix has been modified in order to directly use the shape of the variogram to calculate the eigenvectors. We then apply this approach on a real case study and compare results of the spatial and non spatial models. As a practical example, we used a French national dataset collated for the Black kite (*Milvus migrans*), a diurnal raptor. A particular emphasis was given to the estimation of species distribution and its population size, which are major issues in management and conservation plans.

1) Material and methods

1.1 Survey and datasets

A national survey aiming to estimate the distribution and population size of all diurnal raptors was undertaken between 2000 and 2002, with around 1,600 volunteers. For this study, we used a subset of the available data, consisting in 683 sampling units in France (see Figure 1) with known searching effort. Sampling protocol consisted in counting the number of breeding pairs of diurnal raptors on 25-km² quadrats (5 x 5 km; see Thiollay & Bretagnolle 2004 for details). The time spent on each quadrat was recorded by observers. Each quadrat was also described using environmental variables from a climatic dataset (Hijmans *et al.* 2005, Bioclim, www.worldclim.org/bioclim) and a land cover dataset (CLC: Corine Land Cover, www.eea.europa.eu). The climatic dataset consisted in 19 variables measured between 1960 and 1990, which provided robust estimates of measures such as average temperature, rainfall, temperature variation and rainfall variation at a resolution of approximately 1-km. The land cover dataset had 44 variables giving land use in 2000 on a 1-hectar cell. From these 44 classes, 9 habitat hyper-classes were built from a functional (ecological) point of view for

raptors (see Table A1 in Appendix A). The percentage of coverage per 25-km² quadrat was calculated for each of these habitat hyper-classes. High correlations occurred between several environmental variables, which cause matrix inversion problems (null determinant). In order to overcome multicollinearity, a Principal Component Analysis (hereafter PCA) was performed separately on each dataset (climate and land use) and principal components were used as environmental variables. The label "ClimDim.x" was used to nominate the x^{st} principal component from the climate dataset and the label "ClcDim.x" was used in the same way for the land cover dataset.

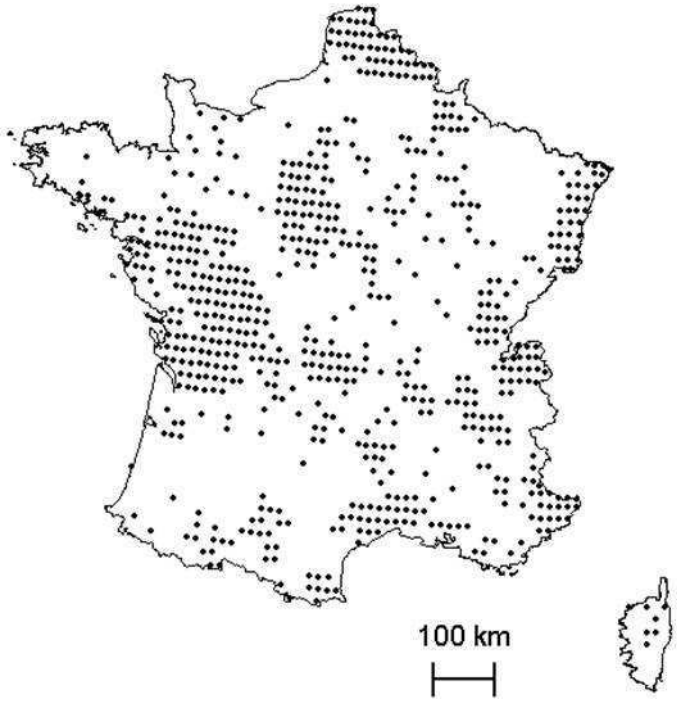


Figure 1: Map of the 683 locations (25-km² quadrats) used for analyses. Each location is represented by a black point.

1.2 Model selection by spatial cross-validation

Model selection consisted in a comparison of candidate models in order to select which predicted best the observed data. As the number of environmental variables k was high (19 climatic and 9 habitat variables), the number of candidate model became oversized (2^k). So a stepwise procedure was used to reduce computation time (Efroymson 1960; Hocking 1976). The stepwise process was implemented in two steps: first, environmental variables with linear effects were selected and then, quadratic terms and interactions. A Poisson distribution was assumed for the number of breeding pairs per quadrat, considering that there was no additional overdispersion, other than that due to spatial autocorrelation (see Griffith &

Haining 2006; Haining *et al.* 2009 for details about the relationship between overdispersion and RSA). The time spent per quadrat was included as an offset.

The error of prediction was considered as a selection criterion because the aim of this model was to predict at unsampled points. Error of prediction was calculated by cross-validation (Allen 1974; Geisser 1975; Stone 1974), a widely used technique for model selection and model validation involving many different splittings (see Arlot & Celisse 2010 for a recent overview of the cross validation procedures for model selection). Here, leave-one-out cross-validation was used, consisting in deleting one observation (the validation set) and use all the others as training dataset, i.e. to estimate model parameters. An overall prediction error can then be calculated using the Root Mean Square Error of Prediction (hereafter RMSEP), see Eq. 1:

$$RMSEP = \sqrt{\mathrm{E}\left[\sum_{i=1}^{n}[y_i - \hat{y}_i]^2\right]} \quad \text{(Eq. 1)}$$

In Eq. 1, n is the sample size (number of quadrats), y_i is the deleted observation (the validation set) and $\hat{y}_i$ is the predicted abundance at this location using parameters estimated from all the others (the training set). When using cross validation, a critical prerequisite is that the training set and the validation set must be independent, thus dependent model residuals may bias the error of prediction (Altman 1990). Several possible alternatives have been proposed to correct the cross-validation procedures (in context of nonparametric regression, see Chu & Marron, 1991; Burman *et al.* 1994). We extended the ”Modified Cross Validation” (Chu & Marron 1991) to a spatial context by using a threshold distance between the validation and the training dataset, guaranteeing these datasets to be spatially independent. This threshold was chosen as the value of the range of the variogram on the residuals from the model including all covariates (125 km in our case). This represented the spatial autocorrelation that could not be accounted for by our environmental variables (see Figure 2). Deviance residuals were chosen because Pearson residuals had some extreme values, which could affect the variogram quality (Cressie & Hawkins 1980a). The selected model was labeled non-spatial model because it did not incorporate an explicit spatial component, unlike the model below.

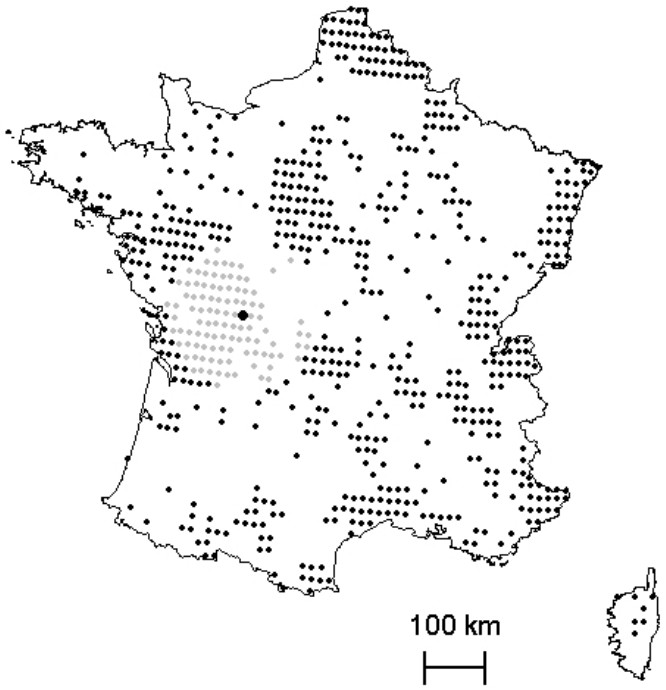


Figure 2: One example of the Modified Leave-One-Out Cross Validation applied in a spatial context with a threshold distance of 125 km. The heavy black point is the point which was left out of the model and where the error of prediction was calculated. Grey points were also excluded due to the residual spatial autocorrelation. Others black points were used in the training data set. This procedure was used iteratively for each observation in order to calculate an overall prediction error.

1.3 Accounting for residual spatial autocorrelation

The spatial structure of residuals can be easily evaluated using a correlogram or a variogram, both based on a measure of the covariance between observations according to the distance between them. A variogram was estimated using the residuals of the previously selected model, i.e. the non-spatial model (see Figure 3a). We then used an approach based on spatial filters (see Diniz-Filho *et al.* 2005; Dray *et al.* 2006; Getis & Griffith 2002; Griffith 2000 and see also Griffith 2006a, 2002, for developments with Poisson regressions, i.e. count data) with a modification of the weight matrix. The weight matrix *W* was defined through the shape of the variogram (as tested in Getis & Aldstadt, 2004) constructed from the deviance residuals of the non-spatial model (see Figure 3a). As in Getis & Griffith (2002), the diagonal of the matrix *W* is composed of zeros, which enables the estimation of the relationship between observations, assuming the relationship with the observation itself is zero (see also Dray *et al.* 2006). Other values matched with the Eq. 2:

$$f(d) = 1 - \left[\frac{\gamma(sill, range, d)}{sill} \right] \qquad \text{(Eq. 2)}$$

In Eq. 2, *sill* and *range* are parameters of the variogram, *d* is the distance between observations and γ is the exponential variogram function. This equation could be interpreted as the degree of connectivity between two observations and was defined between 0, i.e. no connectivity and 1, i.e. maximal connectivity. The nugget effect did not appear in this equation because we expected that $f(d) \sim 1$ when $d \sim 0$.

Eigenvectors were then extracted from the $(I-11'/n)\, W\, (I-11'/n)$ matrix transformation (see Getis & Griffith 2002), where n is the number of observations, I is the n by n identity matrix, 1 is a n by 1 vector of ones and W is the weight matrix. The Moran's I was also calculated for each eigenvector and only those having positive values were retained. This rule led to selecting eigenvectors having only positive spatial autocorrelation (Griffith 2003). Kissling & Carl (2008) recommended that the selection of the spatial term be based on a metric of RSA as well as a metric of fit. Therefore, the set of candidate eigenvectors were included linearly in the selected GLM by two stepwise procedures. The first one minimized the RSA, i.e the sill of the variogram. Relevant eigenvectors were added until sill = 0, which indicated a totally "flat" theoretical variogram. Second, unnecessary eigenvectors were removed by minimizing the AIC. This model was labeled spatial model.

1.4 Distribution and population size

The non spatial and the spatial models were compared with regard to their ability to predict the distribution and the population size of Black kites. Predictions were made over a grid of France constituted by 22500 cells of 25 km² using a partial regression (see Legendre & Legendre 1998) on environmental variables (climatic and habitat) excluding eigenvectors. Thus the non spatial and the spatial model relied, for prediction, on the same environmental variables. However in the non spatial model, RSA was not taken into account whereas it was in the spatial model. Population size and its confidence interval were calculated by running 10,000 Monte Carlo simulations of the sum of abundance predicted at the 22500 cells, i.e., the total abundance expected. At each simulation, the value of each parameter was set randomly, by considering a normal distribution with the mean equal to the regression coefficient, and the standard deviation equal to the standard error. This process allowed the full range of model parameters to be considered.

All analyses were performed using the R software version 2.13.0 (R Development Core Team 2011, www.R-project.org)

2) Results

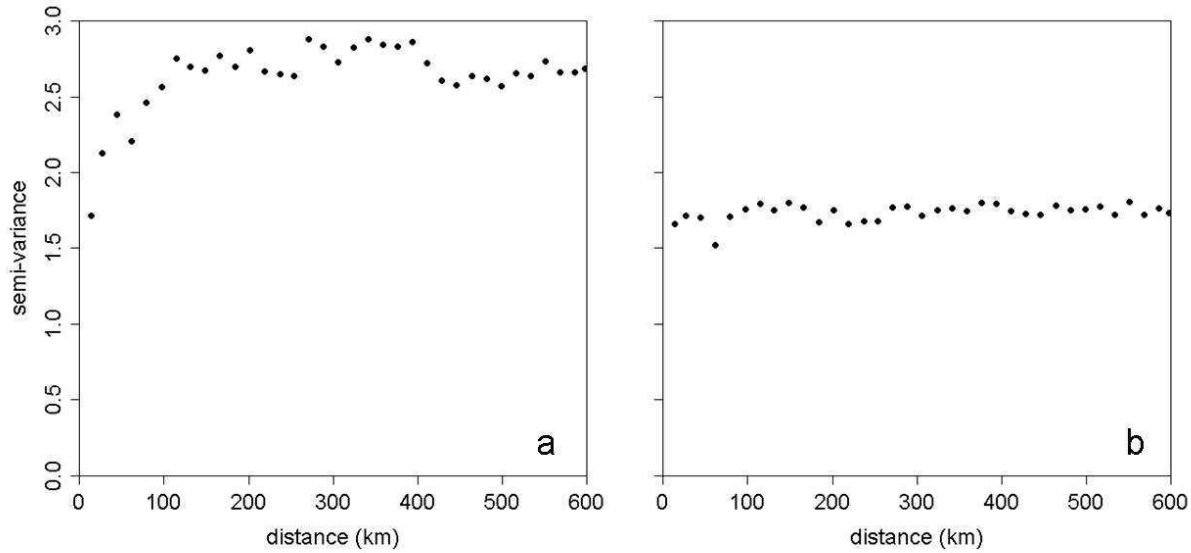


Figure 3: Variogram of the deviance residuals of the non-spatial model (a) and the spatial model (b).

The value of RMSEP for the null model, i.e. the model only including a constant and the offset, was 3.30. The value of RMSEP for the selected model (Table 1) was 3.00, with a total of 11 parameters associated to environmental features (7 linear terms, 3 quadratics terms and 1 interaction). Thus selected environmental variables reduced the RMSEP by about 9 %. The RSA was fully removed by adding 42 eigenvectors to the non-spatial model (see Figure 3). As a consequence, the estimation of the model parameters from the non-spatial and the spatial model were not the same (see Table 1).

Table 1: The coefficients (Coef), standard errors (StdE), t-value (t) and p-value (p) for the non spatial and spatial models.

	Non-spatial model				Spatial Model			
Label	**Coef**	**StdE**	**t**	**p**	**Coef**	**StdE**	**t**	**p**
Intercept	-3.48	0.06	-54.66	**$<1.10^{-324}$**	-4.10	0.12	-34.95	**$<1.10^{-267}$**
ClimDim.2	0.12	0.05	2.59	**0.010**	-0.68	0.11	-6.35	**$<1.10^{-09}$**
ClimDim.3	1.24	0.08	14.84	**$<1.10^{-50}$**	1.91	0.13	14.70	**$<1.10^{-49}$**
ClimDim.3²	-0.69	0.07	-10.21	**$<1.10^{-24}$**	-0.67	0.11	-6.17	**$<1.10^{-09}$**
ClimDim.6	0.05	0.04	1.45	0.148	-0.26	0.04	-6.18	**$<1.10^{-09}$**
ClimDim.6²	0.12	0.03	4.43	**$<1.10^{-05}$**	-0.06	0.04	-1.60	0.111
ClimDim.11	0.31	0.04	7.53	**$<1.10^{-13}$**	0.02	0.06	0.38	0.705
ClcDim.1	-0.48	0.05	-10.05	**$<1.10^{-24}$**	-0.32	0.06	-5.16	**$<1.10^{-06}$**
ClcDim.1²	-0.59	0.05	-11.58	**$<1.10^{-31}$**	-0.46	0.06	-7.37	**$<1.10^{-12}$**
ClcDim.4	0.19	0.05	4.05	**$<1.10^{-04}$**	0.24	0.06	4.13	**$<1.10^{-04}$**
ClcDim.6	0.15	0.04	4.16	**$<1.10^{-04}$**	0.11	0.04	3.04	**0.002**
ClcDim.1:ClimDim.6	-0.15	0.04	-3.62	**$<1.10^{-03}$**	-0.27	0.05	-5.34	**$<1.10^{-07}$**

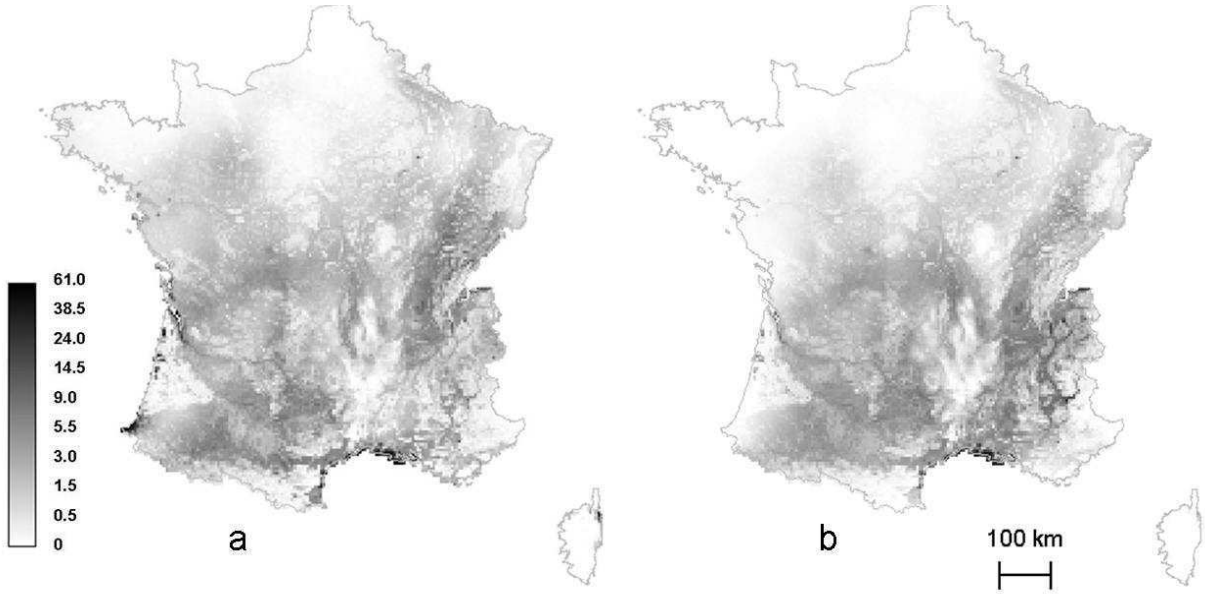


Figure 4: Predicted distribution of the Black kite (in number of pairs per 25 km²) using the non-spatial model (a) and the spatial model (b).

Distribution maps built from the non spatial and spatial model were also different, with a lower predicted abundance of Black kites in western of France using the spatial model compared to the non-spatial model (see Figure 4). Estimation of population size was also different using the non-spatial and the spatial model (see Figure 5). The average population size prediction using the non-spatial model was 36,122 (95% confidence interval 28,780 - 45,683) breeding pairs of Black kite in France, whereas using the spatial model it was 32,133 pairs (21,426 - 47,072).

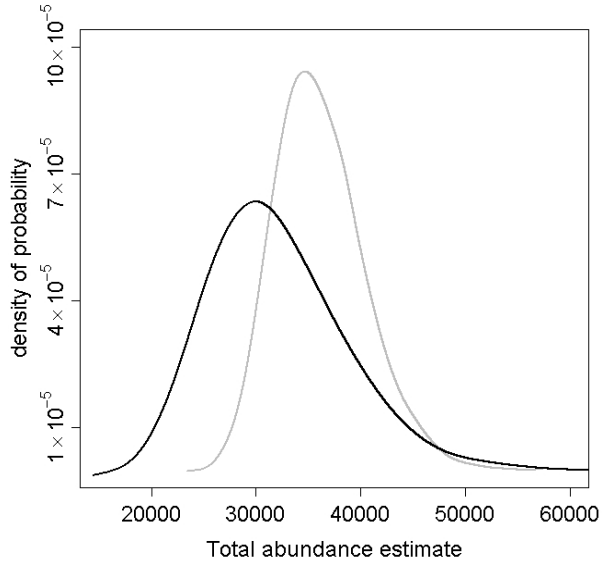


Figure 5: Predicted population size and confidence interval using the non-spatial model (in grey) and the spatial model (in black). The two curves are obtained by dividing the 10,000 Monte Carlo simulations in 100 breaks points and calculating the density of probability for each break point. A cubic smoothing spline was then used (with 15 equivalent degree of freedom) in order to make the figure clearer.

3) Discussion

Differences in coefficients estimation were found between the spatial or the non-spatial model (see Table 1), with some coefficients even changing sign (see ClimDim2 and ClimDim.6). The latter reflects a possible inverse fit of the data when the RSA was not accounted for, a pattern previously observed (see Kühn 2007). Other coefficients showed moderate to rather large reduction of there statistical effect (see coefficients and standard errors in Table 1).

The selection of eigenvectors by a two step procedure minimizing both RSA and AIC (a strategy first suggested by Kissling & Carl 2008) seems promising. The first step (minimizing the RSA), led to select 49 eigenvectors, while the second step (minimizing AIC) allowed the removal of 7 eigenvectors without impacting on the RSA. We also recommend checking for collinearity between the selected environmental variables and the selected eigenvectors, which could artificially increase the standard error of the regression parameters (see Freckleton 2002). Here the correlation matrix between the environmental variables showed only a slight correlation (correlation coefficient never above 0.5, see Table C1 in the Appendix C). These correlations were lower than the critical value of 0.7 proposed by Dormann *et al.* (2012).

The PCA approach was used to overcome a problem of multicollinearity between the environmental variables. However, there is a cost in using PCA, since it raises some difficulties in understanding model coefficients from an ecological point of view. In order to check that selected PCA axes are biologically relevant, one must interpret these axes. In our case, the axe ClcDim.1 represented mainly a natural gradient (see Figure B1 in Appendix B) where positive values indicated a high percentage of forest in the quadrat and negative values indicated a high percentage of intensive farming (i.e. no forest). The ClcDim.1 effect in the model resulted from a linear effect (-0.32, see Table 1) as well as a quadratic effect (-0.46, see Table 1), the latter indicating that this effect was stronger in quadrats dominated by forest. ClcDim.4 mostly represented wetlands (see Figure B1 in Appendix B) where positive values indicated a high percentage of wetlands. The coefficient of ClcDim.4 was 0.24 (see Table 1), underlying strong preference of the Black kite for wetland habitats. These two aspects of the landscape fit well with previous knowledge for this species: Black kites avoid large forests, and prefer anthropic environments (agricultural lands) and wetlands (Thiollay & Bretagnolle, 2004). Therefore, despite interpretation problems, we believe that in our case, the advantages to use PCA outweighed the disadvantages, since our aim was first to provide unbiased parameters estimation in order to predict population distribution and size.

The selection of these PCA axes was done using a leave-one-out cross-validation, which is known to be asymptotically equivalent to AIC (Stone 1977), but allows dealing with RSA using a threshold distance between the subsets. However, as cross validation criterion, we used the RMSEP as it is used in linear models, i.e. without accounting for the fact that we used a Poisson distribution. This means that there is equal penalization between an observed abundance of 0 and a predicted one of 1, and an observed abundance of 50 and a predicted

one of 51, whereas mean equal variance on a Poisson distribution, i.e. high count have more variation. Thus high errors of prediction have more probability to occur on high counts. Thus a better or refined RMSEP should be found, for example using a logarithmic scale, which would select better predictors than those found here.

Moreover, we did not account for overdispersion in our analysis, whereas the analysis of count data often presents overdispersion. This choice was made because there is a strong link between RSA and overdispersion since RSA generates overdispersion (see Griffith & Haining 2006; Haining *et al.* 2009). Here we were interested into the effects of RSA, and therefore we have fixed overdispersion, considering that there was no additional overdispersion than the one due to RSA. We have however checked the overdispersion in the final spatial model, which turned to be about 3.32 units using the Pearson Chi Square statistic (the Poisson distribution fixes it to 1). For comparison, the non spatial model had an overdispersion of 9.99 units, which clearly shows that accounting for RSA also reduces overdispersion. However, since some overdispersion remains in the spatial model, further improvement seems necessary, e.g. by considering a Quasi-Poisson distribution in the final spatial model.

Conclusions

Predicted distribution and population size of the Black kite between the non-spatial and the spatial model were similar, but there were also substantial differences. The spatial model predicted lower abundance in western France compared to the non spatial model (see Figure 4), and hence a lower population size than the spatial model (see Figure 5). The model uncertainty was also larger for the spatial model than for the non-spatial model (see Table 1), which was expected, but may strongly impact model predictions (see Figure 5). Thus in addition to give misleading distribution maps of species, RSA also gave a false feeling of precise predictions, which *a priori* may suggest this model shows a better fit of the data. In terms of conservation applications, a poorly predicted map of abundance may have serious consequences: in our case, not accounting for RSA would lead to the interpretation that western and eastern of France are equal important and suitable breeding habitat for Black kites, whereas in fact eastern France is the main area of breeding for this species.

Acknowledgements

Particular thanks are deserved to an anonymous referee for his very helpful comments (including missed references) and suggestions which greatly improved the paper. We also acknowledge the Editor of this proceeding (Dr Mark O'Connell) for his additional comments. We also thank all voluntaries who have carried out the field survey. We particularly thank Jean Sériot who coordinated the national survey “Rapaces nicheurs de France”, and Fabienne David who is now coordinating the French monitoring raptor scheme. We thank Arzhela Hemery and Maxime Passerault who have participated at the preparation of the dataset used here and particularly for extraction of satellite data. We thank people from “Biostatistics and spatial processes” team (INRA Avignon), and particularly Pascal Monestiez, Joël Chadoeuf and Rachid Senoussi, for their advices and comments about spatial statistics. We also thank Patrick Duncan and Samantha Patrick for the English revision and their very interesting comments. Finally we thank the *Région Poitou-Charentes* and the *Département des Deux-Sèvres* for funding PhD grant.

References

voir page 125.

Appendix A

Table A1: The nine habitat hyper-classes used in our analyses and the Corine Land Cover initial classification. One row corresponds to one initial Corine Land Cover class, which is defined by three distinct labels.

44 Corine Land Cover nomenclatures			**9 Habitat hyper-classes**
Label 1	**Label 2**	**Label 3**	
Artificial surfaces	Urban fabric	Continuous urban fabric	Anthropic areas
Artificial surfaces	Urban fabric	Discontinuous urban fabric	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Industrial or commercial units	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Road and rail networks and associated land	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Port areas	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Airports	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Mineral extraction sites	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Dump sites	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Construction sites	Anthropic areas
Artificial surfaces	Artificial, non-agricultural vegetated areas	Green urban areas	Anthropic areas
Artificial surfaces	Artificial, non-agricultural vegetated areas	Sport and leisure facilities	Anthropic areas
Agricultural areas	Arable land	Non-irrigated arable land	Intensive agriculture
Agricultural areas	Arable land	Permanently irrigated land	Intensive agriculture
Agricultural areas	Arable land	Rice fields	Intensive agriculture
Agricultural areas	Permanent crops	Vineyards	Permanent agriculture
Agricultural areas	Permanent crops	Fruit trees and berry plantations	Permanent agriculture
Agricultural areas	Permanent crops	Olive groves	Permanent agriculture
Agricultural areas	Pastures	Pastures	Extensive farming
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Natural grasslands	Extensive farming
Agricultural areas	Heterogeneous agricultural areas	Annual crops associated with permanent crops	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Complex cultivation patterns	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Land principally occupied by agriculture, with significant areas of natural vegetation	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Agro-forestry areas	Forest areas
Forest and semi-natural areas	Forests	Broad-leaved forests	Forest areas
Forest and semi-natural areas	Forests	Coniferous forest	Forest areas
Forest and semi-natural areas	Forests	Mixed forest	Forest areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Moors and heathland	Transitional areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Sclerophyllous vegetation	Transitional areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Transitional woodland-shrub	Transitional areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Beaches, dunes, sands	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Bare rocks	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Sparsely vegetated areas	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Burnt areas	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Glaciers and perpetual snow	Open areas
Wetlands	Inland wetlands	Inland marshes	Wetlands
Wetlands	Inland wetlands	Peat bogs	Wetlands
Wetlands	Maritime wetlands	Salt marshes	Wetlands
Wetlands	Maritime wetlands	Salines	Wetlands
Wetlands	Maritime wetlands	Intertidal flats	Wetlands
Water bodies	Inland waters	Water courses	Wetlands
Water bodies	Inland waters	Water bodies	Wetlands
Water bodies	Maritime waters	Coastal lagoons	Wetlands
Water bodies	Maritime waters	Estuaries	Wetlands
Water bodies	Maritime waters	Sea and ocean	Not used

Appendix B

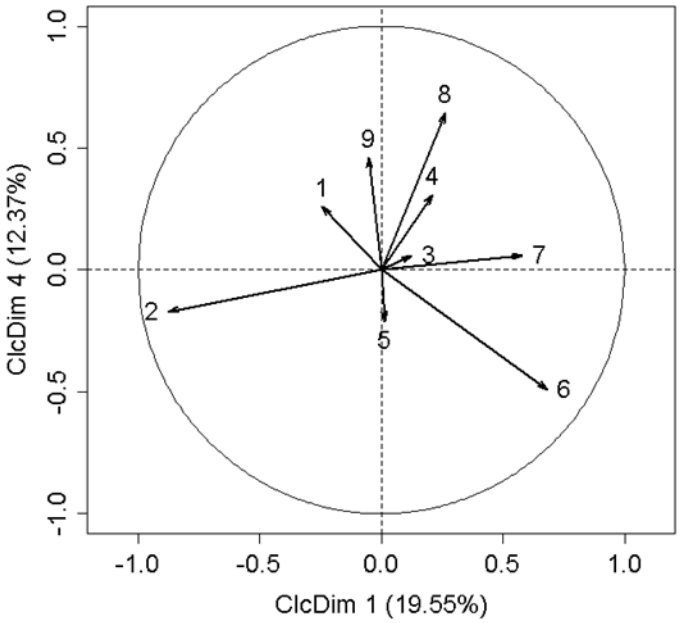


Figure B1: Correlation circle of ClcDim.1 and ClcDim.4 principal components

1: Anthropic areas
2: Intensive agriculture
3: Permanent agriculture
4: Extensive farming
5: Heterogeneous agriculture
6: Forest areas
7: Transitional areas
8: Open areas
9: Wetlands

Appendix C

Table C1: Correlation matrix between environmental variables (columns) and eigenvectors (rows).

Eigenvectors	ClimDim.2	ClimDim.3	ClimDim.6	ClimDim.11	ClcDim.1	ClcDim.4	ClcDim.6
E1	0.50	0.27	-0.14	-0.14	-0.15	-0.03	0.10
E3	0.18	-0.26	-0.06	0.17	-0.04	0.18	-0.05
E6	0.30	-0.08	-0.02	-0.04	-0.01	0.03	0.09
E7	-0.07	0.04	-0.21	-0.12	0.11	-0.01	0.03
E11	0.14	-0.01	-0.11	0.03	0.09	0.01	-0.02
E13	0.12	-0.15	-0.02	-0.10	0.00	0.05	0.00
E15	-0.09	0.17	-0.14	-0.10	0.09	-0.12	-0.04
E17	0.05	-0.08	-0.10	-0.09	-0.04	0.02	0.09
E23	0.01	0.05	0.19	-0.11	-0.04	0.01	-0.01
E27	0.13	-0.02	0.01	0.11	0.11	0.08	0.09
E30	-0.14	-0.01	-0.19	-0.02	0.03	0.05	-0.02
E32	0.04	0.06	-0.13	0.06	-0.07	0.07	-0.03
E33	-0.07	-0.04	-0.01	0.01	-0.04	0.07	0.04
E34	0.05	-0.09	-0.01	0.11	0.01	0.02	0.01
E35	-0.03	0.06	-0.09	-0.06	-0.06	-0.05	0.00
E37	0.18	-0.05	-0.03	-0.10	-0.01	0.09	-0.01
E40	0.07	0.08	0.03	0.00	0.08	0.03	0.07
E43	0.09	-0.01	-0.09	0.09	0.06	0.07	0.01
E56	-0.06	0.03	-0.02	0.04	-0.04	0.01	-0.02
E58	-0.07	0.06	0.03	-0.09	0.04	-0.18	0.01
E65	-0.03	0.02	0.04	0.05	-0.08	-0.08	0.03
E70	0.01	-0.02	-0.03	-0.05	-0.06	0.07	-0.04
E78	0.02	-0.03	-0.02	0.00	0.01	0.00	0.01
E82	-0.01	-0.02	-0.06	-0.01	0.02	0.06	0.04
E87	0.01	0.03	-0.02	0.00	-0.04	-0.05	0.05
E90	-0.01	0.00	0.00	0.02	0.04	-0.01	0.03
E93	0.01	0.01	-0.03	-0.04	0.01	-0.02	-0.02
E96	0.04	0.00	0.01	0.00	0.01	0.03	-0.05
E103	0.02	0.00	0.03	0.00	0.06	-0.01	-0.02
E107	0.00	0.00	0.01	0.01	-0.03	0.03	-0.02
E108	0.00	0.03	0.02	0.00	-0.04	-0.02	0.01
E111	0.02	0.01	-0.01	-0.10	0.03	0.01	0.03
E112	-0.01	0.00	0.00	-0.02	0.01	-0.03	0.06
E115	0.00	0.01	-0.01	0.00	-0.01	0.01	0.02
E121	0.02	0.02	0.01	-0.02	0.06	0.02	-0.06
E123	0.00	0.02	0.02	-0.03	0.01	-0.05	-0.01
E129	-0.02	0.00	-0.01	-0.04	-0.03	-0.04	-0.03
E134	-0.01	-0.01	-0.02	0.03	0.01	0.00	-0.05
E135	0.02	0.01	0.00	-0.07	0.02	-0.01	0.00
E141	0.02	0.02	0.02	-0.01	0.03	-0.08	-0.01
E142	0.03	-0.01	0.03	0.00	-0.01	0.01	0.04
E143	0.05	0.03	0.04	-0.05	-0.02	-0.01	0.03

CHAPITRE 2 : EVALUATION D'UNE METHODE POUR LA SELECTION DE VARIABLE EN PRESENCE D'AUTOCORRELATION SPATIALE : LE SLOO

Dans le premier chapitre de cette thèse, nous avons utilisé une procédure de sélection de variable basée sur une modification du *leave-one-out* qui permet théoriquement de tenir compte du problème d'autocorrélation spatiale résiduelle dans un modèle de régression. Cependant, cette procédure de sélection de variable n'a pas encore fait l'objet d'une évaluation adéquate. De plus, comme discuté dans le premier chapitre, le critère de sélection utilisé, basé sur la somme des erreurs au carré (RMSEP), n'est pas forcément le mieux adapté dans le cas où l'hypothèse de normalité n'est pas respectée (par exemple pour les GLMs). Ce second chapitre va donc s'intéresser plus spécifiquement à l'évaluation de la méthode de sélection de variable proposée dans le premier chapitre tout en adaptant le critère de sélection au cadre plus général des GLMs. Nous utilisons désormais l'acronyme SLOO (*spatial-leave-one-out*) pour faire référence à cette méthode.

L'article présenté ci-après a été accepté pour publication dans le journal international *Global Ecology & Biogeography*. L'éditeur de ce journal n'ayant que récemment donné ses commentaires, la version du papier présentée dans cette thèse est la version initialement soumise au journal.

The spatial-leave-one-out cross-validation (SLOO) for variable selection in the presence of spatial autocorrelation

Author: Le Rest Kévin [1]*, Pinaud David [1], Monestiez Pascal [1, 2, 3], Chadoeuf Joël [3] and Bretagnolle Vincent [1]

[1] Centre d'Etudes Biologiques de Chizé (CEBC), CNRS UPR 1934, 79360 Beauvoir-Sur-Niort, France

[2] INRA (USC 1339), CEBC-CNRS, 79360 Beauvoir-sur-Niort, France

[3] INRA Provence-Alpes-Côte d'Azur, Unité Biostatistique et Processus Spatiaux (BioSP), Domaine Saint-Paul Site, Agroparc, 84914 Avignon Cedex 9

Email addressees: lerest.k@gmail.com; pinaud@cebc.cnrs.fr; monestiez@avignon.inra.fr; chadoeuf@avignon.inra.fr; bretagnolle@cebc.cnrs.fr

*Corresponding author:

Email: lerest.k@gmail.com

Tel: +33 (0)5 49 09 35 13

Fax: +33 (0)5 49 09 65 26

Accepted manuscript

Global Ecology & Biogeography

This version is the initially submitted (August 2013)

Overview

Abstract

Aim Processes and variables measured in ecology are almost always spatially autocorrelated, potentially leading to choosing overly complex models when performing variable selection. One way to solve this problem is to account for residual spatial autocorrelation (RSA) for each subset of variables considered and then use a classical model selection criterion such as the Akaïke Information Criterion (AIC). However, this method can be fastidious and it raises other concerns such as which spatial model to use or how to compare different spatial models. To improve accuracy of variable selection in ecology, this study evaluates an alternative method based on a spatial cross-validation procedure. Such procedure is usually used for model evaluation but can also provide interesting outcomes for variable selection in the presence of spatial autocorrelation.

Innovation We propose to use a special case of spatial cross-validation, the spatial leave-one-out (SLOO), giving a criterion equivalent to AIC in the absence of spatial autocorrelation. SLOO only computes non-spatial models and uses a threshold distance (equal to the range of RSA) to keep each point left out spatially independent from the others. We first provide some simulations to evaluate how SLOO performs compared to AIC. We then assess the performance of SLOO on a large-scale dataset. R software codes are provided for generalized linear models.

Main conclusions The AIC was relevant for variable selection in the presence of RSA if the variables considered were not spatially autocorrelated. It otherwise failed as highly spatially autocorrelated variables were more often selected than others. Conversely, SLOO had similar performances whether the variables were themselves spatially autocorrelated or not. It was particularly useful when the range of RSA was small, which is a common property of spatial tools. SLOO appears to be a promising solution for selecting relevant variables from most ecological spatial datasets.

Key-words: AIC, Common Buzzard *Buteo buteo*, Spatial cross-validation, GLM, Residual spatial autocorrelation, Simulations.

Abbreviations

The following abbreviations are used in this paper:
AIC: **A**kaïke **I**nformation **C**riterion
BIC: **B**ayseian **I**nformation **C**riterion
GLM: **G**eneralized **L**inear **M**odel
GRF: **G**aussian **R**andom **F**ield
LOO: **L**eave-**O**ne-**O**ut
PCA: **P**rincipal **C**omponent **A**nalysis
RSA: **R**esidual **S**patial **A**utocorrelation
SLOO: **S**patial-**L**eave-**O**ne-**O**ut

Introduction

Ecological processes *in natura* are inherently spatial, either for environmental or intrinsic biological reasons (Legendre & Fortin, 1989; Legendre, 1993; Koenig, 1999; Keitt *et al.*, 2002). The data collected in the field are thus usually spatially autocorrelated. Spatial autocorrelation can alter the statistical independence of residuals in regression models, leading to bias such as falsified tests or likelihoods (Lennon, 2000; Bahn *et al.*, 2006; Hoeting *et al.*, 2006; Dormann, 2007 but see Diniz-Filho *et al.*, 2003). Statistical methods able to capture the residual spatial autocorrelation (hereafter RSA), so called 'spatial models', are required to correct for those biases (see Lichstein *et al.*, 2002; Fortin & Dale, 2005; Griffith & Peres-Neto 2006; Dormann *et al.*, 2007; Betts *et al.*, 2009). While such methods were shown to be efficient for the estimation of model parameters from spatial datasets, they are mainly applied after the process of variable selection. This begs the question of the validity of variable selection in the presence of RSA.

