Chapitre 2.2.5.
Prédiction de la localisation des sites archéologiques

Laure NUNINGER(1), Philip VERHAGEN(2)

(1) Chrono-Environnement UMR 6249/MSHE C.N.
Ledoux UAR 3124, CNRS, Université Marie et Louis Pasteur, Besançon,
(2) Faculty of Social Science and Humanities, Vrije Universiteit Amsterdam, Amsterdam

Comme le rappellent Renfrew et Bahn dans la huitième édition de leur manuel "Archaeology : Theories, Methods and Practice » (Renfrew, Bahn 2020), une des tâches majeures de l'archéologue est de localiser et d'enregistrer le lieu où se trouvent les sites ou les vestiges archéologiques. Dans une approche d'archéologie spatiale, cette tâche est d'autant plus importante que les résultats des analyses statistiques, sur la distance par exemple (*cf.* 2.2.4), vont dépendre de ces localisations. Cependant, même si elle est fondamentale et systématiquement prise en compte, la localisation des vestiges archéologiques est une tâche complexe qui dépend de nombreux paramètres techniques, financiers, humains et environnementaux. Ainsi, comme le soulignent Hodder et Orton (Hodder, Orton 1976) les jeux de données que nous manipulons en archéologie spatiale sont incomplets, soit parce que les vestiges n'ont pas été vus, soit parce qu'ils ont disparu ou encore parce qu'ils n'ont jamais existé dans le cas d'une activité ne laissant pas de trace matérielle. Comme nous l'avons vu (*cf.* 2.2.3.), l'incomplétude* des données peut être prise en compte dans l'analyse statistique des distributions spatiales* en jouant sur l'échelle ou la taille du maillage mais cela présuppose néanmoins que la connaissance de l'espace considéré soit homogène à défaut d'être complète. Dans certaines conditions, sous hypothèse de continuité spatiale du processus étudié - dans le cas des sols par exemple - il est également possible d'estimer les valeurs de points inconnus par les valeurs des points connus grâce à des méthodes d'interpolation spatiale* (*cf.* 2.2.6). Enfin, une autre façon de considérer le problème est de produire de l'information archéologique localisée artificiellement à partir de séries de données existantes et de connaissances historiques voire d'hypothèses sur les comportements

humains. C'est un principe de base des démarches dites de modélisation prédictive qui stipule que "le comportement spatial de l'homme est dans une large mesure prévisible, ce qui implique que les lieux où les gens ont vécu et effectué leurs activités quotidiennes peuvent être identifiés sur la base de modèles statistiques et/ou explicatifs" (Verhagen, Whitley 2020). Toutefois, la modélisation prédictive en archéologie mobilise un ensemble de méthodes et d'outils qui font en réalité référence à des démarches variées que l'on peut distinguer en trois approches principales : inductive, déductive ou bayésienne (Tableau 1 a, b et c). La première implique un jeu de données en entrée, ce type d'approche est également qualifiée de "data driven", tandis que la seconde qui s'appuie sur des hypothèses associées à une base de connaissance en entrée est aussi qualifiée de "hypothesis driven". Si conceptuellement, il est possible de créer un modèle soit à partir d'induction pure, soit à partir de déduction pure, en pratique il est important de souligner qu'en archéologie, l'élaboration d'un modèle emprunte forcément aux deux approches (Wheatley, Gillings 2002). La distinction permet ici de montrer si la modélisation est plutôt guidée par les données ou plutôt par la théorie (hypothèses). La troisième approche dite bayésienne repose sur une distribution de probabilité définie *a priori* (*cf.* 12) mais il est important de souligner qu'il ne s'agit pas encore d'une approche considérée comme standard dans le domaine de la modélisation prédictive qui même dans ce cadre emprunte également aux deux approches précédentes.

Tab. 1 : Résumé des approches mobilisées en archéologie prédictive

	a) Approche inductive / "data driven" (*cf.* 5)	**b) Approche déductive / "hypothesis driven" (*cf.* 6 à 11)**	**c) Approche bayésienne (*cf.* 12)**
ENTRÉE	Collecte de données en entrée	Une base de connaissance en entrée (à partir de données observées) et une théorie statistique	Croyance "subjective" (intègre l'incertitude (statistique)*) > construction du modèle de distribution de probabilité *a priori* (structure attendue) par exemple en fonction des connaissances empiriques des experts.

MODELISATION	Analyse statistique appliquée sur l'ensemble ou sur des échantillons* du corpus de données.	Analyses combinées : multicritères (sans donnée en entrée) et/ou selon un modèle mathématique (par exemple modèle linéaire ou logarithmique, distribution normale)	Inférence* > apprentissage, modification du modèle de distribution* *a priori* à partir de données observées afin d'obtenir un modèle de distribution *a posteriori*
SORTIE	Modèle statistique (régression logistique)	Modèle de décision multicritères (méthode mathématique)	Modèle statistique bayésien
VALIDATION	Validation du modèle par la confrontation à d'autres jeux de données ou à une partie du jeu de données initial, ou encore par le calcul d'indicateurs statistiques (indicateurs de performance, de significativité statistique/valeur-p). On peut également valider le modèle par confrontation au jugement des experts (validation qualitative) pour identifier notamment les problèmes de qualité des données en entrée.	Validation du modèle par la confrontation à un ou plusieurs jeux de données et par le calcul d'indicateurs statistiques (indicateurs de performance, de significativité statistique/valeur-p)	Validation du modèle par calcul des indicateurs de performance

1 L'approche inductive et ses limites

Le développement de la modélisation prédictive en archéologie a répondu en premier lieu à une question d'ordre patrimonial face aux politiques d'aménagement du territoire, dans les années 80 aux Etats-Unis, puis dans les années 90 en Europe, en particulier dans les pays densément peuplés comme les Pays-Bas. Ce mouvement fait notamment suite à la révision de la convention européenne pour la protection du patrimoine archéologique de La Vallette (1992). Il s'agissait de trouver une alternative au coût financier que représentait la prospection archéologique sur de larges zones, en proposant un outil d'évaluation non-destructive du potentiel archéologique, soit une mesure informative du sous-sol, afin d'établir des politiques patrimoniales d'intégration, de protection et de recherche.

