LA DIMENSION SPATIALE EN ÉCOLOGIE MOLÉCULAIRE

Science de l'information géographique et conservation des ressources génétiques végétales et animales

par Stéphane Joost

Laboratoire de systèmes d'information géographique
Faculté ENAC, École polytechnique fédérale de Lausanne
Bâtiment GC, Station 18 1015 LAUSANNE, Suisse
E-mail : stephane.joost@epfl.ch

Cet article présente une série de contributions que la science de l'information géographique peut apporter dans le domaine de l'étude des ressources génétiques végétales et animales dans une perspective de conservation de la biodiversité. Le principal apport consiste en une méthode d'analyse spatiale originale, basée sur le concept de coïncidence spatiale et sur le calcul simultané de nombreuses régressions logistiques dont le rôle est de détecter des régions du génome soumises à la sélection naturelle. Son application permet d'identifier les mêmes régions que celles mises en évidence par une approche théorique en génomique des populations et, en outre, d'identifier les paramètres environnementaux responsables de la sélection, ce qui était difficilement faisable jusqu'ici. Cette recherche ouvre des perspectives d'applications pour la science de l'information géographique en écologie moléculaire.

1 Introduction

Les plantes, les animaux et les hommes sont répartis dans l'espace géographique et interagissent avec leur environnement. L'information génétique contenue dans ces organismes vivants peut par conséquent être partiellement décrite par des coordonnées géographiques. Cette mise en relation de l'information génétique avec l'espace géographique est très bien illustrée par un récent projet appelé «Genographic» et initié de concert par la National Geographic Society et IBM Corporation, et dont le but est de collecter et d'analyser plus de 100 000 échantillons d'ADN humain afin de retracer nos origines et de cartographier les mouvements des populations durant les 60 000 dernières années (Behar et al. 2007).

Comme cela a été démontré dans de nombreux cas, l'isolation géographique est à même de déclencher des processus évolutifs divergents, puis la spéciation correspondante (de Duve 2005). L'exemple fameux des pinsons de Darwin (Darwin 1839) met en évidence le fait que les organismes sont directement influencés par les caractéristiques de leur environnement, et que les liens genome–phenome dans la nature ne sont pas dus au hasard, mais structurés et corrélés avec la diversité environnementale (Nevo 2001). Les ressources génétiques sont scellées dans un contexte géographique dont il faut tenir compte lorsque l'on essaie de les comprendre avec les mécanismes qui leur permet d'évoluer.

2 Des méthodes inédites pour comprendre les processus évolutifs

La génétique est une discipline scientifique dont l'on parle beaucoup dans notre société notamment en raison des débats grand public à propos des organismes génétiquement modifiés ou des thérapies géniques. Malgré cela, l'information génétique n'a été jusqu'ici que rarement analysée par les géographes ou les chercheurs en science de l'information géographique, alors que les travaux des généticiens des populations et des écologistes moléculaires ont permis de mettre au point des approches spatiales innovantes. Ce sont principalement les travaux d'Arthur Mourant sur la distribution spatiale des

groupes sanguins en 1954, ou les recherches de Luigi Cavalli-Sforza, Paolo Menozzi et Alberto Piazza, au début des années 1990, qui montrent sur des cartes géographiques la variation de la fréquence de gènes humains (Cavalli-Sforza et al. 1994), qui ont le plus largement contribué à l'émergence progressive de la génétique environnementale (Landscape Genetics, Manel et al. 2003).

La démarche appliquée dans le cadre de cette recherche met clairement l'accent sur l'information géographique et montre comment la science de l'information géographique peut fournir aux sciences de la vie des moyens alternatifs pour exploiter les données produites par les biotechnologies. Beaucoup de projets de recherche, que ce soit sur le séquençage du génome (humain, du poulet, etc.), en génétique des populations ou en biologie de la conservation, participent à la collecte de grandes quantités d'échantillons biologiques, dont une part croissante est géographiquement référencée. Les techniques avancées en biologie moléculaire rendent possible l'analyse et la mesure efficace de ces quantités toujours croissantes de données génétiques. Dans ce contexte, il est souhaitable que différents types d'approches interdisciplinaires soient exploités de manière à aborder cette information si complexe sous différents angles, de manière à procurer des moyens complémentaires pour la comprendre (Woese 2004). La science de l'information géographique est justement une approche qui aborde l'information génétique d'un point de vue original, susceptible de fournir des méthodes inédites dans le but d'aborder quelques-uns des défis liés à la compréhension des processus évolutifs.