Model selection has gained a wide audience in ecology (Johnson & Omland, 2004), with the main aim of selecting pertinent variables by comparing several models with different subsets of variables and choosing which ones are most likely to explain the observed pattern in relation to the studied process. A few metrics have been proposed to help in this process, e.g. the Mallows' Cp (Mallows, 1973), the Akaïke Information Criterion (AIC, Akaike, 1973) or the Bayesian Information Criterion (BIC, Schwarz, 1978). These selection criteria usually reflect a balance between the data fit and the model complexity (George, 2000). The principal difference between them comes from the penalty accorded to the model complexity. For instance, Mallows' Cp and AIC are almost equivalent and correspond to a penalty of 2 whereas BIC corresponds to a penalty of log(n), where n is the number of independent observations (George, 2000). BIC is asymptotically consistent since it will select the true model as $n \rightarrow \infty$ (see Stone, 1979; George, 2000). An implicit assumption is however the existence of a 'true model' in the set of candidate ones (Stone, 1979; Shao, 1997; George, 2000). In biological sciences this assumption is unrealistic because the number of variables affecting the processes can be very high, if not infinite (see Burnham & Anderson, 2002). It will thus be better to allow the dimension of the true model to increase with n (Stone, 1979), which is a fundamental property of Mallows' Cp and AIC.

The presence of RSA invalidates the use of classical model selection criteria such as Mallows' Cp, AIC or BIC since they are based on the overall likelihood assuming independent residuals. However in practice, these criteria were still used without accounting for RSA (see for example Kühn *et al.*, 2009). This may lead to the selection of overly complex models having a much larger number of variables than necessary (Hoeting *et al.*, 2006; Cassemiro *et al.*, 2007; Diniz-Filho *et al.*, 2008). As acknowledged by Dormann *et al.* (2007), variable selection in the presence of RSA has received surprisingly little interest so far in the literature. Identifying the relevant variables in the presence of RSA thus remains challenging. The classical strategy for variable selection in the presence of RSA consists in

first accounting for RSA in all candidate models and then to compare them with a classical model selection criterion. The computed criterion is this time valid since RSA has been removed. For instance, Hoeting *et al.* (2006) underlined the necessity to account for RSA when using AIC for variable selection in a geostatistical modeling framework, as did Diniz-Filho *et al.* (2008) who compared two methods to account for RSA. This approach, however, has three main drawbacks: first, models accounting for RSA need much longer computation time, making model selection very difficult when the number of variables is large; second, the variables finally selected may depend on the method used to account for RSA (see Diniz-Filho *et al.*, 2008); and third, most of 'spatial explicit methods' may lead to a 'spatial confounding' effect between the variables and the spatial term, hiding the importance of some spatially autocorrelated variables (Reich *et al.*, 2006; Betts *et al.*, 2009; Bini *et al.*, 2009; Hodges & Reich, 2010; Paciorek, 2010; Hughes & Haran, 2013). This latter effect is less known but is of primary interest when one wants to perform variable selection with spatially autocorrelated variables, which likely happens in most of real applications.

Yet another method based on a modification of a cross-validation procedure has been frequently used for model evaluation in the presence of RSA and should also be used for variable selection in this context. Cross-validation usually consists in splitting the initial dataset in two subsets (see Arlot & Celisse, 2010 for an overview), one is used to estimate model parameters (the training set) and the other one is used to evaluate the predictive power of the model (the validation set). A critical prerequisite is that training and validation sets be independent (Arlot & Celisse, 2010), at least under the model being evaluated. Otherwise, the difference between the observation and the prediction may be unreliable (Altman, 1990). In a spatial context, most observations are related each others, training and validation sets are thus rarely independent, which highly reduces the power of cross-validation to evaluate a model. An intuitive way to solve this problem consists in splitting the spatial data in several non-overlapping geographical areas that are used as training and validation sets, a technique often referred as spatial cross-validation (see Chung & Fabbri, 2003; Brenning, 2005; Pinkerton *et al.*, 2010; Russ & Brenning, 2010; Bahn & McGill, 2013). It is also necessary that the distance between the training and the validation areas is greater than the range of RSA (i.e. the distance at which a pair of observations are independent) of the model evaluated in order to guaranty a full independence (Brenning, 2005; Russ & Brenning, 2010). Unfortunately this minimal distance between the training and validation sets is almost always ignored (see for example Chung & Fabbri, 2003; Pinkerton *et al.*, 2010; Russ & Brenning, 2010; Bahn & McGill, 2013). The spatial cross-validation is actually a spatial version of the delete-d-cross-validation (Geisser, 1975) where d is the number of observations in the validation set. If there is no true model, which is expected in ecological applications, delete-d-cross-validation is only useful for variable selection when $d = 1$, i.e. when a simple leave-one-out cross-validation (hereafter LOO, Allen, 1974; Stone, 1974) is used (see Shao, 1997). The current form of spatial cross-validation considering $d >> 1$ should thus not be useful for variable selection in this context. The special case of $d = 1$ would however provide a useful criterion

since leave-one-out cross-validation is known to be asymptotically equivalent to AIC (Stone, 1977).

In this paper we thus propose to evaluate the performance of the spatial leave-one-out cross-validation (hereafter SLOO) for variable selection in the presence of RSA. The selection criterion is computed by applying a LOO in a spatial context, i.e. by using a threshold distance between the training and the validation sets that removes some data in order to eliminate the bias due to RSA. This approach has been recently used by Le Rest *et al.* (2013), though these authors did not provide neither a suitable calculus of the selection criterion, nor an evaluation of its performance. We first give a full description of the method and the way to compute the selection criterion. Second, we use a simulation approach to evaluate how SLOO performs compared to AIC in selecting a continuous variable that affects the studied process while avoiding another one that does not affect the process. We quantify in particular the relative effects of i) the RSA, ii) the threshold distance used to calculate SLOO and iii) the spatial autocorrelation in the explanatory variables. In the third section, we use SLOO with different threshold distance on a real data set composed of twenty height environmental variables suspected to influence the abundance of a diurnal raptor species, the Common Buzzard *Buteo buteo* (*Linnaeus*, 1758). This case study illustrates the utilization of SLOO and its performance for the selection of variables in a species distribution model. Finally, Supplementary material (online) provides computing codes and an example based on the simulations showing how to calculate SLOO with R software from generalized linear models (hereafter GLMs).

1) The spatial-leave-one-out (SLOO)

The SLOO method relies in practice on four steps. The first step removes one observation from the initial dataset (the grey cross in Fig. 1). The second step removes all the observations that are spatially correlated with this removed observation, i.e. removing all data inside a buffer of a radius equal to the range of the RSA of the model considered (see for example the grey buffer in Fig. 1). All remaining observations (black points in Fig. 1) constitute the training set and are used in a GLM framework to estimate the parameters. A prediction (step three) is then made at the location of the removed observation (validation set, grey point in Fig. 1) using the estimated parameters of the GLM. The fourth step calculates a score between the observed value and the predicted one. This procedure is repeated for every single observation of the dataset, which allows calculating an overall criterion of fit. Note that LOO is a special case of SLOO when the threshold distance used is null and SLOO is thus asymptotically equivalent to AIC in absence of RSA (Stone, 1977), allowing a direct comparison between these two selection criteria.

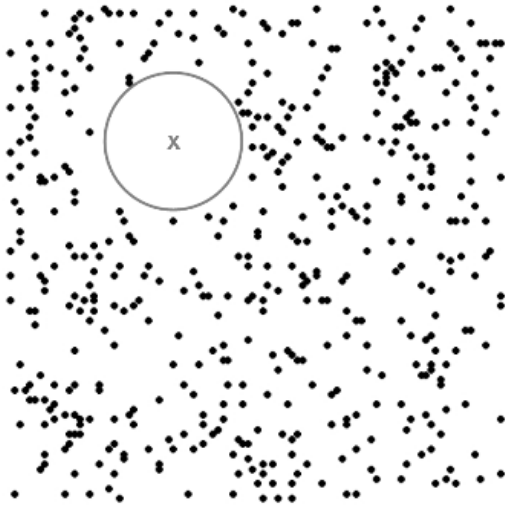


Figure 1: One example of spatial-leave-one-out on a grid of 100 x 100 pixels having 500 observations. Here the threshold distance is set arbitrarily to 15 pixels (radius of the grey buffer). The grey cross is the point leaved out, i.e. the validation set, and the black points are the training set.

The criterion of the SLOO (eqn 1) is based here on likelihood instead of the classical sum of squares of errors, because it is more adapted for non-normally distributed response variables (see Knafl & Grey, 2007 for details one the likelihood *versus* least square cross-validation) and is therefore more suitable for GLMs. In practice, we compute the probability P (for a discrete response variable, the density probability for a continuous one) of the left out observed value y_i according to the predicted one $\hat{y}_i$ by using the training set. This is achieved by using the theoretic distribution of the model (Normal, Binomial, Poisson, etc…). The sum of the logarithm of these probabilities leads to an overall cross-validated log-likelihood for the model, which is the selection criterion to be maximized.

$$SLOO_{\log Lik} = \sum_{i=1}^{n} \log\left[P\left(y_i \middle| \hat{y}_i\right)\right] \qquad \textbf{(eqn 1)}$$

All simulations and analysis were performed using R version 2.13.0 (R Core Team, 2013, www.R-project.org). Full details on how calculate this criterion with R and an example can be found in Supplementary material (online).

2) Simulations

We conducted the first simulation on a 100 x 100 pixel regular grid approximating a continuous field and iterated the following process 10,000 times:

a) Generating three independent Gaussian Random Fields (hereafter GRFs) with a spherical spatial structure with mean equal to zero, variance (sill) equal to one, no nugget effect and a range chosen randomly between 1 and 100 pixels.

b) Generating the response variable *Y* such as:

$$Y = GRF1 + \beta\, GRF2 \qquad \textbf{(eqn 2)}$$

GRF1 was considered as unavailable (unknown process) and was used to generate RSA from its spatial properties. *GRF2* played the role of an available and influential variable, and the parameter β reflected its actual importance. β was taken from *N(0,1)*, which allowed to scan a wide range of values but avoiding too high values. High β (over 2 in absolute) were irrelevant since they always led to select the influential variable (*GRF2*) in the model whatever the selection criterion used. *GRF3* did not affect the response *Y* in eqn 2 and can be considered as an available but non-influential variable. A random sample of 500 observations from the 100 x 100 grid was considered as the available dataset.

c) Running two variable selection procedures (one using AIC and one using SLOO) with *GRF2* and *GRF3* being the candidate variables of the model. Note that SLOO was computed by using the range of *GRF1* as threshold distance.

d) Recording which variables (*GRF2* and/or *GRF3*) were selected for each selection criterion used, i.e. either by minimizing AIC or by maximizing $SLOO_{logLik}$.

Note that the "true model", i.e. holding the two influential variables (*GRF1* and *GRF2*) and avoiding the non-influential one (*GRF3*) could never be selected since *GRF1* was considered as unavailable; thus the term "best model" was used to qualify the model holding the influential variable (*GRF2*) and avoiding the non-influential one (*GRF3*).

Binomial regression models were used to represent graphically either the probability to select the best model, the probability to select the influential variable (*GRF2*) or the probability to select the non-influential one (*GRF3*), by considering alternatively AIC or SLOO and depending on varying levels of RSA. The range of RSA was first used as a factor having 100 modalities (between 1 and 100 pixels) and then smoothed using cubic smoothing splines in order to provide an easier graphical representation.

i) Effect of the residual spatial autocorrelation

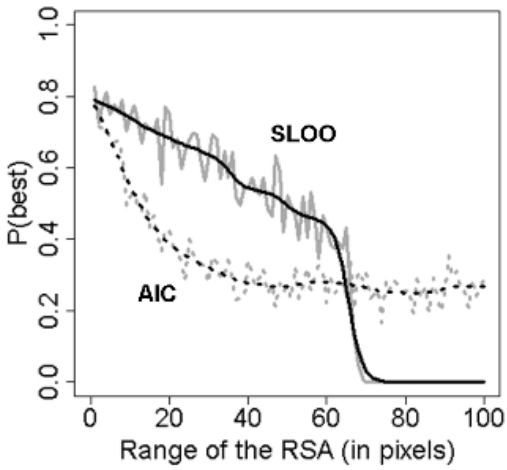


Figure 2: Probability (P) to select the best model depending on the range of the residual spatial autocorrelation (RSA), by using two selection criteria: the Akaïke information criterion (AIC, dotted lines) and the spatial-leave-one-out (SLOO, solid lines). Black lines are cubic smoothing splines and grey lines consider the range of RSA as a factor (measure of the variability).

Fig. 2 shows that the probability to select the best model, i.e. holding the influential variable (*GRF2*) and avoiding the non-influential one (*GRF3*), was always higher using SLOO than AIC except when the range of RSA was higher than 60 pixels (i.e., 60% of the grid wide), a value at which most of the training set was removed (see section I and Fig. 1). Indeed in our simulated area, the farthest distance between two locations was about 140 pixels (diagonal of the grid), which explained why the capacities of SLOO fell down with threshold distance between 60 and 70 pixels. We also found that the probability to select the best model decreased as RSA increased, considering either AIC or SLOO (Fig. 2).

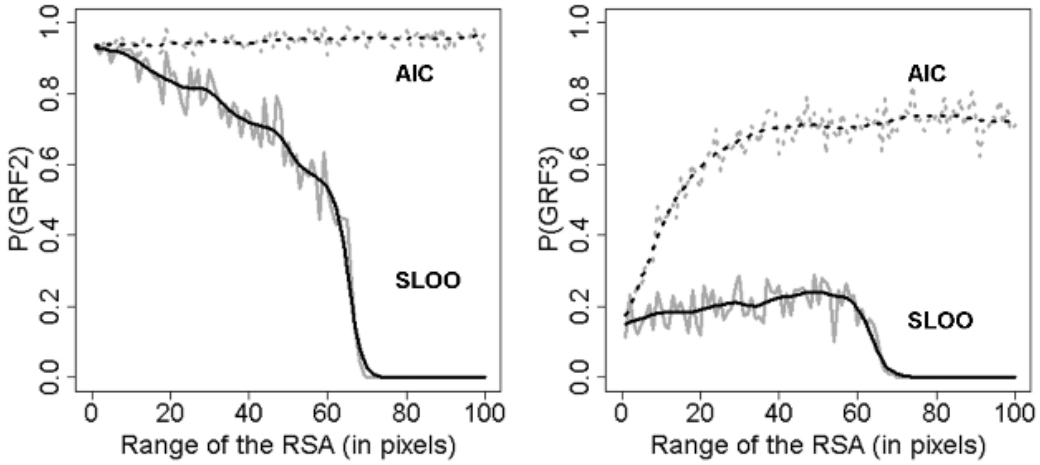


Figure 3: Probability (P) to select the influential variable (*GRF2*) and the non-influential one (*GRF3*) depending on the range of the residual spatial autocorrelation (RSA), by using two selection criteria: the Akaïke information criterion (AIC, dotted lines) and the spatial-leave-one-out (SLOO, solid lines). Black lines are cubic smoothing splines and grey lines consider the range of RSA as a factor (measure of the variability).

Using AIC, the probability to select the influential variable (*GRF2* in Fig. 3) was always high and actually slightly increased with the range of RSA. But on the other hand, the probability to select the non-influential one (*GRF3* in Fig. 3) dramatically increased with the range of RSA. These results were expected since AIC is known to select overcomplex models in presence of RSA (Hoeting *et al.*, 2006; Cassemiro *et al.*, 2007; Diniz-Filho *et al.*, 2008). Conversely using SLOO, the probability to select the influential variable (*GRF2* in Fig. 3) decreased with increasing RSA and the probability to select the non-influential one (*GRF3* in Fig. 3) just slightly increased as RSA increased (up to a certain limit, see comments above) but remained rather low. However it was not possible to determine if these effects were due to RSA because the range of RSA was also the threshold distance used in the SLOO, which caused a decrease in the number of observations of the training set (see section I and Fig. 1), also decreasing the statistical power of the SLOO. The effect of the threshold distance was thus investigated with another simulation.

ii) Effect of the threshold distance used in absence of RSA

The first simulation procedure (section II-i) was modified in order to separately study the threshold distance: here there was no RSA (i.e., *GRF1* had no spatial structure), and varying threshold distances were used for SLOO, chosen randomly between 1 and 100 pixels (as for the range of *GRF1* in the first simulation).

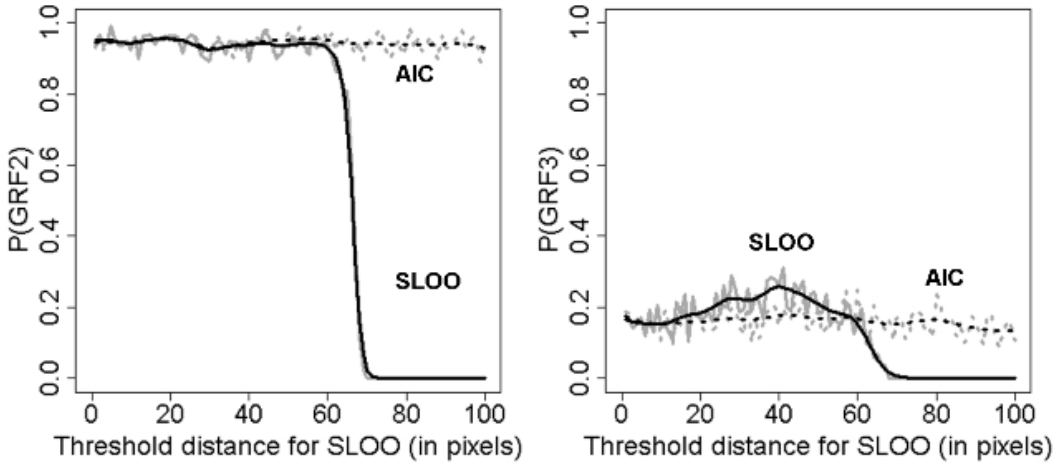


Figure 4: Probability (P) to select the influential variable (*GRF2*) and the non-influential one (*GRF3*) in absence of residual spatial autocorrelation but according to the threshold distance used for spatial-leave-one-out (SLOO, solid lines). Results for Akaïke information criterion (AIC, dotted lines) are given for an easier visual comparison. Black lines are cubic smoothing splines and grey lines consider the threshold distance as a factor (measure of the variability).

The probability of selecting the influential variable in the absence of RSA using SLOO was very close to AIC performances whatever the threshold distance used (see *GRF2* in Fig. 4), except when the threshold distance was higher than half of the extent of the study area (see section II-i for explanations). Moreover the probability to select the non-influential

variable using SLOO (see *GRF3* in Fig. 4) was only slightly increased by the threshold distance used. Overall the threshold distance used in SLOO only slightly affected the probability to select the variables, and was thus not the cause of the important decrease on the probability to select the influential variable when increasing the range of RSA (*GRF2* in Fig. 3). This latter result may be explained by pseudo-replication caused by RSA, leading naturally to a loss of power by decreasing the true number of degree of freedom (see Legendre, 1993).

iii) Effect of the spatial autocorrelation of the variables

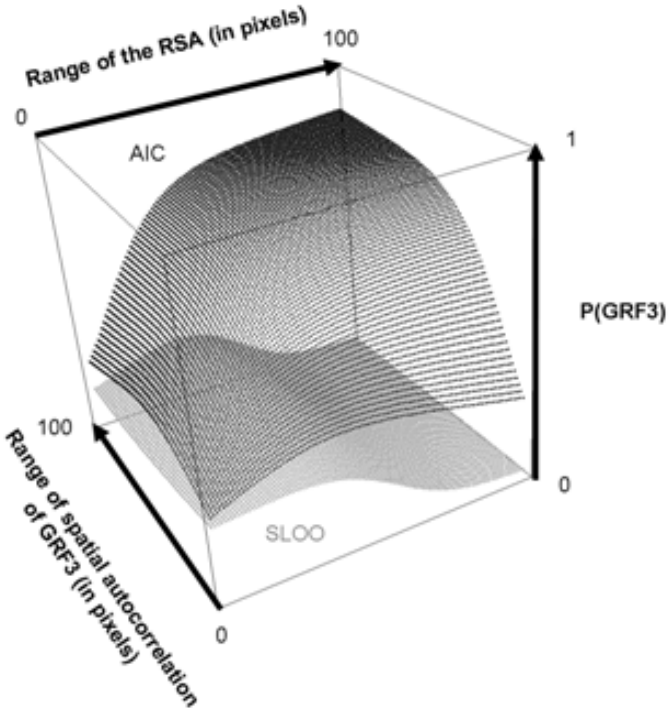


Figure 5 : Probability (P) to select the non-influential variable (*GRF3*) depending on its own range of spatial autocorrelation and the range of residual spatial autocorrelation (RSA) by considering two selection criteria: the Akaïke information criterion (AIC, in black) and the spatial-leave-one-out (SLOO, in grey). The surface plots are obtained from cubic smoothing splines accounting for each dimension and also accounting for the interaction between them.

The first simulation (section II-i) also showed that the probability to select the variables depended on their own spatial autocorrelation. In presence of RSA, the probability to select the non-influential variable using AIC also increased with its own spatial autocorrelation (Fig. 5). Lennon (2000) found a similar result by considering correlations and levels of significance on explanatory variables. Thus in presence of RSA, both the amount of RSA and the amount of the spatial autocorrelation of the explanatory variables could affect the probability to select the variables when using AIC. SLOO conversely showed less sensitivity to RSA; in particular, the probability to select the non-influent variable was not affected by its own spatial autocorrelation (see Fig. 5).

iv) Effect of the sample size and the number of explanatory variables

We also analysed the effect of sample size on the initial simulation set (from 100 to 10 000). 10 000 observations led to a dramatic increase in the probability of selecting the non-influent variable using AIC, which sharply increased the difference between AIC and SLOO in selecting the best model (see Fig. S1 (c) in Appendix S1); conversely, reducing the number of observations (to 100) led to reducing the difference between AIC and SLOO (see Fig. S1 (a) in Appendix S1). This could be explained by the fact that observations were chosen randomly on the grid and were thus far apart for low sample size, which reduced the impact of RSA.

Including higher numbers of explanatory variables (10 influent variables and 10 non-influent ones) did not affect the precedent results (compare Fig. S2 in Appendix S2 versus Fig. 3). It was expected because variables were independent.

3) Application to a real case study

We applied AIC and SLOO to a dataset from a French national survey of breeding diurnal raptors (Thiollay & Bretagnolle, 2004; Le Rest *et al.*, 2013). Our aim was to identify the environmental variables affecting abundance of the most abundant raptor that breeds in France, the Common Buzzard *Buteo buteo*. We used 1206 sampled quadrats of 5 x 5 km (Fig. 6) and twenty heigh environmental variables (nineteen climatic and nine land use variables) suspected to influence raptor abundance. A Principal Component Analysis (PCA) was performed separately on each environmental dataset (climate and land use) because high correlations occurred between initial environmental variables. These principal components were then used in place of the initial environmental variables for analysis. The principal components were labelled as follow: "ClimDim.x" denoted the x^{st} principal component from the climate dataset and "ClcDim.x" was used in the same way for the land cover dataset.

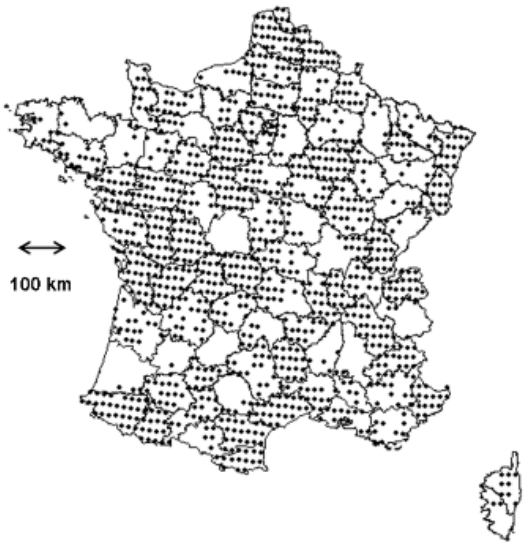


Figure 6: Map of the 1206 sampling quadrats of 5 x 5 km over France (Projection: Lambert Azimuthal Equal Area, ETRS89, EPSG3035). The minimal and maximal distances between observations are respectively 15 km and 1200 km.

Variable selection was performed by assuming a Poisson distribution in a GLM framework. An automated forward step by step algorithm was used in order to reduce the computation time. We first used AIC without accounting for RSA and then, in order to access the performance of the SLOO for this dataset, we tested several threshold distances for SLOO, from 0 to 630 km (respectively, the minimum and the maximum distance at which SLOO could be used) every 15 km step (the minimal distance between actual observations). In each case, only the best model, either minimizing AIC or maximizing $SLOO_{logLik}$, was retained for simplicity.

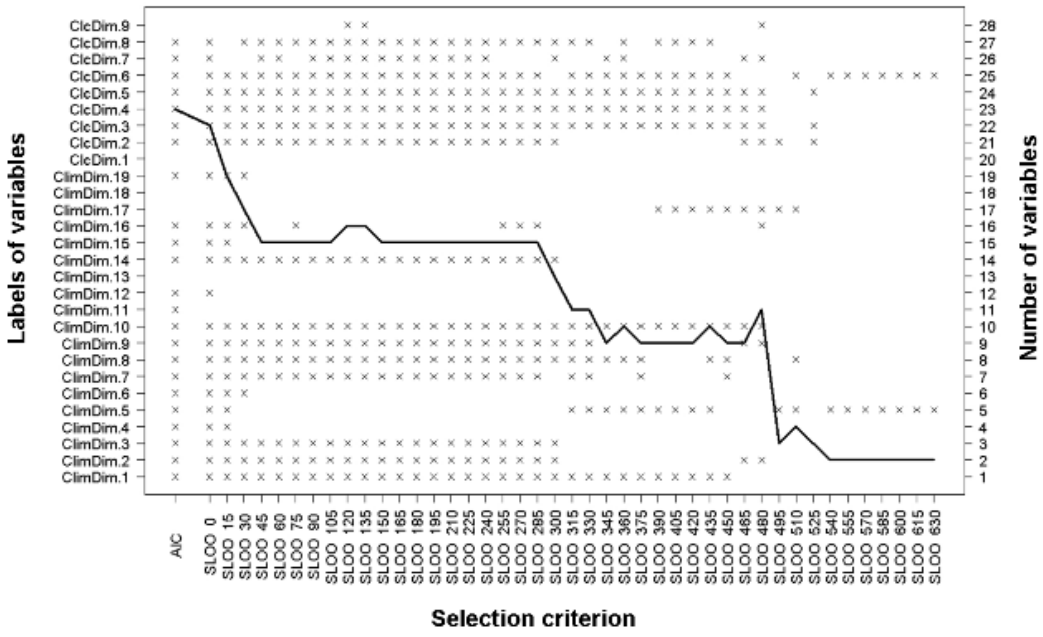


Figure 7: List and number of variables selected in the model by using two selection criteria: the Akaïke information criterion (AIC) and the spatial-leave-one-out used with different threshold distances (SLOO xxx). The abbreviation SLOO xxx denotes the use of SLOO with a threshold distance of xxx km. The black line represents the number of variables selected in the best model depending on the selection criterion considered and the marks identify the labels of the variables selected.

The use of AIC without accounting for RSA led to the selection of twenty three variables in the model (Fig. 7). All but one of these variables were also selected by using LOO (*SLOO 0* in Fig. 7), i.e. without threshold distance. The variable that was not selected by LOO only reduced AIC by 0.14 units, which meant that these two models were almost equivalent. This was expected regarding the asymptotic equivalence between these two criteria (Stone, 1977). Increasing the threshold distance used for the SLOO led then to the selection of fewer variables in the model until reaching fifteen variables when using a threshold distance of 45 km (SLOO 45 in Fig. 7). The expected range of RSA was given by the residuals of the full model since it gave the spatial autocorrelation that could not be accounted for by the available variables. It was between 40 and 50 km (see Fig. 8) emphasizing the fact that SLOO led to select more variables in the model when the threshold distance used was lower than the range of RSA. Conversely when the threshold distance

reached and exceeded the range of RSA (i.e. from 45 to 255 km in Fig. 7), SLOO selected a stable set of variables, with only very marginal differences (differences of $SLOO_{loglik} < 1$). This was in line with the fact that SLOO accounted for RSA when the threshold distance was equal or higher than the range of RSA. Above a threshold distance of 255 km, variables selected in the model became unstable and their number decreased dramatically for threshold distances above 300 km (i.e. 1/4 of the extent of the studied area), which suggested that the SLOO was not efficient with too large threshold distance.

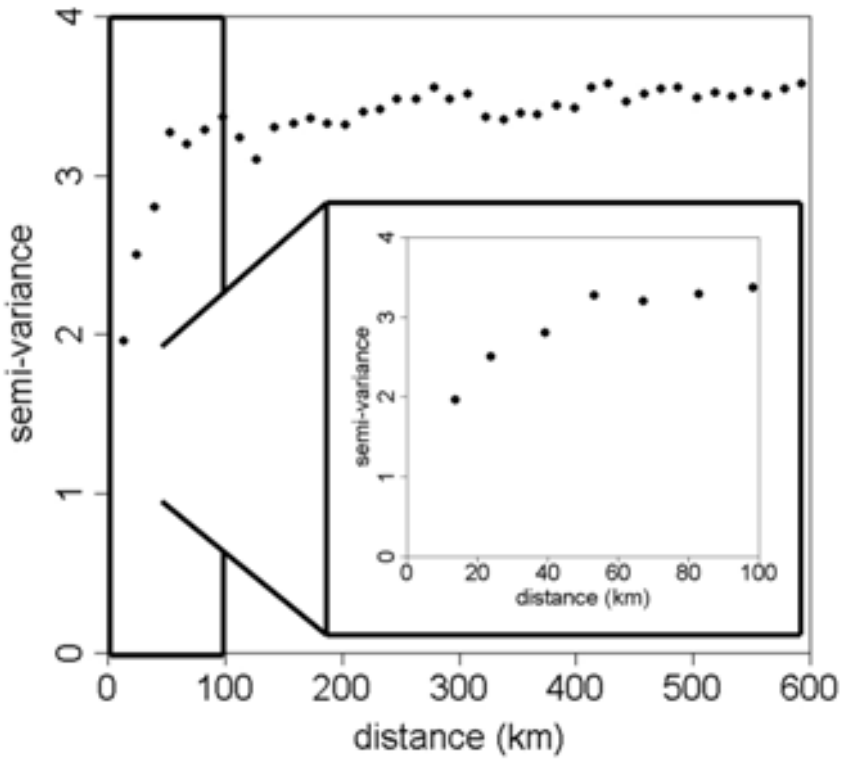


Figure 8: Variogram of the (deviance) residuals of the full model (including all the 28 environmental variables) to explain the abundance of the Common Buzzard *Buteo buteo*.

The Common buzzard is present over the entire territory of France (Thiollay & Bretagnolle, 2004). This species is thus particularly adapted over all climatic constraints of this country. That is why we expected that climatic variables play a rather small role on the explanation of the abundance of this species. However, almost all principal components of climatic variables were selected by using the AIC on this dataset, which could suggest the opposite. These climatic variables were also highly spatially autocorrelated and we showed in the simulations that the spatially autocorrelated variables have more chances to be selected by using AIC in the presence of RSA. The fact that several variables were selected by using the AIC but not selected by using the SLOO with a correct threshold distance should thus be evidence of spatial autocorrelation present in both the residuals and the explanatory variables.

4) Discussion

In the simulations, using AIC for variable selection in the presence of RSA led to select unnecessary variables in the regression models. These results were consistent with the conclusions of Diniz-Filho *et al.* (2008). However, this only occurred when the explanatory variables considered were themselves spatially autocorrelated (see Fig. 5). This could be explained by the fact that two random variables are more likely to be correlated (based on their absolute value of correlation coefficient) when they are both spatially autocorrelated (Liebhold & Sharov, 1998). The correlation between the explanatory variables and the residuals of the regression model was thus inflated in presence of spatial autocorrelation. The highly spatially autocorrelated variables were then more often selected (see Fig. 5). Conversely, SLOO had similar performances whether the variables were themselves spatially autocorrelated or not (see Fig. 5), providing a great alternative to AIC for variable selection in the presence spatial autocorrelation. However, SLOO became less efficient when the threshold distance increased (see Fig 4). This phenomenon resulted from the fact that increasing the threshold distance reduced the number of observations in the training set. When the training set had only a few observations, the estimated parameters were quite unstable between samples and $SLOO_{logLik}$ was improved by chance. For the same reasons, SLOO could not be used when the threshold distance exceeded one half of the extent of the studied area (there were no observations in the training set).

The results of the case study were highly concordant with the simulations despite the important theoretical constraints of the simulations that are never entirely verified with a real dataset, e.g. random sampling in space, stationarity and isotropy. Even if the truth remains unknown for the real dataset, we found evidence that the use of AIC led to keep unnecessary climatic variables for explaining the abundance of the Common Buzzard (see Fig. 7). It was no coincidence that these variables were also highly spatially autocorrelated. Spurious inclusion of meaningless variables in a model may lead to misguided statistical inference (Johnson & Omland, 2004). Using AIC for variable selection in this case study thus reduced the ability of the data collected to bring relevant ecological information on species. Conversely, SLOO appeared useful for variable selection as soon as the threshold distance exceeded the range of RSA. It was however not relevant when the threshold distance was lower than the range of RSA since many unnecessary variables were still selected in the model. Moreover, SLOO became unstable when the threshold distance exceeded 1/4 of the extent of the studied area (about 250 km here, see Fig.6).

The modification of LOO by removing the non-independent data between the training and validation sets has initially been proposed in non-spatial settings (see Chu & Marron, 1991; Burman *et al.*, 1994). It has been already mentioned that removing too much data may impact the effectiveness of the expected prediction error and a limit of 1/4 has also been evoked by Burman *et al.* (1994). SLOO thus appears a safe technique for variable selection when the range of RSA not exceeds 1/4 of the extent of the studied area. This limit is not so restrictive since spatial tools (such as the variogram estimation) become anyway less efficient

when the range of RSA exceeds 1/3 of the extent of the studied area. Moreover, in most of ecological applications, the range of RSA rarely exceeds 1/4 of the extent of the studied area, which allows using SLOO in almost all ecological case studies. It is easy to use since it computes GLMs and only needs the range of RSA as additional spatial information (used as threshold distance). A prerequisite is however to correctly estimate the range of RSA. We propose to use the range of RSA on the residuals of the full model, i.e. the model including all available variables, because it gives the RSA that cannot be accounted for by the available variables. Note that this strategy may underestimate the true range of RSA by including unnecessary spatially autocorrelated variables in the model. Caution must thus be taken in establishing the threshold distance used with SLOO, and one must keep in mind that if SLOO appears robust when estimating the range of RSA at an upper limit, it may not be robust against underestimation (see Fig. 7). The performance of SLOO may also depend on how the space has been sampled. All spatial tools are affected by irregular sampling and we do not expect that SLOO have more concerns than other methods. Further investigation of SLOO performance on irregular spatial datasets remains necessary to confirm this claim.

In most of situations, SLOO can thus be used for variable selection in the presence of RSA instead of computing all candidate models in a spatial explicit framework. Even if it has a real advantage in terms of computation time, it does not address, however the problem of correctly modelling the RSA. This can be seen both as an advantage and an inconvenient: an advantage because it avoids the choice between the many spatial explicit methods that are available, which may give different results (see Diniz-Filho *et al.*, 2008); but also an inconvenient because it prevents to understand the unknown ecological processes having generated the RSA. SLOO is thus only the first step of the statistical analysis. Once the variables are selected, it remains to use a spatially explicit framework to correctly modelling the RSA and make correct inference from the dataset. We advise to use spatial explicit methods able to deal with the 'spatial confounding' effect, e.g. by introducing a spatial term that is orthogonal to the variables considered (see Reich *et al.*, 2006; Hughes & Haran, 2013).

Acknowledgments

We thank the 'Agripop' team of the CEBC for their helpful comments and in particular Luca Borger for its proofreading of the manuscript. We also gratefully thank Creagh Breuner for the English revision. Finally, we thank the 'Région Poitou-Charentes' and the 'Conseil Général des Deux-Sèvres' for funding K. Le Rest PhD grant.

References

voir page 125.

Appendix S1: Extensive simulations for different amount of data

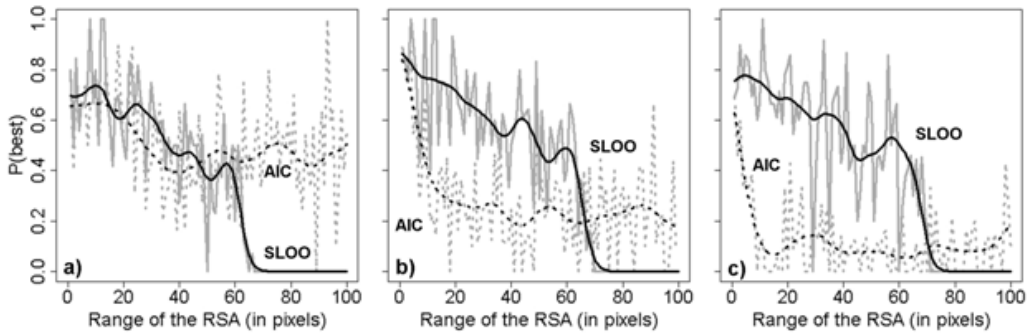


Figure S1: Probability (P) to select the best model depending on the range of the residual spatial autocorrelation (RSA), by using 100 (a), 1000 (b), 10 000 (c) random observations and two selection criteria: Akaïke information criterion (AIC, dotted lines) and spatial-leave-one-out (SLOO, solid lines). Black lines are cubic smoothing splines and grey consider the range of RSA as a factor (measure of the variability). 1000 iterations were each time done.

Appendix S2: Extensive simulations for a higher number of variables.