Pour expliciter la démarche et le questionnement spécifique dans ce type de contexte, nous présenterons l'exemple du modèle IKAW (Indicative Map of Archaeological Values of the Netherlands). Ce modèle a été créé au milieu des années 90 pour répondre aux demandes de la convention de La Valette en matière d'archéologie, à partir des données numériques disponibles à cette époque, ce qui explique nombre de ses lacunes. Il a été vivement critiqué et a fait l'objet d'un programme de recherche afin d'évaluer et améliorer l'utilisation de la modélisation prédictive pour la gestion du patrimoine (Kamermans 2007 ; Van Leusen *et al.* 2005). Il a donc fait l'objet de recommandations tant du point de vue thématique que des techniques statistiques mobilisées. Nous le présentons ici comme un contre-exemple pédagogique afin de mettre en évidence les éléments de problématique essentiels, qui sont souvent négligés par les novices, dans ce type d'approche. Aux Pays-Bas, la nécessité d'un modèle prédictif était motivée par un contexte géographique fortement impacté par l'anthropisation impliquant l'invisibilité en surface de l'essentiel des vestiges archéologiques généralement enfouis sous plusieurs mètres de sédiments. L'objectif du modèle IKAW était donc de produire une carte où l'on peut supposer la présence de matériel archéologique. La construction de cette carte reposait sur l'hypothèse de base qu'il existe une relation entre la répartition des vestiges archéologiques dans l'espace territorial d'une part, et les aspects physiques du paysage moderne d'autre part. Le modèle cherchait donc à identifier des corrélations* statistiques à partir d'un jeu de données existant et à les visualiser sous la forme d'une carte, en utilisant quatre classes ou valeurs se référant à la probabilité de rencontrer des vestiges archéologiques, sans préciser la densité spatiale*, l'âge ou le type de site concerné (Kamermans 2007). En entrée, on retrouvait donc d'une part la base de données nationale des sites et monuments archéologiques nationaux connus et géo-référencés (ARCHIS), d'autre part des informations spatiales sur l'environnement moderne (carte des sols, carte géologique) ou passé (carte de la paléogéographie du delta de la Meuse et du Rhin (région centrale des Pays-Bas)). Dans ce modèle, l'approche était principalement inductive et reposait sur des analyses statistiques simples de corrélation (Deeben *et al.* 2002) (*cf.* 5). Ainsi, il s'agissait dans une région donnée de calculer le nombre de sites archéologiques observés et d'estimer le nombre de sites potentiels pour une catégorie environnementale (variable*). La catégorie environnementale (concernant parfois plusieurs entités spatiales) pouvait être définie par une combinaison de critères. Dans le cas du modèle IKAW 2nd génération[1], il s'agissait,

[1] Une première version du modèle IKAW a été livrée en 1997, puis une seconde en 2001.

par exemple, du type de sol combiné au niveau de la nappe phréatique. Pour définir le nombre de sites estimés par catégorie environnementale (n'_e), les auteurs se sont appuyés sur une hypothèse de proportionnalité entre le nombre de sites observés et la proportion de surface occupée par chacune des catégories environnementales dans l'ensemble de la région analysée, soit :

$n'_e = n \, . \, (S_e/S)$
où n est le nombre sites archéologiques observés
S_e est la surface occupée par une catégorie environnementale spécifique
S est la surface totale de la région analysée

Ils obtenaient ainsi une valeur indicative (ou prédictive) par catégorie (I_e) à partir du ratio entre le nombre de sites observés par catégorie environnementale (n_e) et le nombre de sites estimés par catégorie environnementale (n'_e), soit :

$I_e = n_e / n'_e$

Cette valeur était ensuite classée en 3 catégories de valeurs indicatives (ou prédictives) :

- Faible : corrélation négative (pour des valeurs de I_e inférieure à 0,6)
- Moyenne : corrélation plus ou moins neutre (avec des valeurs de I_e entre 0,6 et 1,5)
- Forte : corrélation positive (quand la valeur de I_e est supérieure à 1,5)

Ce modèle quantitatif séduisant par sa simplicité permet de pointer plusieurs problèmes archéologiques, spatiaux et statistiques afin d'introduire des modèles prédictifs plus évolués. Le modèle IKAW a été fortement critiqué car la démarche inductive implique une dépendance forte aux données existantes dans la base ARCHIS. Comme Patriarche[2]. en France, ces données présentent énormément de biais en fonction de l'intensité des recherches et du fait des problèmes de taphonomie (érosion ou recouvrement des vestiges par exemple). Ainsi, dans une zone alluviale, la quasi absence de diagnostics ou de fouilles, combinée à une forte sédimentation se traduira par un faible nombre de sites observés. Dans ces conditions, la valeur indicative du modèle sera faible et d'autant plus que la zone concernée est grande. Autrement dit, dans une région occupée par une large surface de zone alluviale, la conclusion sera qu'il y a de faibles chances de trouver des sites archéologiques… Pour tenter de remédier à ce problème, il est possible d'inclure une carte taphonomique[3] dans le processus de modélisation. Il s'agit, par unités géomorphologiques reclassées en catégories taphonomiques et par périodes d'occupation, d'attribuer une valeur qualitative à la prédiction. Ce type de modèle a été développé dans la Vallée du Rhône pour caractériser l'impact de la sédimentation fluviale ou des colluvions sur la carte archéologique réalisée à partir d'investigation de surface (Verhagen, Berger 2001, Verhagen 2007b). Dans ce cas particulier, la région avait fait l'objet de recherches préventives intensives avec des

[2]Patriarche pour Patrimoine archéologique : système numérique de la carte archéologique nationale française qui couple une base de données (SGBD), un système d'informations géographiques (SIG) et un logiciel de requêtes

[3]La taphonomie, en archéologie, renvoie à l'étude des processus d'enfouissement, de conservation et d'altération des vestiges matériels et organiques. La carte taphonomique renvoie ici à des catégories qui combinent des informations géologiques, géomorphologiques et pédologiques pour évaluer les conditions de sédimentation ou d'érosion des sols.

sondages permettant de collecter un grand nombre de données archéologiques totalement masquées par les sédiments. Il a ainsi été possible d'analyser la relation entre les catégories taphonomiques et les sites visibles par rapport à l'ensemble des sites découverts. Dans un premier temps, la dépendance entre chaque catégorie taphonomique (par exemple « dépots colluviaux », « terrasse ou cône alluvial Pléistocène instable »…) et la densité de sites visibles a été testée à l'aide d'une analyse de Chi2*[4] (*cf.* 5). On compare ainsi la fréquence des données observées dans les conditions de recherche donnée (investigation de surface) aux fréquences des données estimées dans des conditions de distribution uniforme (*cf.* 7). La même opération a été réalisée sur la fréquence de l'ensemble des sites visibles et invisibles ce qui permet de qualifier l'apport des sondages archéologiques. Prenons l'exemple d'une relation d'indépendance* entre une unité taphonomique donnée et la densité spatiale de tous les sites observés. Si pour la même unité, il y a dépendance avec les seuls sites visibles, alors la valeur indicative du modèle initial (tel IKAW), en particulier si elle est forte, devra être considérée avec précaution car il s'agit sans doute d'un biais de la recherche archéologique.

Au-delà des conditions de la recherche et des problèmes taphonomiques, les modèles prédictifs génériques de type IKAW ne tiennent pas compte des différents types de sites et de leur appartenance chrono-culturelle. Or selon les périodes chronologiques, les habitats par exemple ne se répartissent pas de la même manière dans l'environnement car les besoins diffèrent en fonction des techniques agraires par exemple. En outre, les critères environnementaux ne sont pas les seuls déterminants de leur localisation, la proximité d'un habitat existant ou d'une route par exemple peut avoir un rôle important dans les choix d'implantation. Cet aspect déterministe de certains modèles prédictifs a été largement critiqué dès la fin des années 80 (Ingold 1993, Tilley 1994, Gaffney, van Leusen 1995).