Cette perspective est développée pour évaluer les contributions potentielles que la science de l'information géographique peut apporter à la biologie de la conservation en particulier. Le projet de recherche européen ECONOGENE (http://www.-econogene.eu) consacré à la conservation durable des ressources génétiques des chèvres et des moutons a fourni la plus grande partie des données. Des cas d'étude complémentaires sur la plante brassicacée (Biscutella laevigata), sur la grenouille rousse, l'ours brun de Scandinavie ou le grand charançon du pin ont pu montrer comment la dimension spatiale de l'information génétique pouvait être utilisée pour évaluer des mesures de conservation concernant des espèces sauvages menacées.

3 Différents apports de la science de l'information géographique

Dans ce travail, trois aspects distincts relatifs au traitement de l'information génétique ont été traités grâce à différents outils développés en science de l'information géographique. Premièrement, l'application de l'analyse exploratoire de données spatiales (ESDA) permet de traiter de grandes quantités d'informations afin d'en révéler les structures spatiales sous-jacentes (MacEachren 1994). Cette approche rend possible l'intégration de nombreux paramètres, permet de distinguer des pistes nouvelles pour la compréhension des phénomènes de dispersion des ressources génétiques, et d'en extraire des lignes directrices en vue de recherches plus approfondies.

D'autre part, le recours à la cartographie thématique et aux règles de la sémiologie graphique a rendu possible la production de cartes efficaces pour représenter clairement la distribution géographique de la diversité génétique des races étudiées (voir fig. 1,2 et 3). La carte thématique est un outil important afin de rendre compte de l'état de la biodiversité. Elle joue également un rôle dans le cadre des mesures de surveillance des ressources génétiques des animaux d'élevage, mises en œuvre par la FAO (via la convention sur la biodiversité) afin de juguler le phénomène de disparition progressive de races, voire d'espèces entières.

Finalement, une méthode d'analyse spatiale (Sam) originale basée sur le concept de coïncidence spatiale et sur le calcul simultané de nombreuses régressions logistiques a été développée afin de détecter des signatures de sélection naturelle au sein du génome (Joost et al. 2007 ; Joost et al. 2008a ; Joost et al. 2008b). Des localisations précises sur des chromosomes choisis peuvent être mises en évidence grâce à leur niveau d'association avec des paramètres environnementaux (variables climatiques ou pédologiques, altitude, etc.), qui caractérisent les zones dans lesquelles les races étudiées sont élevées. Les deux premières contributions ont une portée relativement modeste : elles facilitent l'exploration des données génétiques et proposent des règles afin de les représenter de manière optimale respectivement. Par contre, la troisième est majeure et fournit un outil qui permet à la génétique des populations de progresser, de faire des découvertes. Au-delà de l'identification de régions génomiques potentiellement soumises à la sélection, la méthode proposée permet d'identifier les facteurs environnementaux qui exercent la pression de sélection.

4 Aborder la sélection du point de vue géo-environnemental

La combinaison de l'analyse spatiale et de la biologie moléculaire permet d'étudier les bases géné-

tiques de l'adaptation à différents environnements (écologie moléculaire). La compréhension de ce processus est au centre des préoccupations des chercheurs en biologie de l'évolution (Storz 2005). La mise en évidence de traces de sélection naturelle au sein du génome est fondamentale dans la mesure où elle peut permettre de comprendre quels sont les gènes qui sont fixés par la sélection naturelle. La découverte de telles signatures génomiques est susceptible de favoriser le développement d'applications dans des domaines aussi variés que la médecine, l'élevage des animaux domestiques (sélection naturelle et artificielle), ou la conservation d'espèces en voie de disparition.