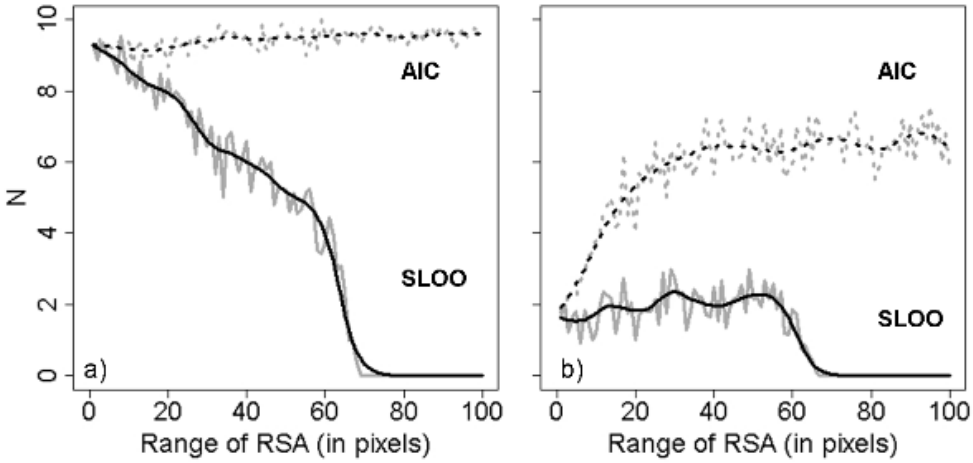


Figure S2: Number (N) of influential variables (a) and non-influential ones (b) selected in the model depending on the range of the residual spatial autocorrelation (RSA), by using two selection criteria: the Akaïke information criterion (AIC, dotted lines) and the spatial-leave-one-out (SLOO, solid lines). Black lines are cubic smoothing and grey lines consider RSA as a factor (measure of the variability). 1000 iterations were done.

Supporting information (online only)

Three R files:

- ‘split.r’ splits the initial data in training sets by using a threshold distance. Return a ‘list’ of training sets.
- ‘sloo.r’ computes SLOO logLikelihood from a model of class ‘lm’ or ‘glm’. Currently, only the Gaussian (family = ‘gaussian’), the Bernouilli (family = ‘binomial’ with y in 0/1 coding scheme) or the Poisson (family = ‘poisson’) distributions are allowed. It needs the object return by ‘split.r’.
- ‘example.r’ compares SLOO and AIC performances from a simulated gaussian data. It allows finding the results of the simulations.

CHAPITRE 3 : UTILISATION DU SLOO EN PRESENCE DE SURDISPERSION

Le deuxième chapitre de cette thèse a démontré la capacité du *spatial-leave-one-out* (SLOO) à sélectionner les bonnes variables explicatives même en présence d'autocorrélation spatiale résiduelle, du moins lorsque celle-ci n'avait pas une trop grande portée. Cependant, ni le premier, ni le second chapitre ne s'intéressent à la possible présence de surdispersion, en plus de l'autocorrélation résiduelle. Nous avons déjà souligné dans le premier chapitre que la présence d'autocorrélation spatiale résiduelle (positive) peut générer de la surdispersion lors de l'analyse de données de comptages. Corriger l'autocorrélation résiduelle devrait dans ce cas corriger la surdispersion. Néanmoins la surdispersion peut aussi se produire en absence d'autocorrélation résiduelle. Ces deux caractéristiques, bien que liées dans une certaine mesure, peuvent donc être conjointement présentes dans un modèle de régression, ce qui nécessite d'en tenir compte simultanément.

Ce troisième chapitre s'intéresse à la manière d'utiliser le SLOO pour la sélection de variables dans le cas où l'analyse des données de comptages par un modèle de régression de Poisson conduit à de la surdispersion. L'article présenté ci-après sera bientôt soumis à un journal mais des modifications ne sont pas exclues.

Dealing with overdispersion and spatial autocorrelation during the variable selection step.

Authors: Le Rest Kévin*, Pinaud David and Bretagnolle Vincent
Centre d'Etudes Biologiques de Chizé (CEBC), CNRS UPR 1934, 79360 Beauvoir-Sur-Niort, France
Email addressees: lerest.k@gmail.com, pinaud@cebc.cnrs.fr and breta@cebc.cnrs.fr
*Corresponding author:
Email: lerest.k@gmail.com
Tel: +33 (0)5 49 09 35 13
Fax: +33 (0)5 49 09 65 26

Manuscript in progress / March 2014

Overview

Abstract

Variable selection is often needed to draw reliable statistical inference from a regression model. However, the analysis of spatial count data are very often confronted to important problems such as overdispersion, i.e. a variance higher than expected by the model, and residual spatial autocorrelation, i.e. the presence of strong spatial dependences in model residuals. Both of these features may lead selecting unnecessary variables in the final model, artificially increasing model complexity. Performing variable selection within a generalized linear model (GLM) framework in the presence of overdispersion and spatial autocorrelation is still challenging. Difficulties arise due to the high computation time needed for computing spatial explicit GLMs. An alternative method for variable selection in the presence of spatial autocorrelation consists in using a spatial-leave-one-out cross-validation (SLOO, Le Rest *et al.* 2014) where autocorrelated data are removed from the training set. It has the main advantage to compute only classical GLMs, alleviating the model estimation difficulties during the variable selection step. We propose extending such method to the case where overdispersion also occurs. We first evaluated some ways to compute the selection criteria in the presence of overdispersion by using simulations. We then applied the SLOO on a real dataset involving both strong spatial autocorrelation and high overdispersion, by whether accounting for overdispersion or not. Results showed that adding an overdispersion parameter to the model, e.g. using a negative binomial model instead of a Poisson model, is not needed for computing the SLOO criterion. However, the model must account for the non-stationnary effects of variables affecting the studied process in order to guaranty a reliable variable selection.

Introduction

When building a regression model for studying an ecological process, a critical step concerns the choice of a relevant set of explanatory variables. These variables must be chosen *a priori*, i.e. one must suspect their influence on the studied process. However, using all of them in the same regression model may lead to useless statistical inference, inflating the variance of the model so much that the influence of variables can no longer be detected (i.e. the so-called bias/variance trade-off, see Burnham & Anderson 2002). Variable selection is thus often necessary to draw reliable statistical inference from a dataset. Many metrics have been proposed to help in this task (see Kohavi 1995; Shao 1997; McQuarrie & Tsai 1998; Burnham & Anderson 2002; Arlot & Celisse 2010). They usually consist in calculating a balance between the model fitting and its complexity, i.e. following the principle of parsimony or Ockham's razor (Forster 2000, 2001; Burnham & Anderson 2002; Johnson & Omland 2004). The Akaïke information criterion (AIC, Akaïke 1974) is the most often used criterion for regression models. It is actually an approximation of the Kullback-Leibler divergence (KLD, Kullback & Leibler 1951), i.e. the divergence between the true (but unknown) distribution of the studied process and the one provided by the statistical model used (see Burnham & Anderson 2001; Richards 2008). This KLD is reliable for model comparison without having to assume that the true model is available from the set of candidates (see Burnham & Anderson 2001). It is thus well suited for variable selection in biological sciences where the true model is usually unavailable (Burnham & Anderson 2001, 2002).

When analysing count data with a Poisson regression model, one assumes that the variance is equal to the mean. But it rarely happens in practice since such model is almost always confronted to a problem of overdispersion, i.e. the observed variance is greater than the one assumed by the model (see Hinde & Demétrio 1998; Ver Hoef & Boveng 2007; Cameron & Trivedi 2013). Overdispersion may come from the fact that important but unobserved variables are not in the model, which produces unobserved heterogeneity (Hinde & Demétrio 1998; Richards 2008; Cameron & Trivedi 2013). It leads to an overconfident statistical inference, which can in turn invalidate ecological conclusions (Burnham & Anderson 2002). Moreover, using AIC or other related metrics for variable selection in the presence of overdispersion may lead to the selection of overly complex models (Anderson *et al.* 1994; Richards 2008). Several methods were proposed to account for overdispersion, such as the quasi-likelihood approach (see Wedderburn 1974) or the negative binomial model (see Greene 2008 for a recent overview). The former involves dividing the estimated log-Likelihood of the Poisson regression model by a correction term (the overdispersion parameter), which is derived from the Poisson model and is equal to the ratio between the residual deviance and its residual degree of freedom. Its value is expected to be close to 1 in absence of overdispersion. Model comparison can then be performed by using the quasi-AIC (QAIC, see Lebreton *et al.* 1992; Anderson *et al.* 1994), which has been proved efficient in

many different situations (see for example Richards 2008). Switching from the Poisson model to a more suitable one, such as the negative binomial model, is a widely used alternative to the quasi-Likelihood framework. The dispersion parameter is this time estimated during the model estimation as other model parameters. This allows calculating a true log-Likelihood and then using AIC for model comparison (see Ver Hoef & Boveng 2007; Richards 2008 for a comparison between these two approaches).

In addition to overdispersion, spatial autocorrelation is another frequent characteristic of ecological count data, leading to unreliable parameter estimation when the residuals of the model are not independent (see Griffith 2006a). The presence of positive residual spatial autocorrelation (spatially close locations having more similar residuals than remote locations) leads to a similar outcome as the presence of overdispersion (Haining *et al.* 2009). For instance, it leads to the selection of overly complex models when using AIC for variable selection (Hoeting *et al.* 2006; Cassemiro *et al.* 2007; Diniz-Filho *et al.* 2008). Again, there are several methods available to deal with the residual spatial autocorrelation in regression models, often called 'spatial explicit models' (see Dormann *et al.* 2007; Betts *et al.* 2009; Beale *et al.* 2010). For generalized linear models, such as the Poisson or negative binomial regression, two main methods are claimed to be useful: the first one consists in using additional artificial spatially autocorrelated variables, known as spatial filters, in order to remove the residual spatial autocorrelation (Griffith & Peres-Neto 2006; Dray *et al.* 2006); and the second one consists in adding a spatially structured random effect accounting for the residual spatial autocorrelation (see Diggle *et al.* 1998). With the first method, the user still keeps a generalized linear model framework whereas the second method uses a generalized linear mixed model framework, the latter being computationally challenging hence relying usually on approximations (see Bolker *et al.* 2009).

Analysing count data in spatially explicit context thus requires, accounting simultaneously for both overdispersion and residual spatial autocorrelation. This may be done by combining methods for each problem, e.g. by using a spatial explicit negative binomial model (see Gschlößl & Czado 2008; Haining *et al.* 2009; Neyens *et al.* 2012). The latter might be convenient for estimating model parameters, but it is however less convenient at the variable selection step due to the high computation time needed for spatially explicit generalized linear models. To reduce the computation time, Diniz-Filho *et al.* (2008) proposed to fix the spatial term across all candidate models. As acknowledged by the authors themselves however, this is not optimal since the amount of residual spatial autocorrelation may depend on the variables included in the regression model. Recently, Le Rest *et al.* (2014) have proposed using an alternative technique for variable selection in the presence of residual spatial autocorrelation. It consists in applying a leave-one-out (LOO, see Stone 1974) cross-validation in a spatial context, by removing the part of the data that is not independent from the point leaved out. This method, called spatial leave-one-out (SLOO), is very much related to AIC since these metrics are asymptotically equivalent in the absence of residual spatial autocorrelation (Stone 1977; Le Rest *et al.* 2014). SLOO seemed thus a promising tool to

perform model selection in the presence of residual spatial autocorrelation but its efficiency was not assessed yet when used in the presence of both residual spatial autocorrelation and overdispersion. This is precisely the aim of the present study, i.e. extend the use of SLOO to perform variable selection in the presence of both overdispersion and residual spatial autocorrelation.

An intuitive modification of the SLOO allowing for overdispersion involves computing the variable selection criterion with a negative binomial model instead of the Poisson one. We first perform a simulation studying the performances of the classical LOO in the presence of overdispersion but in the absence of residual spatial autocorrelation. AIC and LOO are compared as variable selection criteria with a Poisson or a negative binomial regression model and the probability in selecting a variable against its true effect is computed. A similar simulation comparing SLOO and AIC in the presence of both overdispersion and residual spatial autocorrelation is provided in Appendix A. These two simulations allow discrediting in one hand the effect of overdispersion and in other hand the effect of residual spatial autocorrelation. Second we compare the performances of AIC, LOO and SLOO on a real count dataset presenting both important overdispersion and spatial autocorrelation when analysed with a non-spatial Poisson regression model. The dataset reports on the measured abundance of a diurnal raptor over the entire country of France breeding in France, the Montagu's Harrier *Circus pygargus*. Satellite data on environmental variables are extracted in order to explain the abundance of this species in France and the aim is to identify a relevant subset of such variables.

1) Simulations

We simulated 10 000 count data with 100 observations each, obtained from a negative binomial distribution (NB) such as:

$$log(\lambda) = \beta_1 X_1 \; ; \; Y \sim NB(\lambda , k) \qquad \textbf{Equation 1}$$

where Y was the count data, λ was the mean of the count and k was the dispersion parameter. This latter was parameterized such as the variance of the count was equal to $\lambda + \lambda^2 / k$ (the default parameterization in R software, see Bolker 2008 p.124). $NB(\lambda , k)$ is thus equivalent to $Poisson(\lambda)$ when $k \rightarrow + \infty$. Here we set k equal to 0.5, i.e. a rather high overdispersion. $X_1 \sim N(0,1)$ was a random variable influencing λ and its importance depended on β_1, which was chosen randomly between 0 and 0.99 each 0.01, giving 100 possibilities. When $\beta_1 = 0$, X_1 had no effect on λ, whereas when $\beta_1 = 0.99$, it had a strong effect increasing λ by exp(0.99), i.e. 169%, when X_1 increased by one.

For each 10 000 datasets, four variable selection procedures were run: two procedures using the AIC considering either a Poisson or a negative binomial regression model; and two using the LOO instead of AIC. Note that the selection criterion of the LOO was a cross-validated log-Likelihood (as in Le Rest *et al.* 2014), which was computed with either a Poisson or a negative binomial model distribution. The probability to select X_1 was then

computed for each β_1 value as the number of times (divided by the number of simulations done for the β_1 value) when adding X_1 in the model reduced the selection criterion (either AIC or LOO).

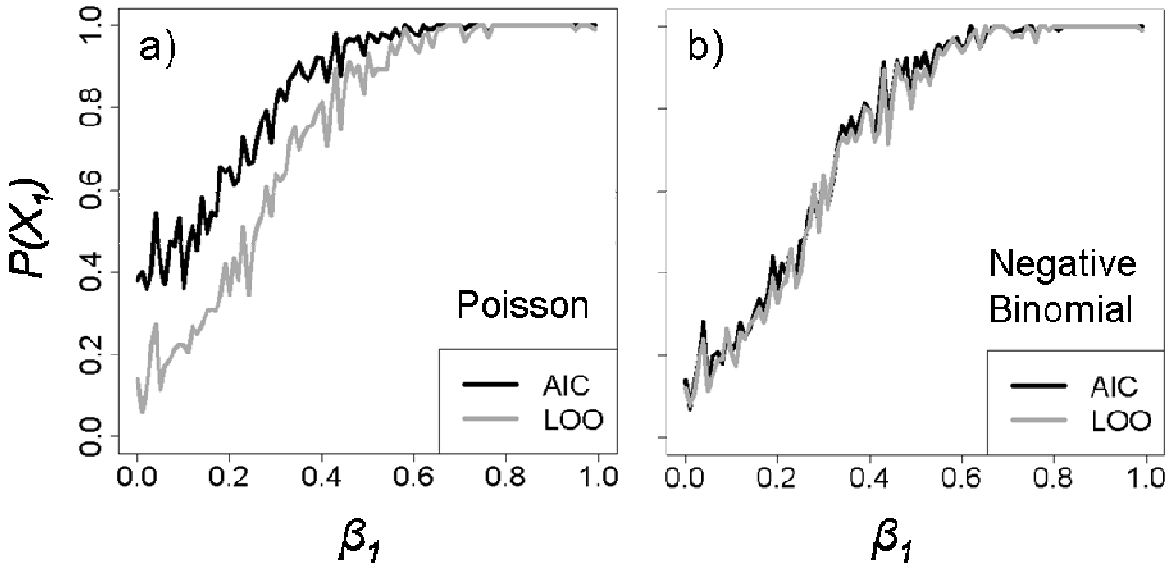


Figure 1: Probability (P) to select X_1 depending on its regression coefficient (β_1), in the presence of overdispersion. The Akaïke information criterion (AIC, in black) and the leave-one-out cross-validation (LOO, in grey) are confronted by using two different models: the Poisson model (a) and the negative binomial one (b). The lines represent the mean probability for each of the β_1 modalities (from 0 to 0.99 each 0.01). This allows giving a graphical representation of the mean but also of its variability. A similar Figure but in the additional presence of residual spatial autocorrelation is given in Figure A1 of Appendix A.

Considering a negative binomial regression model (Fig.1b) led to the same capacities in selecting X_1 by using either AIC or LOO. Since the negative binomial model accounted for overdispersion, the equivalence between these two selection criteria was confirmed for count data in the absence of overdispersion (see also Stone 1977 for the mathematical proof of equivalence on a Gaussian data). Conversely, outcomes were different with the Poisson model, which did not accounted for overdispersion (Fig.1a). Indeed, using LOO with a Poisson model gave always a lower probability than using AIC in selecting X_1, which outlined that LOO and AIC were not equivalent in the presence of overdispersion. Especially, using the AIC with a Poisson model inflated the importance of X_1 when it had actually no effect ($P(X_1)$ was about 0.4 when $\beta_1 = 0$, see Fig.1a). This was expected since AIC is known overselecting variables in the presence of overdispersion (see Anderson *et al.* 1994). However, using LOO with a Poisson model gave a reliable probability to select X_1, i.e. a similar probability than by using a negative binomial model (compare Fig.1a and Fig.1b).

2) A case study

The data used came from the French national survey of breeding diurnal raptors (Thiollay & Bretagnolle 2004; Le Rest *et al.* 2013). A total of 1206 quadrats of 5 x 5 km were sampled between 2000 and 2002 over the country (see Fig.2). Here we studied the spatial distribution of a raptor species, the Montagu's Harrier, which showed a strong spatial autocorrelation (about 100 kilometres, see Figure B1 in Appendix B) and a high overdispersion (negative binomial dispersion parameter about 0.45; see also the count frequency, Figure B2 in Appendix B) when analysed with a non-spatial Poisson regression model. A set of 28 environmental variables (19 climatic and 9 land use variables) were extracted in order to explain the abundance of this raptor (additional details on these datasets can be found in Le Rest *et al.* 2013). A Principal Component Analysis (PCA) was performed separately on each environmental dataset (climate and land use) because high correlations occurred between initial environmental variables. These principal components were then used in place of the initial environmental variables for analysis. The principal components were labelled as follow: "ClimDim.x" denoted the x^{st} principal component from the climate dataset and "ClcDim.x" was used in the same way for the land cover dataset.

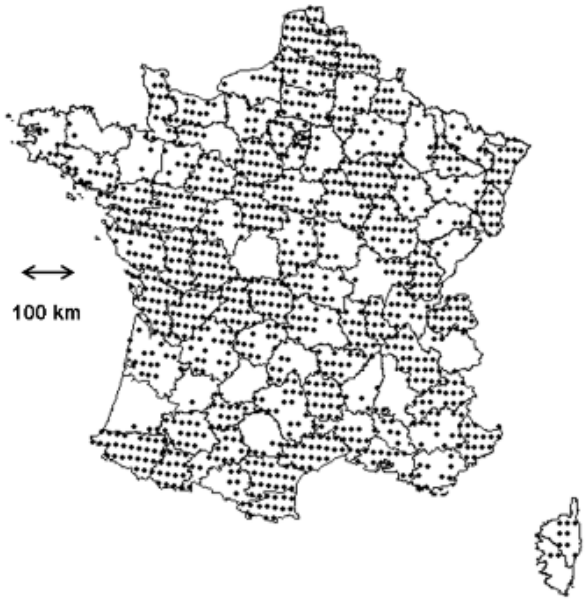


Figure 2: Map of the 1206 quadrats sampled across France between 2000 and 2002. Each black point represents a quadrat.

Variable selection was performed by using AIC, LOO and SLOO. AIC and LOO were used by ignoring residual spatial autocorrelation whereas SLOO (see Le Rest *et al.* 2014) was used with a threshold distance of 100 km between the points left out and the remaining data, i.e. the observed range of residual spatial autocorrelation (see Figure B1). AIC, LOO and SLOO were used by considering either a Poisson or a negative binomial model. An automated forward step by step algorithm was used in order to reduce the computation time. For each case, only the best model, either minimizing AIC or maximizing LOO and SLOO log-Likelihood, was retained for simplicity.

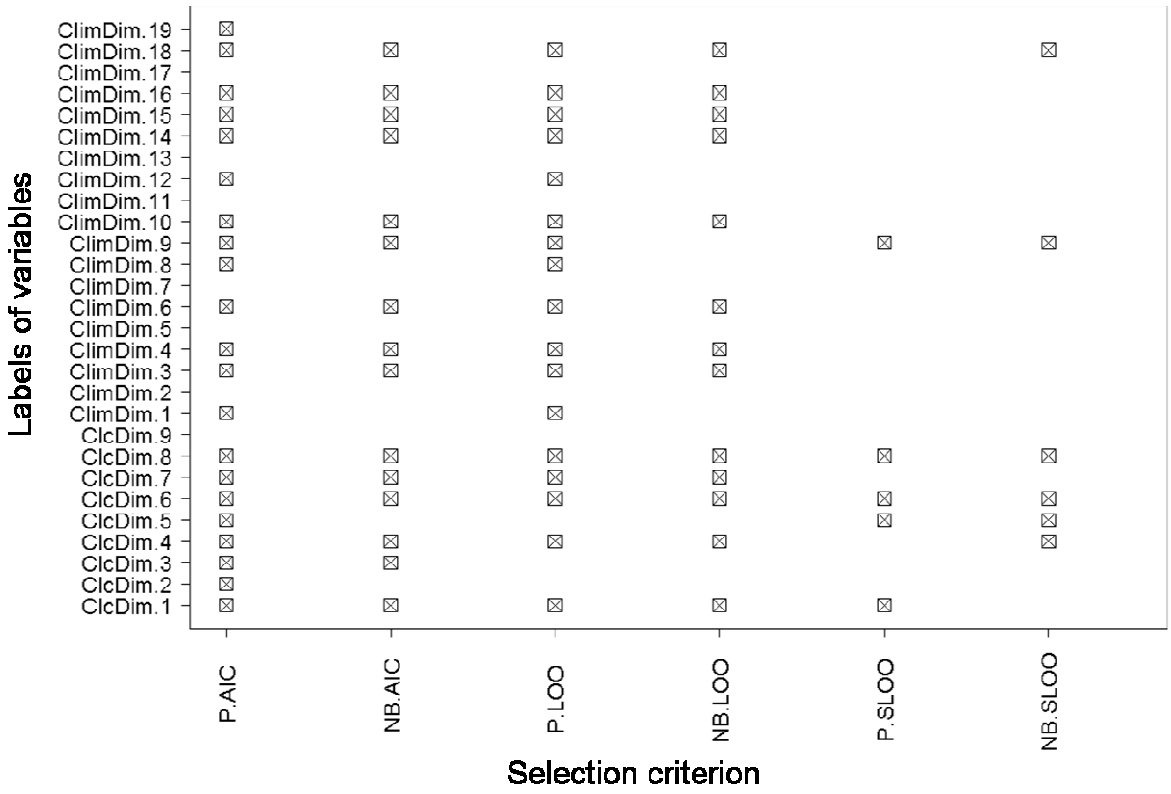


Figure 3: Variables selected among the 28 candidate variables according to the different model selection criteria used: the AIC, LOO or SLOO with either the Poisson (P) or the negative binomial (NB) distribution. 'P.AIC' denotes the use of AIC with a Poisson model. The boxes represent the variable selected for each selection criterion considered.

This case study underlined how much the variable selection criterion used may influence the complexity of the model retained in the presence of both overdispersion and spatial-autocorrelation. Using AIC with a Poisson regression model for variable selection (P.AIC in Fig.3), i.e. not accounting for either overdispersion or spatial autocorrelation, led to a very high number of selected variables (21). Though using LOO with a Poisson model (P.LOO in Fig.3) not explicitly accounted for overdispersion nor spatial autocorrelation, it slightly decreased the number of selected variable (17). Using AIC or LOO with a negative binomial model (respectively NB.AIC and NB.LOO in Fig.3), which accounted for overdispersion but not for spatial autocorrelation, resulted in respectively 15 and 13 selected variables. Computing an AIC (assuming a negative binomial regression model) from the set of variable select by NB.LOO showed that these two sets of variables had actually a similar AIC (ΔAIC = 0.52). The SLOO criteria (P.SLOO and NB.SLOO in Fig.3), the only ones accounting for spatial autocorrelation, led decreasing drastically the number of variable selected since only 5 variables were selected with a Poisson model and 6 with a negative binomial model.

3) Discussion

Simulations showed that AIC should be avoided for variable selection with a Poisson regression model in the presence of overdispersion. It led selecting non-influential variables with a quite high probability, which in turn may led considering overcomplex or even false models for statistical inference, a problem already known (see Anderson *et al.* 1994; Richards 2008). The negative binomial regression model accounted for overdispersion and thus corrected for this undesirable effect. Using either LOO or AIC criteria led to very similar outcomes when using this model. This implied that if applied with a negative binomial regression model, the modification of the LOO in a spatial context (SLOO, Le Rest *et al.* 2014) will be able to perform variable selection while accounting for both overdispersion and residual spatial autocorrelation. This assertion was confirmed by the additional simulations shown in Appendix A1. Using LOO with a Poisson model provided an alternative way to perform variable selection in the presence of overdispersion. Indeed, simulations showed that using LOO with a Poisson model gave similar results than using it with a negative binomial model (see Fig.1). This was especially valuable in term of computation time since computing Poisson models is much easier and faster than computing negative binomial models. Using SLOO with a classic Poisson regression model seems therefore another relevant way to perform variable selection in the presence of both overdispersion and residual spatial autocorrelation. This assertion was also confirmed by the additional simulations shown in Appendix A1.

The case study presented here mainly confirmed the results obtained with simulations. Indeed using AIC with a Poisson regression model led keeping a very high number of variables in the final model (21, see Fig.3), which decreased to some extant when using AIC with a negative binomial model (15) or when using LOO with a Poisson model (17). Using SLOO with a Poisson or a negative binomial regression model led keeping the minimal number of variables (5 and 6 variables respectively). Apart the varying number of variables kept in final models according to selecting procedure, another difference between the models concerned the identity of the variables selected. Especially, a single variable was selected with all metrics, though not when using SLOO with a negative binomial model (ClcDim.1, see Fig.3). This variable reflected a natural gradient (low values indicating intensive farming and high values, large forests) and was supposed to affect the abundance of Montagu's Harrier, which is a raptor of open landscapes. The explanation on why it was not selected by using SLOO with a negative binomial model likely came from its heterogeneous effect. Indeed, in some part of France, this variable has a negative effect, i.e. Montagu's Harrier mainly breeds in intensive agricultural landscapes. But in others part of France, it breeds in more natural lands and even sometimes in forest cuts. This variable had thus apparently not a sufficient homogeneous effect to be detected when using SLOO with a negative binomial model. One way to account for this non-homogeneity would be to include some relevant interactions in the model.

In conclusion one should retain that even if computing SLOO with either the Poisson or the negative binomial model give very similar outcomes for variable selection in simple cases (see simulations), these two approaches are not strictly equivalent with real data (see the case study). This may come from variables having non-stationnary effects along the study area (such as ClcDim.1 in our case study). A critical step is thus to allow for these complex effects to be incorporated in the model. If this is done properly, then computing SLOO with either the Poisson or the negative binomial regression model will give the same outcome (see Appendix A). The Poisson model keeps an advantage in term of computation time.

Acknowledgements

Not yet provided.

References

voir page 125.

Appendix A

We simulated 10 000 count data with 500 observations each, obtained from a negative binomial distribution (NB) such as:

$$log(\lambda) = \beta_1 X_1 + \varepsilon \; ; \; Y \sim NB(\lambda, k) \qquad \textbf{Equation A1}$$

Equation A1 is similar to Equation 1 but with an additional error term; $\varepsilon \sim N(0,1)$ have a short-range spherical spatial structure, which generated the residual spatial autocorrelation. The range was set at 1/5 of the width of the spatial domain. Moreover, X_1 was computed with a long-range spherical spatial structure (the range was the width of the spatial domain).

For each 10 000 datasets, four variable selection procedures were run: two procedures using the AIC considering either a Poisson or a negative binomial regression model; and two using the SLOO instead of AIC.

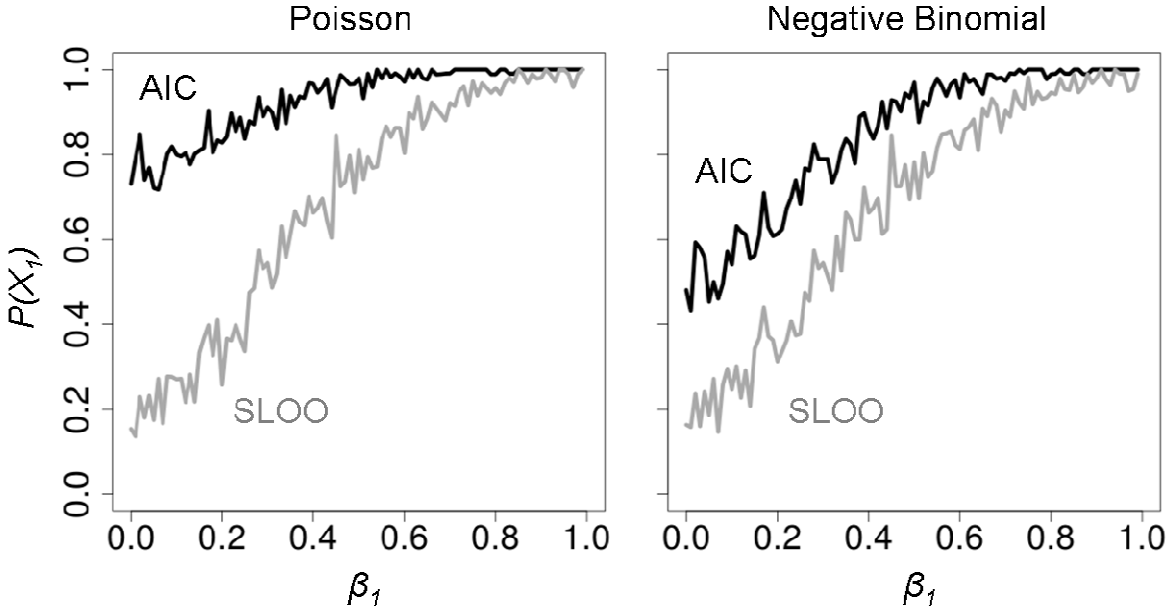


Figure A1: Probability (P) to select X_1 depending on its regression coefficient (β_1), in the presence of overdispersion and residual spatial autocorrelation. The Akaïke information criterion (AIC, in black) and the spatial-leave-one-out cross-validation (SLOO, in grey) are confronted by using two different models: the Poisson model (left) and the negative binomial one (right). The lines represent the mean probability for each of the β_1 modalities (from 0 to 0.99 each 0.01). This allows giving a graphical representation of the mean but also of its variability.

The SLOO allowed selecting X_1 with similar probabilities whatever the model used and gave a suitable probability when $\beta_1 = 0$.

Appendix B

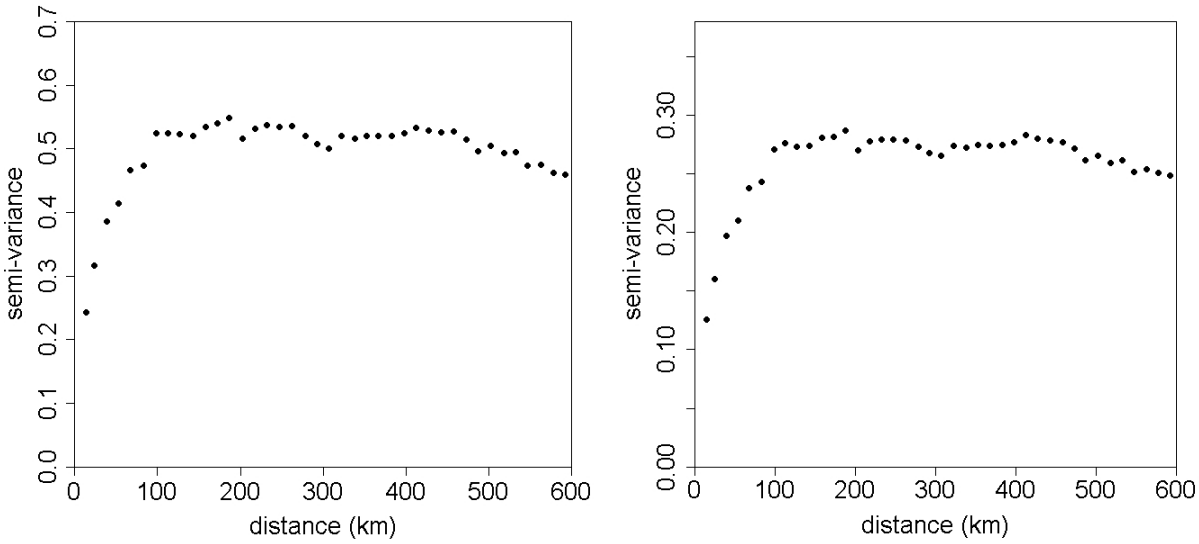


Figure B1: Variogram of the residuals of the model including all variables, i.e. the spatial autocorrelation that can not been account for by the available variables. A Poisson regression model (left) and a negative binomial one (right) are used, giving the same shape of residual spatial autocorrelation.

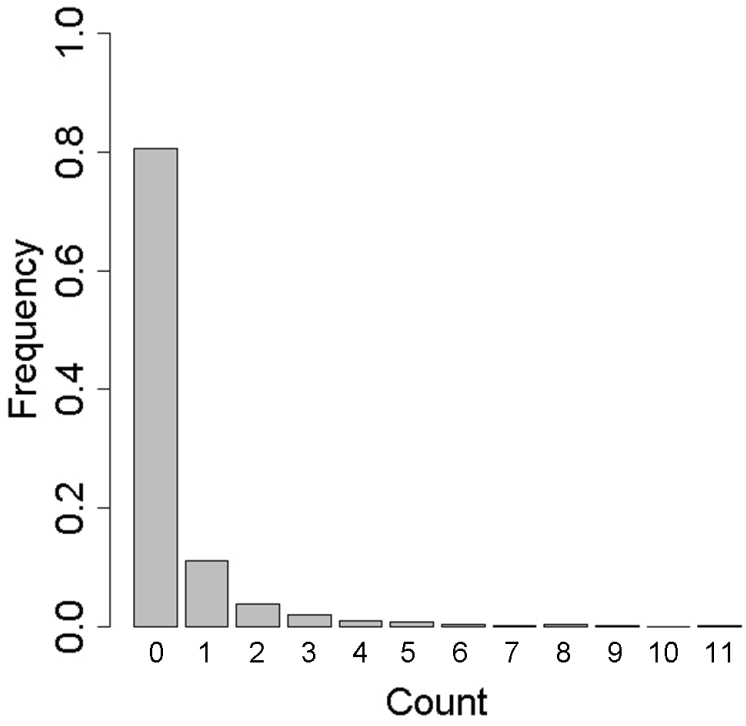


Figure B2: Frequency of observed counts.

CHAPITRE 4 : LIMITES D'UTILISATION DES MODELES ZERO-ENFLES A MELANGE

Dans le troisième chapitre, nous avons abordé les problèmes liés à la présence de surdispersion dans un modèle de régression et discuté la manière d'utiliser le *spatial-leave-one-out* (SLOO) dans un tel cadre. Un aspect encore ignoré jusqu'alors est la possible présence d'un cas particulier de surdispersion, lié à la présence d'un surnombre de zéros : l'inflation en zéro. Ce quatrième chapitre s'intéresse à la relation entre surdispersion et inflation en zéro, et aux possibles confusions que cela peut générer dans un modèle de régression. Afin de se focaliser sur ces aspects, l'autocorrélation résiduelle sera cette fois absente. Les métriques évaluées concernent d'une part la probabilité de sélectionner les variables et d'autre part la qualité d'estimation des paramètres. L'ajout d'une métrique s'intéressant à la qualité d'estimation cherche à répondre au fait que cet aspect n'est pas beaucoup étudié en présence de surdispersion et d'inflation en zéro alors qu'il a déjà été largement étudié en présence d'autocorrélation résiduelle. Les procédures de sélection de variables ne nécessitant pas l'utilisation du SLOO en l'absence d'autocorrélation résiduelle, il ne sera pas utilisé dans ce chapitre.

L'article présenté ci-après apporte un regard critique sur l'utilisation des modèles zéro-enflés à mélange (Lambert 1992), qui sont de plus en plus utilisés pour modéliser des données ayant beaucoup de zéros. Il sera soumis très prochainement à un journal international.

Zero-inflated mixture models behave wrongly in the presence of overdispersion

Authors: Le Rest Kévin*, Pinaud David and Bretagnolle Vincent
Centre d'Etudes Biologiques de Chizé (CEBC), CNRS UPR 1934, 79360 Beauvoir-Sur-Niort, France
Email addressees: lerest.k@gmail.com, pinaud@cebc.cnrs.fr and breta@cebc.cnrs.fr
*Corresponding author:
Email: lerest.k@gmail.com
Tel: +33 (0)5 49 09 35 13
Fax: +33 (0)5 49 09 65 26

Manuscript in progress / March 2014

Overview

Abstract

Zero-inflated mixture models (ZIM), such as the zero-inflated Poisson (ZIP), are models claimed to account for an excess of zeros in count data, i.e. the zero-inflation issue. They are increasingly used in ecological applications when data have many zeros because of supposed better performance over more conventional models, such as the Poisson and the negative binomial (NB) models. However, ZIM make a critical assumption: one or some processes must affect the probability of having a zero without affecting other count values. If properly tested, this assumption would likely restrict the use of such model to specific cases, yet ZIMs are currently used in situations where this assumption is unlikely to be matched. By using simulations, we critically examine the performance of ZIMs over traditional models when this assumption matched or not. We especially envisage the case where classical overdispersion occurring (a very common case in ecological count data), and consider its effect on variable selection and on parameter estimation. Our simulation results showed that using ZIP in the presence of overdispersion led to a wrong variable selection, misleading zero-inflation, and inaccurate parameter estimates. Using ZIP was even worse than using the conventional Poisson model in some cases. Using the zero-inflated negative binomial model (ZINB) corrected for most of these problems but may still led selecting overcomplex models by including unnecessary terms in the excess-zero-part. The conventional NB performed equally to ZINB in most of situations. Our results imply that the ZIP would be inappropriate for analysing an ecological count data and that the ZINB is only useful when a mechanism can be identified that split the studied population in two subpopulations, one leading to only zeros and one leading to both zeros and positive counts. Such situation may be uncommon in ecological data.