Toutes ces remarques montrent que l'analyse quantitative inductive peut être conduite à titre exploratoire pour mieux connaître les données sur lesquelles on travaille. Cependant, elle n'a aucune capacité explicative d'un point de vue archéologique en termes de présence-absence (fortement dépendante de l'intensité des recherches) et encore moins pour la caractérisation des choix d'implantation. En d'autres termes, un modèle prédictif doit être construit à partir d'une connaissance experte de la région étudiée (taphonomie, niveau d'investigation) et inclure de multiples variables à la fois : environnementales, sociales, comportementales…

2 L'approche déductive

Cette seconde approche repose sur une démarche par hypothèse qui consiste d'abord à définir à partir de nos connaissances ce qui potentiellement constitue une zone attractive pour la découverte des sites archéologiques. L'entrée ne se fait donc pas par les données archéologiques observées mais par des hypothèses d'ordre anthropologiques et cognitives de comportement. Il s'agit de prédire la localisation de ressources ou de

[4] Il s'agit d'une procédure standard dans la modélisation inductive mais qui n'avait pas été mise en œuvre dans le modèle IKAW d'où sa faiblesse et les critiques dont il a fait l'objet. Sur ce point nous renvoyons à un autre modèle développé dès le milieu des années 90 au Minnesota (Etats-Unis), appelé Mn/model. Bien que critiqué, ce modèle présente l'un des meilleurs cas de modélisation inductive du point de vue méthodologique (Hudak *et al.* 2002, Verhagen, Whitley 2020).

contraintes environnementales plutôt que de prédire la localisation des sites en fonction de catégories géoenvironnementales données, pour ensuite tester leur relation avec la densité de sites archéologiques observés. Ce type de modèle a été élaboré, notamment pour répondre aux critiques de modèles fondés sur des données biaisées, sur une approche déterministe et sur l'utilisation trompeuse d'indicateurs de performance du modèle (*cf.* 2.2.5.4) qui n'a pas de valeur explicative d'un point de vue thématique. Pour illustrer cette approche, nous utiliserons le cas d'un modèle développé sur la côte de Georgie aux Etats-Unis par Thomas Whitley (Whitley 2013, Verhagen, Whitley 2020). Ce modèle concerne les choix d'implantation de populations préhistoriques (4500-300 BP) en fonction de ressources de subsistance potentielles et du coût énergétique que représente leur collecte (temps de déplacement, coût technique…). Il repose sur une théorie économique dite de maximisation (ou d'optimisation) impliquant un équilibre admis entre les coûts et les dépenses énergétiques. Dans ces conditions, le modèle présuppose que la sélection d'un lieu (un site) par une population préhistorique est le résultat d'un processus cognitif qui évalue les avantages et les risques pour sa subsistance. L'objectif n'est pas de créer un modèle prédictif pour évaluer un potentiel archéologique mais d'utiliser l'analyse spatiale* guidée par une théorie pour appréhender les facteurs de choix d'implantation de populations chasseurs-cueilleurs. Les variables créés s'appuient d'une part sur la localisation de ressources animales et végétales en fonction du potentiel énergétique qu'elle représente (qualité et quantité de ressource en fonction d'hypothèses sur le régime alimentaire), d'autre part sur les contraintes liées à ces localisations en termes de dépense d'énergie (accès, investissement technique et humain pour la collecte par exemple). Pour chaque type de ressource animale ou végétale et pour chaque mois de l'année une variable (couche d'information spatiale) a été créée avec un degré d'intensité variant de 0 à 1 reflétant le potentiel de calories disponibles, par pixel de 30 m de côté, dans la zone étudiée. Ce potentiel repose sur des estimations de taille des populations, de taux de reproduction… Une série de cartes (couches d'information spatiale) est ainsi produite représentant, par exemple, le potentiel de calories disponibles de la ressource en coquillages pour le mois de septembre autour de 500 BP (Withley 2013, fig. 13.2). Toutes les variables (couches d'information spatiale) sont ensuite combinées (additionnées) pour obtenir une carte de la valeur calorifique de chaque pixel de la zone étudiée. Il s'agit des surfaces de calories "disponibles". Elles sont ensuite modifiées en surface de calories "restituées" par la soustraction des variables de coût d'extraction estimées en fonction de l'accès (carte de friction terrestre et aquatique) ou du traitement nécessaire à la collecte de la ressource. Pour en savoir plus sur les différentes variables nous renvoyons le lecteur à la publication originale (Whitley 2013).

Ce type d'approche fournit un modèle paléoéconomique de localisation potentielle qui tout proportion gardée pourrait s'apparenter à un modèle météorologique. Il n'induit pas la réalisation de la prédiction mais permet d'expliciter la répartition de sites selon certaines conditions à la fois spatiales et temporelles qui sont propres au modèle.

3 Faire face à l'incertitude : approche bayésienne

Que la démarche soit inductive ou déductive, la qualité des hypothèses et des données en entrée est fondamentale pour obtenir de bons résultats. Or, en archéologie

nous sommes le plus souvent face à des données lacunaires ou incertaines. Pour tenter d'intégrer cette incertitude une nouvelle génération de modèle a été développé dans les années 2000 en s'appuyant sur le cadre théorique bayésien (*cf.* 12). Dans cette approche, il s'agit d'une certaine manière d'inverser le processus en définissant la quantité *a priori* de présence ou d'absence de site sous la forme d'une distribution de probabilité, que l'on appelle connaissance *a priori*. Cette connaissance *a priori* est ensuite modifiée à l'aide d'un modèle mathématique (*cf.* 12) en utilisant des données quantitatives pour obtenir une connaissance *a posteriori*[5]. L'intérêt du modèle bayésien est de fournir un cadre formel pour améliorer les modèles qui permet de quantifier explicitement l'incertitude sans approximation (*cf.* 12). Nous illustrerons notre propos à partir d'un modèle bayésien développé sur la région de Rijssen-Wierden au Pays-Bas où un premier modèle prédictif de l'habitat fondé sur une approche statistique classique avait été réalisé dans les années 90 (Verhagen *et al.* 2011, Van Leusen *et al.* 2009). Pour les auteurs, l'objectif était d'explorer l'approche bayésienne pour gérer la question de l'incertitude au niveau des données mais aussi au niveau de l'avis des experts.

3.1 Connaissance *a priori*

Concrètement, pour définir la connaissance *a priori*, il a été demandé aux experts de reprendre les 6 catégories environnementales définies dans le premier modèle et d'évaluer leur probabilité relative de contenir des sites archéologiques. Par exemple, dans la catégorie des sols et pour la variable* texture du sol (limoneux, sableux, caillouteux…), l'expert A jugera selon ses connaissances si une modalité*, par exemple "sols sableux", est plus attractive qu'une autre, c'est-à-dire si l'on a plus ou moins de chance d'y trouver des sites. Cette appréciation qualitative est traduite quantitativement sous la forme de proportion, par exemple si on considérait trois modalités de texture de sol, toute chose égale par ailleurs, c'est-à-dire considérant que toutes les modalités sont idéalement représentées de manière équivalente dans la région, l'expert A pourrait juger que l'on a deux fois moins de chance de trouver des sites dans des sols sableux (soit 0,5) que dans les sols limoneux (soit 1) et dix fois moins de chance dans les sols caillouteux (soit 0,1) par exemple. Ces appréciations qualitatives d'une modalité de texture de sol à l'autre peuvent être résumées sous la forme d'un tableau (tableau 2 a) puis traduite quantitativement sous la forme de proportion (tableau 2 b) où chaque modalité de texture de sol est codée de 0 à 3.