4.1 Modèles d'association

La méthode mise au point tire profit de la récente évolution d'un pan de la génétique vers la génomique. Grâce aux progrès des biotechnologies, il est désormais possible de séquencer des zones qui couvrent de très grandes parties du génome. Alors qu'il y a quelques temps on devait se satisfaire de connaître la localisation de quelques marqueurs moléculaires sur un chromosome, ce sont maintenant des milliers de marqueurs qui peuvent être détectés en une seule opération. Ces marqueurs sont des points de repère fixes dont on va mesurer la fréquence en fonction de la variation de variables environnementales choisies.

Le principe de l'approche est basé sur l'un des six concepts d'analyse spatiale distingués par Michael Goodchild (1996), la coïncidence spatiale. Celle-ci permet de mettre en relation les caractéristiques génétiques des organismes étudiés avec des valeurs de paramètres environnementaux grâce à des coordonnées géographiques communes. Par conséquent, pour fonctionner la SAM utilise deux jeux de données géoréférencés : l'un qui contient le génotype des organismes étudiés, et l'autre avec les valeurs d'une ou de plusieurs variables environnementales. La régression logistique (Hosmer and Lemeshow 2000) fournit la mesure du niveau d'association entre la fréquence des marqueurs moléculaires et les valeurs des paramètres environnementaux. On calcule la significativité des modèles constitués par toutes les paires possibles [marqueur / paramètre environnemental] afin d'identifier les marqueurs impliqués dans les modèles les plus significatifs : ces marqueurs jouent probablement un rôle dans les processus d'adaptation. Le nombre de modèles calculés simultanément peut être très important et c'est justement un point fort de la méthode. Par exemple, dans le cas du mouton avec des marqueurs de type microsatellites, c'est le niveau d'association de plus de 110 paramètres environnementaux avec environ 750 marqueurs, soit plus de 80 000 modèles, qui a été calculé en bloc. Le processus de calcul a été automatisé dans un logiciel basé sur Matlab® et en particulier la fonction GLMfit - Generalized Linear Model Fitting (MacCullagh and Nelder 1989). Le programme Sam résout les équations de vraisemblance afin d'estimer les paramètres, calcule les p valeurs associées aux tests statistiques (Wald et G), génère les graphes des courbes de probabilité, et exporte les résultats. Des macros Excel complémentaires ont été développées afin de traiter les matrices de résultats dans des tables dynamiques qui permettent d'ajuster le niveau de significativité et de sélectionner les marqueurs intéressants (Joost et al. 2008a).

4.2 Validation par la génomique des populations

En génomique des populations, des modèles théoriques ont été développés afin d'identifier des régions du génome potentiellement sous sélection. Ils font appel à des modèles théoriques en génomique des populations, développés dès les années 1920 déjà (Nielsen 2005). En tenant compte d'un taux de mutation donné, ils comparent la diversité génétique au sein des populations étudiées avec la diversité entre ces populations. Afin de valider la méthode SAM, une approche de ce type développée par Beaumont et Balding (1996), puis améliorée par Beaumont et Nichols (2004), a été systématiquement appliquée sur les mêmes jeux de données afin de comparer les résultats (logiciels Fdist2 et Dfdist).

5 Résultats

Les analyses ont été effectuées sur une plante (Biscutella laevigata), sur différentes races de chèvres et de moutons, sur la grenouille rousse et l'ours brun de Scandinavie, avec deux types de marqueurs génétiques (microsatellites et AFLP). Les résultats obtenus avec les deux méthodes (analyse spatiale et génomique des populations) sont concordants : les régions du génome détectées par la méthode Sam sont confirmées par Fdist2 ou Dfdist, avec en prime l'identification de la source potentielle de pression de sélection, qui n'est pas identifiée par les approches théoriques en génétique des populations. Les résultats les plus intéressants ont été obtenus chez le mouton avec les deux types de marqueurs. En effet, la méthode Sam a permis d'une part de révéler l'association d'un marqueur de type microsatellite impliqué dans un processus de résistance à un parasite avec des variables liées à l'humidité (Joost et al. 2007). D'autre part, elle a mis en évidence un marqueur AFLP potentiellement lié

au processus de sécrétion de cire de laine, associé avec de grandes quantités de précipitations et un fort taux d'humidité.