Introduction

Many ecological processes of interest are actually described by a discrete variable, e.g. the abundance of a species. In such cases, it is highly recommended to switch from the use of a classic linear regression model to a generalized linear model, acknowledging both the discrete and positive nature of the response variable (O'Hara & Kotze 2010). However, the classic Poisson regression model may not be appropriate for such data since it assumes equal mean and variance (see Hinde & Demétrio 1998). This assumption is overwhelmingly violated in ecological count data analysis since the observed variance is almost systematically higher than the mean. This phenomenon is known as overdispersion (Hinde & Demétrio 1998; Richards 2008). When modelling a dependent variable using a set of independent variables, overdispersion usually comes from the fact that important but unobserved independent variables (often unavailable variables) are lacking in the model, producing unexplained heterogeneity (Cameron & Trivedi 2013; Hinde & Demétrio 1998; Berk & MacDonald 2008; Richards 2008; Boes 2010). If overdispersion is ignored, e.g. by using a Poisson model, an overconfident statistical inference arises, which may in turn invalidate ecological conclusions (Burnham & Anderson 2002). Unfortunately, the number of unobserved variables is expected to be high in ecological studies, either because mechanisms governing ecological processes are complex or because the measurements are prone to errors, e.g. heterogeneous detection probability for species abundance data. The presence of overdispersion should thus be largely expected when analysing an ecological count data by using a Poisson model. Statistical methods able to correct for overdispersion are available for a long time. For instance the widely used negative binomial distribution (hereafter NB), which use a dispersion parameter to account for overdispersion, have already been described by Fisher (1941) as an extension of the Poisson series for imperfect data (see also Greene 2008 for a recent overview).

When the unobserved variables generate only zeros, overdispersion is referred to zero-inflation (Lambert 1992; Tu 2002). Zero-inflation is thus a special case of overdispersion (see Ridout *et al.* 1998). It can be corrected for by using a model with an additional parameter such as the zero-inflated Poisson model (hereafter ZIP, Lambert 1992). The ZIP assumes that the studied process comes from a mixture of two processes, one generating only zeros with proportion ρ (the excess-zero-part) and another one generating both zeros and positive counts with mean λ (the count-part) (Lambert 1992; Tu 2002). It is also possible to deal with the presence of both overdispersion and zero-inflation by considering that the count-part of the ZIP model is a NB instead of a Poisson distribution, which leads to the zero-inflated negative binomial model (hereafter ZINB, see Greene 1994). Zero-inflated 'mixture' models (hereafter ZIMs), such as ZIP and ZINB, have become fairly popular in ecological literature (Welsh *et al.* 1996; Warton 2005; Martin *et al.* 2005; Rathbun & Fei 2006) and are now widely used (see for example Wenger & Freedman 2008; Sileshi *et al.* 2009; Linder & Lawler 2012). In particular, they have attracted many ecologists who have ever been interested in methods able

to deal with many zeros, a frequent characteristic of ecological count data (Martin *et al.* 2005).

The presence of zero-inflation implies that one or some processes must affect the probability of having a zero without affecting other count values (Lambert 1992; Tu 2002). In some situations, this assumption seems realistic, for instance when studying the number of offspring in a species, one should account for the fact that some individuals are sterile. If an individual is sterile, it has inevitably no offspring whatever the biological variables able to influence its reproductive success. The excess-zero-part of ZIMs should then determine the probability that an individual is sterile. However in most ecological situations, this assumption is questionable: does an unobserved process truly affect the number of zeros without affecting others counts values? For example, several studies on species abundance data have proposed to used ZIMs in order to distinguish the zeros due to detection failure, i.e. false zeros, from the zeros due to ecological processes, i.e. true zeros (Martin *et al.* 2005; Linder & Lawler 2012). However, it is very unlikely that a low detection generates only false zeros without affecting others counts values, leading to the presence of false ones, false twos, *et cetera*. A low detection probability is more likely to affect the overall count data by underestimating the true counts whatever their values (Lancia *et al.* 1994; Schmidt 2003). This rather suggests that ZIMs should not be used in such situations and that a NB is likely more relevant.

In practice, ZIMs are often used by justifying an important number of zeros occurring in the data collected; however this justification is spurious (Warton 2005; Wenger & Freeman 2008). Over and unduly use of ZIMs has recently been criticized (Paul Allison, www.statisticalhorizons.com/zero-inflated-models). It is well admitted that overdispersion and/or zero-inflation lead to erroneous results when not accounted for but actually present (Hinde & Demétrio 1998; Tu 2002; Berk & MacDonald 2008), but the consequences of inadequately correcting for overdispersion and/or zero-inflation have received less interest. Especially the consequences of correcting for zero-inflation whereas there is actually overdispersion or *vice versa*, has not been addressed so far. Further evaluations of ZIMs are thus needed in such situations in order to identify when they truly improve ecological conclusions or conversely, when they may lead to further confusion.

This study explores how ZIMs perform *versus* their non zero-inflated versions in the presence of different levels of overdispersion and zero-inflation. We used a controlled framework by using simulations since it was the only way to fully manage the count data, by using specific unobserved variables producing overdispersion and/or zero-inflation. We focused on the performance of the four commonly used models for the analysis of count data, the Poisson, the NB, and their zero-inflated versions, respectively the ZIP and the ZINB. These models were evaluated by using two statistical measurements: i) their ability to select a variable of interest and ii) the reliability on the estimation of the regression coefficients. Implications of our results for ecological data are then discussed.

1) Materials & Methods

A) The simulated count datasets

The simulated count datasets were generated from a mixture of a Poisson (Pois) distribution and a Bernoulli (*Ber*) distribution, i.e. a zero-inflated count data.

$$Y \sim ZIP(\lambda, \pi) \text{ or equivalently } Y \sim Pois(\lambda) \times Ber(1-\pi) \quad \text{(Equation 1)}$$

count-part (Poisson):	*excess-zero-part (Bernoulli):*
$log(\lambda) = \beta_1 X_1 + X_2$	$\pi = c$

λ was the mean of the count-part and π the probability of excess-zero. For the sake of simplicity, the excess-zero-part had only a constant term, i.e. a constant proportion of excess-zero. This excess-zero proportion was considered either null ($\pi = 0$), low ($\pi = ¼$), moderate ($\pi = ½$) or high ($\pi = ¾$). The count-part resulted from the combination of two random variables: $X_1 \sim N(0, \mathbf{1})$ and $X_2 \sim N(0, \boldsymbol{\sigma^2})$. β_1 gave the importance of X_1 and was chosen randomly between 0 and 0.495 each 0.005, giving 100 possibilities. If $\beta_1 = 0$, X_1 had no effect on λ whereas if $\beta_1 = 0.495$, λ was multiplied by $\exp(0.495) \approx 1.64$ per unit of X_1, i.e. increasing λ by 64% when X_1 increase by one. X_2 was set as an unknown variable and thus produced unobserved heterogeneity if $\boldsymbol{\sigma^2} > 0$. The higher $\boldsymbol{\sigma^2}$ was, the higher the unobserved heterogeneity was; $\boldsymbol{\sigma^2}$ was thus used as a measure of the expected overdispersion. Four levels of expected overdispersion were considered, null ($\boldsymbol{\sigma^2} = 0$), (low ($\boldsymbol{\sigma^2} = 0.5$), moderate ($\boldsymbol{\sigma^2} = 1$) and high ($\boldsymbol{\sigma^2} = 2$). Overdispersion could also been handled by using a NB distribution but in ecological applications it is more likely that overdispersion comes from unobserved variables (Richards 2008), which motivated this alternative parameterization.

X_1 was the variable of interest and the aim was to reliably estimate β_1, i.e. determinate the importance of this variable. For each level of overdispersion and zero-inflation considered, 10,000 datasets having each time 500 observations were generated from Eq.1. We chose to consider 500 observations, which is a quite high number, in order to allow detecting complex mixing effects such as zero-inflation. This cannot be done for low sample sizes because different distributions such as NB or ZINB, may give the same observed values, making impossible the identification of a zero-inflation component. The NB model is highly flexible and will thus be the most appropriate for low sample sizes (see Vaudor *et al.* 2011).

B) Evaluation of model performances

i) Probability to select the variable of interest: X_1

The probability to select a variable of interest was evaluated by the probability to select X_1 in Eq.1, which naturally depended on β_1: high β_1 should increase the probability to select X_1. We used the Akaïke information criterion (hereafter AIC, Akaïke 1974) for variable selection: the AIC indicated whether adding X_1 improved the bias/variance trade-off or not (see Burnham & Anderson 2002). The probability to select X_1 for a given β_1 value was thus the number of times (divided by the number of simulations done for the β_1 value) when adding X_1 in the model reduced the AIC.

In ZIMs, X_1 could be selected in the count-part and/or in the excess-zero-part of the model, either indicating that it influenced the count values or the probability of having an excess of zeros. However, one could also consider that X_1 should not affect the excess-zero-part due to prior knowledge. We had thus used two classes of ZIMs, the first one fixed the excess-zero-part to a constant (intercept only) while the second one allowed X_1 to be select in both parts of ZIMs. The probability to select X_1 in the count-part of ZIMs reported in Figures (*ZIP.count* and *ZINB.count*) was estimated by using the former, alleviating the effect of trying to select X_1 in the excess-zero-part. Conversely, the probability to select X_1 in the excess-zero-part of ZIMs reported in Figures (*ZIP.zero* and *ZINB.zero*) was estimated by using the latter.

ii) Reliability on the estimation of the parameter of interest: β_1

The reliability on the estimation of the parameter of interest was evaluated by computing the probability density of β_1 from $N(b_1, se.b_1)$, where b_1 is the estimation of β_1 from the model and $se.b_1$ is its standard error. This measure represented actually the likelihood to estimate correctly β_1. We also provided the bias of the estimation of β_1 in Appendix A and the AIC values of the four models used in Appendix B, expecting that lower was the AIC, better was the reliability on the estimation of β_1. These three measures (β_1 likelihood, bias and AIC) were obviously done only for the models including X_1, i.e. by assuming that X_1 was known to influence λ (see Eq.1). For ZIMs, we included X_1 only in the count-part, i.e. assuming that X_1 was known to influence only this part.

2) Results

i) Probability to select a variable of interest: X_1

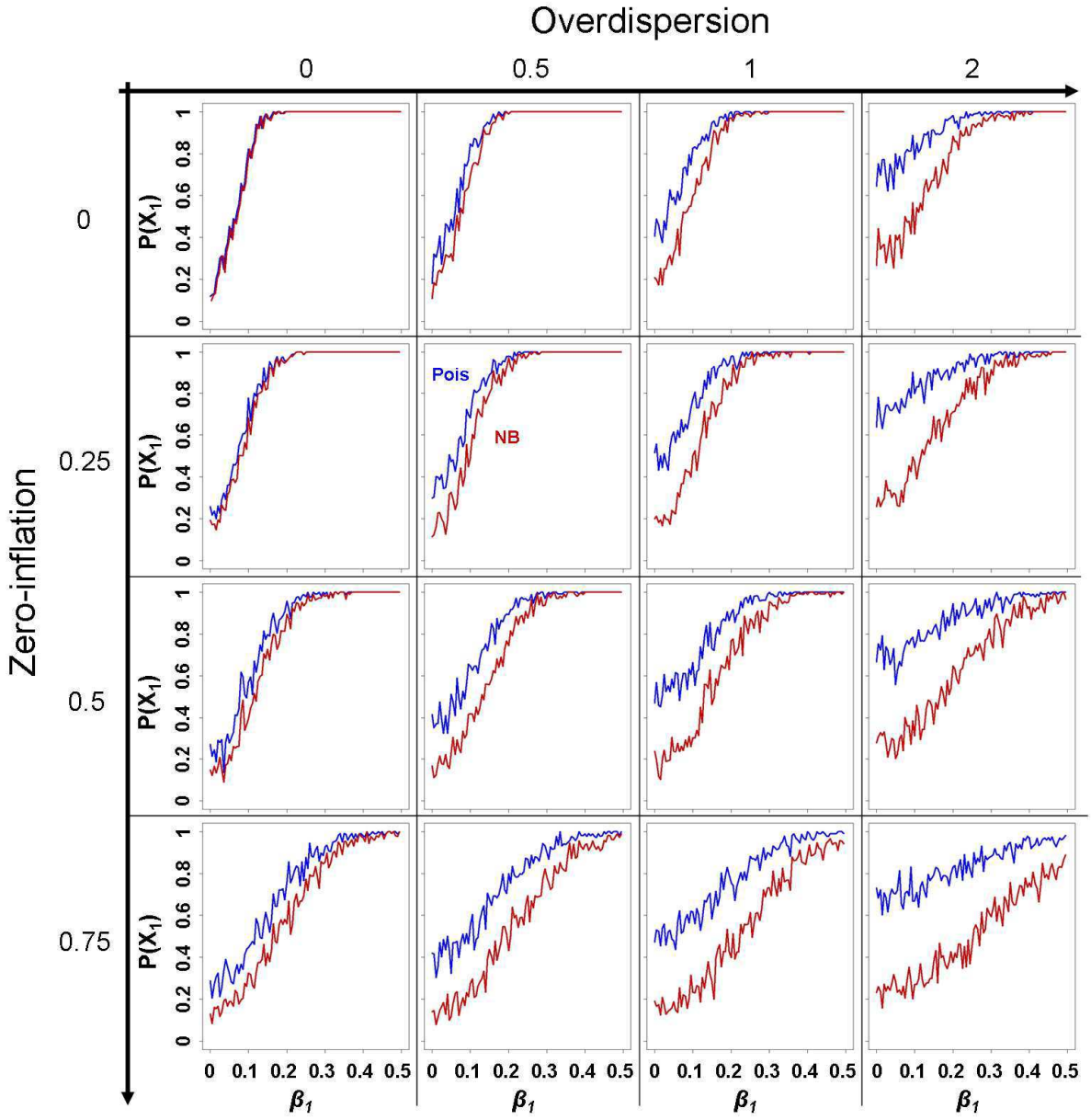


Figure 1: Probability (*P*) to select X_1 depending on its true regression coefficient (β_1), by using a Poisson (*Pois*, in blue) or negative binomial (*NB*, in red) model. Sixteen combinations of overdispersion and zero-inflation levels are here considered. Overdispersion switches from 0 (null) to 2 (high) and zero-inflation switches from 0 (null) to 0.75 (high). The lines represent the mean of probability for each of the β_1 modalities (from 0 to 0.495 each 0.005). This allows giving a graphical representation of the mean but also of its variability.

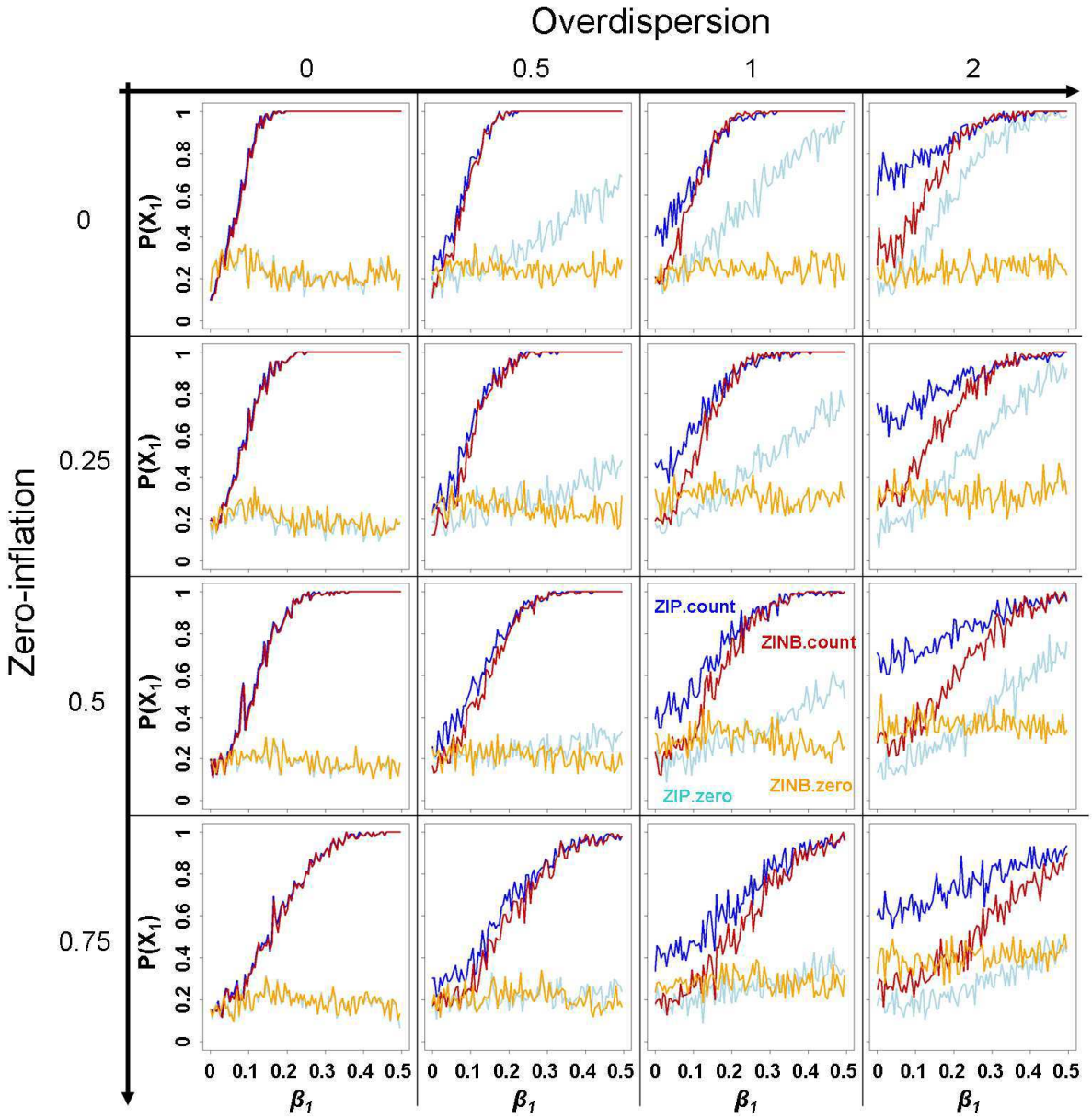


Figure 2: Probability (*P*) to select X_1 depending on its true coefficient (β_1), by using the zero-inflated Poisson (ZIP) or the zero-inflated negative binomial (ZINB) model. ZIP.count (in dark blue) refers to the probability to select X_1 in the count-part of the ZIP whereas ZIP.zero (in light blue) refers to the one in the excess-zero-part. The same goes for ZINB.count (in dark red) and ZINB.zero (in orange). Sixteen combinations of overdispersion and zero-inflation levels are here considered. Overdispersion switches from 0 (null) to 2 (high) and zero-inflation switches from 0 (null) to 0.75 (high). The lines represent the mean of probability for each of the β_1 modalities (from 0 to 0.495 each 0.005). This allows giving a graphical representation of the mean but also of its variability.

In the absence of both overdispersion and zero-inflation, using either the Poisson or the NB model for variable selection was almost equivalent (Fig.1). However, using the Poisson model (blue line in Fig.1) in the presence of overdispersion and/or zero-inflation led to highly overselecting X_1. Conversely, the use of the NB (red line in Fig.1) led selecting X_1 with almost the same probability than by using a ZINB (red line in Fig.2).

Concerning ZIMs performances in the absence of overdispersion, using either the ZIP or the ZINB for variable selection gave almost the same results results (see Fig.2). However in the presence of overdispersion, using a ZIP led to highly overselected X_1 in the count part when it had actually no or low effect (see 'ZIP.count' dark blue line in Fig.2). This led to probabilities in selecting X_1 similar to those obtained by using the Poisson model (blue line in Fig.1), even though being always lower. Note also that allowing X_1 to be selected in the excess-zero-part of the ZIP (see 'ZIP.zero' light blue line in Fig.2) led to highly overselect X_1 in this model-part ; a phenomenon which increased as β_1 increased. Conversely, using the ZINB corrected for this undesirable phenomenon (see 'ZINB.zero' orange line in Fig.2), i.e. X_1 was selected with the same probability whatever β_1 value. Note however that X_1 was selected with a high probability in the excess-zero-part when overdispersion was high (about 0.4).

ii) Reliability on the estimation of the parameter of interest: β_1

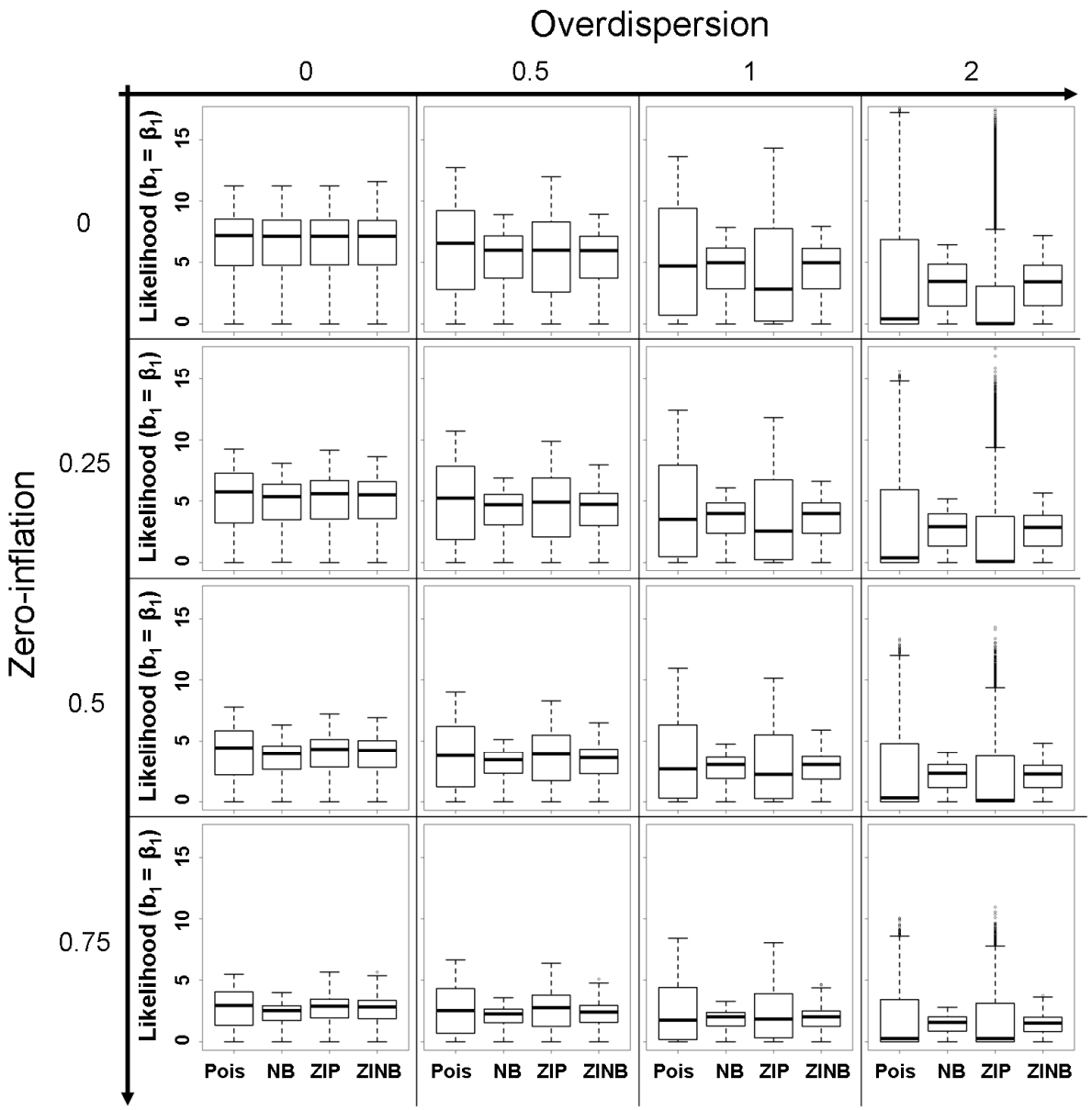


Figure 3: Likelihood of a correct estimation of β_1 by using the Poisson (Pois), the negative binomial (NB), the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) models. Sixteen combinations of overdispersion and zero-inflation levels are here considered. Overdispersion switches from 0 (null) to 2 (high) and zero-inflation switches from 0 (null) to 0.75 (high).

In the absence of both overdispersion and zero-inflation, the likelihood of b_1 were identical whatever the model used (see Fig.3). In the presence of overdispersion, using either NB and ZINB was almost equivalent and gave the better β_1 estimation. Conversely, using either the Poisson or the ZIP gave less suitable β_1 estimation, with the ZIP being even worst than using the Poisson model. In the additional presence of zero-inflation, the likelihood of b_1 decreased with the same order of magnitude whatever the model used. The ZIP kept a slight

advantage over others models in the presence of high zero-inflation and in the absence of overdispersion.

3) Discussion

Based on our simulations, using AIC with a Poisson model was not relevant for variables selection in the presence of overdispersion and/or zero-inflation. Indeed, it led overselecting X_1 when it had no or low effect (see blue line in Fig.1). Moreover, the reliability on the estimation of β_1 (b_1) was weak (see *Pois* in Fig.3). This occurred because the estimated standard errors of b_1 were low, while on the other hand b_1 was not improved (see Figure A1 in Appendix A), leading to precise but erroneous estimates. These results thus fully confirm previous findings showing that the Poisson model is inefficient for the analysis of count data in the presence of overdispersion and/or zero-inflation due to a severe underestimation of the variance (Hinde & Demétrio 1998; Ridout *et al.* 1998; Richards 2008).

The ZIP model, yet performing well in the presence of zero-inflation, was poorly efficient in the presence of overdispersion, even at low rate. It was actually less efficient than the Poisson model in most situations. Indeed, it increased the probability to select X_1 in the wrong part (the excess-zero-part) as β_1 increased, while leading to overselect X_1 in the count part when it had actually no or low effect. Moreover, the reliability on the estimation of β_1 was worst than by using a Poisson model: the estimation of the coefficient β_1 (b_1) was more biased (see Figure A1) and its standard error was still low. Ridout *et al.* (2001) also outlined the poor performance of ZIP in the presence of overdispersion. But intriguingly, comparing AIC-values (see Figure B1 in Appendix B) showed that ZIP gave lower AIC-values than the Poisson model. One could think that it outlined a better fit to the data but it was actually due to confusion between zero-inflation and overdispersion during the model estimation. Actually the excess-zero-part of the ZIP wrongly compensated for the presence of overdispersion, which decreased the AIC even if the model performed worse in term of variable selection capacities and regression coefficient estimation. Comparing Poisson and ZIP models by using AIC in the presence of overdispersion thus leads to wrong conclusions, e.g. supporting the presence of zero-inflation when there is actually only classical overdispersion.

We suggested that Poisson and the ZIP models should not be used for analysing real ecological count datasets since they do not account for overdispersion, despite unobserved variability is likely the rule in all biological datasets (Burnham & Anderson 2002). Recently however, Linder & Lawler (2012) used ZIMs in order to discriminate true and false zeros in a count data of several primate species. They interpret the excess-zero part of ZIMs as the result of false zeros, i.e. the species is present but not detected. They found a significant effect of two variables in the excess-zeros part when using a ZIP model whereas they did not find any effect when using a ZINB. They interpreted the excess-zero-part of the ZIP as a result of species cryptic behaviour. Based on our results, we propose that their patterns result rather from the presence of an additional overdispersion which is not accounted for when using a

ZIP. Models accounting for overdispersion, such as the NB and the ZINB, are thus needed to draw relevant ecological conclusions.

The performances of NB and ZINB were almost always equivalent even in the presence of zero-inflation. They provided a better alternative than using a Poisson model or a ZIP. For instance, they were the more reliable for estimate β_1 in the presence of both overdispersion and zero-inflation. ZINB did not perform much better than NB when focusing on the estimation of the count part, even in the presence of high zero-inflation and low overdispersion (see Fig. 3). The NB model thus remains as relevant as ZINB in the presence of both overdispersion and zero-inflation, at least when focusing on the part affecting count values. The NB is a highly flexible model able to account for overdispersion (Greene 2008), and, as zero-inflation is a special case of overdispersion, it is not surprising to see that, in a certain measure, it also dealt with zero-inflation (see Fig.1 and 3). However, NB does not specifically account for zero-inflation since it does not distinguish between processes producing only zeros and others. If one wants separating these processes, the ZINB seems more appropriated. When choosing between NB and ZINB, it is possible to use a likelihood ratio-test (Vuong 1989; Greene 1994), a score test (Ridout *et al.* 2001) or the AIC (Miaou 1994; Warton 2005).

However, the ZINB needs careful considerations because it has some important drawbacks. For instance, performing variable selection when using a ZINB instead of a NB increases the number of candidate model from 2^k to 4^k (Jochmann 2013), with k the number of candidate variables. The probability to select a non-influential variable in the model is thus inflated. In our simulations, we also showed that the use of a ZINB instead of a NB allowed selecting X_1 in the wrong part (the excess-zero-part) with a rather high probability (about 0.2 to 0.4 depending one the amount of overdispersion, see Fig. 2). This may lead to choose unnecessary overcomplex models, and thus hiding some important biological variables due to an inflated variance (Burnham & Anderson 2002). This overparameterization may be exacerbated if the type of NB is not the most suited to account for overdispersion. Indeed, the common type of NB implies that the variance of the count was equals $\lambda + \lambda^2 / \alpha$ (the current default in R software (R Core Team 2013, see also Bolker 2008 p.124) where α was the dispersion parameter. $NB(\lambda , \alpha)$ was thus equivalent to $Poisson(\lambda)$ when $\alpha \rightarrow + \infty$. Even if this type of NB has been shown useful in most of situations, it may be sometimes not adapted (see for example Ver Hoef & Boveng 2007). In such case the NB model would not be able to fully correct for overdispersion and the ZINB should tend to compensate this by adding parameters in the excess-zero-part, as would do the ZIP in the presence of overdispersion. It will be thus relevant to compare different type of NB (see Ismail & Jemain 2007; Greene 2008) or using aleternative models such as the Generalized Poisson (see Famoye 1993 and references therein), before concluding zero-inflation is occurring, guarantying that the apparent zero-inflation is not due to a misspecification of the variance on the count-part.

To conclude, the use of ZINB over the NB is only relevant when the studied population can be truly divided in two subpopulations, one leading to only zeros and one

leading to both zeros and positive counts (see above example on the number of offspring produced by a species where sterile and fertile individuals are present in the same dataset). One of the most important issues is likely giving a reliable interpretation of the excess-zero-part. Preisser *et al.* (2012) shows that, in epidemiological studies, the interpretation of ZIMs is 'often imprecise or inadvertently misleading'. The same conclusion may apply in ecology.

Acknowledgments

Not yet provided

References

voir page 125.

Appendix A

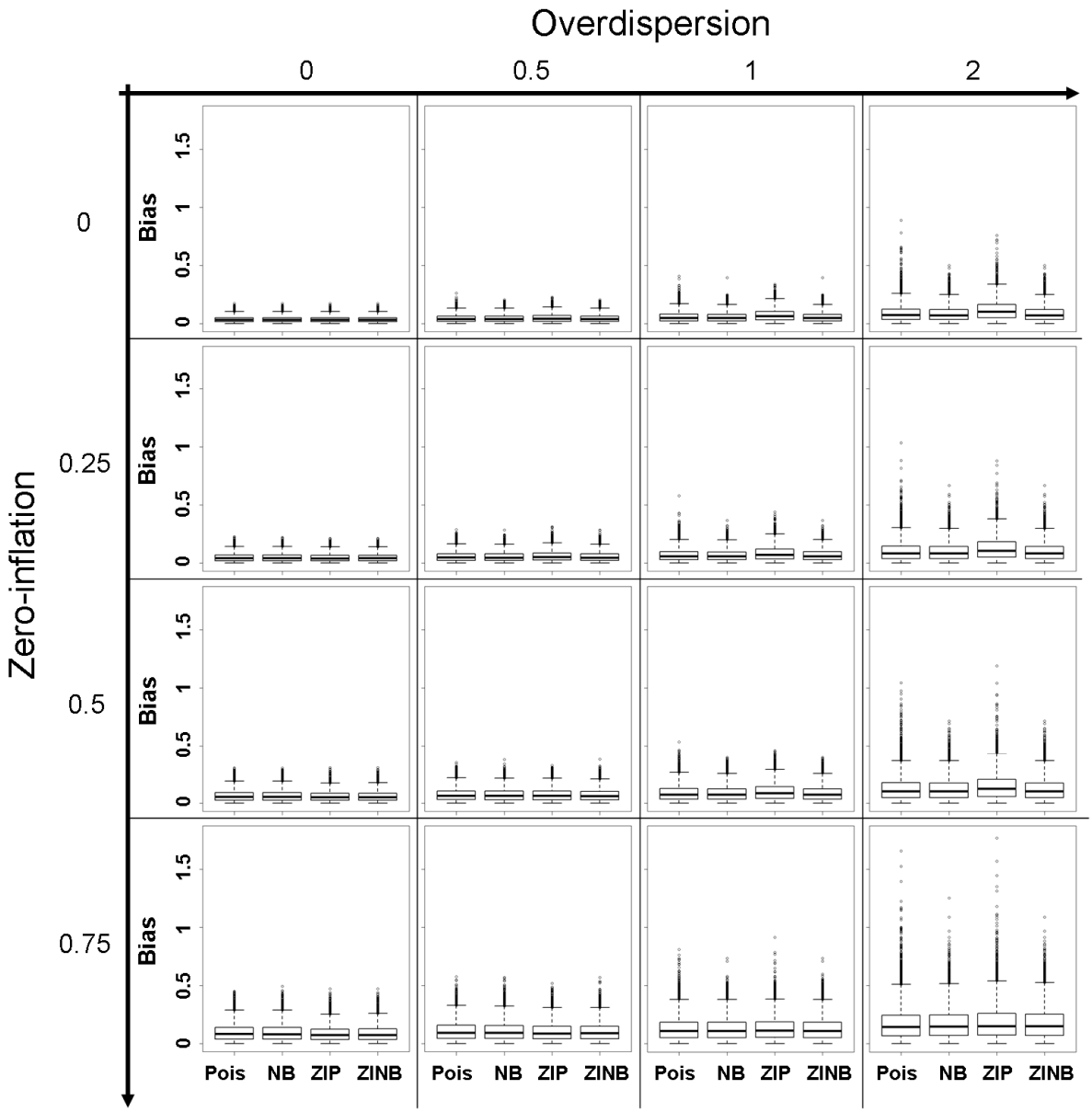


Figure A1: Absolute difference between β_1 and its estimate b_1 (i.e., bias) by using the Poisson (Pois), the negative binomial (NB), the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) models. Sixteen combinations of overdispersion and zero-inflation levels are here considered. Overdispersion switches from 0 (null) to 2 (high) and zero-inflation switches from 0 (null) to 0.75 (high).

Appendix B

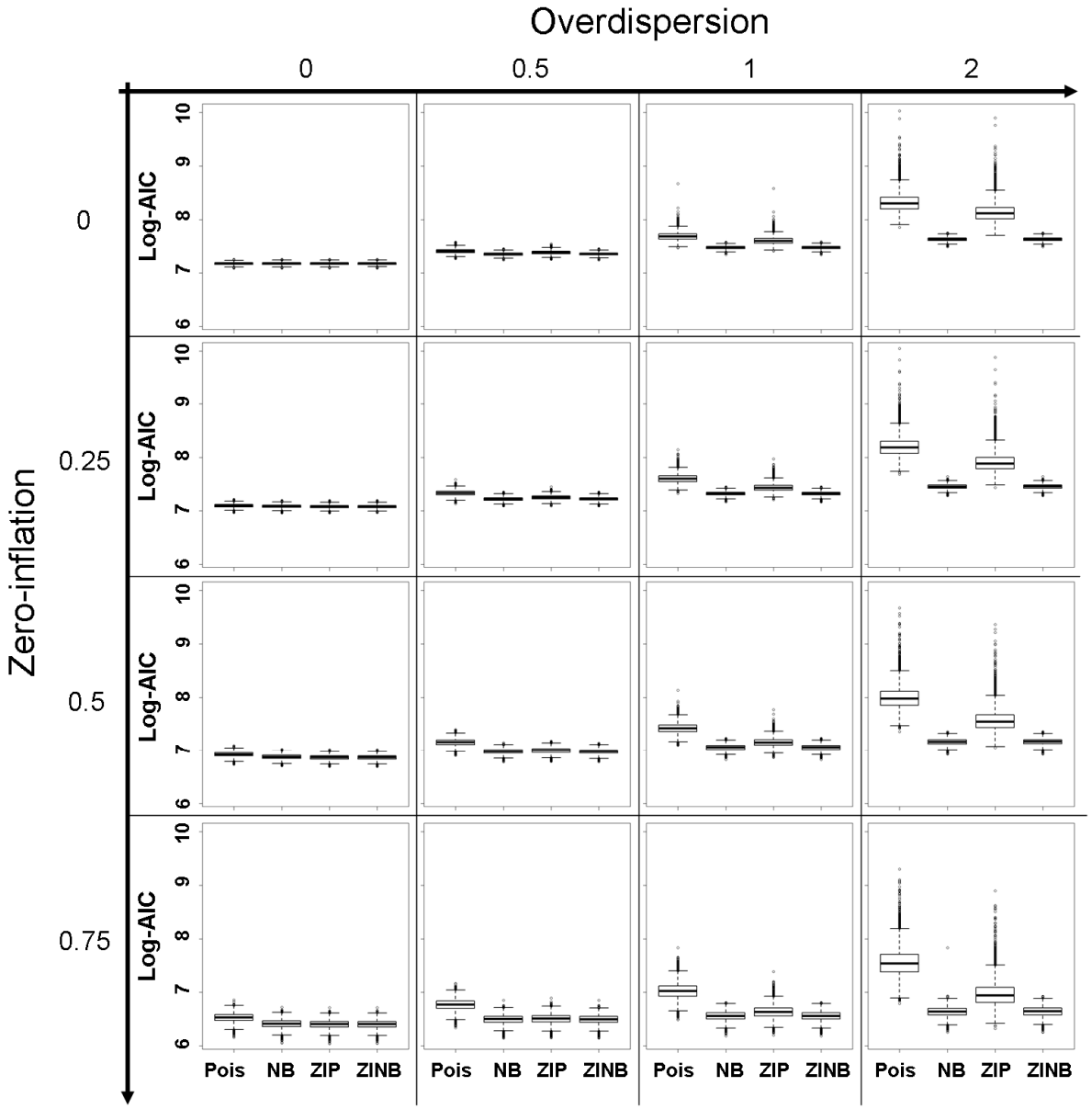


Figure B1: AIC (at log scale) by using the Poisson (Pois), the negative binomial (NB), the zero-inflated Poisson (ZIP) and the zero-inflated negative binomial (ZINB) models. Sixteen combinations of overdispersion and zero-inflation levels are here considered. Overdispersion switches from 0 (null) to 2 (high) and zero-inflation switches from 0 (null) to 0.75 (high).