Cela permet de voir que les estimations, selon les combinaisons, ne concordent pas toujours et qu'une marge d'incertitude existe dans l'appréciation de cet expert. Par exemple, selon notre expert A, la modalité de texture 2 devrait contenir entre 17 et 53% de sites avec une moyenne* de 29% et un coefficient de variation* (C.V.) de 56% (tableau 2b, ligne de la modalité 2). Autrement dit, l'estimation de l'expert A pour cette modalité est relativement incertaine, tandis qu'elle paraît beaucoup mieux évaluée pour la modalité 0 avec un coefficient de variation de 11%, même si la prédiction est en moyenne faible (environ 6% de sites). S'il existe une marge d'incertitude dans

[5] Il est important de noter que ce type de démarche visant à acquérir de la connaissance est également possible avec une approche "inductive" fondée sur des algorithmes d'apprentissage (machine learning) tel que le modèle de forêt d'arbres décisionnels (Breiman 2001)

l'appréciation d'un seul expert, cela est aussi possible dans l'appréciation donnée par plusieurs experts avec une marge de variation entre chacune d'elles pour une même modalité. Ainsi, en comparant la moyenne des appréciations de chacun, il est possible d'estimer le degré relatif d'incertitude pour chaque modalité de texture de sol (tableau 3).

Tab. 2a et b - Exemple d'appréciation d'un expert (expert A), du potentiel archéologique en fonction des combinaisons de modalités de texture de sol : a) par exemple, par rapport à la modalité 3 (en colonne) l'expert juge que l'on a deux fois moins de chance de trouver des sites dans la modalité de texture de sol 1 et dix fois moins de chance dans la modalité 0 ; b) traduction des appréciations par modalité sous forme de probabilité en divisant par la somme des colonnes (tab. 2a), moyenne et coefficient de variation (C.V. %) par modalité de texture de sol. Source Van Leusen *et al.* 2009.

modalité	**0**	**1**	**2**	**3**
0	1	0,5	0,25	0,1
1	2	1	0,25	0,5
2	4	4	1	0,33
3	10	2	3	1
Σ	17	7,5	4,5	1,93

modalité	**0**	**1**	**2**	**3**	**moyenne**	**C.V.**
0	0,059	0,067	0,056	0,052	0,058	0,109
1	0,118	0,133	0,056	0,259	0,141	0,602
2	0,235	0,533	0,222	0,172	0,291	0,564
3	0,588	0,267	0,667	0,517	0,510	0,340
Σ	1	1	1	1		

Tab. 3 - Appréciations préalables des experts exprimées sous forme de probabilités (*cf.* tableau 2b) par modalité de texture de sol et moyenne de toutes les appréciations par modalité et écart-type*. Source Van Leusen *et al.* 2009.

modalité	**0**	**1**	**2**	**3**	**Σ**
expert A	0,058	0,141	0,291	0,510	1
expert B	0,027	0,051	0,217	0,704	1
expert C	0,150	0,050	0,300	0,500	1
moyenne	0,078	0,081	0,269	0,571	1
écart-type	0,064	0,052	0,046	0,115	

Dans la réalité l'avis des experts est plus complexe et dépend de la combinaison de plusieurs facteurs, aussi dans l'expérimentation menée sur la région de Rijssen-Wierden, chaque expert a eu la possibilité d'évaluer l'attractivité des modalités de texture par exemple selon plusieurs combinaisons de critères, liées à leur connaissance du terrain, soit des conditions différentes. La combinaison de toutes les appréciations des experts, permet d'obtenir une évaluation du jugement moyen et sa variabilité (coefficient de variation) pour ensuite estimer "l'équivalent en données", c'est-à-dire le nombre de sites observés qu'il faudrait pour respecter le jugement des experts sur leur distribution par modalité. Pour calculer "l'équivalent en données", les auteurs ont utilisé deux types d'approche. La première (méthode A, tableau 4) consiste à ce

que chaque expert vaut pour une observation, soit un total de 3 sites pour 3 experts, distribué ensuite par classe de texture de sol en fonction de la probabilité moyenne du jugement de tous les experts. Selon cette méthode, on aura donc l'équivalent en nombre de sites de 0,81 pour la modalité 2 (tableau 4). La seconde approche (méthode B1 et B2, tableau 4) prend en compte la variabilité du jugement des experts. Statistiquement le calcul intègre l'écart-type pour estimer la valeur α_0[6], définie comme $(\mu(1 - \mu)/\sigma^2)$ - 1. Si les experts sont d'accord, l'écart-type est faible et cela se traduit par une valeur α_0 élevé. Si les experts sont en désaccord, l'écart type est fort et cela se traduit par une valeur α_0 faible. Dans notre exemple, la méthode B1 utilise la moyenne de ces valeurs α_0 pour obtenir l'"équivalent en données" en données de la variable "texture de sol", soit 38,6 sites. La méthode B2 est plus prudente en s'appuyant sur le minimum de ces valeurs α_0 pour obtenir l'"équivalent en données" en données de la variable "texture de sol", soit 16,8 sites. Finalement, les auteurs ont fait le choix de retenir la valeur de 30 sites (dites méthode C) comme un compromis entre les méthodes B1 et B2, et représentatif de la gamme des valeurs obtenues sur les 6 variables environnementales.

Tab. 4 - Calcul de "l'équivalent en données" à l'aide des vecteurs de distribution de Dirichlet et selon plusieurs méthodes. Source Van Leusen *et al.* 2009.

modalité	**0**	**1**	**2**	**3**	**équivalent en données**
méthode A	0,24	0,24	0,81	1,71	3
α_0	16,8	26,1	94	17,4	
méthode B1	3,0	3,1	10,4	22,1	38,6
méthode B2	1,3	1,4	4,5	9,6	16,8
méthode C	2,4	2,4	8,1	17,1	30

3.2 Connaissance *a posteriori*

À partir de cette connaissance *a priori*, la probabilité relative de trouver un site en fonction du jugement des experts a été calculée pour l'ensemble des 6 variables (figure 1). Cette carte qui résume le jugement des experts a été confrontée au modèle de prédiction initial, fondé sur 80 sites réels observés, et montre des divergences. Par exemple, dans certaines zones les sites sont absents alors que la zone présentait une probabilité relativement élevée selon le jugement des experts.