Les résultats fournis par la méthode Sam ont été également exploités de manière originale pour un autre type d'application. En effet, lorsque l'on établit des cartes d'habitat potentiel pour des espèces végétales ou animales, il est nécessaire de sélectionner un certain nombre de prédicteurs pertinents afin de les inclure dans un modèle (GLM, GAM, ENFA, etc., voir Guisan et Zimmermann 2000). Il est raisonnable de formuler une hypothèse selon laquelle les paramètres environnementaux les plus significativement associés avec certaines régions du génome sont des constituants de l'habitat de l'espèce étudiée. La méthode Sam permet de sélectionner ces variables environnementales et de les inclure dans un modèle capable de produire une carte prédictive qui repose sur la réalité de l'empreinte laissée par l'environnement sur l'organisme étudié. Ce raisonnement a été appliqué avec succès à des populations d'ours brun en Scandinavie (fig. 4).

6 Conclusion et perspectives

La méthode proposée est parfaitement adaptée pour pleinement exploiter les données produites en génomique des populations et disponibles pour beaucoup d'espèces (Luikart et al. 2003). La mise en relation des données environnementales avec des données génétiques issues de larges balayages du génome permet de compléter les apports de la génétique environnementale (Manel et al. 2003). Elle met à disposition de la génomique fonctionnelle un outil capable de fournir des points de départ pour le repérage des gènes et la compréhension de leur(s) fonction(s). Elle fournit également des pistes de recherche grâce à l'identification de la source de pression environnementale. Grâce aux progrès des biotechnologies, aux améliorations constantes des équipements informatiques, et surtout au renforcement des efforts en vue d'une recherche interdisciplinaire (Morange 2005), la génomique environnementale va certainement faire progresser les connaissances que nous avons des mécanismes génétiques de l'évolution.

Bibliographie

Avise J.C., 2004, *Molecular Markers, Natural History, and Evolution,* Sinauer, Sunderland.

Beaumont M.A., Nichols, R.A., 1996, "Evaluating loci for use in the genetic analysis of population structure", *Proceedings of the Royal Society of London Series B-Biological Sciences*, 263 (1377), p. 1619-1626.

Behar D. M., Rosset S., Blue-Smith J., Balanovsky O., Tzur S., Comas D., Mitchell R. J., Quintana-Murci L., Tyler-Smith C., Wells R. S., and G. Consortium, 2007, "The genographic project public participation mitochondrial DNA database", *Plos Genetics,* 3 (6), p.1083-1095.

Cavalli-Sforza et al., 1994, *The history and geography of human genes*, Princeton, Princeton University Press.

Darwin C., 1839 (1989), *The voyage of the Beagle, Charles Darwin's journal of researches*, London, Penguin Books.

De Duve C., 2005, *À l'écoute du vivant*, Paris, Odile Jacob.

Goodchild M.F., 1996, "Geographic Information Systems and spatial analysis in the social sciences", dans Aldenderfer, M. and Maschner, H.D.G. (eds), *Anthropology, space, and Geographic Information Systems*, New York, Oxford University Press, p. 214-250.

Guisan A., Zimmermann N., 2000, "Predictive habitat distribution models in ecology", *Ecological Modelling*, 135, p.147-186.

Hosmer D.W., Lemeshow S., 2000, *Applied logistic regression*, New York, John Wiley & Sons.

Joost S., Kalbermatten M., Bonin A., 2008a**,** "Spatial Analysis Method (SAM): a software tool combining molecular and environmental data to identify candidate loci for selection", *Molecular Ecology Resources*, 8, p.957–960.

Joost S., Bonin A., Taberlet P., Caloz R., 2008b, « Un rôle pour la science de l'information géographique en écologie moléculaire: la détection de régions du génome soumises à la sélection naturelle », *Revue internationale de géomatique*, 2, p. 215-237.