CHAPITRE 5 : APPLICATION AUX RAPACES NICHEURS DE FRANCE

Les différents aspects méthodologiques abordés dans les précédents chapitres permettent maintenant de faire face aux problèmes statistiques rencontrés lors de l'analyse d'un jeu de données de comptages à large-échelle géographique. En particulier le SLOO va aider à choisir les variables explicatives à inclure dans le modèle (voir Chapitres 1-3). De plus, nous avons montré que l'utilisation d'un modèle zéro-enflé à mélange n'est pas toujours souhaitable pour modéliser l'abondance d'une espèce, même en présence de nombreux zéros (voir Chapitre 4).

Ce dernier chapitre utilise les méthodes discutées auparavant dans un cadre plus général qui va jusqu'à l'estimation des paramètres de populations : distributions, effectifs et tendances. Pour estimer ces paramètres nous proposons d'utiliser la boite à outil R-INLA (au lieu des filtres spatiaux utilisés dans le premier chapitre), ce qui permet une estimation rapide des modèles spatialement explicites. Cet outil permet également d'interpoler aux sites non-échantillonnés en utilisant à la fois l'information des variables explicatives et l'information sur l'autocorrélation spatiale (seules les variables explicatives sont utilisées dans le premier chapitre). Nous utilisons le jeu de données complet sur l'abondance des rapaces en France entre 2000 et 2012, et donnons quelques clés sur l'état des populations de rapaces en France. L'article présenté ci-après sera soumis à une revue internationale mais nécessite encore quelques modifications. En particulier, les données de l'année 2013 seront ajoutées et une partie sur les relations entre espèces sera ajoutée.

Volunteers-based surveys and new statistical tools offer great opportunities for biodiversity monitoring across broad spatial extent.

Authors: Le Rest Kévin*, Pinaud David and Bretagnolle Vincent
Centre d'Etudes Biologiques de Chizé (CEBC), CNRS UPR 1934, 79360 Beauvoir-Sur-Niort, France
Email addressees: lerest.k@gmail.com, pinaud@cebc.cnrs.fr and breta@cebc.cnrs.fr
*Corresponding author:
Email: lerest.k@gmail.com
Tel: +33 (0)5 49 09 35 13
Fax: +33 (0)5 49 09 65 26

Manuscript in progress / November 2013

Overview

Abstract

The growing public interest in biodiversity projects provides great opportunities to monitor biodiversity across broad geographic areas, which would be helpful for collecting global biodiversity data at low cost. Such volunteers based surveys should however need careful consideration during statistical analysis since the presence of residual spatial autocorrelation and over-heterogeneity is highly expected. The recent development of some statistical tools now allows accounting for these problems in all steps of the statistical analysis. Especially, the spatial-leave-one-out method allows accounting for spatial autocorrelation in the variable selection step while the R-INLA tool box provides a useful way to estimate complex spatial hierarchical models in a minimum computation time. We applied such approaches on a dataset collected by volunteers between 2000 and 2013 giving the relative abundance of 12 raptors breeding in France. We then extracted highly valuable information for conservation, i.e. trends, spatial distribution and population sizes of these species. Three raptor species had significant positive population trend, the Black Kite, the Short-toed Snake Eagle and the Eurasian Hobby while one species had a significant negative trend, the Common Kestrel. Overall it appeared that raptors breeding in agricultural landscapes were declining whereas raptors breeding in natural areas such as forest were more stable or increasing. R-codes are provided with the dataset for one species allowing reproducing the work easily.

Introduction

The '2010 biodiversity target' aiming at a significant reduction in the rate of biodiversity loss by 2010 has not been achieved (Butchart *et al.* 2010; Convention on Biological Diversity Secretariat 2010). This objective has thus been renewed by 2020 and 20 important biodiversity targets have been proposed to help in this task (Mace *et al.* 2010; Perrings *et al.* 2010, 2011; Rands *et al.* 2010). Evaluating whether these targets are fulfilled relies on the existence of valid indicators reflecting, as likely as possible, the truth on biodiversity condition (Butchart *et al.* 2010; Jones *et al.* 2011). While existing global biodiversity indicators were shown to be efficient, they still should be improved by 'collecting data in a way that reduce existing bias' (Jones *et al.* 2011). Global indicators are usually obtained by gathering data from different taxonomic groups and their quality is then highly dependent on the relevance of the primary data used (Butchart *et al.* 2010; Jones *et al.* 2011). A critical way to improve the reliability of global biodiversity indicators is thus to improve the reliability of these primary data (Jones *et al.* 2011). In particular, they should be collected at broad spatial extent, which is the correct scale for both population functioning and policy decision making (Jones 2011; Jones *et al.* 2011). Moreover they need to be achieved during a long time period, insuring that detected trends are not just due to natural fluctuations of populations (Magurran *et al.* 2010). However long-term and broad-scale species monitoring is very costly while funds to protect biodiversity are very limited, highlighting the need to maximize the cost-effectiveness of monitoring programs (see Jones 2011; Jones *et al.* 2011).

A promising solution to reduce the cost in collecting primary biodiversity data arises with the growing public interest in biodiversity projects. Especially, the participation of volunteers in such projects has rapidly grown over the past decade (see Dickinson *et al.* 2010). Concomitant to this citizen science wave, new communication tools such as internet and the free availability of remote sensing databases (Kerr & Ostrovsky 2003) have marked the emergence of new quantitative approaches able to address questions on the species distribution across very broad geographic areas (Dickinson *et al.* 2010). However, broad scale data involving volunteers also raises other concerns with regard to the statistical analysis to use since such data may have potentially higher heterogeneity than expected by conventional models (overdispersion, see Hinde & Demétrio 1998) and in addition will present strong spatial autocorrelation (Beale *et al.* 2010; Dickinson *et al.* 2010; Hothorn *et al.* 2011). In this paper, we argue that combining broad scale volunteers based survey and appropriate statistical analyses is highly valuable to estimate species parameters, such as distribution, abundance and trends, at spatial scales that have not been addressed so far. We promote the use of the recent statistic tool box R-INLA (Rue *et al.* 2009) and its spatial module SPDE (Lindgren *et al.* 2011) as an easy way to build powerful statistical model while accounting for both overdispersion and spatial autocorrelation.

We analysed a dataset collected on bird top predators in terrestrial ecosystems. Top predators have a particular role in ecosystems since they depend on underlying trophic levels,

i.e. they are affected by bottom-up forces (see Power 1992). However, in a resource limited world, they may also themselves influence the lower trophic levels, i.e. acting as top-down forces (see Power 1992; Bretagnolle & Gillis 2010). For instance, predators may regulate directly prey number, and they may also regulate non-prey's species indirectly, e.g. they can regulate plants by regulating herbivores (see Schmitz 2003). In both cases, top predators may provide critical information on the distribution and abundance of many organisms, explaining why they are often used as biodiversity indicator species for conservation purposes (Carroll *et al.* 2001; Gittleman *et al.* 2001; Sergio *et al.* 2005 but see Andelman & Fagan 2000). Global biodiversity indicators should collate as many taxonomic groups as possible, hence top predators appear as a case in point. Raptors are particularly well suited and several studies have already demonstrates strong links between raptors presence and biodiversity (see Sergio *et al.* 2004, 2005, 2006, 2008a, b). Considering several top predator species is however highly recommended (Carroll *et al.* 2001).

We analysed a national volunteers-based survey on the abundance of 24 diurnal raptors breeding in France (about 1000 by 1000 km) over 13 years. We additionally used free remote sensing climatic and habitat (land cover) datasets to link observed abundance with environmental variables suspected to directly or indirectly influence the raptor abundance. As said before, a particular emphasis concerns the statistical analyses done, which should be conducted accounting for both spatial autocorrelation and overdispersion in each step of the modelling scheme, from variables selection to model predictions. To do so during the variables selection step, we used the spatial-leave-one-out method (SLOO, Le Rest *et al.* 2014), which allows choosing relevant variables while avoiding undesirable effect of spatial autocorrelation. Variables selected were then used in a spatial explicit negative binomial model, i.e. a model accounting for both spatial autocorrelation and overdispersion. Predictions were made from this model over the entire studied area. Spatial analyses and predictions are conducted using the promising tool box R-INLA and all R-codes are made available for further uses.

1) Materials & Methods

Study models

Raptors (birds of prey) are predators that belong principally to families *Accipitridae* and *Falconidae*. Raptors have long been used as biological indicators in terrestrial ecosystems (see Newton 1979; Sergio *et al.* 2004, 2005, 2006, 2008a, b). There are 24 breeding species of diurnal raptors in France (Thiollay & Bretagnolle 2004), many of them being present in tiny numbers either because their breeding habitat is restricted, or because their breeding distribution is very limited. We focus here on the first 12 most abundant species of France, which are (in decreasing order of abundance) the Common Buzzard *Buteo buteo*, the Common Kestrel *Falco tinnunculus*, the Eurasian Sparrowhawk *Accipiter nisus*, the Black Kite *Milvus migrans*, the European Honey Buzzard *Pernis apivorus*, the Hen Harrier *Circus*

cyaneus, the Eurasian Hobby *Falco subbuteo*, the Northern Goshawk *Accipiter gentilis*, the Montagu's Harrier *Circus pygargus*, the Red Kite *Milvus milvus*, the Short-toed Snake Eagle *Circaetus gallicus* and the Marsh Harrier *Circus aeruginosus*.

Survey & datasets

The dataset used concerned a national survey aiming to monitor diurnal raptors breeding in the whole country of France. Field surveys were carried out by volunteer ornithologists (under the supervision of the National NGO *Ligue pour la Protection des Oiseaux*, LPO) and data were analysed by a scientific research lab (the Centre d'Etudes Biologiques de Chizé, CEBC). The field protocol as standardised as much as possible but remained simple. It consists in counting the total number of breeding pairs of each raptor species on a 25-km2 quadrats during several field sessions in the whole breeding season (5×5 km; see Thiollay & Bretagnolle 2004; Le Rest *et al.* 2013 for details). The survey began in 2000 by three years of intensive field work (2000, 2001 and 2002) with the main aim to obtain an accurate starting point about the distribution and population size of raptors in France. From then a yearly monitoring program was set up to estimate trends and distribution shifts, but based on a much lighter sampling scheme surveying about one hundred quadrats by year. However, several years were needed before enough volunteers were involved to reach the objective of one hundred quadrats by year (see Table 1).

Table 1: Number of quadrats surveyed per year between 2000 and 2012.

Year	**2000**	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**
Number of quadrats surveyed	440	630	190	**5**	**7**	21	40	71	95	96	83	91	79

Since the years 2003 and 2004 had only 5 and 7 quadrats surveyed respectively, they were assigned to 2002 and 2005 respectively, which led to 195 and 28 quadrats instead of 190 and 21 for these two years (see Table 1). In total, 1848 quadrats were surveyed between 2000 and 2012, corresponding of 1367 distinct locations covering widely the studied area (see Fig.1).

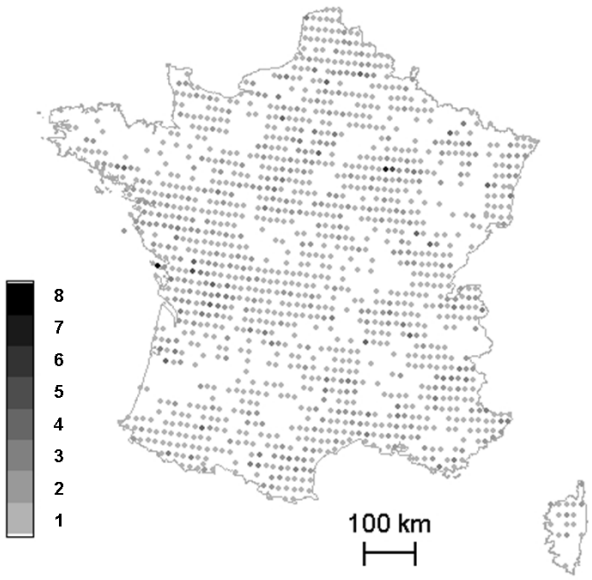


Figure 1: Position of the 1848 quadrats surveyed between 2000 and 2012, the grey scale indicates the number of time each quadrat were surveyed during this period, from 1 in light grey to 8 in black.

Each quadrat was thus described using environmental variables using climatic (Hijmans *et al.* 2005, Bioclim, www.worldclim.org/bioclim) and a land cover (CLC: Corine Land Cover, www.eea.europa.eu) remote sensing datasets. The climatic dataset consisted in 19 variables measured between 1960 and 1990, which provided strong estimates of measures such as average temperature, rainfall, temperature variation and rainfall variation at a resolution of approximately 1-km. Not surprisingly, high correlations occurred between several climatic variables. In order to overcome multicollinearity, a principal component analysis (PCA) was performed on this dataset and principal components were used as climatic variables. The label "ClimDim.x" was used to nominate the xst principal component from the climate dataset. The land cover dataset had 44 variables depicting land use in 2000 on a 100 x 100 m cell resolution. From these 44 classes, 9 habitat hyper-classes were built from a functional (ecological) point of view for raptors (see Table A1 in Appendix A). The percentage of coverage per 25-km2 quadrat was calculated for each of these habitat hyper-classes.

Statistical modelling

Climatic and land cover variables are useful for modelling abundance of raptor species and have largely been used already (see for example Seoane *et al.* 2003; Bustamante & Seoane 2004; Hothorn *et al.* 2011; Le Rest *et al.* 2013). However, all of them may not be always necessary. We thus performed a variable selection for each species, able to select the more relevant variables assuming a trade-off between model complexity and data fitting (see Burnham & Anderson 2002). Since residual spatial autocorrelation was present for all studied species, we used the spatial-leave-one-out (SLOO) method instead of the AIC (Le Rest *et al.* 2014). SLOO consists in doing a leave-one-out cross-validation but where the spatially autocorrelated observations between the validation and the training sets are removed. SLOO allows performing variables selection in an easy way, accounting for both the residual spatial

autocorrelation and the autocorrelation present in the variables (see Le Rest *et al.* 2014 for full details).

Once a set of variables had been selected for each species, we computed negative binomial generalised linear mixed models (hereafter GLMM) with a distance-based spatially structured random effect (again for each species). A negative binomial model was used to account for an (over)heterogeneity in the data (i.e. overdispersion, see Hinde & Demétrio 1998), mainly expected due to heterogeneous detection capacities between observers. The count data thus did not represent the true abundance but the abundance relative to the observer's detection. The aim was to extract the mean of this relative abundance, i.e. how many breeding pairs a medium observer will likely detect in a given location? The negative binomial model allowed high stochastic variations around the mean, which recognizes a high heterogeneity between counts values (see Greene 2008). As the detection probability for each observer was not available in this survey, it was not possible estimating the true abundance (see Royle & Nichols 2003).

The negative binomial models were computed with a distance-based spatially structured random effect in order to account for residual spatial autocorrelation (see Beale *et al.* 2010; Saas & Gosselin 2014). We used the R-INLA tool box for estimating these models, which allows fast Bayesian inference using the integrated nested Laplace approximation (INLA, Rue *et al.* 2009). R-INLA tool box proposes an easy way to compute continuous spatial processes by using the stochastic partial differential equation (SPDE) method with a Matérn covariance matrix (see Lindgren *et al.* 2011). Note that the Matérn covariance function is a flexible form of covariance (Lindgren *et al.* 2011) and have been shown to perform well with the SPDE method (see Lindgren *et al.* 2011; Beguin *et al.* 2012; Camelleti *et al.* 2012; Lindgren 2013). There has been a recent and flourishing number of papers and tutorials about this method, making possible the estimation of complex spatial models possible in a minimum computation time (Beguin *et al.* 2012; Camelleti *et al.* 2012; Lindgren 2013; Lindgren & Rue unpublished; Krainski & Lindgren unpublished). Using R-INLA with SPDE necessities to construct a constrained refined Delaunay triangulation (Krainski & Lindgren unpublished), called 'mesh'. Since our data locations were rather regularly spaced, we defined the 'mesh' from data locations with only very weak refinements (see Fig. B1 in Appendix B). For more details on this step, see the 'work in progress' SPDE tutorial (Krainski & Lindgren unpublished).

For each species, we considered only the best subset of variables selected for statistical inference because model averaging is not always straightforward with spatial hierarchical models. For instance, the random spatially structured term may compensate the omission of an important ecological variable if this variable was itself spatially structured (see Hodges & Reich 2010), but one should prefer the information given by a variable than the one given by a random term.

For each raptor species, we estimated the linear population trend between 2000 and 2012, the relative abundance for every year, the spatial distribution (i.e. the relative number of

pairs per 25 km² quadrat) and the population size (i.e. the number of pairs in France). Linear trends were obtained by considering the year effect as a continuous variable and were represented as the average rate of change per year (in pairs). Relative abundances per year were obtained by considering the year as a factor (fixed effect). We presented these results as the slope for abundance between 2000 and 2012 (except for 2003 and 2004, see survey & datasets section). The spatial distribution was obtained by using the prediction of count mean over a grid of France (22 363 quadrats of 25 km²), expressed in a log-scale for easier view. The maps of standard error of these distributions were also given (see Appendix C and D). Finally, the relative population size was estimated by sampling 10'000 times from the approximated posterior distribution of the model and summing the predicted values for the whole country. This process allowed estimating a 95% confidence interval (2.5% and 97.5% quantiles) for the estimated population size. Note that the function used for sampling from the approximated posterior is a function in progress (see inla.posterior.sample in R).

2) Results

Trends

Table 2: Linear trends of raptors between 2000 and 2012. Trends are done in average rate by year, i.e. rates inferior to 1 indicates a decreasing population trend.

Quantiles / Species	2.5%	50%	97.5%
Buteo buteo	0.994	1.002	1.009
Falco tinnunculus	**0.979**	**0.987**	**0.994**
Accipiter nisus	0.985	0.995	1.004
Milvus migrans	**1.008**	**1.023**	**1.039**
Pernis apivorus	0.990	1.003	1.017
Circus cyaneus	0.976	0.991	1.005
Falco subbuteo	**1.002**	**1.016**	**1.031**
Accipiter gentilis	0.986	1.002	1.019
Circus pygargus	0.968	0.988	1.008
Milvus milvus	0.981	1.019	1.059
Circaetus gallicus	**1.013**	**1.032**	**1.052**
Circus aeruginosus	0.966	0.989	1.014

Most species showed non significant linear trends at the 0.05 level (see Table 2) between 2000 and 2012, indicating either stable population size or undetected trends. Three raptor species had a positive trend, the Black Kite *Milvus migrans*, the Eurasian Hobby *Falco subbuteo* and the Short-toed Snake Eagle *Circaetus gallicus*. One species had a negative trend, the Common Kestrel *Falco tinnunculus*.

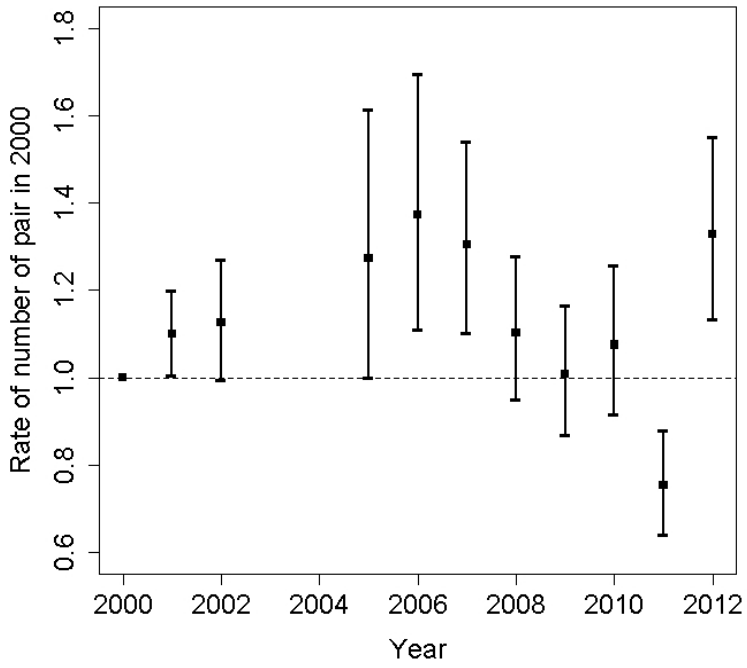


Figure 2: Relative abundance per year for Common Buzzard *Buteo buteo*. Points are the mean and lower and upper bounds represent respectively the 2.5% and the 97.5% confidence intervals. Results for other species are shown in Appendix C.

In spite of non-significant linear trends, many raptors, such as Common Buzzard, had significant inter-annual variations (see Fig. 2 and Appendix C), suggesting a high variability in the numbers of breeding pairs between years. These fluctuations were likely due to pairs not breeding every year (i.e. floaters), which was known to occur frequently for raptor species (see Sergio *et al.* 2009).

Spatial distribution

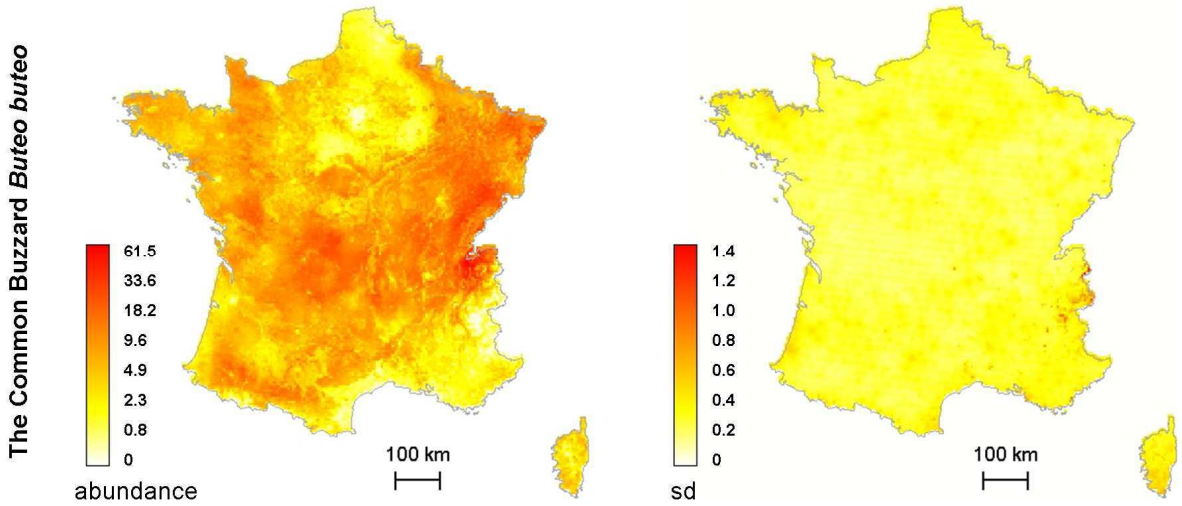


Figure 3: Spatial distribution (left: relative number of pairs, *abundance*; right: standard error of predictions at log-scale, *sd*) of Common Buzzard *Buteo buteo*. Maps for other species are shown in Appendix D.

Distribution maps gave a accurate estimation of the relative number of pairs over a large part of France (see Fig. 3 and Appendix D). However some areas remained often more imprecisely estimated (see Appendix D), which mainly occurred for less sampled habitats such as high mountains, cities or for less sampled area such as borders or islands (e.g. Corsica).

Population size

Table 3: Relative number of breeding pairs in France for the year 2000. The 2.5%, 50% and 97.5% quantiles are given.

Quantiles / Species	2.5%	50%	97.5%
Buteo buteo	148 970	159 173	170 363
Falco tinnunculus	101 870	108 535	115 679
Accipiter nisus	43 101	47 088	51 462
Milvus migrans	25 671	30 448	36 243
Pernis apivorus	19 298	21 993	25 082
Circus cyaneus	13 694	15 818	21 843
Falco subbuteo	11 661	13 447	15 611
Accipiter gentilis	7 128	8 606	10 434
Circus pygargus	5 600	7 075	9 062
Milvus milvus	3 487	4 992	7 303
Circaetus gallicus	3 295	4 370	5 775
Circus aeruginosus	2 887	4 108	6 470

All models allowed estimating fairly precisely the population size (see Table 3). However for *Circus cyaneus*, six quadrats (over the 22 363 used, i.e. the whole prediction grid) were ignored because of non-consistent estimations (sometimes several millions of pairs on these 25 km² quadrats).

3) Discussion

Methodological considerations

Recent advances in spatial statistics now offer the opportunity to account for major problems in the analysis of spatial count data, such as spatial autocorrelation and overdispersion, in a completed framework. We accounted for spatial autocorrelation during the variable selection step by selecting the variables with the SLOO method instead of AIC (see Le Rest *et al.* 2014), which avoided the selection of overly complex models (see Diniz-Filho *et al.* 2008; Le Rest *et al.* 2014). Then, the statistical model combined environmental variables and a spatial explicit random term. Environmental variables were extracted from remote sensing datasets, which allows predicting the spatial distribution of species across broad spatial extent while accounting for basic environmental variability (Elith & Leathwick

2009). On the other side, the spatial explicit random term was helpful for two purposes: first correcting for the estimation of the variables effects, which otherwise will be affected by the residual spatial autocorrelation (see Beale *et al.* 2010); and second improving spatial predictions by interpolating the spatial term at unsampled locations. Such model was thus especially useful for prediction purposes. The recent development of the statistical tool box R-INLA (Rue *et al.* 2009) allowed estimating such complex spatial hierarchical models in several minutes while it could need several hours or days with MCMC (see Beguin *et al.* 2012).

Some studies had already proposed general framework in order to analyse broad scale count data. For instance, Hothorn *et al.* (2011) recently provided an approach decomposing the environmental, the spatial and the spatio-temporal effect in a single model. They applied such method in order to predict the presence/absence of Red Kite and the number of Orthoptera species in Germany. The same set of remote sensing datasets than here was used as environmental variables. However the spatial term used was a non-linear function of the geographic coordinates, which only removed large-scale spatial trends but not accounted for fine scale residual spatial autocorrelation (Dormann *et al.* 2007; Beale *et al.* 2010). Another suited method to account for residual spatial autocorrelation in count data regression is the spatial filtering technique (Dray *et al.* 2006; Griffith & Peres-Neto 2006). It consists in adding several spatially autocorrelated eigenvectors in the model, which are calculated from a weight matrix reflecting connectivity between locations. Haining *et al.* 2009 used a negative binomial regression model adding some spatial filters to deal with residual spatial autocorrelation. However, spatial filtering, yet being convenient to correct for undesirable effect of residual spatial autocorrelation, seems less adapted for spatial prediction on new locations (Bivand 2002). The approach used here based on a spatial GLMM seemed thus having high advantages over the other methods, especially for prediction purposes, which was also supported by Beguin *et al.* (2012).

Beyond these spatial aspects, there is also another problem to address when using volunteers based surveys. It concerns the high heterogeneity amongst observer performances (Dickinson *et al.* 2010). To overcome this issue, usually a great care is imposed on field protocols, with recording of observer performances to be taken into account in the statistical analysis (see Royle & Nichols 2003; Royle 2004). But at of broad-scale, such information are difficult to obtain for most studies, either either because protocols correcting for observer detection were not yet available when the survey began or because it involves too much constraint for volunteers participating to the survey. The least to do in such case is to account for the expected over-heterogeneity by using statistical models accounting for overdispersion (see Hinde & Demetrio 1998), e.g. by using a negative binomial model. Overdispersion would otherwise affect the estimation of regression coefficients by artificially reducing their uncertainty (see Richards 2008). It is usually due to the omission of important information in the model (Hinde & Demetrio 1998) and here likely occurred due to the absence of information on the observer performances. Models accounting for overdispersion usually

assume that the over-heterogeneity is random in space and time, i.e. constant overdispersion. Here, we cannot rule out that volunteers may increase their performances through time but the random sampling of surveyed quadrats prevented to a large extent such learning effect (see Jiguet 2009). In addition, over the years different observers surveyed the same quadrats, providing some random variability in time. The main drawback in using such approach was that it did not allow addressing the estimation of true abundance but rather a relative abundance. Even if relative abundance may sometimes be convenient for population monitoring (Engeman 2003), it was however not suited for monitoring rare species because the true density becomes a measure of great interest, likely reflecting a risk of extinction (Courchamp *et al.* 1999). We thus encourage using some measure of the detection probability in the field protocol, e.g. by using the replicated count methodology (Royle 2004).

Trends, distribution and abundance of diurnal raptors

Three raptors species out of the 12 studied showed a significant increasing population trend in France: the Black Kite, the European Hobby and the Short-toed Snake Eagle (see Table 2). The Black Kite is an opportunistic species able to adapt its diet according to food availability and is able to live sometimes quite closely to humans activities (e.g. cities neighbours, open waste). Despite that the IUCN Red List (www.iucnredlist.org) not yet provides trend for this species, it was not surprising that the population of this species increased in France, which was also support by another independent breeding bird survey in France (STOC, see Jiguet *et al.* 2012). In addition, as for Short-toed Snake Eagle, such species may currently benefit from global warming. The species range does not yet entirely cover France (see Appendix D) whereas suitable breeding habitat are present elsewhere, e.g. on the northern-west coast. France is a major country for this species since breeding population size was estimated about 30 000 pairs (see Table 3) whereas European population size is estimated between 64 000 and 100 000 pairs (BirdLife International, www.birdlife.org). Moreover we shown an increasing population trend (see Table 2), which additionally outlined that the Black Kite population is fine in France. The Short-toed Snake Eagle is a species mainly eating reptiles and is present in the southern-east of France (see Appendix D). The IUCN outlined a stable population trend whereas STOC supported an increasing population trend yet not significant. Our results showed a marked increasing population trend, which must be set in relation to the high potential of northern colonisation for this species. The main limitation of northern colonisation was likely the prey's resource but as temperatures warmed, one could hypothesize that the Short-toed Snake Eagle should find more suitable habitats in north of France and especially more reptiles. It remains to check if a northern shift was observed, which has been shown for some other raptor species, e.g. in the wintering distribution of Northern American raptors (La Sorte & Thompson 2007). France is again a major country for this species since breeding population size was here estimated about 7 000 pairs (see Table 3) whereas European population size is estimated between 8 400 and 13 000 pairs (BirdLife International). Note however that these estimates are not directly

comparable and only give a vague idea of the importance of France for this species. The increasing population trend (see Table 2) assessed that the Short-toed Snake Eagle population is also fine in France. The European Hobby likes the presence of water and heterogeneous agriculture providing large trees for nesting. Its increasing population trend was less expected than for two others since the IUCN suggested a decreasing population size and STOC a non-significant decrease. However it is a poorly studied raptor species in France, difficult to detect due to its behaviour (mainly evening activity). This raptor monitoring scheme thus provided important new knowledge about the trend of this species in France. With a population size about 13 000 pairs in France (see Table 3), which was likely still underestimated due its cryptic behaviour, and a slightly increasing population trend (see Table 2), the European Hobby seems fine in France.

Only a single species showed a significant decreasing population trend (see Table 2), the Common Kestrel, which is the second most abundant diurnal raptor species in France (see Table 3). Although this decrease was not seem alarming given the current population size (over 100 000 pairs, see Table 3), it is actually however significant, and incidentally, the IUCN and STOC also proposed the same population trend. The Common Kestrel decline is likely to be related to farmland habitat degradation due to intensification of agriculture (see Chamberlain *et al.* 2000). Others species breeding in agricultural landscapes, such as the three Harrier species, Hen, Montagu's and Marsh, are also declining even if their estimated trends were not significant (see Table 2). A particular attention would thus be given for the raptors species breeding in agricultural landscapes. Conversely, most raptors breeding in natural habitats showed no specific trends or increasing trends. It was not surprising to see that raptors breeding in forest, such as the Northern Goshawk and the European Honey Buzzard, had a quite stable population trend since the percentage of forest in France remained stable between years. These population trends were elsewhere similar to the information given by the IUCN. Finally, species breeding in semi-natural areas or at the junction between natural and anthropized areas, such as the Common Buzzard, the Eurasian Sparrowhawk or the Red Kite, had quite uncertain population trends (see Table 2 and Appendix C) even if they looked stable at first seen. The IUCN and STOC gave contradictory information on the population trends of these species. Future monitoring will thus be helpful for determinate the status of these raptors in France.

Acknowledgments

Not provided for the moment.

References

voir page 125.

Appendix A

Table A1: The nine habitat hyper-classes used in our analyses and the Corine Land Cover initial classification. One row corresponds to one initial Corine Land Cover class, which is defined by three distinct labels.

44 Corine Land Cover nomenclatures			**9 Habitat hyper-classes**
Label 1	**Label 2**	**Label 3**	
Artificial surfaces	Urban fabric	Continuous urban fabric	Anthropic areas
Artificial surfaces	Urban fabric	Discontinuous urban fabric	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Industrial or commercial units	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Road and rail networks and associated land	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Port areas	Anthropic areas
Artificial surfaces	Industrial, commercial and transport units	Airports	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Mineral extraction sites	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Dump sites	Anthropic areas
Artificial surfaces	Mine, dumps, and construction sites	Construction sites	Anthropic areas
Artificial surfaces	Artificial, non-agricultural vegetated areas	Green urban areas	Anthropic areas
Artificial surfaces	Artificial, non-agricultural vegetated areas	Sport and leisure facilities	Anthropic areas
Agricultural areas	Arable land	Non-irrigated arable land	Intensive agriculture
Agricultural areas	Arable land	Permanently irrigated land	Intensive agriculture
Agricultural areas	Arable land	Rice fields	Intensive agriculture
Agricultural areas	Permanent crops	Vineyards	Permanent agriculture
Agricultural areas	Permanent crops	Fruit trees and berry plantations	Permanent agriculture
Agricultural areas	Permanent crops	Olive groves	Permanent agriculture
Agricultural areas	Pastures	Pastures	Extensive farming
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Natural grasslands	Extensive farming
Agricultural areas	Heterogeneous agricultural areas	Annual crops associated with permanent crops	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Complex cultivation patterns	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Land principally occupied by agriculture, with significant areas of natural vegetation	Heterogeneous agriculture
Agricultural areas	Heterogeneous agricultural areas	Agro-forestry areas	Forest areas
Forest and semi-natural areas	Forests	Broad-leaved forests	Forest areas
Forest and semi-natural areas	Forests	Coniferous forest	Forest areas
Forest and semi-natural areas	Forests	Mixed forest	Forest areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Moors and heathland	Transitional areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Sclerophyllous vegetation	Transitional areas
Forest and semi-natural areas	Scrub and/or herbaceous vegetation associations	Transitional woodland-shrub	Transitional areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Beaches, dunes, sands	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Bare rocks	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Sparsely vegetated areas	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Burnt areas	Open areas
Forest and semi-natural areas	Open spaces with little or no vegetation	Glaciers and perpetual snow	Open areas
Wetlands	Inland wetlands	Inland marshes	Wetlands
Wetlands	Inland wetlands	Peat bogs	Wetlands
Wetlands	Maritime wetlands	Salt marshes	Wetlands
Wetlands	Maritime wetlands	Salines	Wetlands
Wetlands	Maritime wetlands	Intertidal flats	Wetlands
Water bodies	Inland waters	Water courses	Wetlands
Water bodies	Inland waters	Water bodies	Wetlands
Water bodies	Maritime waters	Coastal lagoons	Wetlands
Water bodies	Maritime waters	Estuaries	Wetlands
Water bodies	Maritime waters	Sea and ocean	Used as offset

Appendix B

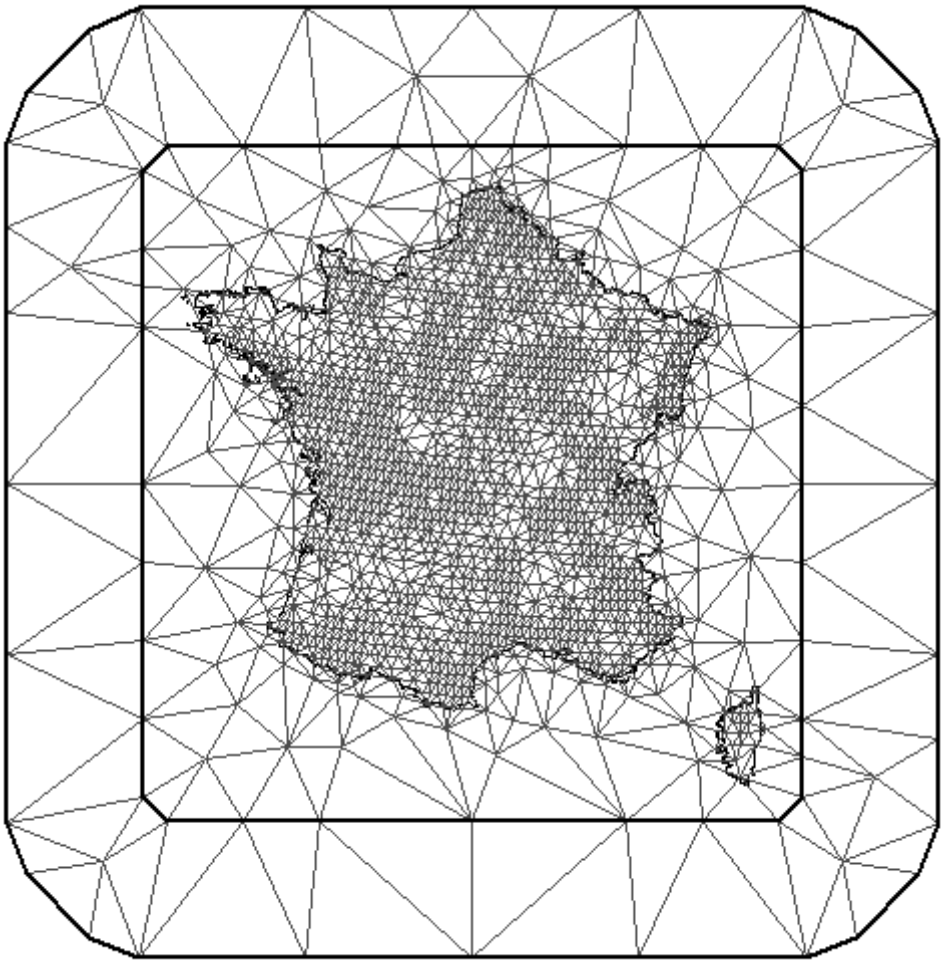
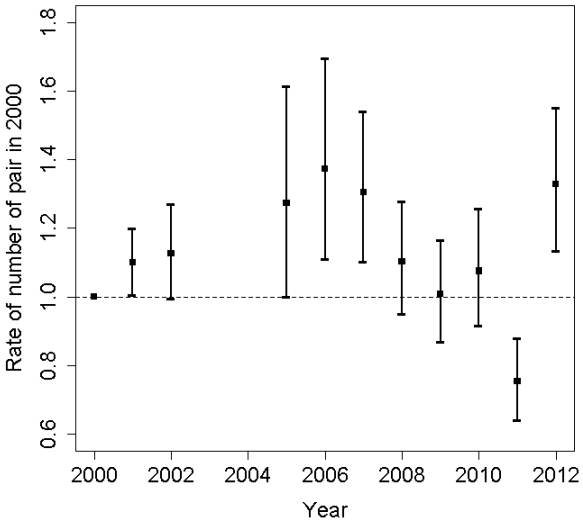


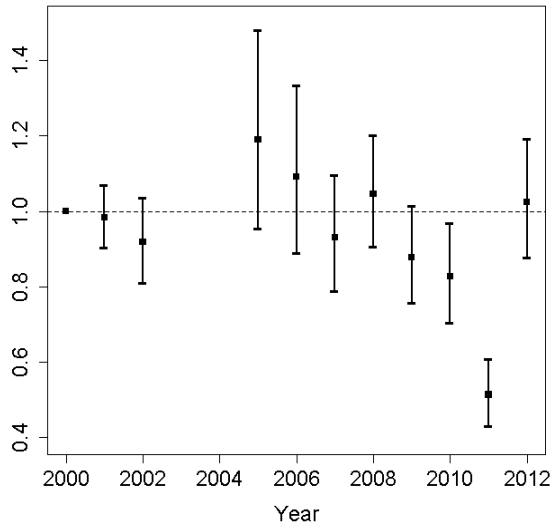
Figure B1: Constrained refined Delaunay triangulation used for building spatial GLMM. Triangles vertices in the studied area (France, black contour) are data locations.