[6]Le calcul de l'équivalent en données est réalisé en utilisant les vecteurs de distribution de Dirichlet. La distribution de Dirichlet est une famille de distribution de probabilité paramétrée par un vecteur de réels positifs que l'on nomme alpha (α). Cette distribution représente la distribution *a priori* qui conjugue les distributions discrètes non paramétriques, c'est-à-dire les distributions par catégories exprimées par la moyenne et la variance* de chacune.

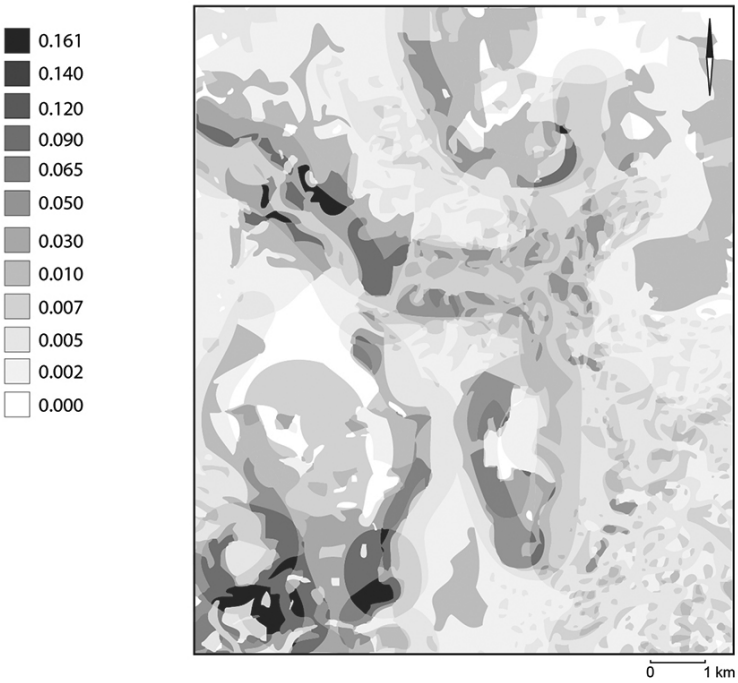


FIG. 1 : Carte résumant l'avis des experts sous forme de densité relative de sites potentiels : une cellule avec une valeur de 0,12 a deux fois plus de chance de contenir un site qu'une cellule d'une valeur de 0,06. Source Van Leusen *et al.* 2009, Figure 9.3.

Ce résultat est en partie normal puisque la découverte des sites est également liée aux conditions de découvertes (visibilité, travaux…). Inversement, si une forte proportion de sites est trouvée dans une zone de faible prédiction par les experts, cela permet de questionner la fiabilité du jugement des experts, ou celle des données, voire les deux. Pour affiner la connaissance *a priori* (jugement des experts), le corpus de 80 sites observés est ajouté au 30 sites "équivalent en données" et le modèle est exécuté de nouveau pour calculer les densités *a posteriori* sur les 110 observations[7] . La carte obtenue (figure 2) permet d'affiner la valeur de prédiction des sites en combinant les données réelles à l'expertise.

[7]Dans ces conditions, on peut dire que le jugement des experts pèse d'environ un tiers dans les paramètres du modèle. Quand les données sont lacunaires sur une région, mais l'expertise très bonne, il est possible d'augmenter le poids du jugement des experts en incluant plus de sites "équivalent en données", et inversement de diminuer le poids de ce jugement si les données sont en quantité et qualité suffisantes.

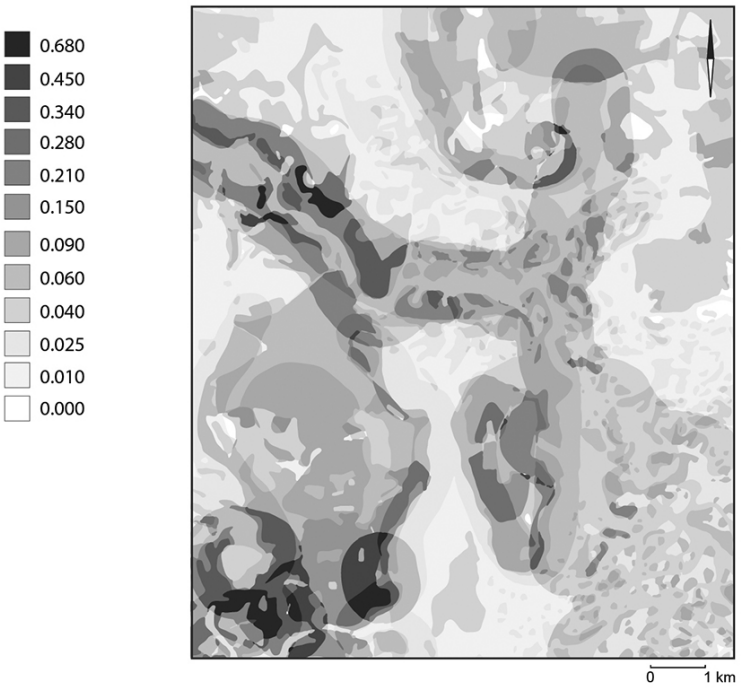


Fig. 2 : Carte de densité relative de sites potentiels *a posteriori*, après avoir ajouté les données observées. Les valeurs de densité spatiale sont exprimées en nombre de sites par hectare. Source Van Leusen *et al.* 2009, Figure 9.10.

On voit par exemple que certaines zones sont marquées par une densité relative plus élevées que celle observées sur la carte de jugement des experts (passage du blanc au gris, entre les figures 1 et 2). Mais on peut également observer que certaines zones à forte valeur prédictive selon les experts (en noir sur la figure 1) ont tendance à se généraliser (zone grise à noire plus homogène). La différence[8], entre les densités *a priori* et les densités *a posteriori*, montre comment l'introduction des données fait varier les prédictions : soit de façon positive (en noir, figure 3), c'est-à-dire avec plus de sites prédits que ne le laissait présager le jugement des experts ; soit de façon négative (en blanc, figure 3) c'est-à-dire avec moins de sites prédit qu'attendu par les experts. L'intérêt de la méthode est qu'elle permet d'ajouter progressivement des sites observés pour améliorer le modèle et les prédictions *a posteriori*, ou de réintroduire des éléments de connaissance experte *a priori* pour affiner le modèle, tout en contrôlant l'ensemble des paramètres du modèle d'un point de vue quantitatif.

[8]La différence entre les deux cartes est calculée par une simple soustraction des valeurs de chaque cellule de l'image (méthode d'algèbre de cartes), soit (valeur de densité de la cellule *a posteriori*) - (valeur de densité de la cellule *a priori*).