Joost S., Bonin A., Bruford M.W., Després L., Conord C., Erhardt G., Taberlet P., 2007, A Spatial Analysis Method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation, *Molecular Ecology*, 18, p. 3955–3969.

Luikart G. et al., 2003, “The power and promise of population genomics: From genotyping to genome typing”, *Nature Reviews Genetics*, 4 (12), p. 981-994.

MacCullagh P., Nelder J.A., 1989, *Generalized Linear Models*, London, Chapman & Hall/CRC.

MacEachren A.M., Taylor D.R.F., 1994, *Visualization in modern cartography*, Oxford, Pergamon.

Manel S. et al., 2003, “Landscape Genetics: combining Landscape Ecology and Population Genetics”, *Trends in Ecology & Evolution*, 18 (4), p.189-197.

Morange M., 2005, *Les secrets du vivant, contre la pensée unique en biologie*, Paris, La Découverte.

Nevo E., 2001, “Evolution of genome-phenome diversity under environmental stress”, *Proceedings of the National Academy of Sciences of the United States of America*, 98, p. 6233-6240.

Nielsen R., 2005, “Molecular Signatures of Natural Selection”, *Annual Review of Genetics*, 39, p.197-218.

Storz J.F., 2005, “Using genome scans of DNA polymorphism to infer adaptive population divergence”, *Molecular Ecology,* 14, p. 671-688.

Woese C.R., 2004, “A new biology for a new century”, *Microbiology and Molecular Biology Reviews*, 68 (2), p. 173-186.