Appendix C: Relative abundance per year of diurnal raptors in France

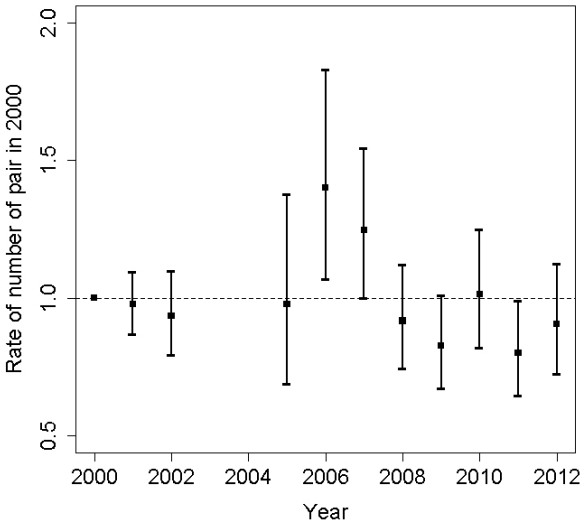
The Common Buzzard *Buteo buteo*



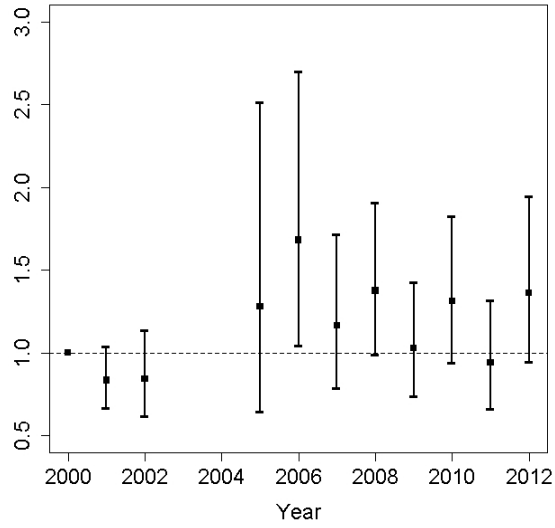
The Common Kestrel *Falco tinnunculus*



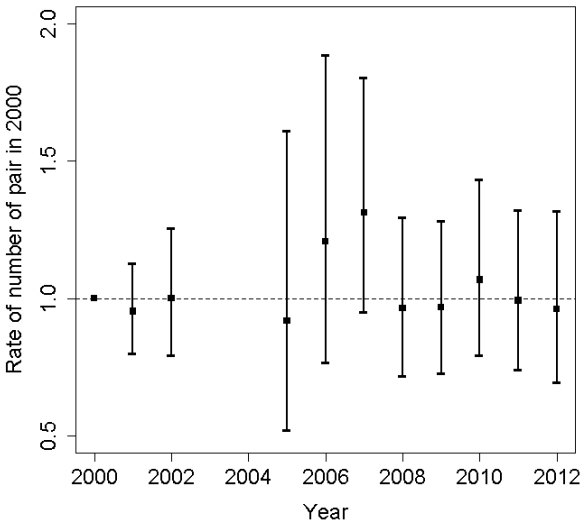
The Eurasian Sparrowhawk *Accipiter nisus*



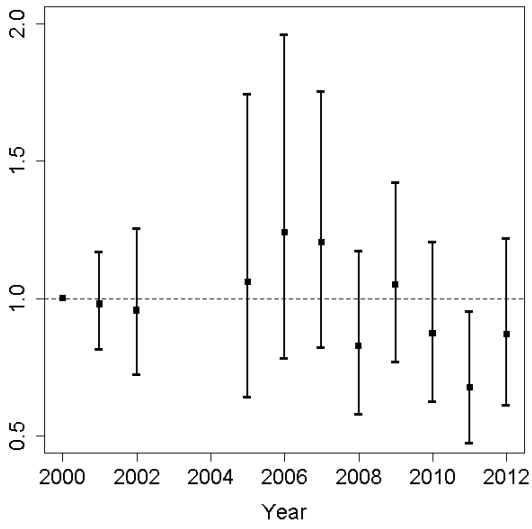
The Black Kite *Milvus migrans*



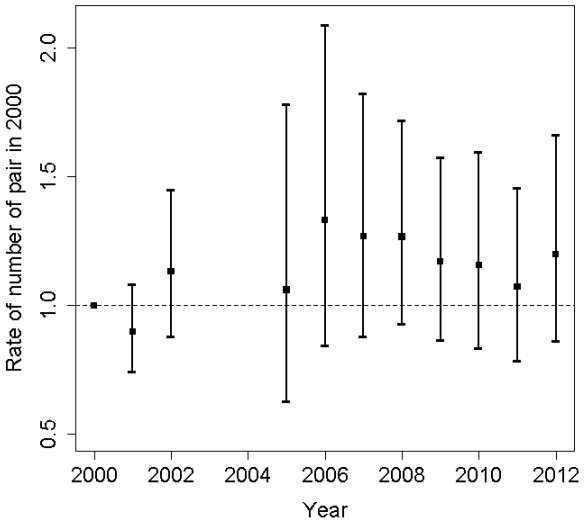
The European Honey Buzzard *Pernis apivorus*



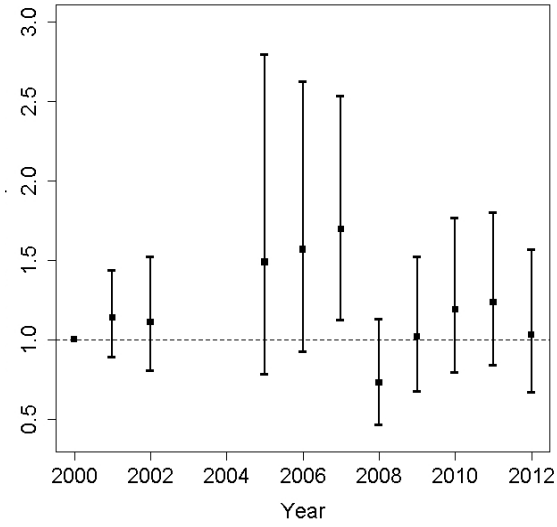
The Hen Harrier *Circus cyaneus*



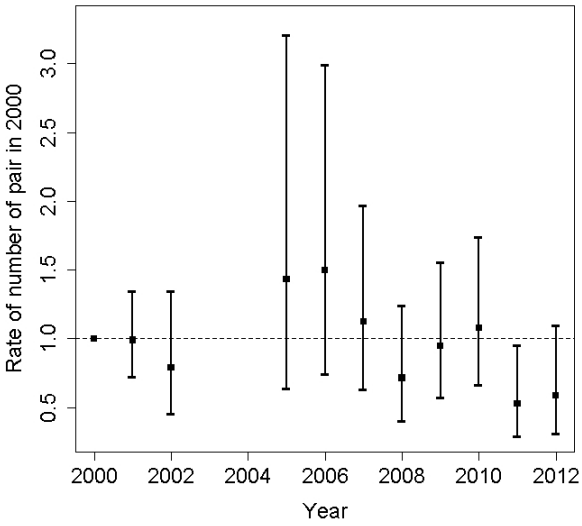
The Eurasian Hobby *Falco subbuteo*



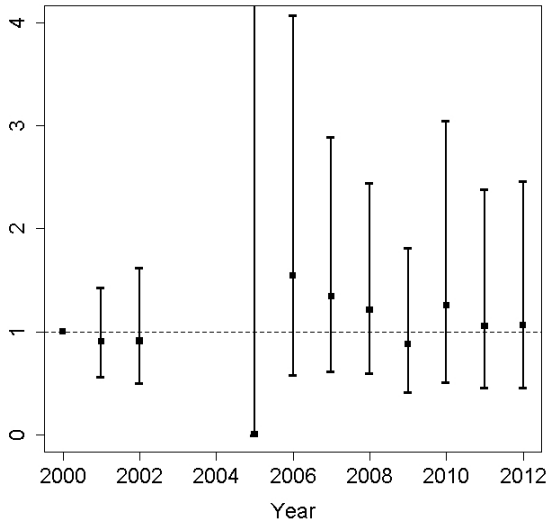
The Northern Goshawk *Accipiter gentilis*



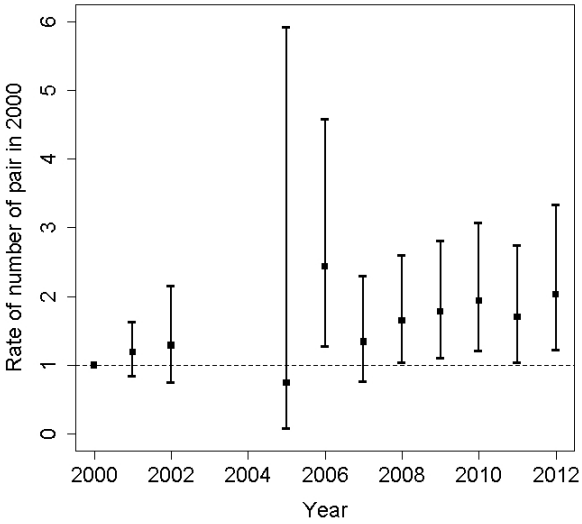
The Montagu's Harrier *Circus pygargus*



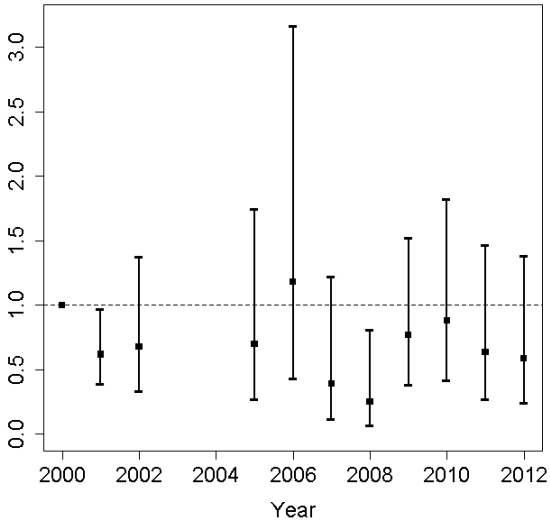
The Red Kite *Milvus milvus*



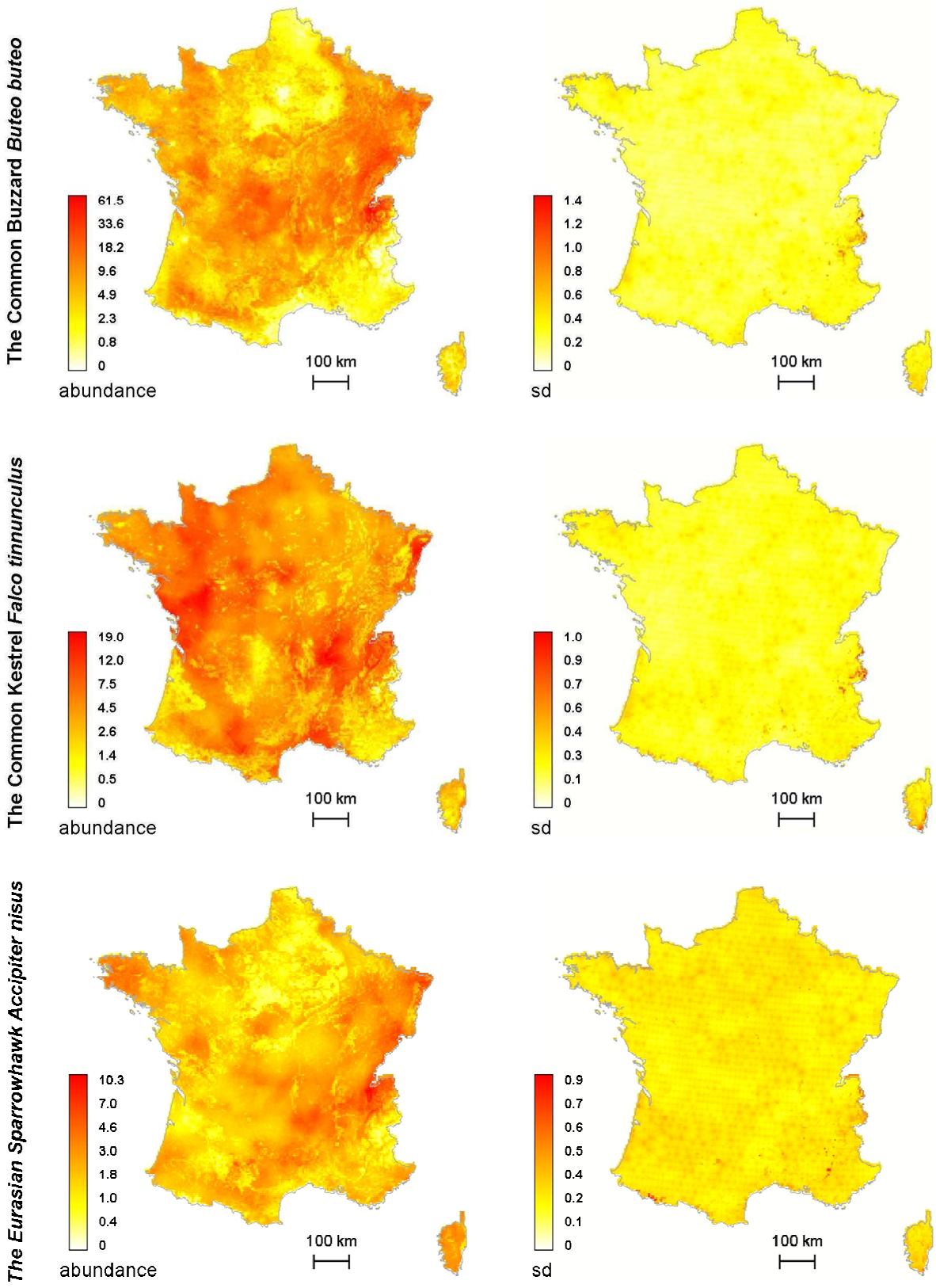
The Short-toed Snake Eagle *Circaetus gallicus*



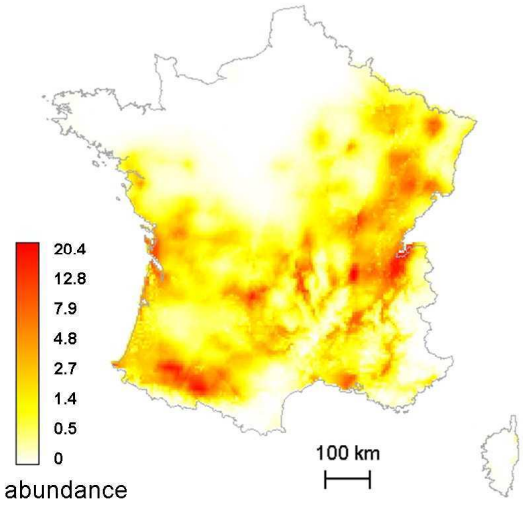
The Marsh Harrier *Circus aeruginosus*

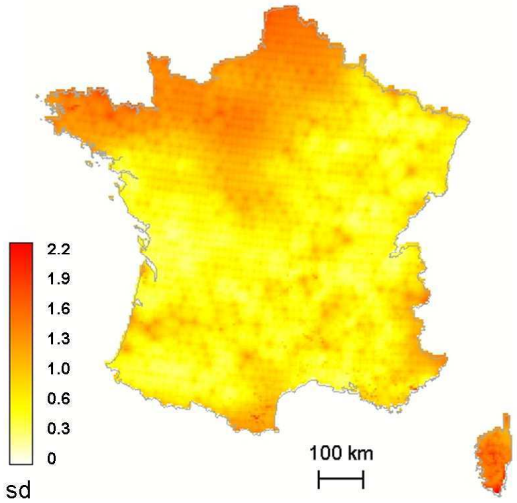


Appendix D: Distribution maps of diurnal raptors in France

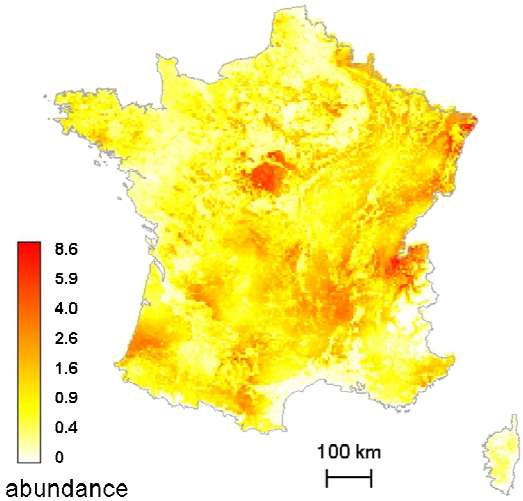


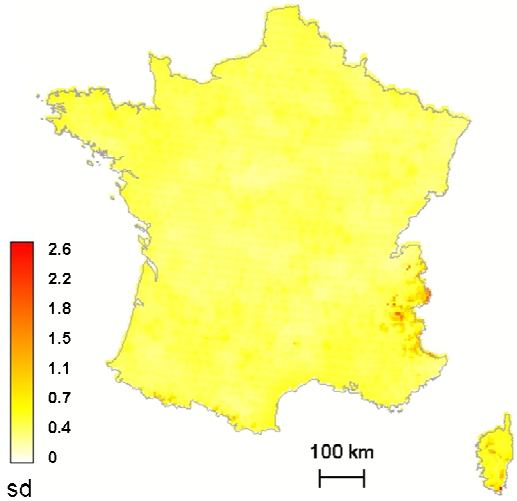
The Black Kite Milvus migrans



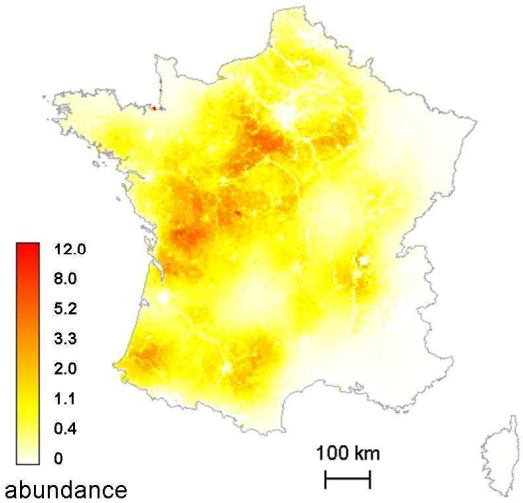


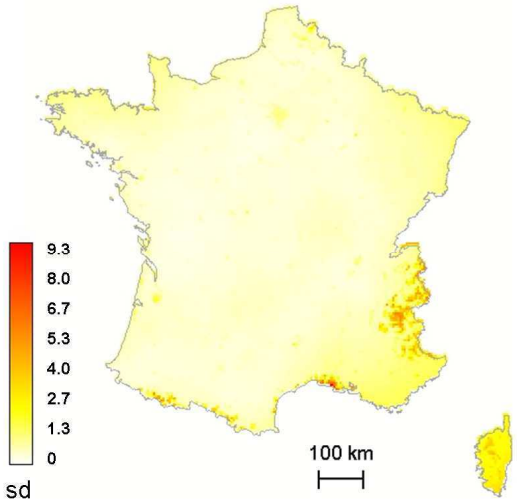
The European Honey Buzzard *Pernis apivorus*

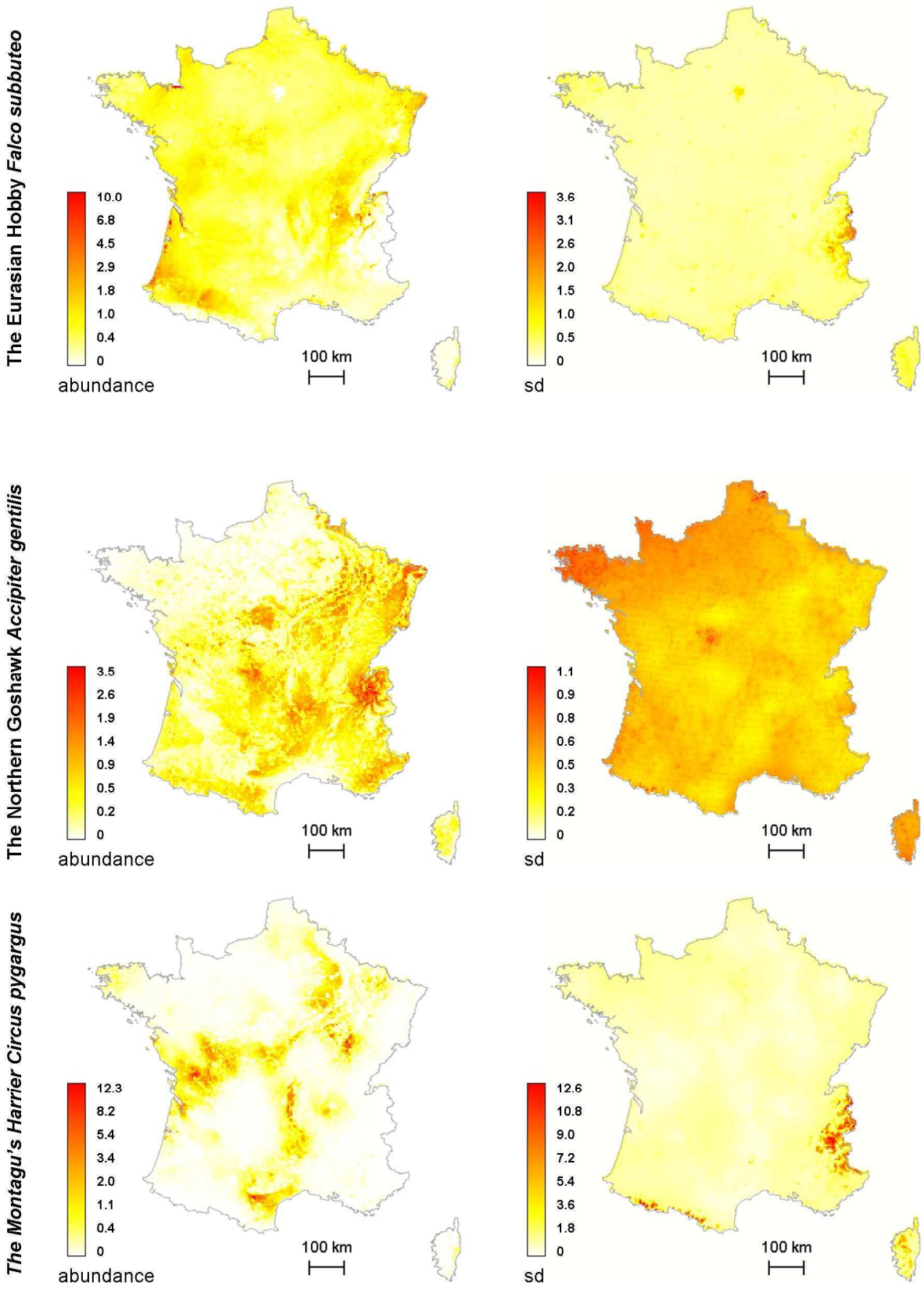




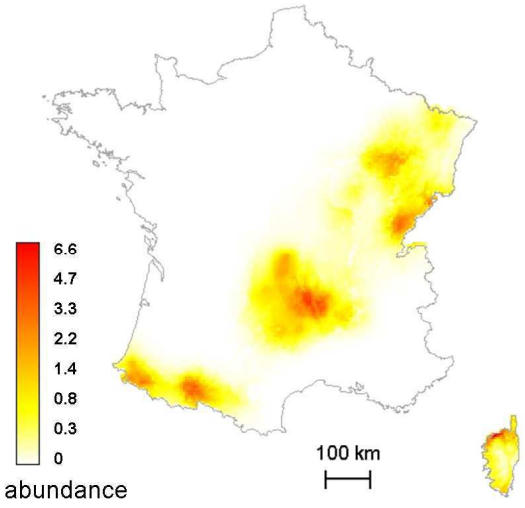
The Hen Harrier *Circus cyaneus*

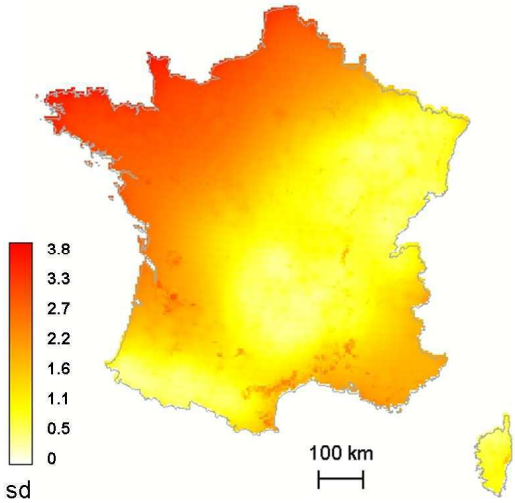




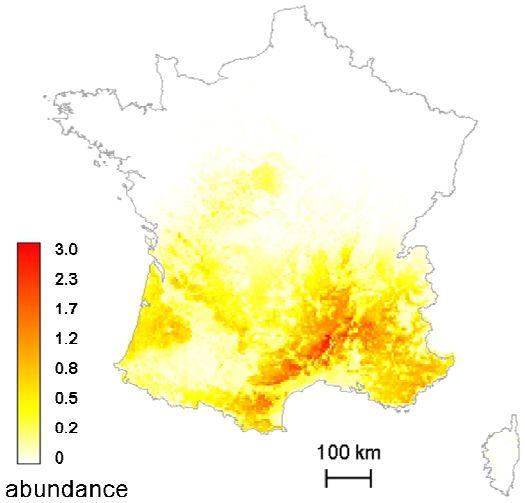

The Eurasian Hobby Falco subbuteo
10.0
6.8
4.5
2.9
1.8
1.0
0.4
0
abundance
100 km
3.6
3.1
2.6
2.0
1.5
1.0
0.5
0
sd
100 km
The Northern Goshawk Accipiter gentilis
3.5
2.6
1.9
1.4
0.9
0.5
0.2
0
abundance
100 km
1.1
0.9
0.8
0.6
0.5
0.3
0.2
0
sd
100 km
The Montagu's Harrier Circus pygargus
12.3
8.2
5.4
3.4
2.0
1.1
0.4
0
abundance
100 km
12.6
10.8
9.0
7.2
5.4
3.6
1.8
0
sd
100 km

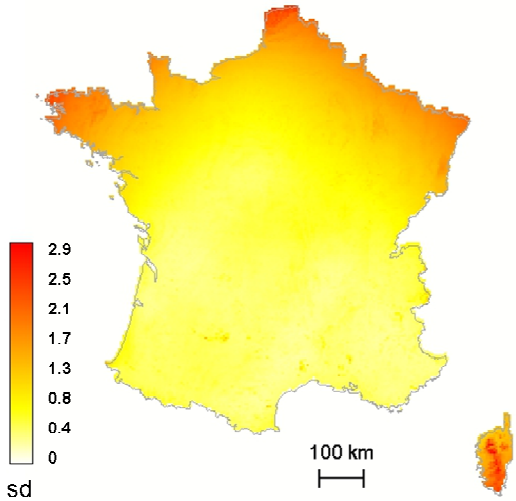
The Red Kite *Milvus milvus*



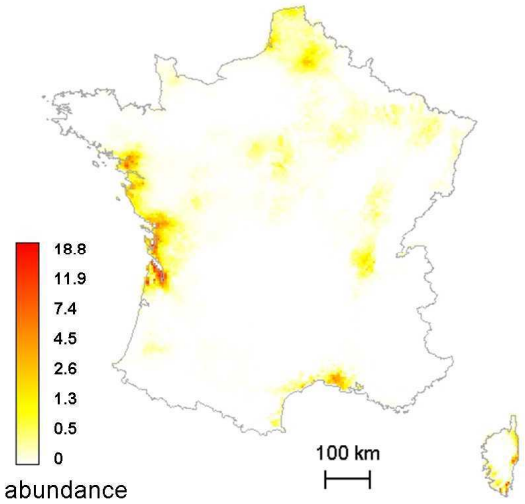


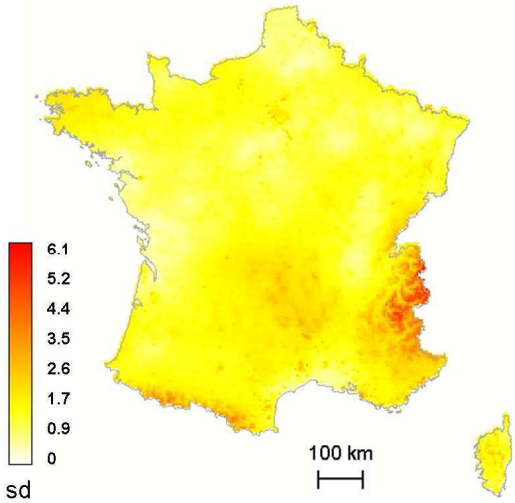
The Short-toed Snake Eagle *Circaetus gallicus*





The Marsh Harrier *Circus aeruginosus*





DISCUSSION GENERALE

Le problème de confusion spatiale

Le principal problème des modèles spatialement explicites actuels est que la plupart négligent une possible redondance (collinéarité) entre le terme spatial et certaines variables qui sont elles mêmes spatialement autocorrélées (Reich *et al.* 2006 ; Hughes & Haran 2013). Ainsi, l'utilisation d'un terme spatial peut conduire à cacher l'effet d'autres variables (Hodges & Reich 2010). Ce problème de confusion entre le terme spatial et les variables présentes dans le modèle a conduit à penser que le fait de ne pas tenir compte de l'autocorrélation spatiale résiduelle pouvait affecter grandement l'estimation des coefficients de régression (c'est à dire la moyenne des effets), voir même inverser l'effet de certaines variables (Kühn 2007). Le premier chapitre de cette thèse montre d'ailleurs lui aussi que l'ajout d'un terme spatial (dans ce cas un ensemble de filtres spatiaux) conduit à inverser l'effet de certaines variables. Il est pourtant bien connu que d'un point de vue théorique, la présence d'autocorrélation résiduelle affecte principalement la variance des estimations (voir Cressie 1993 p. 1-26). Dans ce premier chapitre, nous avons donc négligé l'influence de la collinéarité entre les filtres spatiaux et les variables utilisées en évoquant le fait que la corrélation était toujours inférieure à 0.5. Le problème est que même si ces corrélations étaient peu élevées, chaque variable était souvent corrélée à un grand nombre de filtres spatiaux, ce qui conduisait tout de même à une multicollinéarité importante (voir Table C1 dans l'Annexe C du Chapitre 1). Sans surprise ce sont les variables qui présentaient la plus forte multicollinéarité avec les filtres spatiaux qui conduisaient à un changement important des coefficients (*shift*) entre le modèle spatial et le modèle non-spatial. Des études en écologie ont déjà essayé d'élucider la raison de ce '*shift*' des coefficients mais n'y sont pas parvenues (Hawkins *et al.* 2007 ; Bini *et al.* 2009), concluant même qu'il n'est pas possible de l'expliquer (Bini *et al.* 2009). Il est donc nécessaire d'alerter rapidement les utilisateurs de ces méthodes de la raison de ce '*shift*', qui est en fait simplement lié à un problème de multicollinéarité lors de l'utilisation de certains modèles spatialement explicites.

Validation croisée spatialisée et régression spatiale restreinte

Malgré le nombre important de méthodes existantes pour modéliser des données spatialement structurées, certains problèmes n'ont pas encore de solution clairement définie. C'est particulièrement le cas pour l'étape de sélection de variables en présence d'autocorrélation résiduelle et de surdispersion. En pratique, ces problèmes sont la plupart du temps ignorés durant cette étape (voir Kühn *et al.* 2009 ; Mellin *et al.* 2010 ; Vaclavik *et al.* 2011), bien que cela puisse conduire à considérer des modèles trop complexes (Diniz-Filho *et al.* 2008 ; Richards 2008). La solution proposée et discutée dans les chapitres 2 et 3, basée sur une procédure de validation croisée spatialisée, est une méthode relativement simple pour y remédier. Son principal atout provient de sa capacité à traiter le problème d'autocorrélation résiduelle sans faire appel aux méthodes d'estimation complexes normalement nécessaires

(voir Diggle *et al.* 1998 ; Beale *et al.* 2010). De plus, elle permet de tenir compte de l'autocorrélation présente dans les variables elles-mêmes, chose que les modèles spatialement explicites actuels négligent (voir le problème de confusion spatiale, Hodges & Reich 2010), rendant ces méthodes douteuses pour la sélection des variables.

Une solution proposée au problème de confusion spatiale consiste à utiliser une régression spatiale restreinte, où le terme spatial est orthogonal aux variables présentes dans le modèle (Reich *et al.* 2006 ; Hodges & Reich 2010 ; Hughes & Haran 2013). Cette méthode vient d'ailleurs tout juste (en Août 2013) de donner lieu à la création d'un package pour le logiciel R nommé 'ngspatial' (Hughes in press), qui reprend les modèles de Julian Besag (Besag 1972 ; Besag *et al.* 1991), mais en corrigeant le problème de confusion spatiale. Le problème de la sélection des variables reste néanmoins toujours problématique dans le cas d'une régression spatiale restreinte. En effet, considérer un modèle spatialement explicite restreint pour chaque ensemble de variables possible conduirait à estimer des termes spatiaux différents à chaque fois (car orthogonaux aux variables présentes), rendant la comparaison entre les modèles douteuse. Par ailleurs, l'utilisation d'un terme spatial unique, orthogonal à toutes les variables disponibles, est difficilement envisageable car à chaque fois qu'un modèle évalué n'intègrera pas toutes les variables autocorrélées, la structure résiduelle sera affectée et donc le terme spatial commun ne sera plus adapté. Une solution pourrait être de considérer l'ensemble des variables dans un même modèle mais en contraignant l'estimation des paramètres par une méthode de régularisation tel que la régression Lasso (présentée dans l'introduction), ce qui donnerait la régression Lasso spatiale restreinte. Néanmoins, à ma connaissance, il n'existe pas encore d'outils pour estimer ce genre de modèle. La procédure de sélection par validation croisée spatialisée, pourra quant à elle permettre de sélectionner les variables dans un tel cas de figure, variables qui devront ensuite être intégrées dans une régression spatiale restreinte.

Application de la validation croisée spatialisée dans d'autres cadres

La validation croisée spatialisée, détaillée dans les chapitres 2 & 3, peut être modifiée pour traiter le cas d'autocorrélation résiduelle temporelle ou même spatio-temporelle. Il suffit de retirer les données autocorrélées dans le temps qui invalident la validation croisée standard. Par analogie, la méthode peut aussi être utilisée dans des cas non-spatiaux où l'autocorrélation résiduelle est délimitée à des entités bien définies, comme par exemple des individus. Il suffit alors d'exclure toutes les observations de l'entité en question. Ce genre de procédure a déjà été largement utilisée, portant parfois le nom de *leave-one-individual-out* (LaFrance *et al.* 2003 ; Sakar & Kursun 2010). Cette procédure est surtout mise en place pour évaluer le pouvoir prédictif du modèle choisi plutôt que pour sélectionner les variables affectant l'ensemble des individus. Néanmoins, il a été récemment montré que ce genre de procédure est asymptotiquement équivalent à utiliser l'AIC dans le cadre de modèles mixtes (Fang 2011), plus précisément équivalent à l'AIC marginal, mAIC (voir Vaida & Blanchard 2005), ce qui renforce sa validité pour la sélection des variables.

Par ailleurs, la validation croisée spatialisée pourrait aussi être utilisée pour l'estimation des paramètres. On parle alors plutôt de *jackknife* (Quenouille 1949, 1956 ; Tukey 1958) dont la procédure est similaire au leave-one-out (voir Miller 1974) mais où l'objectif est cette fois l'estimation des paramètres. Un tel procédé a déjà été proposé (voir Lele 1991 ; Cressie 1993 ; Kramer *et al.* 2001) mais a été rapidement supplanté par l'apparition des modèles de régression spatialement explicites. L'avantage de ces derniers est qu'en plus de tenir compte de l'autocorrélation résiduelle, ils permettent d'utiliser cette information pour la prédiction aux sites non-échantillonnés (comme cela a été réalisé dans le chapitre 5), du moins si le point à prédire se situe à une distance inférieure à la portée de l'autocorrélation résiduelle.