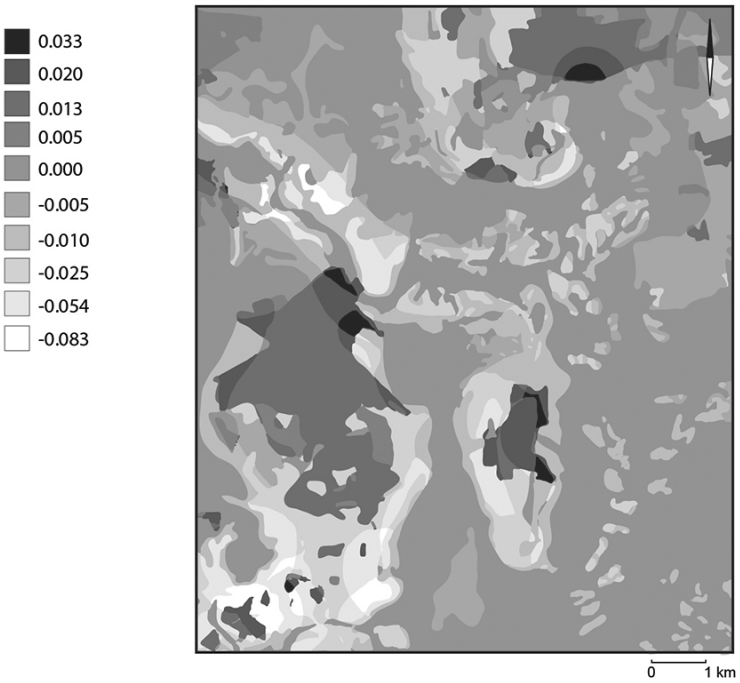


Fig. 3 : Carte résumant la différence entre les connaissances *a posteriori* et *a priori* : en noir, ce sont les zones où le modèle a posteriori à fait croître la densité de sites prédits ; en blanc, ce sont les zones où le modèle *a posteriori* à fait décroître la densité de sites prédits par les experts. Source Van Leusen *et al.* 2009, Figure 9.5.

4 Questions d'évaluation : notions de performance, de validation et de test

Dès lors qu'on élabore un modèle prédictif, la question de son évaluation se pose. C'est un problème non résolu pour lequel, là encore, on peut faire appelle à plusieurs méthodes. Sans être exhaustif, nous donnons ici un cadre très général de la démarche pour faire un lien avec les outils statistiques mais sans entrer dans les détails. Dans le processus d'évaluation du modèle prédictif, il faut d'abords distinguer trois éléments :

- la performance du modèle dont l'objectif est d'estimer le degré de prédiction correct de la présence ou de l'absence d'un site archéologique.

- la validation du modèle qui consiste à comparer un ensemble de données (jeu de données test) afin d'établir les performances de ce dernier. Il s'agit d'évaluer la cohérence interne du modèle et sa mesure ne nécessite pas forcément de nouvelles données.
- le test du modèle qui permet enfin de comparer plusieurs ensembles de données de test à l'aide du modèle, afin d'accepter ou de rejeter ce dernier. Dans ce cas, de nouvelles données sont nécessaires.

4.1 Performance du modèle

L'une des méthodes la plus connue et la plus simple pour estimer la performance d'un modèle est le calcul du *Gain* (G) statistique de Ken Kvamme (Kvamme 1988) selon la formule suivante :

$G = 1 - (\frac{p_a}{p_s})$ où

p_a = proportion de surface des zones où le modèle prédit des sites

p_s = proportion de surface des zones où on observe des sites là où ils ont été prédits

D'un point de vue pratique, ce modèle de validation simple est généralement établi à partir de la base de données renseignée initialement pour élaborer le modèle prédictif. Pour des zones bien cernées avec un grand nombre de sites présents, si on a un gain élevé, le modèle est très précis. D'autres mesures de performance privilégient l'exactitude par rapport à la **précision** (figure 4). C'est le cas, par exemple, de la mesure Kj développée par Wansleeben et Verhart (1992) qui se présente sous la forme :

$K_j = \sqrt{p_s \frac{p_s - p_a}{p_w}}$

où p_w correspond à la proportion de surface des zones sans sites.

Alors, p_w est un paramètre qui permet de corriger en la relativisant la valeur du gain maximal[9].

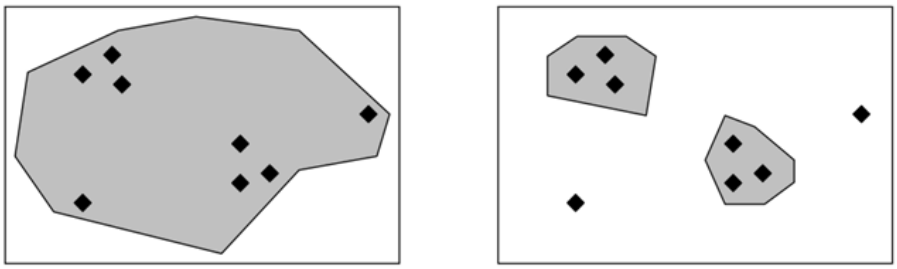


Fig. 4 : Différence entre exactitude et **précision** : à gauche le modèle est parfaitement exact puisque la zone de prédiction haute (en gris) englobe tous les sites ; à droite le modèle est moins exact car deux sites n'entrent pas dans la zone de prédiction haute mais il est beaucoup plus précis. Source Verhagen 2008.

D'autres modèles ont également été développés pour mesurer les erreurs de classification ou encore pour optimiser les mesures de performance sur plusieurs variables (Verhagen 2007a). Il est important de noter néanmoins que l'usage d'indices composites est de moins en moins courant et la plupart des études récentes comparent désormais les faux et les vrais positifs à l'aide de matrices de confusion et de courbes ROC[10], et parfois aussi les faux et les vrais négatifs.

[9] Le gain maximum correspond à la différence entre la proportion d'établissements situés dans une classe X et la proportion de cette classe X dans l'ensemble de l'espace étudié. Certains auteurs parlent également de « gain relatif » (Wansleeben, Verhart 1992, Verhagen 2007a).

[10] ROC signifie Receiver Operating Characteristic (caractéristique de fonctionnement du récepteur ou caractéristique de performance). La courbe ROC est une représentation graphique des performances d'un

Par exemple, il est désormais assez fréquent de calculer ce que l'on appelle la matrice de confusion (figure 5). Issue des méthodes d'apprentissage machine, la matrice de confusion permet de mesurer la qualité d'un système de classification et d'évaluer en faux ou vrai positif les valeurs de prédiction.

		Classe réelle (connue ou observée)	
		Négatif (0)	Positif (1)
Classe prédite par le modèle	Négatif (0)	vrais négatifs (TN)	faux négatifs (FN)
	Positif (1)	faux positifs (FP)	vrais positifs (TP)

Fig. 5 : Matrice de confusion, d'après C-A Azencott (2022), où les TP (true positive) sont les exemples positifs correctement classifiés tandis que FP sont les exemples négatifs classés positivement par le modèle et inversement pour TN et FN.

Dans la modélisation archéologique prédictive, ce sont les faux négatifs qui représentent le problème le plus grave. En effet, les faux négatifs font référence aux cas où les prédictions sont négatives alors que les valeurs réelles sont positives.