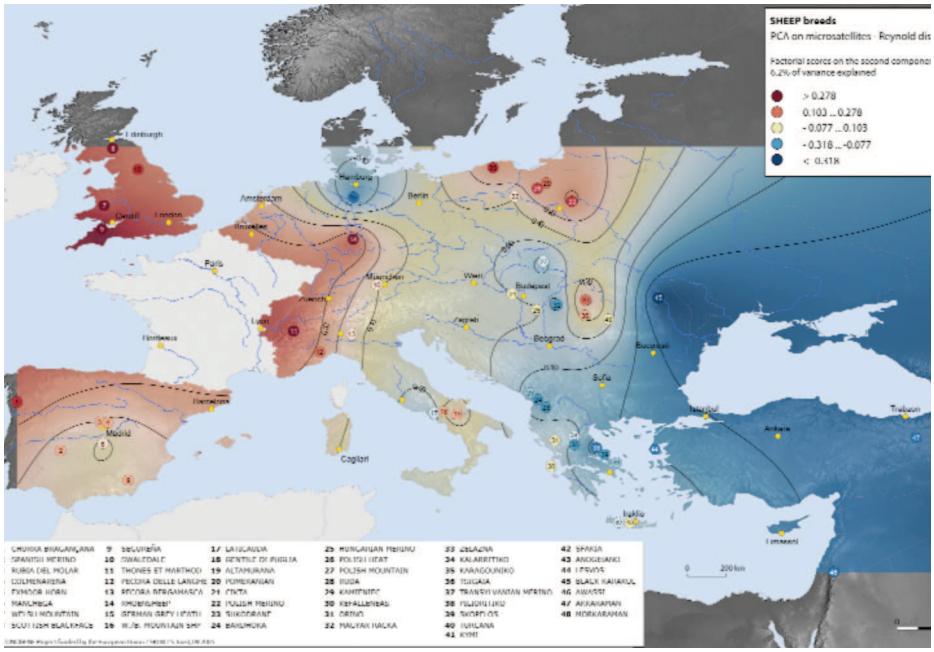

figure 1 : Analyse en composantes principales sur les distances génétiques (distance de Reynolds calculée à l'aide de marqueurs microsatellites) entre des individus constituant différentes races de moutons, étudiées dans le cadre du projet de recherche européen ECONOGENE sur la conservation des ressources génétiques des chèvres et des moutons en Europe. La valeur interpolée est le score factoriel de chaque race sur la deuxième composante qui représente 6.2% de la variance totale. Cette composante a été choisie dans la mesure où elle montre un gradient qui s'étend du sud-est vers le nord-ouest, et qui semble corroborer une hypothèse concernant un grand mouvement de migration des agriculteurs-éleveurs du Néolithique depuis les régions du Croissant Fertile vers l'Europe occidentale. Les races proches géographiquement sont également proches sur le plan génétique : il existe une relation entre distance géographique et distance génétique. La carte montre également une différence nette entre la race German Grey Heath (GGH, #15) et les autres races situées à proximité. Cette race allemande a conservé beaucoup d'authenticité : les tentatives de croisement avec d'autres races ont toujours échoué dans le passé en raison des conditions environnementales sévères de la Lande au nord de l'Allemagne, et des grandes capacités d'adaptation requises pour survivre. Il faut préciser que l'interpolation n'a pas été effectuée dans le but de prédire des valeurs, mais dans celui de mettre en évidence la structure spatiale de l'information.]

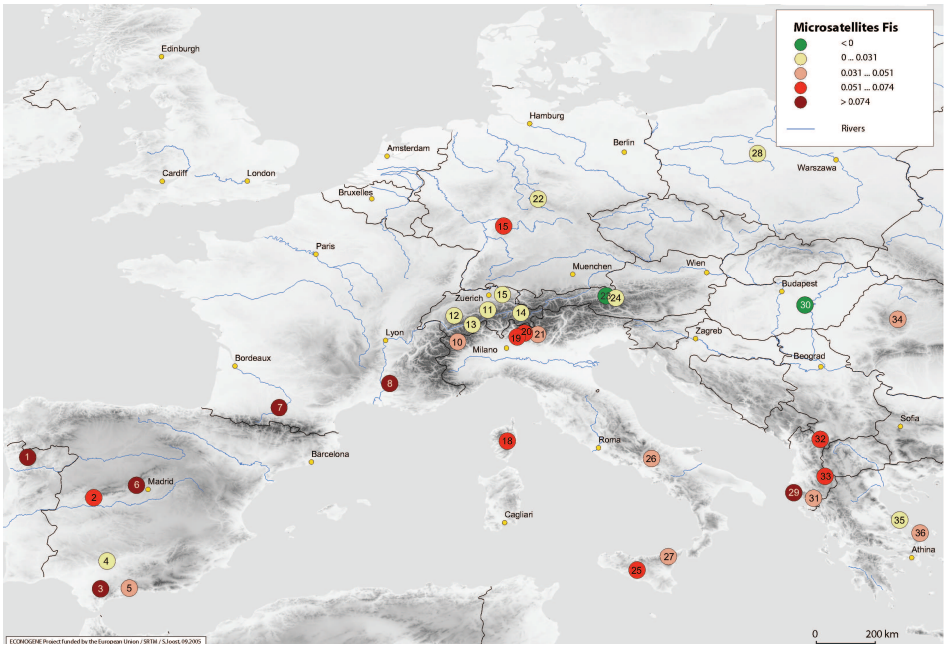

figure 2 : Cette carte montre une mesure de consanguinité (FIS) dans les races de chèvres étudiées dans le cadre du projet ECONOGENE. Un indice FIS positif met en évidence la consanguinité (peu de diversité génétique). Seules les races Pinzgauer et Hungarian Native ont un indice FIS négatif, signe de diversité génétique. L'équilibre dans les races de chèvres en Suisse est bien mis en évidence, ce qui est probablement la conséquence d'un petit marché dynamique qui favorise les échanges entre éleveurs.]