Le postulat de stationnarité

Un aspect important des modèles pour les données à large échelle qui n'a pas été détaillé au cours de cette thèse concerne le respect du postulat de stationnarité. Il convient cependant de bien différencier la stationnarité de l'effet des variables et la stationnarité du terme spatial. Dans le premier cas, cela signifie que la relation entre la variable réponse (par exemple l'abondance d'une espèce) et les variables utilisées dans le modèle (par exemple la quantité de forêt) est constante dans l'espace (voir Osborne *et al.* 2007 ; Hothorn *et al.* 2011 ; Hawkins 2012 ; Miller 2012). Dans le second cas, cela signifie que la structure des résidus a les mêmes propriétés quelque soit la position dans l'espace (voir Myers 1989). On suppose en général une stationnarité de second ordre, c'est à dire que la moyenne et variance sont invariantes par translation. Supposer la stationnarité de l'effet des variables pour l'ensemble de la zone d'étude lorsque celle-ci est grande peut sembler irréaliste (voir Hothorn *et al.* 2011 ; Miller 2012). Néanmoins il est relativement facile de corriger ce cas de non-stationnarité. La solution la plus simple consiste à ajouter des interactions entre les variables qui sont elles mêmes spatialement structurées (Miller 2012). Une autre solution similaire est d'ajouter des interactions entre les variables utilisées et la position des observations dans l'espace (interactions possiblement non-linéaires, Hastie & Tibshirani 1993 ; Osborne *et al.* 2007 ; Hothorn *et al.* 2011). Le postulat de stationnarité fait sur les propriétés de la structure résiduelle est lui aussi peu réaliste pour des données à large échelle (Haas 1990 ; Fuentes 2002), et il existe également des outils simples pour en tenir compte, comme une distorsion artificielle de l'espace pour recouvrer des propriétés homogènes (Sampson & Guttorp 1992) ou considérer seulement des sous ensembles homogènes (Haas 1990, 1995). L'écart à la stationnarité (au moins du second ordre) est supposée faible si la composante fixe du modèle est bien spécifiée car il reste alors un terme résiduel vraisemblablement homogène. Notons que ces principes s'appliquent aussi pour des données temporelles ou spatio-temporelles (voir Cressie & Wikle 2011 pour plus de détails).

La validation croisée spatialisée, présentée dans les chapitres 2 & 3, nécessite le respect du postulat de stationnarité des variables. Si ce n'est pas le cas la méthode risque de ne pas détecter des variables importantes, comme cela se produit pour le Busard cendré dans

le chapitre 3. Il faut donc inclure les interactions suspectées parmi les ensembles de variables à traiter ou éventuellement inclure des interactions avec l'espace si cela n'est pas suffisant. La validation croisée spatialisée ne nécessite par contre pas nécessairement de respecter le postulat de stationnarité sur l'intensité (la variance) de l'autocorrélation résiduelle puisqu'elle se contente de retirer les données qui posent des problèmes. Il reste cependant nécessaire de respecter ce postulat de stationnarité sur le paramètre de portée de l'autocorrélation résiduelle. Une petite modification de la méthode pourrait néanmoins permettre de traiter ce cas de non-stationnarité, par exemple en faisant varier dans l'espace la taille du buffer visant à retirer les données autocorrélées. Il en va de même pour le phénomène d'anisotropie.

Liens entre autocorrélation résiduelle et surdispersion

Nous avons pu voir que les causes de l'autocorrélation résiduelle et de la surdispersion étaient en fait généralement liées à une mauvaise spécification du modèle (voir chapitre 3 & 4, Hinde & Demétrio 1998 ; Dormann *et al.* 2007). Cette mauvaise spécification se produit généralement en raison de l'omission de variables importantes mais inconnues, non mesurables ou non mesurées. A partir de ce constat, il est possible d'identifier deux cas de figures pour mieux comprendre les différences et les ressemblances entre l'autocorrélation résiduelle et la surdispersion.

1er cas de figure - Lorsque ces variables non-observées ne sont pas autocorrélées ni dans le temps ni dans l'espace, leur omission conduit simplement à augmenter l'hétérogénéité non-expliquée, due au fait que la variance de ces variables n'est pas prise en compte par le modèle. Cela ne pose pas de problème particulier dans le cadre du modèle linéaire classique puisque celui-ci estime un paramètre de variance σ^2 qui reflète justement celle non prise en compte par les différents termes du modèle. Un problème se produit par contre lorsqu'il n'y a pas ce paramètre de variance dans le modèle, comme c'est le cas pour la régression de Poisson ou la régression binomiale.

2eme cas de figure - Lorsque ces variables non-observées sont cette fois autocorrélées dans le temps et/ou dans l'espace, leur omission conduit à la fois à de l'hétérogénéité non-observée et à de l'autocorrélation résiduelle. L'autocorrélation résiduelle n'est pas un problème en soit mais conduit à surévaluer le nombre de données indépendantes (dans le cas de l'autocorrélation positive) et donc de surévaluer aussi le nombre de degrés de liberté. Cela pose problème même dans le cadre du modèle linéaire puisque l'estimation de la précision des effets dépend du nombre de degrés de liberté. Si celui-ci est surévalué, la précision des effets l'est aussi.

Tenir compte de la surdispersion ne permet donc pas de corriger le problème d'autocorrélation résiduelle. Par contre, tenir compte de l'autocorrélation résiduelle peut permettre de résoudre le problème de surdispersion. En effet, les méthodes visant à corriger le problème d'autocorrélation résiduelle consistent souvent à incorporer une composante de variance, structurée dans l'espace et /ou dans le temps (GLMM, voir par exemple Diggle *et al.* 1998). Ainsi les modèles de régression sans composante de variance s'en voient attribuer

une, ce qui corrige potentiellement les deux problèmes d'un seul coup. Il n'est donc pas étonnant de voir que les approches combinant le traitement de l'autocorrélation résiduelle et de la surdispersion conduisent à résultats proches que la surdispersion soit ou non prise en compte (Gschlößl & Czado 2008 ; Haining *et al.* 2009 ; Neyens *et al.* 2012). Ne pas en tenir compte conduit néanmoins à un nombre de paramètres 'effectifs' souvent plus important.

Au final, quel modèle utiliser pour modéliser la distribution de populations à large échelle géographique ?

Tout au long de cette thèse, il a été souligné que si l'on souhaite modéliser la distribution d'une population à large échelle, il convient d'utiliser des outils adéquats. Je propose, dans cette dernière partie, de résumer en 3 points les principales caractéristiques que devrait présenter un modèle de régression dans un tel cadre. Pour cela, je suppose que les variables sont déjà sélectionnées ou qu'une méthode de régularisation (comme par exemple le Lasso) sera utilisée pour l'estimation des paramètres. Je n'aborde donc pas de nouveau l'étape de sélection de variables, qui a déjà été traitée en détail auparavant.

1) Le type de distribution du modèle utilisé doit être en accord avec les propriétés de la variable mesurée. Par exemple, pour des comptages, il convient d'utiliser un modèle de régression adéquat comme le modèle de Poisson ou celui binomial négatif (voir O'Hara & Kotze 2010). Si le comptage provient, qui plus est, d'un mélange de deux processus et que l'un d'eux ne génère que des zéros, les modèles zéro-enflés à mélange (ZIMs, voir Lambert 1992) sont adaptés (mais voir aussi le chapitre 4 de cette thèse).

2) Le modèle utilisé doit rendre compte du fait que toutes les variables explicatives ne sont pas prises en compte dans le modèle et qui plus est, certaines d'entre elles sont vraisemblablement autocorrélées dans l'espace et /ou dans le temps. Cela implique la correction de la précision des estimations, de préférence par l'ajout d'un terme spatial et /ou temporel explicite, et ce en particulier si l'objectif est l'interpolation aux sites et /ou aux périodes non-échantillonnés. Il faut par contre éviter que ce terme n'entre pas en concurrence avec les variables explicatives (voir Hodges & Reich 2010), en utilisant par exemple un terme orthogonal à ces variables. Il est aussi la plupart du temps nécessaire d'utiliser une composante de variance supplémentaire dans le cadre de la régression de Poisson ou celle de la régression binomiale, rendant compte d'une hétérogénéité non-observée additionnelle (surdispersion), par exemple lié au processus inhomogène de récolte des données.

3) Le modèle doit être construit de manière à tenir compte d'éventuels effets de non stationnarité. Aussi, il ne faut pas négliger les interactions importantes dans le modèle (Hothorn *et al.* 2011). De plus, il convient de vérifier la stationnarité du terme d'autocorrélation résiduelle (si utilisé), par exemple en découpant l'espace et /ou le temps en morceaux disjoints et de voir si la structure observée est semblable ou non. Dans le cas contraire, des solutions doivent être envisagées pour en tenir compte (voir Cressie & Wikle 2011 pour un aperçu récent).

Pour conclure, il faut retenir que de nombreuses méthodes statistiques existent pour modéliser les facteurs influençant les paramètres d'une population et que le choix de l'une d'entre elle va conditionner les résultats et les conclusions qui vont être données. Il convient de rester critique face aux nombreux outils proposés et de bien identifier les caractéristiques des méthodes finalement utilisées. La présence d'autocorrélation résiduelle et de surdispersion peut réellement conduire à de fausses conclusions, en raison de la sur-estimation de la précision des effets. Mais en tenir compte de manière inadéquat peut être encore pire puisque les coefficients eux même peuvent être affectés. Le comble serait de montrer l'effet bénéfique d'un type d'habitat sur une espèce alors qu'il lui est en faite défavorable. Cela peut se produire en raison du phénomène de confusion spatial détaillé auparavant ou en raison de la présence de surdispersion lors de l'utilisation d'un modèle zéro-enflé (voir Chapitre 4). Il faut donc garder à l'esprit que l'autocorrélation résiduelle et la surdispersion n'affectent généralement pas de manière importante l'estimation des coefficients mais plutôt leur précision. Si l'on constate une forte différence entre les coefficients estimés par un modèle tenant compte de ces problèmes (un modèle plutôt complexe) *versus* un autre n'en tenant pas compte (un modèle plutôt simple), il faut d'abord vérifier la validité du modèle complexe avant de conclure que le modèle simple conduit à une mauvaise estimation des coefficients.

RÉFÉRENCES GENERALES

Aitken A. (1935) On Least Squares and Linear Combinations of Observations. Proceedings of the Royal Society of Edinburgh, vol. 55, p. 42-48.

Akaike H. (1973) Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (B. N. Petrov and F. Csàki, eds.). Academiai Kiàdo, Budapest, p. 267-281

Akaïke H. (1974) A new look at statistical model identification. IEEE Transaction on automatic control, vol. 19, n° 6, p. 716-723.

Alldredge M. W., Pollock K. H. & Simons T. R. (2006) Estimating detection probabilities from multiple-observer point counts. The Auk, vol. 123, n° 4, p. 1172-1182.

Allee W. C. (1931) *Animal aggregations, a study in general sociology*. The University of Chicago Press, Chicago, Illinois, U.S.A.

Allee W. C. (1938) *The social life of animals.* W. W. Norton & Company Inc., New York, U.S.A.

Allee W. C., Emerson A. E., Park O., Park T. & Schmidt K. P. (1949) *Principles of Animal Ecology.* W. B. Saunders, Philadelphia, PA.

Allen D. M. (1974) The relationship between variable selection and data augmentation and a method for prediction. Technometrics, vol. 16, n° 1, p. 125-127.

Altman N. S. (1990) Kernel smoothing of data with correlated errors. Journal of the American Association, vol. 85, p. 749-759.

Andelman S. J. & Fagan W.F. (2000) Umbrellas and flagships: Efficient conservation surrogates or expensive mistakes? Proceeding of the National Academy of Science, vol. 97, n° 11, p. 5954-5959.

Anderson D. R. (2001) The need to get the basics right in wildlife field studies. Wildlife Society Bulletin, vol. 29, n° 4, p. 1294-1297.

Anderson D. R. (2003) Response to Engeman: index values rarely constitute reliable information. Wildlife Society Bulletin, vol. 31, n° 1, p. 288-291.

Anderson D. R. & Burnham K. P. (1999) Understanding information criteria for selection among capture-recapture or ring recovery models. Bird Study, vol. 46:S1, p. S14-S21.

Anderson D. R., Burnham K. P. & White G. C. (1994) AIC model selection in overdispersed capture-recapture data. Ecology, vol. 75, n° 6, p. 1780-1793.

Anderson R. L. (1954) The problem of autocorrelation in regression analysis. Journal of the American Statistical Association, vol. 49, n° 265, p. 113-129.

Anselin L. (1988) *Spatial Econometrics: Methods and Models*. Kluwer Academic, Dordrecht.

Arlot S. & Celisse A. (2010). A survey of cross-validation procedures for model selection. Statistics Surveys, vol. 4, p. 40-79.

Austin M. P. (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. Ecological Modelling, vol. 157, p. 101-118.

Austin M. (2007) Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling, vol. 200, p. 1-19.

Bahn V. & McGill B. J. (2007) Can niche-based distribution models outperform spatial interpolation? Global Ecology and Biogeography, vol. 16, n° 6, p. 733-742.

Bahn V. & McGill B. J. (2013) Testing the predictive performance of predictive models. Oikos, vol. 122, p. 321-331.

Bahn V., O'Connor R. J. & Krohn, W. B. (2006) Importance of spatial autocorrelation in modeling bird distributions at a continental scale. Ecography, vol. 29, p. 835-844.

Baillie S. R. (1990) Integrated population monitoring of breeding birds in Britain and Ireland. Ibis, vol. 132, n° 2, p. 151-166.

Baillie S. R. (1991) Monitoring terrestrial breeding bird populations. In *Monitoring for Conservation and Ecology* (B. Goldsmith, ed.). Chapman & Hall, London, p. 112-132.

Baillie S. R., Sutherland W.J., Freeman S. N., Gregory R.D. & Paradis E. (2000) Consequences of large-scale processes for the conservation of bird populations. Journal of Applied Ecology, vol. 37, suppl. 1, p. 88-102.

Barry S. & Elith J. (2006) Error and uncertainty in habitat models. Journal of Applied Ecology, vol. 43, p. 413-423.

Bartlett M. S. (1935) Some aspects of the time-correlation problem in regard to tests of significance. Journal of the Royal Statistical Society, vol. 98, n°3, p. 536-543.

Beale C. M., Lennon J. J., Yearsley J. M., Brewer M. J. & Elston D. A. (2010) Regression analysis of spatial data. Ecology Letters, vol. 13, p 246-264.

Beguin J., Martino S., Rue H. & Cumming S. G. (2012) Hierarchical analysis of spatially autocorrelated ecological data using integrated nested Laplace approximation. Methods in Ecology and Evolution, vol. 3, p. 921-929.

Berk R. & MacDonald J. M. (2008) Overdispersion and Poisson regression. Journal of Quantitative Criminology, vol. 24, n° 3, p. 269-284.

Berkson J. (1944) Application of the logistic function to bio-assay. Journal of the American Statistical Association, vol. 39, n° 227, p. 357-365.

Berkson J. (1953) A statistically precise and relatively simple method of estimating the bioassay with quantal response, based on the logistic function. Journal of the American Statistical Association, vol. 48, n° 263, p. 565-599.

Berkson J. (1955) Maximum likelihood and minimum |chi2 estimates of the logistic function. Journal of the American Statistical Association, vol. 59, n° 269, p. 130-162.

Besag J. E. (1972) Nearest-neighbour systems and the auto-logistic model for binary data. Journal of the Royal Statistical Society. Series B (Methodological), vol. 34, n° 1, p. 75-83.

Besag J. (1974) Spatial interaction and the statistical analysis of lattice systems. Journal of the Royal Statistical Society, Series B (Methodological), vol. 36, n° 2, p. 192-236.

Besag J. E., York J. & Mollié A. (1991) Bayesian image restoration with applications in spatial statistics. Annals of the Institute of Statistical Mathematics, vol. 43, n° 1, p. 1-20.

Betts M. G., Ganio L. M., Huso M. P., Som N. A., Huettmann F., Bowman J. & Wintle B.A. (2009) Comment on "Methods to account for spatial autocorrelation in the analysis of species distributional data: a review". Ecography, vol. 32, p. 374-378.

Bini L. M., Diniz-Filho J. A. F., Rangel T. F. L. V. B., Akre T. S. B., Albaladejo R. G., Albuquerque F. S., Aparicio A., Araújo M. B., Baselga A., Beck J., Isabel Bellocq M., Böhning-Gaese K., Borges P. A. V., Castro-Parga I., Khen Chey V., Chown S. L., De Marco Jr P., Dobkin D. S., Ferrer-Castán D., Field R., Filloy J., Fleishman E., Gómez J. F., Hortal J., Iverson J. B., Kerr J. T., Kissling D. W., Kitching I. J., León-Cortés J. L., Lobo J. M., Montoya D., Morales-Castilla I., Moreno J. C., Oberdorff T., Olalla-Tárraga M. Á., Pausas J. G., Qian H., Rahbek C., Rodríguez M. Á., Rueda M., Ruggiero A., Sackmann P., Sanders N. J., Carina Terribile L., Vetaas O. R. & Hawkins B. A. (2009) Coefficient shifts in geographical ecology: an empirical evaluation of spatial and non-spatial regression. Ecography, vol. 32, p. 193–204.

Bishop Y.M. M., Fienberg S. E. & Holland P.W. (1975) *Discrete multivariate analysis: theory and practice.* M.I.T. Press, Cambridge, Massachusetts.

Bivand R. (2002) Spatial econometrics functions in R: Classes and methods. Journal of Geographical Systems, vol. 4, n°4, p. 405-421.

Bliss C. I. (1934) The method of probits--A correction. Science, New Series, vol. 79, n° 2053, p. 409-410.

Boes S. (2010) Count data models with correlated unobserved heterogeneity. Scandinavian Journal of Statistics, vol. 37, p. 382–402.

Bolker B. M. (2008) *Ecological models and data in R.* Princeton University Press, Princeton.

Bolker B. M., Brooks M. E., Clark C. J., Geange S. W., Poulsen J. R., Stevens M. H. & White J-S. S. (2009) Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution, vol.24, n° 3, p. 127-135.

Borcard D. & Legendre P. (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. Ecological Modelling, vol. 153, p. 51- 68.

Box G. E. P. & Draper N. R. (1987) *Empirical Model-Building and Response Surfaces*. Wiley, New York.

Box J. E. P. & Pearce D. A. (1970) Distribution of residual autocorrelations in autoregressive-Integrated moving average time series models. Journal of the American Statistical Association, vol. 65, n° 332, p. 1509-1526.

Breiman L. (1995) Better subset regression using the nonnegative garrote. Technometrics, vol. 37, n° 4, p. 373-384.

Breiman L. & Freedman D. (1983) How many variables should be entered in a regression equation? Journal of the American Statistical Association, vol. 78, n° 381, p. 131-136.

Brenning A. (2005) Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Sciences, vol. 5, p. 853-862.

Breslow N. E. & Clayton D. G. (1993) Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, vol. 88, n° 421, p. 9-25.

Bretagnolle V. & Hanneke G. (2010) Predator–prey interactions and climate change. In: *Effect of climate change on birds* (A. P. Moller, F. Wolfgand, P. Berthold, eds). Oxford University Press, Oxford, p. 227-248.

Browne W. J. & Draper D. (2006) A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis, vol. 1, n° 3, p. 473–514.

Buckland S. T., Anderson D. R., Burnham K. P. & Laake J. L. (1993) *Distance sampling: Estimating abundance of biological populations.* Chapman and Hall, London.

Buckland S. T., GoudieI. B. J. & Borcherset D. L. (2000) Wildlife population assessment: past developments and future directions. International Biometric Society, vol. 56, n° 1, p. 1-12.

Buckland S. T., Anderson D. R., Burnham K.P., Laake J. L., Borchers D. L. & Thomas L. (2001) *Introduction to distance sampling.* Oxford University Press, Oxford.

Buckland S. T., Anderson D. R., Burnham K.P., Laake J. L., Borchers D. L. & Thomas L. (2007) *Advanced distance sampling.* Oxford University Press, New York.

Buckland S. T., Magurran A. E., Green R. E. & Fewster R. M. (2005) Monitoring change in biodiversity through composite indices. Philosophical Transactions of the Royal Society B, vol. 360, p. 243–254.

Buckley L. B., Urban M. C., Angilletta M. J., Crozier L. G., Rissler L. J. & Sears M. W. (2010) Can mechanism inform species distribution models? Ecology Letters, vol. 13, p. 1041–1054.

Burman P., Chow E. & Nolan D. (1994) A cross-validatory method for dependent data. Biometrika, vol. 81, p. 351-358.

Burnham K. P. & Anderson D.R. (2001) Kullback- Leibler information as a basis for strong inference in ecological studies. Wildlife Research, vol. 28, p.11-119.

Burnham K. P., Anderson D. R. (2002) *Model selection and multimodel inference: A practical information-theoretic approach.* Second edition, Springer, New York.

Burnham K. P. & Anderson D. R. (2004) Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods Research, vol. 33, p. 261-304.

Burnham K. P., Anderson D. R., Laake J. L. (1980) Estimation of density from line transect: Sampling of biological populations. Wildlife Monographs, vol. 72, p. 3-202.

Bustamante J. & Seoane J. (2004) Predicting the distribution of four species of raptors (Aves: Accipitridae) in southern Spain: statistical models work better than existing maps. Journal of Biogeography, vol. 31, p. 295-306.

Butchart S. H. M., Walpole M., Collen B., van Strien A., Scharlemann J. P. W., Almond R. E. A., Baillie J. E. M., Bomhard B., Brown C., Bruno J., Carpenter K. E., Carr G. M., Chanson J., Chenery A. M., Csirke J., Davidson N. C., Dentener F., Foster M., Galli A., Galloway J. N., Genovesi P., Gregory R. D., Hockings M., Kapos V., Lamarque J-F., Leverington F., Loh J., McGeoch M. A., McRae L., Minasyan A., Hernández Morcillo M., Oldfield T. E. E., Pauly D., Quader S., Revenga C., Sauer J. R., Skolnik B., Spear D., Stanwell-Smith D., Stuart S. N., Symes A., Tierney M., Tyrrell T.D., Vié J-C., Watson R. (2010) Global biodiversity: Indicators of recent declines. Science, vol. 328, p. 1164-1168.

Cameletti M., Lindgren F., Simpson D. & Rue H. (2012) Spatio-temporal modeling of particulate matter concentration through the SPDE approach. Advances in Statistical Analysis, vol. 97, n° 2, p. 109-131.

Cameron A. C. & Trivedi P. K. (2013) *Regression Analysis of Count Data*. Second edition. Cambridge University Press, Cambridge.

Carignan V. & Villard M. A. (2002) Selecting indicator species to monitor ecological integrity: a review. Environmental Monitoring and Assessment, vol. 78, p. 45–61.

Carl G. & Kühn I. (2010) A wavelet-based extension of Generalized Linear Models to remove the effect of spatial autocorrelation. Geographical analysis, vol. 42, p. 323-337.

Caro T. M. & O'Doherty G. (1999) On the use of surrogate species in conservation biology. Conservation Biology, vol. 13, n°4, p. 805–814.

Carroll C., Noss R. F. & Paquet P. C. (2001) Carnivores as focal species for conservation planning in the rocky mountain region. Ecological Applications, vol. 11, p. 961-980.

Cassemiro F. A. S., Diniz-Filho, J. A. F., Rangel T. F. L. V. B. & Bini L. M. (2007). Spatial autocorrelation, model selection and hypothesis testing in geographical ecology: Implications for testing metabolic theory in New World amphibians. Neotropical Biology and Conservation, vol. 2, p. 119-126.

Chamberlain D. E., Fuller R. J., Bunce R. G. H., Duckworth J. C. & Shrubb M. (2000) Changes in the abundance of farmland birds in relation to the timing of agricultural intensification in England and Wales. Journal of Applied Ecology, vol. 37, p.771-788.

Champernowne D. G. (1948) Sampling theory applied to autoregressive sequences. Journal of the Royal Statistical Society. Series B (Methodological), vol. 10, n°2, p. 204-242.

Chatfield C. & Goodhardt G. J. (1970) The beta-binomial model for consumer purchasing behaviour. Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 19, n° 3, p. 240-250.

Chu C. K. & Marron J. S. (1991) Comparison of two bandwidth selectors with dependent errors. The Annals of Statistics, vol. 19, p. 1906-1918.

Chung C-J. F. & Fabbri A. G. (2003) Validation of spatial prediction models for landslide hazard mapping. Natural hazards, vol. 30, p. 451-472.

Clayton D. (1996) Generalized linear mixed models. In *Markov Chain Monte Carlo in Practice* (W. R. Gilks, S.Richardson and D. J. Spiegelhalter, eds.). Chapman and Hall, London, p. 275-301.

Cliff A. & Ord K. (1972) Testing for spatial autocorrelation among regression residuals. Geographical Analysis, vol. 4, n° 3, p. 267-284.

Clifford P., Richardson S. & Hemonet D. (1989) Assessing the significance of the correlation between two spatial processes. International Biometric Society, vol. 45, n° 1, p. 123-134.

Cochran W. G. (1963) *Sampling Techniques*. Second Edition. Wiley, New York.

Cochrane D. & Orcutt G. H. (1949) Application of least squares regression to relationships containing auto- correlated error terms. Journal of the American Statistical Association, vol. 44, n° 245, p. 32-61.

Cohn J. P. (2008) Citizen Science: Can volunteers do real research? American Institute of Biological Sciences, vol. 58, n° 3, p. 192-197.

Conlisk E., Conlisk J. & Harte J. (2007) The impossibility of estimating a negative binomial clustering parameter from presence-absence data: a comment on He and Gaston. The American Naturalist, vol. 170, n° 4, p. 651-654.

Conlisk E., Conlisk J., Enquist B., Thompson J. & Harte J. (2009) Improved abundance prediction from presence-absence data. Global Ecology and Biogeography, vol. 18, p. 1-10.

Conn P. B., Arthur A.D., Bailey L. L. & Singleton G. R. (2006) Estimating the abundance of mouse populations of known size: promises and pitfalls of new methods. Ecological Applications, vol. 16, p. 829-837.

Convention on Biological Diversity (2010) *Global Biodiversity Outlook 3.* Secretariat of the Convention on Biological Diversity, Montréal.

Cook R. D. & Jacobson J. O. (1979) A design for estimating visibility bias in aerial surveys. Biometrics, vol. 35, n° 4, p. 735-742.

Courchamp F., Berec L. & Gascoigne J. (2008) Allee effects in ecology and conservation. Oxford University Press, Oxford.

Courchamp F., Clutton-Brock T & Grenfell B. (1999) Inverse density dependence and the Allee effect. Trends in Ecology & Evolution, vol. 14, n° 10, p. 405-410.

Cox D. R. (1983) Some Remarks on Overdispersion. Biometrika Trust, vol. 70, n°1, p. 269-274.

Cox D. R. & Snell E. J. (1989) *Analysis of Binary Data.* Second Edition. Chapman and Hall New York

Craig M. H., Sharp B. L., Mabaso M. L. H. & Kleinschmidt I. (2007) Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. International Journal of Health Geographics, vol. 6: 44.

Cressie N. & Hawkins D. (1980a) Robust estimation of the variogram: I. Mathematical Geology, vol. 12, p. 115-125.

Cressie N. & Hawkins D. M. (1980b) Robust estimation of the variogram: II. Mathematical Geology, vol. 12, n° 2, p.115-125.

Cressie N. & Huang H. C. (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. Journal of the American Statistical Association, vol. 94, n°448, p. 1330-1340.

Cressie N. A. C. (1993) *Statistics for spatial data: Revised edition.* Wiley, New York.

Cressie N. A. C. & Wikle C. K. (2011) *Statistics for spatio-temporal data.* Wiley, New York.

Crowder M. J. (1978) Beta-Binomial Anova for Proportions. Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 27, n° 1, p. 34-37.

Davidian M. & Gallant A. R. (1992) Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. Journal of Pharmacokinetics and Biopharmaceutics, vol. 20, n° 5, p. 529-556.

Dempster A. P., Laird N. M. & Rubin D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), vol. 39, n° 1, p.1-38.

Devictor V., Whittaker R. J. & Beltrame C. (2010) Beyond scarcity: citizen science programmes as useful tools for conservation biogeography. Diversity and Distributions, vol. 16, p. 354–362.

Dickinson J. L., Zuckerberg B. & Bonter D. N. (2010) Citizen science as an ecological research tool: challenges and benefits. Annual Review of Ecology, Evolution, and Systematics, vol. 41, p. 149-172.

Diggle P. J. (2003) *Statistical analysis of spatial point patterns.* Second edition. Arnold, London.

Diggle P. J., Tawn J. A. & Moyeed R. A. (1998) Model-based geostatistics. Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 47, n° 3, p. 299-350.

Diniz-Filho J. A. F. & Bini L. M. (2005) Modelling geographical patterns in species richness using eigenvector-based spatial filters. Global Ecology and Biogeography, vol. 14, p. 177-185.

Diniz-Filho J. A. F., Bini L. M. & Hawkins B. A. (2003) Spatial autocorrelation and red herrings in geographical ecology. Global Ecology and Biogeography, vol. 12, p. 53-64.

Diniz-Filho J. A. F., Nabout J. C., Telles M. P. C, Soares T. N. & Rangel T. F. L. V. B. (2009) A review of techniques for spatial modeling in geographical, conservation and landscape genetics. Genetics and Molecular Biology, vol. 32, n° 2, p. 203-211.

Diniz-Filho J. A. F., Rangel T. F. L. V. B. & Bini L. M. (2008) Model selection and information theory in geographical ecology. Global Ecology and Biogeography, vol. 17, p. 479-488.

Dormann C. F. (2007) Effects of incorporating spatial autocorrelation into the analysis of species distribution data. Global Ecology and Biogeography, vol. 16, p. 129-138.

Dormann C. F., Elith J., Bacher S., Buchmann C., Carl G., Carré G., García Marquéz J. R., Gruber B., Lafourcade B., Leitão P. J., Münkemüller T., McClean C., Osborne P. E., Reineking B., Schröder B., Skidmore A. K., Zurell D. & Launtenbach S. (2012) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, vol. 35, p. 001-020.

Dormann, C. F., McPherson J. M., Araújo M. B., Bivand R., Bolliger J., Carl G., Davies R. G., Hirzel A., Jetz W., Kissling W. D., Kühn I., Ohlemüller R., Peres-Neto P. R., Reineking B., Schröder B., Schurr F. M. & Wilson R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography, vol. 30, p. 609-628.

Dray S., Legendre P. & Peres-Neto P. R. (2006) Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecological Modelling, vol. 196, p. 483-493.

Durbin J. & Watson G. S. (1950) Testing for serial correlation in least squares regression: I. Biometrika Trust, vol. 37, n° 3 & 4, p. 409-428

Durbin J. & Watson G. S. (1951) Testing for serial correlation in least squares regression: II. Biometrika Trust, vol. 38, n° 1 &2, p. 159-177.

Durbin J. & Watson G. S. (1971) Testing for serial correlation in least squares regression: III. Biometrika Trust, vol. 58, n° 1, p. 1-19.

Durbin J. (1960) The fitting of time-series models. Review of the International Statistical Institute, vol. 28, n° 3, p. 233-244.

Dutilleul P., Clifford P., Richardson S. & Hemonet D. (1993) Modifying the t test for assessing the correlation between two spatial processes. International Biometric Society, vol. 49, n°1, p. 305-314.

Eberhardt L. L. & Simmons M. A. (1987) Calibrating population indices by double sampling. Journal of Wildlife Management, vol. 51, n° 3, p. 665-675.

Efroymson M. A. (1960) Multiple Regression Analysis, In: *Mathematical methods for digital computers* (A. Ralston & H. S. Wilf, eds.). Wiley, New York.

Eisenhart C. (1947) The assumptions underlying the analysis of variance. International Biometric Society, vol. 3, n° 1, p. 1-21.

Elith J. & Leathwick J. R. (2009) Species distribution models: ecological explanation and prediction across space and time. Annual Review of Ecology, Evolution, and Systematics, vol. 40, p. 677-699.

Engeman R. M. (2003) More on the need to get the basics right: population indices. Wildlife Society Bulletin, vol.31, n°1, p.286-287.

Faddy M. J. & Bosch R. J. (2001) Likelihood-based modeling and analysis of data underdispersed relative to the Poisson distribution. Biometrics, vol. 57, p. 620-624.

Famoye F. (1993) Restricted generalized poisson regression model. Communications in Statistics - Theory and Methods, vol. 22, p. 1335-1354.

Fang Y. (2011) Asymptotic equivalence between cross-validations and Akaike information criteria in mixed-effects models. Journal of Data Science, vol. 9, p. 15-21.

Finney D. J. (1947) *Probit analysis: a statistical treatment of the sigmoid response curve.* Cambridge University Press, Cambridge.

Fisher R. A. (1941) The negative binomial distribution. Annals of Eugenics, vol. 11, n° 1, p. 182-187.

Forster M. R. (2000) Key concepts in model selection: performance and generalizability. Journal of Mathematical Psychology, vol. 44, p. 205-231.

Forster M. R. (2001) The new science of simplicity. In: *Simplicity, Inference and Modelling* (A. Zellner, H. A. Keuzenkamp & M. McAleer, eds.). Cambridge University Press, p. 83-119.

Fortin M.-J. & Dale M.R.T. (2005) *Spatial analysis: a guide for ecologists*. Cambridge University Press, Cambridge.

Franck I. E. & Friedman J. H. (1993) A statistical view of some chemometrics regression tools. Technometrics, vol. 35, n° 2, p. 109-135.

Franklin J. (2009) *Mapping Species Distributions: Spatial Inference and Prediction.* Cambridge University Press, Cambridge.

Freckleton R.P. (2002) On the misuse of residuals in ecology: regression of residuals vs. multiple regression. Journal of Animal Ecology, vol. 71, p. 542-545.

Fretwell S. D. & Lucas H. L. (1969) On territorial behavior and other factors influencing habitat distribution in birds. Acta Biotheoretica, vol. 19, n° 1, p. 16-36.

Fretwell S. D. (1972) *Populations in a Seasonal Environment*, Princeton University Press, Princeton.

Frome E. L., Kutner M. H. & Beauchamp J. J. (1973) Regression analysis of poisson-distributed data. Journal of the American Statistical Association, vol. 63, n° 244, p. 935-940.

Fu W. J. (1998) Penalized regressions: The Bridge versus the Lasso. Journal of Computational and Graphical Statistics, vol. 7, n° 3, p. 397-416.

Fuentes M. (2002) Spectral methods for nonstationary spatial processes. Biometrika, vol. 89, n° 1, p. 197-210.

Fuller R. J., Gregory R. D., Gibbons D. W., Marchant J. H., Wilson J. D., Baillie S. R. & Carter N. (1995) Population declines and range contractions among lowland farmland birds in Britain. Society for Conservation Biology, vol. 9, n° 6, p. 1425-1441.

Galton F. (1986) Regression towards mediocrity in hereditary stature. The Journal of the Anthropological Institute of Great Britain and Ireland, vol. 15, p. 246-263.

Gaston K. J. & Fuller R. A. (2008) Commonness, population depletion and conservation biology. Ecology and Evolution, vol. 23, n° 1, p.14-19.

Geary R. C. (1954) The contiguity ratio and statistical mapping. Royal Statistical Society, vol. 5, n° 3, p. 115-127+129-146.

Geisser S. (1975) The predictive sample reuse method with applications. American Statistical Association, vol. 70, n° 350, p. 320-328.

Gelman A., Carlin J. B., Stern H. S. and Rubin Donald B. (2004) *Bayesian Data Analysis*. Second edition. Chapman & Hall/CRC, Boca Raton.

George E. I. (2000) The variable selection problem. Journal of the American Statistical Association, vol. 95, n° 452, p. 1304-1308.

Getis A. (1990) Screening for spatial dependence in regression analysis. Regional Science Association, vol. 69, p. 69-81.

Getis A. & Aldstadt J. (2004) Constructing the spatial weights matrix using a local statistic. Geographical analysis, vol. 36, p. 90-104.

Getis A. & Griffith D.A. (2002) Comparative spatial filtering in regression analysis. Geographical Analysis, vol. 34, p. 130-140.

Gittleman J. L., Funk S. M., MacDonald D. W. & Wayne R. K. (2001) *Carnivore Conservation.* Cambridge University Press.

Gilks W. R., Richardson S. & Spiegelhalter D. J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Goldstein H. (1991) Nonlinear multilevel models, with an application to discrete response data. Biometrika, vol. 78, n° 1, p. 45-51.

Green P. J. (1987) Penalized likelihood for general semi-parametric regression model. International Statistical Review, vol. 55, n° 3, p.245-259.

Greene W. H. (1994) Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper No. EC-94-10, Department of Economics, Stern School of Business, NYU.

Greene W. (2008) Functional forms for the negative binomial model for count data. Economics Letters, vol. 99, p. 585-590.

Greenwood J. J. D., Baillie S. R., Gregory R. D., Peach W. J. & Fuller R. J. (2008) Some new approaches to conservation monitoring of British breeding birds. British Trust for Ornithology, vol. 137, p. 16-28.

Gregory R. D., Noble D. G. & Custance J. (2004) The state of play of farmland birds: population trends and conservation status of lowland farmland. British Ornithologists' Union, vol. 146, p. 1-13.

Griffith, D. A. (1987) Spatial autocorrelation: a primer. Association of American Geographers.

Griffith D. A. (2000) A linear regression solution to the spatial autocorrelation problem. Journal of Geographical Systems, vol. 2, p. 141-156.

Griffith D. A. (2002) A spatial filtering specification for the auto-Poisson model. Statistics & Probability Letters, vol. 58, p. 245-251.

Griffith D. A. (2003) Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer-Verlag, Berlin Heidelberg.

Griffith D. A. (2006a) Assessing spatial dependence in count data: winsorized and spatial filter specification alternatives to the auto-Poisson model. Geographical Analysis, vol. 38, p. 160-179.

Griffith D. A. (2006b) Hidden negative spatial autocorrelation. Journal of Geographical System, vol. 8, p. 335-355.

Griffith D.A. (2010). Spatial Filtering. In: *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications* (Fischer M. M. & Getis A., eds.). Springer-Verlag Berlin, Heidelberg, p. 301-318.

Griffith D. A. & Haining R. (2006) Beyond Mule Kicks: The Poisson Distribution in Geographical Analysis. Geographical Analysis, vol. 38, p. 123-139.

Griffith D. A. & Peres-Neto P. R. (2006) Spatial modeling in ecology: the flexibility of eigenfunction spatial analysis. Ecology, vol. 87, n° 10, p. 2603-2613.

Gschlößl S. & Czado C. (2008) Modelling count data with overdispersion and spatial effects. Statistical Papers, vol. 49, p.531-552

Guillera-Arroita G., Ridout M., Morgan B. J. T., Linkie M. (2012) Models for species-detection data collected along transects in the presence of abundance-induced heterogeneity and clustering in the detection process. Methods in Ecology and Evolution, vol. 3, p. 358-367.

Guisan A. & Thuiller W. (2005) Predicting species distribution: offering more than simple habitat models. Ecology Letters, vol. 8, p. 993-1009.

Guisan A. & Zimmermann N. E. (2000) Predictive habitat distribution models in ecology. Ecological Modelling, vol. 135, p. 147–186.

Gurevitch J. & Padilla D. K. (2004) Are invasive species a major cause of extinctions? Ecology and Evolution, vol. 19, n° 9, p. 470-474 & p. 619-620.