À partir de cette matrice, il est possible de dériver plusieurs critères d'évaluation tels que :

- Le rappel (recall) ou taux de vrais positifs (TPR)

  $\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$

- La spécificité ou taux de vrais négatifs (FPR)

  $\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$

  Toutefois, le critère TPR ne peut pas être utilisé seul car il est assez facile de prédire des vrais positifs et d'avoir un bon rappel. Généralement on le complète donc avec le critère de précision.

- La précision (precision) ou valeur positive prédictive (PPV) est la proportion de prédictions correctes parmi les prédictions positives

  $\mathrm{PPV} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$

  Enfin pour résumer ces deux valeurs TPR et PPV en un seul indicateur, on calcule le F-Mesure (F1-score en anglais). Il s'agit de la moyenne harmonique (*cf.* 4) de TPR et PPV :

  $\text{F-mesure} = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$

modèle qui permet de tester à différents seuils, le taux de vrais positifs et le taux de faux positifs (Saporta 2011)

Quand le modèle prédictif renvoie non pas à une classe positive ou négative mais à une probabilité d'appartenance à une classe positive ou négative alors, on utilise des courbes ROC et leur aire sous courbe (AUC)[11] construite à partir de valeurs seuils successives dans une fonction de décision (*cf.* Azencott 2022). L'utilisation de tous ces indicateurs est bien adaptée pour établir la performance des modèles, lors de l'application de méthodes statistiques de classification et de régression.

Pour terminer sur la question de performance, il est tout de même important de noter que la catégorisation des variables en entrée a également un impact important sur les résultats du modèle, et que la performance de ce dernier peut également être testé en travaillant sur la manière de classer les données environnementales en entrée, par exemple, qui va nécessairement modifier la localisation et la taille des zones de prédiction.

4.2 Validation du modèle

La validation du modèle vise à vérifier que le modèle n'est pas dépendant de la distribution initiale des données. Il s'agit donc de le tester à partir de plusieurs jeux de données. Idéalement il faudrait extraire de la base de données initiale, deux jeux de données indépendants, l'un utilisé pour construire le modèle, l'autre utilisé pour tester et valider les résultats du modèle. Dans la pratique, les conditions inhérentes à la recherche archéologique ne permettent généralement pas de produire suffisamment de données qui plus est de qualité satisfaisante pour donner la possibilité de créer des jeux de données indépendants. Ainsi, on le plus souvent recours à un rééchantillonnage statistique des données par un tirage aléatoire ou par permutation, par exemple.

Un second type test peut être réalisé sur des jeux de données simulés (produits selon un processus de distribution aléatoire), par exemple en créant une distribution aléatoire de 900 points (sites simulés). Il s'agit ici de reproduire ce que fait le test de Chi2 avec un jeu de données fictif pour mesurer l'indépendance des variables (soit des différentes catégories utilisées dans le modèle). En d'autres termes, si une relation statistiquement significative est observée entre une catégorie et la fréquence des sites, que ce soit pour l'échantillon observé ou pour l'échantillon simulé, alors d'un point de vue thématique cela ne pourra pas être interprété comme une préférence ou un choix de localisation. Dans la même perspective et comme une alternative aux méthodes traditionnelles de test statistique, les jeux de données peuvent être simulés par la méthode de Monte Carlo (ou simulation de probabilités multiples). Il s'agit d'une technique mathématique qui permet de produire des séries d'évènements (données). Le calcul est réalisé à l'aide d'un algorithme utilisant un échantillonnage aléatoire répété afin d'obtenir la probabilité d'occurrence d'une série d'évènements (données). Pour en savoir plus sur les techniques de rééchantillonnage et de simulation utilisées pour la validation des modèles nous renvoyons le lecteur à la littérature existante (Verhagen 2007a). Actuellement, le développement rapide des techniques d'apprentissage automatique a entraîné une augmentation considérable des méthodes visant à corriger le sur- et le sous-échantillonnage de catégories spécifiques de données (*cf.* Remy *et al.* 2020) mais jusqu'à présent, ces méthodes n'ont guère trouvé leur place dans la pra-

[11] Receiver Operating Characteristic and Area Under Curve

tique archéologique. Néanmoins, ce type de méthode de validation dite “interne” est discutée et certains préconisent d'élaborer le modèle de *performance* en testant le modèle prédictif sur un jeu de données indépendant, pour assurer une double validation et éviter les raisonnements circulaires. Pour ce faire, une méthode complémentaire repose sur le jugement d'expert après un retour sur le terrain (ground truth), mais ce dernier n'est pas toujours possible et nécessite des moyens en temps difficilement compatibles avec les délais de l'analyse par modélisation prédictive.

4.3 Test du modèle

Tester le modèle a pour objectif de vérifier sa valeur informative d'un point de vue archéologique. Il implique d'utiliser des jeux de données différents et/ou d'appliquer le modèle sur des zones différentes. Pour être valable le test doit reposer sur une collecte de données respectant les règles de l'échantillonnage probabiliste. En d'autres termes, l'échantillon test ne doit pas être nécessairement aléatoire mais au moins représentatif du fait archéologique étudié. Dans le domaine de l'archéologie, il s'agit d'une condition difficile à mettre en œuvre, même lorsque les principes théoriques sont bien connus (voir par exemple Orton 2000, Verhagen *et al.* 2009, Waagen 2022). Comme la population initiale des sites archéologiques n'est pas connue, il est difficile de préciser le nombre de sites nécessaires à collecter pour le test, aussi certains modèles favorisent une approche incluant les “non-sites” ou fondés sur des estimations de surface occupées plutôt que sur des densités de sites. Enfin, il est important de préciser que des estimations fiables des proportions dans les zones à faible probabilité nécessite une collecte de données plus importantes que dans les zones à forte probabilité car les estimations réalisées sur de faibles effectifs peuvent varier fortement par l'ajout de quelques découvertes seulement.

Depuis quelques années, une attention particulière est dévolue aux méthodes de validation et de test des modèles prédictifs pour estimer la validité des méthodes de prédiction traditionnelle et pour proposer de nouvelles méthodes de modélisation prédictive, fondée par exemple sur le cadre théorique bayésien. En effet, comme nous avons pu le constater ce type d'approche permet de construire le modèle progressivement en ajoutant de nouvelles données, tout en donnant la possibilité d'affiner la connaissance des experts. Il est ainsi possible de monitorer l'incertitude et de la cartographier (figure 6). Le cadre bayésien est de fait compatible avec une approche par test rétrospectif à partir de bases de données de grande qualité que l'on dégrade volontairement (c'est-à-dire pour lequel on supprime sciemment un partie des sites par exemple) pour construire un modèle prédictif et ensuite tester sa robustesse avec l'échantillon indépendant qui n'a pas servi à l'élaboration du modèle.