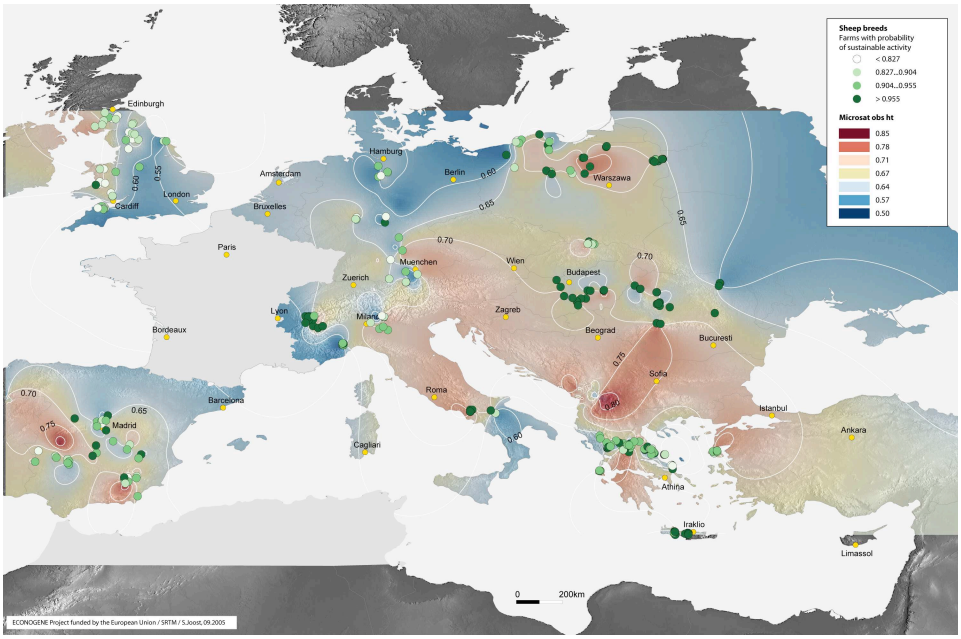

figure 3 : Dans ce cas, un indice de probabilité de durabilité de l'activité des fermes est superposé à une variable de diversité génétique interpolée en une grille continue. La carte met principalement en évidence les conséquences possibles de l'épidémie de fièvre aphteuse qui a touché les moutons en Grande-Bretagne en 2001 : la probabilité pour que les éleveurs de moutons y poursuivent leur activité est basse, à de rares exceptions près. Il faut préciser que l'interpolation n'a pas été effectuée dans le but de prédire des valeurs, mais dans celui de mettre en évidence la structure spatiale de l'information.]

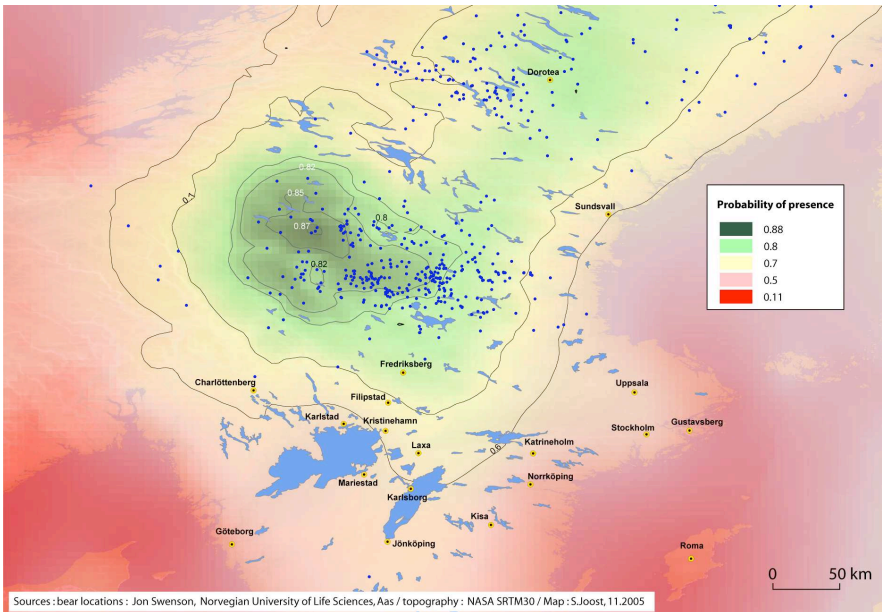

figure 4 : Carte d'habitat potentiel pour l'ours brun de Scandinavie calculé avec l'aide d'un modèle linéaire généralisé (GLM) sur la base de 4 variables environnementales significatives identifiées par la méthode Sam. Les points rouges indiquent les lieux où les 728 ours ont été échantillonnés. Les isolignes noires indiquent les classes de probabilité de présence.]