Haas T. C. (1990) Lognormal and moving window methods of estimating acid deposition. Journal of the American Statistical Association, vol. 85, n° 412, p. 950-963.

Haas T. C. (1995) Local prediction of a spatio-temporal process with an application to wet sulfate deposition. Journal of the American Statistical Association, vol. 90, n° 432, p. 1189-1199.

Haight F. A. (1967) *Handbook of the Poisson Distribution*. John Wiley & Sons, New York.

Haining R. (1990) *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge University Press, Cambridge.

Haining R., Law J. & Griffith D. (2009) Modelling small area counts in the presence of overdispersion and spatial autocorrelation. Computational Statistics and Data Analysis, vol. 53, n° 8, p. 2923-2937.

Halada L., Jongman R. G. H., Gerard F., Whittaker L., Bunce R. G. H., Bauch B. & Schmeller D. S. (2009) The European Biodiversity Observation Network - EBONE. In: *European conference of the Czech Presidency of the Council of the EU TOWARDS eENVIRONMENT (Challenges of SEIS and SISE: Integrating Environmental Knowledge in Europe)* (J. Hřebíček, J. Hradec, E. Pelikán, O. Mírovský, W. Pilmmann, I. Holoubek, R. Legat, eds.*).* Masaryk University, Brno, p. 177-188.

Hansen M. and Yu B. (1999) Bridging AIC and BIC: An MDL model selection criterion. In: *Proceedings of IEEE Information Theory Workshop on Detection, Estimation, Classication and Imaging.* Santa Fe, NM.

Hansen M. H., & Yu B. (2001) Model Selection and the Principle of Minimum Description Length. Journal of the American Statistical Association, vol. 96, n° 454, p. 746 -774.

Hanski I. & Gilpin M. (1991) Metapopulation dynamics: brief history and conceptual domain. Biological Journal of the Linnean Society, vol. 42, p. 3-16.

Hanski I., Moilanen A., & Gyllenberg M. (1996) Minimum viable metapopulation size. The American Naturalist, vol. 147, n° 4, p. 527-541.

Harville D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association, vol. 72, n° 358, p. 320-338.

Hastie T. & Tibshirani R. (1986) Generalized additive models. Statistical Science, vol. 1, n° 3, p. 297-318.

Hastie T. & Tibshirani R. (1993) Varying-coefficient models. Journal of the Royal Statistical Society. Series B (Methodological), vol. 55, n° 4, p. 757-796.

Hawkins B. A. (2012) Eight (and a half) deadly sins of spatial Analysis. Journal of Biogeography, vol. 39, p. 1-9.

Hawkins B. A., Diniz-Filho J. A. F., Bini L. M., De Marco P. & Blackburn T. M. (2007) Red herrings revisited: spatial autocorrelation and parameter estimation in geographical ecology. Ecography, vol. 30, n° 3, p. 375-384.

He F. & Gaston K. J. (2000) Estimating species abundance from occurrence. The American Naturalist, vol. 156, n° 5, p. 552-559.

He F. & Gaston K. J. (2007) Estimating abundance from occurrence: An underdetermined problem. The American Naturalist, vol. 170, n° 4, p. 654-659.

Heilbron D. C. (1994) Zero-altered and other regression models for count data with added zeros. Biometrical Journal, vol. 36, n° 5, p. 531–547.

Hijmans R. J., Cameron S. E., Parra J. L., Jones P. G. & Jarvis A. (2005) Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, vol. 25, p. 1965-1978.

Hinde J. (1982) Compound Poisson regression models. Lecture Notes in Statistics, vol. 14, p. 109-121.

Hinde J. & Demétrio C. G. B. (1998) Overdispersion: Models and estimation. Computational Statistics & Data Analysis, vol. 27, p. 151-170.

Hocking R. R. (1976) The analysis and selection of variables in linear regression. Biometrics, vol. 32, p. 1-49.

Hodges J. S. & Reich B. J. (2010) Adding spatially-correlated errors can mess up the fixed effect you love. The American Statistician, vol. 64, p. 325-334.

Hoerl A. E. & Kennard R. W. (1970a) Ridge regression: biased estimation for nonorthogonal problems. Technometrics, vol. 12, n° 1, p. 55-67.

Hoerl A. E. & Kennard R. W. (1970b) Ridge regression: Applications to nonorthogonal problems. Technometrics, Vol. 12, n° 1, p. 69-82.

Hoeting J. A., Davis R. A., Merton A. A. & Thompson S. E. (2006) Model selection for geostatistical models. Ecological Applications, vol. 16, p. 87-98.

Hoeting J. A., Madigan D., Raftery A. E. & Volinsky C. T. (1999) Bayesian model averaging: A tutorial. Statistical Science, vol. 14, n° 4, p.382-417.

Holt A. R., Gaston K. J., He F. (2002) Occupancy-abundance relationships and spatial distribution: A review. Basic and Applied Ecology, vol. 3, p. 1-13.

Hooker R. H. (1905) On the correlation of successive observations illustrated by corn prices. Journal of the Royal Statistical Society, vol.68, n°4, p.694-703

Hothorn T., Müller J., Schröder B., Kneib T. & Brandl R. (2011) Decomposing environmental, spatial, and spatiotemporal components of species distributions. Ecological Monographs, vol. 81, n° 2, p. 329-347.

Huete A., Didan K., Miura T., Rodriguez E. P., Gao X. & Ferreira L. G. (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment, vol. 83, p. 195-213.

Hughes J. (in press) ngspatial: An R package for spatial data. Journal of Statistical Software.

Hughes J. & Haran M. (2013) Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 75, p. 139-159.

Hurlbert S. H. (1984) Pseudoreplication and the design of ecological field experiments. Ecological Monographs, vol. 54, n° 2, p. 187-211.

Hutchinson G. E. (1957) Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology, vol. 22, p. 415-427.

Ishii G. & Hayakawa R. (1960) On the compound binomial distribution. Annals of the Institute of Statistical Mathematics, vol. 12, n° 1, p. 69-80.

Ismail N. & Jemain A. A. (2007) Handling overdiseprsion with negative binomial and generalized poisson regression models. In: *Casualty Actuarial Society Forum*. Casualty Actuarial Society, Arlington, Virginia, p. 103-158.

Jhala Y., Qureshi O. & Gopal R. (2011) Can the abundance of tigers be assessed from their signs ? Journal of Applied Ecology, vol. 48, p. 14-24.

Jiguet F. (2009) Method learning caused a first-time observer effect in a newly started breeding bird survey, Bird Study, vol. 56, n° 2, p. 253-258.

Jiguet F., Devictor V., Julliard R. & Couvet D. (2012) French citizens monitoring ordinary birds provide tools for conservation and ecological sciences. Acta Oecologica, vol. 44, p. 58-66.

Jiménez-Valverde A., Lobo J. M. & Hortal J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions, vol. 14, n° 6, p. 885-890.

Jochmann M. (2013) What belongs where? Variable selection for zero-inflated count models with an application to the demand for health care. Computational Statistics, vol. 28, n° 5, p. 1947-1964.

Johnson D. H. (1980) The comparison of usage and availability measurements for evaluating resource preference. Ecology, vol. 61, p. 65-71.

Johnson J. B. & Omland K. S. (2004) Model selection in ecology and evolution. Trends in Ecology & Evolution, vol. 19, p. 101-108.

Jolly G. M. (1965) Explicit estimates from capture-recapture data with both death and immigration-stochastic model. Biometrika, vol. 52, n° 1/2, p. 225-247.

Jones J. P. G. (2011) Monitoring species abundance and distribution at the landscape scale. Journal of Applied Ecology, vol. 48, p. 9-13.

Jones J. P. G., Collen B., Atkinson G., Baxter P. W. J., Bubb P., Illian J. B., Katzner T. E., Keane A., Loh J., McDonald-Madden E., Nicholson E., Pereira H. M., Possingham H. P., Pullin A. S., Rodrigues A. S. L., Ruiz-Guttierez V., Sommerville M. & Milner-Gulland E. J. (2011), The why, what, and how of global biodiversity indicators beyond the 2010 target. Conservation Biology, vol. 25, n° 3, p. 450-457.

Joseph L. N., Field S. A., Wilcox C. & Possingham H. P. (2006) Presence–Absence versus abundance data for monitoring threatened species, Conservation Biology, vol. 20, n° 6, p. 1679-1687.

Kearney M. & Porter W. (2009) Mechanistic niche modelling: Combining physiological and spatial data to predict species' ranges. Ecology Letters, vol. 12, p. 334-350.

Keitt T. H., Ottar N., Bjørnstad O. N., Dixon P. M. & Citron-Pousty S. (2002) Accounting for spatial pattern when modeling organism-environment interactions. Ecography, vol. 25, p. 616-625.

Kendall W. L. (1999) Robustness of closed capture–recapture methods to violations of the closure assumption. Ecology, vol. 80, n° 8, p. 2517-2525.

Kerr J. T. & Ostrovsky M. (2003) From space to species: ecological applications for remote sensing. Ecology and Evolution, vol. 18, n° 6, p. 299-305.

Kish L. (1965) *Survey sampling.* Wiley, New York.

Kissling W. D. & Carl G. (2008) Spatial autocorrelation and the selection of simultaneous autoregressive models. Global Ecology and Biogeography, vol. 17, p. 59-71.

Knafl G. J. & Grey M. (2007) Factor analysis model evaluation through likelihood cross-validation. Statistical Methods for Medical Research, vol. 16, p. 77-102.

Koenig W. D. (1999) Spatial autocorrelation of ecological phenomena. Trends in Ecology & Evolution, vol. 14, n° 1, p. 22-26.

Kohavi, R. (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (C. S. Mellish, ed.). Morgan Kaufman, San Mateo, p. 1137-1143.

Krainski E. T. & Lindgren F. (unpublished) The R-INLA tutorial: SPDE models. Tutorial, October 8, 2013.

Kramer M. G., Hansen A. J., Taper M. L. & Kissinger E. J. (2001) Abiotic controls on long-term windthrow disturbance and temperate rain forest dynamics in southeast Alaska. Ecology, vol. 82, p. 2749-2768.

Krebs C. J. (1999) *Ecological methodology.* Second edition. Addison Wesley Longman, New York.

Kremen C. (1992) Assessing the indicator properties of species assemblages for natural areas monitoring. Ecological Applications, vol. 2, n° 2, p. 203-217.

Kühn I. (2007) Incorporating spatial autocorrelation may invert observed patterns. Diversity and Distributions, vol. 13, p. 66-69.

Kühn I., Nobis M. P. & Durka W. (2009) Combining spatial and phylogenetic eigenvector filtering in trait analysis. Global Ecology and Biogeography, vol. 18, p. 745-758.

Kullback S. & Leibler R. A. (1951) On information and sufficiency. The Annals of Mathematical Statistics, vol. 22, n° 1, p. 79-86.

Laird N. M. & Ware J. H. (1982) Random-effects models for longitudinal data. Biometrics, vol. 38, n° 4, p. 963-974.

Lambert D. (1992) Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics, vol. 34, n° 1, p. 1-14.

Lancia R. A., Nichols J. D & Pollock K. H. (1994) Estimating the number of animals in wildlife populations. In *Research and management techniques for wildlife and habitats* (T. A. Bookhout, ed.). Fifth edition. The Wildlife Society, Bethesda, Md, p. 215-253

Lawless J. F. (1987) Negative binomial and mixed poisson regression. The Canadian Journal of Statistics, vol. 15, n°3, p. 209-225.

La Sorte F. A. & Thompson F. R. (2007) Poleward shifts in winter ranges of North American birds. Ecology, vol. 88, n° 7, p. 1803-1812.

LaFrance D., Lands L. C. & Burns D. H. (2003) Measurement of lactate in whole human blood with near-infrared transmission spectroscopy. Talanta, vol. 60, p. 635-/641.

Le Rest K., Pinaud D. & Bretagnolle V. (2013) Accounting for spatial autocorrelation from model selection to statistical inference: Application to a national survey of a diurnal raptor, Ecological Informatics, vol. 14, p. 17-24.

Le Rest K ., Pinaud D., Monestiez P., Chadoeuf J. & Bretagnolle V. (2014) The spatial-leave-one-out (SLOO) cross-validation for variable selection in the presence of spatial autocorrelation. Global Ecology & Biogeography, in press.

Lebreton J. D., Burnham K. P., Colbert J. & Anderson D. R. (1992) Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecological Monographs, vol. 62, n° 1, p. 67-118.

Legendre P. (1993) Spatial autocorrelation: Trouble or new paradigm? Ecology, vol. 74, n° 6, p. 1659-1673.

Legendre P. & Fortin M. J. (1989) Spatial pattern and ecological analysis. Vegetatio, vol. 80, p. 107-138.

Legendre P. & Legendre L. (1998) *Numerical ecology*. Elsevier, Amsterdam.

Legendre P., Dale M. R. T., Fortin M. J., Gurevitch J., Hohn M. & Myers D. (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography, vol. 25, p. 601-615.

Legg C. J. & Naggy L. (2006) Why most conservation monitoring is, but need not be, a waste of time. Journal of Environmental Management, vol. 78, p. 194-199.

Lele S. (1991) Jackknifing linear estimating equations : Asymptotic theory and applications in stochastic processes. Journal of the Royal Statistical Society Series B (Methodological), vol. 53, n° 1, 253-267.

Lennon J. J. (2000) Red-shifts and red herrings in geographical ecology. Ecography, vol. 23, p. 101-113.

Leslie P. H. & Chitty D. (1951) The estimation of population parameters from data obtained by means of the capture-recapture method: I. The maximum likelihood equations for estimating the death-rate. Biometrika, vol. 38, n° 3/4 p. 269-292.

Leslie P. H. (1952) The estimation of population parameters from data obtained by means of the capture-recapture method: II. The estimation of total numbers. Biometrika, vol. 39, n° 3/4 p. 363-388.

Leslie P. H., Chitty D. & Chitty H. (1953) The estimation of population parameters from data obtained by means of the capture-recapture method: III. An example of the practical applications of the method. Biometrika, vol. 40, n° 1/2 p. 137-169.

Levins R. (1969) Some demographic and genetic consequences of environmental heterogeneity for biological control. Bulletin of the Entomological Society of America, vol. 15, n° 3, p. 237-240.

Levrel H., Fontaine B.,Henry P. Y., Jiguet F., Julliard R., Kerbiriou C. & Couvet D. (2010) Balancing state and volunteer investment in biodiversity monitoring for the implementation of CBD indicators: A French example. Ecological Economics, vol. 69, n° 7, p.1580-1586.

Liang K-Y. & Zeger S. L. (1986) Longitudinal data analysis using generalized linear models. Biometrika, vol. 73, n° 1, p. 13-22.

Lichstein J. W., Simons T. R., Shriner S. A.,& Franzreb K. E. (2002) Spatial autocorrelation and autoregressive models in ecology. Ecological Monographs, vol. 72, n° 3, p. 445-463.

Liebhold A. M. & Sharov A. A. (1998) Testing for correlation in the presence of spatial autocorrelation in insect count data. In: *Population and Community Ecology for Insect Management and Conservation* (J. Baumgartner, P. Brandmayr & B.F.J. Manly, eds). Balkema, Rotterdam, p. 11-117.

Lin X. & Zhang D. (1999) Inference in generalized additive mixed models by using smoothing splines. Journal of the Royal Statistical Society Series B (Methodological), vol. 61, n° 2, p. 381-400.

Linder J. M. & Lawler R. R. (2012) Model selection, zero-inflated models, and predictors of primate abundance in Korup national park, Cameroon American Journal of Physical Anthropology , vol. 149, p.417-425.

Lindgren F. (2013) Continuous domain: Spatial models in R-INLA. The ISBA Bulletin, vol.19, n° 4, updated version.

Lindgren F. & Rue H. (unpublished) Bayesian Spatial and Spatio-temporal Modelling with R-INLA. Tutorial.

Lindgren F., Rue H. & Lindström J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society Series B (Statistical Methodology), vol. 73, n° 4, p. 423-498.

Mace G. M., Cramer W., Díaz S., Faith D. P., Larigauderie A., Le Prestre P., Palmer M., Perrings C., Scholes R. J., Walpole M., Walther B. A., Watson J. E. M. &, Mooney H. A. (2011) Biodiversity targets after 2010. Current Opinion in Environmental Sustainability, vol. 2, p. 3-8.

MacKenzie D. I., Nichols J. D., Lachman G. B., Droege S., Royle J. A.& Langtimm C. A. (2002) Estimating site occupancy rates when detection probabilities are less than one. Ecology, vol. 83, n° 8, p. 2248-2255.

Magurran A. E., Baillie S. R., Buckland S. T., Dick J. McP., Elston D. A., Scott E. M., Smith R. I., Somerfield P. J. & Watt A. D. (2010) Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. Trends in Ecology & Evolution, vol. 25, n° 10, 574-582.

Mallow C. L. (1973) Some comments on Cp, Technometrics, vol.15, n°4, p. 661-675.

Martin T. G., Wintle B. A., Rhodes J. R., Kuhnert P. M., Field S. A., Low-Choy S. J., Tyre A. J. & Possingham H. P. (2005) Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. Ecology Letters, vol. 8, p. 1235-1246.

McGill B. J., Etienne R. S., Gray J. S., Alonso D., Anderson M. J., Kassa Benecha H., Dornelas M., Enquist B. J., Green J. L., He F., Hurlbert A. H., Magurran A. E, Marquet P. A., Maurer B. A., Ostling A., Soykan C. U., Ugland K. I. & White E. P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecology Letters, vol. 10, p. 995-1015.

McQuarrie A. D. R. & Tsai C.-L. (1998) *Regression and time series model selection.* World Scientific, Singapore.

Mellin C., Bradshaw C. J. A., Meekan M. G. & Caley M. J. (2010) Environmental and spatial predictors of species richness and abundance in coral reef fishes. Global Ecology and Biogeography, vol. 19, p. 212-222.

Miaou S.-P. (1994) The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention, vol. 26, n° 4. p. 471-482.

Miller A. J. (1984) Selection of subsets of regression. Journal of the Royal Statistical Society, Series A (General), vol. 147, n° 3, p. 389-425.

Miller J., Franklin J. & Aspinall R. (2007) Incorporating spatial dependence in predictive vegetation models. Ecological modelling, vol. 202, p. 225-242.

Miller J. A. (2012) Species distribution models: Spatial autocorrelation and non-stationarity. Progress in Physical Geography, vol. 36, n° 5, p. 681-692.

Miller R. G. (1974) The jackknife—A review. Biometrika, vol. 61, n° 1, p.1-15.

Millstein R. L. (2010) The concepts of population and metapopulation in evolutionary biology and ecology. In: Evolution since Darwin: the first 150 years (M. A. Bell, D. J. Futuyma, W. F. Eanes & J. S. Levinton, eds.). Sinauer.

Mooney H. A. & Cleland E. E. (2001) The evolutionary impact of invasive species. Proceedings of the National Academy of Sciences, vol. 98, n° 10, p. 5446-5451.

Moran P. A. P. (1950) A test for the serial independence of residuals. Biometrika, vol. 37, n° 1, p. 178-181.

Mullahy J. (1986) Specification and testing of some modified count data models. Journal of Econometrics, vol. 33, p. 341-365.

Mullahy J. (1997) Heterogeneity, excess zeros, and the structure of count data models. Journal of Applied Econometrics, vol. 12, n° 3, 337-350.

Myers D. E. (1989) To be or not to be... stationary? That is the question. Mathematical Geology, vol. 21, n° 3, p. 347-362.

Nelder J. A. & Wedderburn R. W. M. (1972) Generalized linear models. Journal of the Royal Statistical Society Series A (General), vol. 135, n° 3, p. 370-384.

Nerlove M. & Press S. J. (1973) *Univariate and multivariate log-linear and logistic models.* Rand Corporation, Santa Monica.

New L. F., Buckland S. T., Redpath S. & Matthioppoulos J. (2011) Hen harrier management: insights from demographic models fitted to population data. Journal of Applied Ecology, vol. 48, p. 1187-1194.

Newton I. (1979). *Population ecology of raptors.* Poyser. Berkhamstead.

Neyens T., Faes C. & Molenberghs G. (2012) A generalized Poisson-gamma model for spatially overdispersed data. Spatial and Spatio-temporal Epidemiology, vol. 3, vol. 185-194.

Nichols J. D., Hines J. E., Sauer J. R., Fallon F. W., Fallon J. E. & Heglund P. J. (2000) A double-observer approach for estimating detection probability and abundance from point counts. The Auk, vol. 117, n° 2, p. 393-408.

Nielsen S. E., Johnson C. J., Heard D. C. & Boyce M. S. (2005) Can models of presence absence be used to scale abundance? Two case studies considering extremes in life history. Ecography, vol. 28, p. 197-208.

Nishii R. (1984) Asymptotic properties of criteria for selection of variables in multiple regression. The Annals of Statistics, vol. 12, n° 2, p. 758-765.

O'Hara R. B. & Kotze D. J. (2010) Do not log-transform count data. Methods in Ecology and Evolution, vol. 1, n° 2, p. 118-122.

Osborne P. E., Foody Giles M. & Suárez-Seoane S. (2007) Non-stationarity and local approaches to modelling the distributions of wildlife. Diversity and Distributions, vol.13, p. 313-323.

Paciorek C. J. (2010) The importance of scale for spatial confounding bias and precision of spatial regression estimators. Statistical Science, vol. 25, p. 107-125.

Patterson H. D. & Thompson R. (1971) Recovery of inter-block information when block sizes are unequal. Biometrika, vol. 58, n° 3, p. 545-554.

Patuelli R., Griffith D. A., Tiefelsdorf M. & Nijkamp P. (2006) The use of spatial filtering techniques : the spatial and space-time structure of German unemployment data. Tinbergen Institute Discussion Papers 06-049/3.

Pearce J. & Ferrier S. (2001) The practical value of modelling relative abundance of species for regional conservation planning: A case study. Biological Conservation, vol. 98, p. 33-43.

Pearson R. G. & Dawson T. P. (2003) Predicting the impacts of climate change on the distribution of species: Are bioclimate envelope models useful? Global Ecology and Biogeography, vol. 12, p. 361-371.

Pereira H. M. & Cooper H. D. (2006) Towards the global monitoring of biodiversity change. Trends in Ecology and Evolution, vol. 21, n° 3, p.123-129.

Pereira H. M., Belnap J., Brummitt N., Collen B., Ding H., Gonzalez-Espinosa M., Gregory R. D., Honrado J., Jongman R. H. G., Julliard R., McRae L., Proença V., Rodrigues P., Opige M., Rodriguez J. P., Schmeller D. S, Van Swaay C. & Vieira C. (2010) Global biodiversity monitoring. Frontiers in Ecology and the Environment, vol. 8, p. 459-460.

Perrings C., Naeem S., Ahrestani F., Bunker D. E., Burkill P., Canziani G., Elmqvist T., Ferrati R., Fuhrman J., Jaksic F., Kawabata Z., Kinzig A., Mace G. M., Milano F., Mooney H., Prieur-Richard A.-H., Tschirhart J. & Weisser W. (2010) Ecosystem Services for 2020. Science, vol. 330, p. 323-324.

Perrings C., Naeem S., Ahrestani F., Bunker D. E., Burkill P., Canziani G., Elmqvist T., Ferrati R., Fuhrman J., Jaksic F., Kawabata Z., Kinzig A., Mace G. M., Milano F., Mooney H., Prieur-Richard A.-H., Tschirhart J. & Weisser W. (2011) Ecosystem services, targets, and indicators for the conservation and sustainable use of biodiversity. Frontiers In Ecology And The Environment, vol. 9, n° 9, p. 512-520.

Pierce D. A. & Sands B. R. (1975) *Extra-Bernoulli Variation in Binary Data*, Technical Report 46, Department of Statistics, Oregon State University.

Pinheiro J. C. & Bates D. M. (1995) Approximations to the log-likelihood function in the nonlinear mixed-effects model. Journal of Computational and Graphical Statistics, vol. 4, n° 1, p. 12-35.

Pinheiro J. C. & Chao E. C. (2006) Efficient Laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed. Journal of Computational and Graphical Statistics, vol. 15, n° 1, p. 58-81.

Pinkerton M. H., Smith A. N. H., Raymond B., Hosie G. W., Sharp B., Leathwick J. R. & Bradford-Grieve J. M. (2010) Spatial and seasonal distribution of adult *Oithona similis* in the southern ocean: predictions using boosted regression trees. Deep-Sea Research I, vol. 57, p. 469-485.

Pledger S., Pollock K. H. & Norris J. L. (2010) Open capture–recapture models with heterogeneity: II. Jolly–Seber model. Biometrics, vol. 66, p. 883-890.

Power M. E. (1992) Top-down and bottom-up forces in food webs: Do plants have primacy. *Ecology*, vol. 73, n° 3, p. 733-746.

Preisser J. S., Stamm J. W., Long D. L. & Kincade M. E. (2012) Review and recommendations for zero-inflated count regression modeling of dental caries indices in epidemiological studies. Caries Research, vol. 46, n° 4, p. 413-423.

Pulliam H. R. (1988) Sources, sinks, and population regulation. The American Naturalist, vol. 132, n° 5, p. 652-661.

Quenouille M. H. (1949) Approximate tests of correlation in time-series. Journal of the Royal Statistical Society Series B (Methodological), vol. 11, n° 1, p. 68-84.

Quenouille M. H. (1956) Notes on Bias in Estimation. Biometrika, vol. 43, n° 3/4, p. 353-360.

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. URL: http://www.R-project.org/.

Rands M. R. W., Adams W. M., Bennun L., Butchart S. H. M., Clements A., Coomes D., Entwistle A., Hodge I., Kapos V., Scharlemann J. P. W., Sutherland W. J. & Vira B. (2010) Biodiversity conservation: Challenges beyond 2010. Science, vol. 329, p.1298-1303.

Rao C. R. & Wu Y. (1989) A strongly consistent procedure for model selection in a regression problem. Biometrika, vol. 76, n° 2, p. 369-374.

Rathbun S. L. & Fei S. (2006) A spatial zero-inflated Poisson regression model for oak regeneration. Environmental and Ecological Statistics, vol. 13, n°4, p. 409-426.

Raudenbush S. W., Yang M.-L. & Yose M. (2000) Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. Journal of Computational and Graphical Statistics, vol. 9, n° 1, p. 141-157.

Reich B. J., Hodges J. S. & Zadnik V. (2006) Effect of residual smoothing on the posterior of the fixed effects in disease-mapping models. Biometrics, vol. 62, p. 1197-1206.

Richards S.A. (2008) Dealing with overdispersed count data in applied ecology. Journal of Applied Ecology, vol. 45, p. 218-227.

Richards S. A., Whittingham M. J. & Stephens P. A. (2011) Model selection and model averaging in behavioural ecology: the utility of the IT-AIC framework. Behavioral Ecology and Sociobiology, vol. 65, p. 77-89.

Ridout M. & Besbeas P. (2004) An empirical model for underdispersed count data. Statistical Modelling, vol. 4, p.77-89.

Ridout M., Demétrio C. G. B. & Hinde J. (1998) Model for count data with many zeros. In: *Proceedings of the 19th International Biometrics Conference*. Cape Town, p. 179-192.

Ridout M., Hinde J. &. Demétrio C. G. B. (2001) A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics, vol. 57, n° 1, p. 219-223.

Rissanen J. (1978) Modeling by shortest data description. Automatica, vol. 14, p. 465-471.

Robinson R. A., Green R. E., Baillie S. R., Peach W. J. & Thomson D. L. (2004) Demographic mechanisms of the population decline of the song thrush *Turdus philomelos* in Britain. Journal of Animal Ecology, vol. 73, p. 670-682.

Royle J. A. & Nichols J. D. (2003) Estimating abundance from repeated presence absence data or points counts. Ecology, vol. 84, n° 3, p. 777-790.

Royle J. A. (2004) N-mixture models for estimating population size from spatially replicated counts. Biometrics, vol. 60, n° 1, p. 108-115.

Rue H., Martino S. & Chopin N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of Royal Statistical Society Series B (Methodological), vol. 71, n° 2, p. 319-392.

Russ G. & Brenning A. (2010) Data mining in precision agriculture: Management of spatial information. In: *Computational Intelligence for Knowledge-Based Systems Design* (E. Hüllermeier, R. Kruse & F. Hoffmann, eds.). Springer-Verlag, Berlin-Heidelberg, p. 350-359.

Saas Y. & Gosselin F. (2014) Comparison of regression methods for spatially-autocorrelated count data on regularly- and irregularly-spaced locations. Ecography, in press.

Sakar C. O. & Kursun O. (2010) Telediagnosis of parkinson's disease using measurements of dysphonia. Journal of Medical Systems, vol. 34, n° 4, p. 591-599.

Sampson P. D. & Guttorp P. (1992) Nonparametric Estimation of Nonstationary Spatial Covariance Structure. Journal of the American Statistical Association, vol. 87, n° 417, p. 108-119.

Schall R. (1991) Estimation in generalized linear models with random effects. Biometrika, vol. 78, n° 4, p. 719-727.

Scheiner S. M., Cox S. B., Willig M. R., Mittelbach G. G., Osenberg C. & Kaspari, M. (2000) Species richness, species-area curves and Simpson's paradox. Evolutionary Ecology Research, vol. 2, p. 791-802.

Schmidt B. R. (2003) Count data, detection probabilities, and the demography, dynamics, distribution, and decline of amphibians. Comptes Rendus Biologies, vol. 326, p. S119–S124.

Schmitz O. J. (2003) Top predator control of plant biodiversity and productivity in an old-field ecosystem. Ecology Letters, vol. 6, p. 156-163.

Schwarz C. J. & Seber G. A. F. (1999) Estimating animal abundance: review III. Statistical Science, vol.14, p. 427-456.

Schwarz G. (1978) Estimating the dimension of a model. The Annals of Statistics, vol. 6, n° 2, p. 461-464.

Seal H. L. (1967**)** Studies in the history of probability and statistics. XV: The historical development of the Gauss linear model. Biometrika, vol. 54, n° 1, p. 1-24.

Seber G. A. F. (1965) A note on the multiple recapture census. Biometrika, vol. 52, p. 249-259.

Seber G. A. F. (1982) *The Estimation of Animal Abundance and Related Parameters*. Second edition. Macmillan, New York.

Seber G. A. F. (1986) A review of estimating animal abundance. Biometrics, vol. 42, n° 2, p. 267-292.

Seber G. A. F. (1992) A review of estimating animal abundance II. International Statistical Review, vol. 60, n° 2, p. 129-166.

Segurado P. & Araújo M. B. (2004) An evaluation of methods for modelling species distributions. Journal of Biogeography, vol. 31, p. 1555-1568.

Sekercioоglu C. H., Primack R. B. & Wormworth J. (2012) The effects of climate change on tropical birds. Biological Conservation, vol. 148, p. 1-18.

Seoane J., Vinuela J., Diaz-Delgado R. & Bustamante J. (2003) The effects of land use and climate on red kite distribution in the Iberian peninsula. Biological Conservation, vol. 111, p. 401-414.

Sergio F., Marchesi L. & Pedrini P. (2004) Integrating individual habitat choices and regional distribution of a biodiversity indicator and top predator. Journal of Biogeography, vol. 31, p. 619-628.

Sergio F., Newton I. & Marchesi L. (2005) Top predators and biodiversity. Nature, vol. 436, p. 192.

Sergio F., Newton I., Marchesi L. & Pedrini P. (2006) Ecologically justified charisma: preservation of top predators delivers biodiversity conservation. *Journal of Applied Ecology*, vol. 43, p. 1049-1055.

Sergio F., Caro T., Brown D., Clucas B., Hunter J., Ketchum J., McHugh K. & Hiraldo F. (2008a) Top predators as conservation tools: ecological rationale, assumptions, and efficacy. Annual Review of Ecology, Evolution, and Systematics, vol. 39, p. 1-19.

Sergio F., Newton I. & Marchesi L. (2008b) Top predators and biodiversity: much debate, few data Journal of Applied Ecology, vol. 45, p. 992-999.

Sergio F., Blas J. & Hiraldo F. (2009) Predictors of floater status in a long-lived bird: A cross-sectional and longitudinal test of hypotheses. Journal of Animal Ecology, vol. 78, p. 109-118.

Sessions D. N. & Stevans N. K. (2006) Investigating omitted variable bias in regression parameter estimation: A genetic algorithm approach. Computational Statistics and Data Analysis, vol. 50, p. 2835-2854.

Shaffer M. L. (1981) Minimum population sizes for species conservation. BioScience, vol. 31, n° 2, p. 131-134.

Shao J. (1993) Linear model selection by cross-validation. Journal of the American Statistical Association, vol. 88, n° 422, p. 486-494.

Shao J. (1997) An asymptotic theory for linear model selection. Statistica Sinica, vol. 7, p. 221-264.

Shibata R. (1981**)** An optimal selection of regression variables. Biometrika, vol. 68, n° 1, p.45-54.

Shibata R. (1984**)** Approximate efficiency of a selection procedure for the number of regression variables. Biometrika, vol. 71, n° 1, p. 43-49.

Sileshi G., Hailu G. & Nyadzi G. I. (2009) Traditional occupancy–abundance models are inadequate for zero-inflated ecological count data. Ecological Modelling, vol. 220, p. 1764-1775.

Simberloff D. (1986) The proximate causes of extinction. In: *Patterns and processes in the history of life* (D. M. Raup & D. Jablonski, eds.). Springer-Verlag Berlin, Heidelberg, p. 259-276.

Sinclair S. J., White M. D. & Newell G. R. (2010) How useful are species distribution models for managing biodiversity under future climates? Ecology and Society, vol. 15, article 8.

Sokal R. R. & Oden N. L (1978a) Spatial autocorrelation in biology 1. Methodology. Biological Journal of the Linnean Society, vol. 10, p. 199-228.

Sokal R. R. & Oden N. L. (1978b) Spatial autocorrelation in biology 2. Some biological implications and four applications of evolutionary and ecological interest. Biological Journal of the Linnean Society, vol. 10, p. 229-249

Stephens P. A., Sutherland W. J. & Freckleton R. P. (1999) What is the Allee effect? Oikos, vol. 87, n° 1, p. 185-190.

Stone M. (1974) Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society Series B (Methodological), vol. 36, n° 2, p. 111-147.

Stone M. (1977) An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society Series B (Methodological), vol. 39, n° 1, p. 44-47.

Stone M. (1979) Comments on model selection criteria of Akaike and Schwarz. Journal of the Royal Statistical Society Series B (Methodological), vol. 41, n° 2, p. 276-278.

Student - Gosset W. (1914) The elimination of spurious correlation due to position in time or space. Biometrika, vol. 10, n° 1, p. 179-180.

Thayn J. B. & Simanis J. M. (2013) Accounting for spatial autocorrelation in linear regression models using spatial filtering with eigenvectors. Annals of the Association of American Geographers, vol. 103, n° 1, p. 47-66.

Tibshirani R. (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society Series B (Methodological), vol. 58, n° 1, p. 267-288

Thierney L. & Kadane J. B. (1986) Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, vol. 81, n° 393, p. 82-86.

Thiollay J. M. & Bretagnolle V. (2004) *Rapaces nicheurs de France: distribution, effectifs et conservation.* Delachaux et Niestlé, Paris.

Tiefelsdorf M., Griffith D. A. & Boots B. (1999) A variance-stabilizing coding scheme for spatial link matrices. Environment and Planning A, vol. 31, n° 1, p. 165-180.

Tu W. (2002) Zero-inflated data, Encyclopedia of Environmetrics, vol. 4, p. 2387–2391.

Tukey J. W. (1958) Bias and Confidence in Not-quite Large Samples. In: *Abstracts of papers.*The Annals of Mathematical Statistics, vol. 29, n° 2, p. 614.

Vaclavik T., Kupfer J. A. & Meentemeyer R. K. (2011) Accounting for multi-scale spatial autocorrelation improves performance of invasive species distribution modelling (iSDM). Journal of Biogeography, vol. 39, p. 42-55.

Vaida F. & Blanchard S. (2005) Conditional Akaike information for mixed-effects models. Biometrika, vol. 92, n° 2, p. 351–370.

Vaudor L., Lamouroux N., Olivier J.-M. (2011) Comparing distribution models for small samples of overdispersed counts of freshwater fish. Acta Oecologica, vol. 37, p. 170-178.

Ver Hoef J. M. & Boveng P. L. (2007) Quasi-Poisson vs. negative binomial regression: How should we model overdispersed count data? Ecology, vol. 88, n°11, p. 2766–2772.

Von Neumann J., Kent R. H., Bellinson H. R. & Hart B. I. (1941) The mean square successive difference. The Annals of Mathematical Statistics, vol. 12, n° 2, p. 153-162.

Vuong Q. H. (1989) Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, vol. 7, n° 2, p. 307-333.

Ware J. H. (1985) Linear models for the analysis of longitudinal studies. The American Statistician, vol. 39, n° 2, p. 95-101.

Warton D. I. (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariateabundance data. Environmetrics, vol. 16, p. 275-289.

Wedderburn R. W. M. (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, vol. 61, n° 3, p. 439-447.

Welsh A. H., Cunningham R. B., Donnelly C. F. & Lindenmayer D. B. (1996) Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological Modelling, vol. 88, p. 297-308.

Wenger S. J. & Freeman M. C. (2008) Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. Ecology, vol. 89, n° 10, p. 2953-2959.

White G. C. (2001) Why take calculus? Rigor in wildlife management. Wildlife Society Bulletin, vol. 29, n° 1, p. 380-386.

Whittle P. (1953) The analysis of multiple stationary time series. Journal of the Royal Statistical Society Series B (Methodological), vol. 15, n° 1, p. 125-139.

Whittle P. (1954) On stationary processes in the plane. Biometrika, vol. 41, n° 3/4, p. 434-449.

William D. A. (1982) Extra-binomial variation in logistic linear models. Journal of the Royal Statistical Society Series C (Applied Statistics), vol. 31, n° 2, p. 144-148.

Williams D. A. (1996) Overdispersion in logistic-linear models. In: *Statistics in Toxicology* (B.J.T. Morgan ed.). Clarendon Press, Oxford.

Wold H. (1938) *A study in the analysis of stationary time series.* Uppsala.

Wolfinger R. & O'Connell M. (1993) Generalized linear mixed models: a pseudo-likelihood approach. Journal of Statistical Computation and Simulation, vol. 48, p. 233-243.

Yates E. (1949) *Sampling methods for censuses and surveys.* Hafner, New York.

Yule G. U. (1921) On the time-correlation problem, with especial reference to the variate-difference correlation method. Journal of the Royal Statistical Society, vol. 84, n° 4 p. 497-537.

Zou H. & Hastie T. (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B (Methodological), vol. 67, p. 301-320.