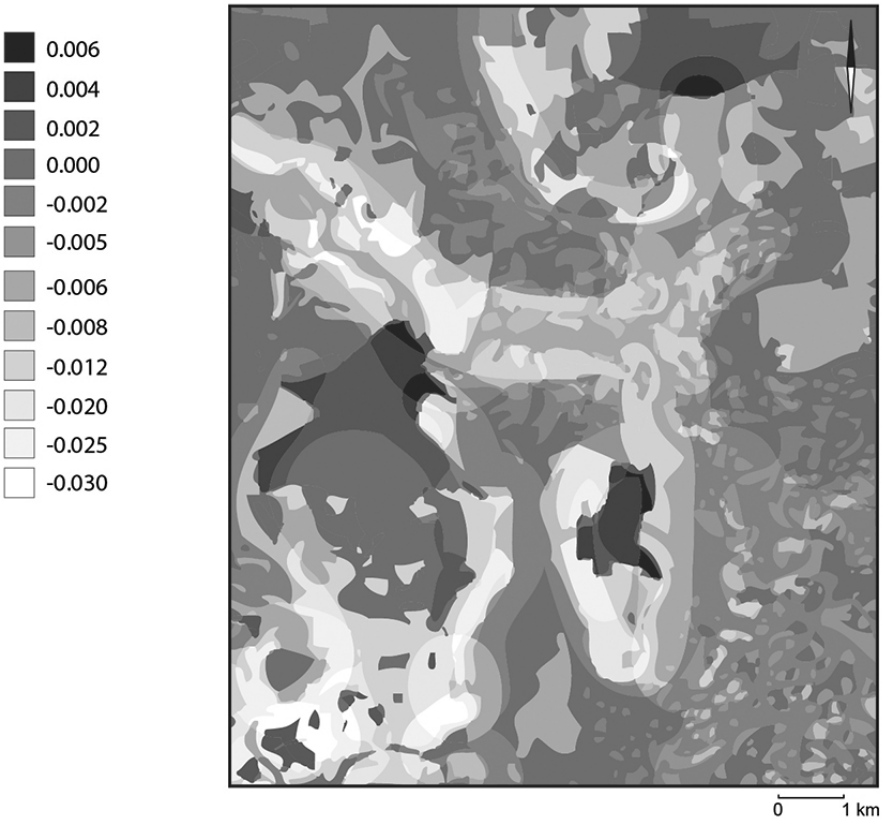


Fig. 6 : Carte résumant la différence entre l'incertitude des connaissances des experts (variation dans le jugement) et l'incertitude lié au poid (quantité) des données observées et incluses dans le modèle *a posteriori* : en noir, ce sont les zones où l'incertitude a augmenté entre la connaissance *a priori* et *a posteriori* ; en blanc, ce sont les zones où l'incertitude à diminué entre la connaissance *a priori* et *a posteriori*. Source Van Leusen *et al.* 2009, Figure 9.8.

Conclusion

Tout en présentant les différentes étapes d'une démarche de modélisation prédictive, ce chapitre a introduit plusieurs manières d'aborder le problème de la prédiction de la localisation de sites archéologiques en s'appuyant sur les deux domaines de la statistique descriptive et de la statistique mathématique. Chacune des approches présentent des limites qu'il est important de maîtriser pour éviter tout contresens, surtout quand cela engage des coûts directs et indirects importants pour la réalisation des opérations archéologiques. Néanmoins, dans une perspective de recherche, même les méthodes inductives les plus simples demeurent intéressantes dans une approche exploratoire qui vise à constituer et expérimenter un corpus de données. Elle permet, par exemple, de travailler sur plusieurs variables en entrée pour ensuite, par exemple, tester la redondance de certaines qui peuvent artificiellement renforcer la prédiction (comme dans les données topographiques, les catégories de pente nulle ou faible et la catégorie sans orientation dominante). Les approches déductives sont en revanche beaucoup plus adaptées quand il s'agit de tester plusieurs hypothèses et de comparer des situations régionales dans un même cadre de référence. C'est notamment, le choix qui a été fait dans les programmes Archaeomedes et IHAPMA[12] pour analyser la localisation des sites en fonction de variables socio-environnementales (Tourneux 2000, Nuninger *et al.* 2012c, Nüsslein *et al* 2020). Néanmoins, cela implique d'avoir des jeux de données archéologiques de qualité pour valider le modèle. Enfin, le cadre bayésien ouvre des perspectives, en ce sens qu'il s'appuie sur une démarche d'apprentissage fondé sur la connaissance préalable (ou la croyance) des experts, c'est-à-dire de ceux qui ont une expérience suffisante pour estimer les chances de découverte, que vient ensuite conforter ou nuancer l'apport de données. Le processus étant itératif, il permet d'affiner progressivement le modèle. L'avantage de ce processus, est qu'il est possible de quantifier l'expertise qualitative ainsi que sa variabilité d'un expert à l'autre (incertitude). Ainsi l'expertise peut être évaluée de manière conjointe à celle des jeux de données intégrés dans le modèle. Pour être pédagogique notre exemple portait sur une expertise de terrain, mais le cadre bayésien peut tout à fait être envisagé dans un contexte de recherche avancé impliquant des connaissances et/ou des hypothèses sur les comportements sociaux ou cognitifs par exemple. Dans ces conditions, le modèle peut-être utilisé comme un outil heuristique qui permet de spécifier progressivement un système complexe de localisation, de manière explicite et formalisée. En outre, il permet de quantifier l'incertitude épistémique qui est liée à la connaissance et non à la donnée. Ce type de modèle ne constitue pas un résultat en soi mais c'est un nouvel outil de connaissance qui peut être utilisé dans des modèles de simulation.

[12] Archaeomedes II (1996-1999) : « Policy relevant models of the natural and anthropogenic dynamics of degradation and desertification and their spatio-temporal manifestations » (Program Environment of the European Commission DGXII), coordination Sander van der Leeuw (University of Cambridge) ; IHAPMA : « Introducing the human (f)actor in predictive modelling for archaeology », PHC Van Gogh, MAE – FRNL, 2010-2011, coordination Ph. Verhagen, L. Nuninger.

En résumé :

- La modélisation prédictive en archéologie est une démarche qui repose généralement sur plusieurs approches imbriquées ainsi que sur des méthodes et outils relevant de plusieurs domaines statistique, géomatique voir de l'intelligence artificielle (machine learning). Il est donc fondamental d'expliciter toutes les étapes de modélisation et les choix retenus afin d'être en mesure d'analyser les résultats obtenus.
- Que ce soit en amont, pour construire le modèle, ou en aval, pour tester le modèle, les jeux de données archéologiques utilisés doivent être de bonne qualité et de taille suffisante pour être significatifs d'un point de vue statistique.
- De la même manière, il est fondamental d'interroger la pertinence des données environnementales intégrées en entrée dans le modèle en fonction des données archéologiques dont on dispose et de la problématique que l'on traite
- La modélisation prédictive peut-être menée à des fins d'analyse exploratoire des données dans un cadre de recherche et n'a pas nécessairement de vocation opérationnelle. Il est donc important de considérer le modèle comme un laboratoire permettant de tester des hypothèses et des variables pour observer la manière dont elles impactent notre perception du jeux de données en terme de biais et d'incertitude
- Il est important de noter que depuis les années 80-90 les méthodes et les outils mobilisés en modélisation prédictive ont considérablement évolué comme les paradigmes liés à leur usage ainsi que les problématiques abordées.