Modélisation prédictive du risque archéologique : application de la méthode "Weights of Evidence " à la plaine du Roussillon. Premiers résultats

Jean-Michel Carozza, Mélanie Pous, Thierry Odiot, Laurent Carozza

TEMPS ET ESPACES DE L'HOMME EN SOCIÉTÉ, ANALYSES ET MODÈLES SPATIAUX EN ARCHÉOLOGIE
XXV^e rencontres internationales d'archéologie et d'histoire d'Antibes
Sous la direction de J.-F. Berger, F. Bertoncello, F. Braemer, G. Davtian, M. Gazenbeek
Éditions APDCA, Antibes, 2005

Modélisation prédictive du risque archéologique : application de la méthode « Weights of Evidence » à la plaine du Roussillon. Premiers résultats

Jean-Michel CAROZZA*, Mélanie POUS*, Thierry ODIOT**, Laurent CAROZZA***

Résumé
La modélisation prédictive apparaît aujourd'hui être un outil indispensable à la gestion du risque archéologique sur le long terme. Cette étude présente l'application de la robuste méthode statistique « Weights of Evidence » à la plaine du Roussillon. Elle permet la pondération inter et intra-couche de paramètres environnementaux, considérés comme des indicateurs de la localisation de sites. Les résultats obtenus sont encourageants et pourraient être validés au cours du tracé de la future ligne LGV.

Abstract
Currently, predictive modelling appears to be an essential tool for the long-range management of the archaeological hazard. This paper presents the application of the Weights of Evidence method to GIS data from the Roussillon basin. It enables the inter- and intra-lay weighing of environmental parameters, considered as indicators of archaeological site locations. The results are encouraging and the LGV layout may validate the models.

***

Le risque archéologique peut être défini, par analogie avec le risque naturel, comme la rencontre d'un aléa – la présence d'un site – et d'une vulnérabilité – le type d'aménagements réalisés. Dans cette optique, la simple présence ou absence d'un site ne suffit pas à définir le risque archéologique. En effet, pour un aléa de même nature, le risque sera différent selon la nature des aménagements : profondeurs des fondations, emprise réellement destructive… De la même manière, la nature stratifiée ou non du site détermine également le risque archéologique. La vulnérabilité est relativement facile à définir. Elle est en effet en partie codifiée

* Faculté de géographie et d'aménagement, Université de Strasbourg, 6, rue de l'Argonne, F-67000 Strasbourg. Mél. : carozza@geographie.u-strasbg.fr

** Service régional de l'archéologie Languedoc-Roussillon.

*** Collège de France.

dans le Plan local d'urbanisme (PLU) qui détermine le type d'aménagements réalisables dans différentes unités spatiales. Le PLU permet également de planifier les enjeux d'aménagement d'un secteur déterminé. Il n'en va pas de même de l'aléa archéologique. Celui-ci ne peut être considéré comme une variable aléatoire. Différentes études depuis les années 1950 ont montré qu'il existait soit des stratégies d'implantations plus ou moins faciles à décrypter et à formaliser, soit des déterminants de la localisation.

Afin de mettre en place une politique de gestion du risque archéologique sur le long terme, suivant le modèle des Pays-Bas (Verhagen, 1995), il est donc nécessaire de se doter d'outils permettant de localiser *a priori* les secteurs offrant les plus forts potentiels archéologiques. Ces secteurs, loin de devenir des sanctuaires intouchables, doivent pouvoir faire l'objet de négociations en amont entre aménageurs et archéologues, afin par exemple de permettre leur étude sur un temps plus long. Mais encore faut-il pouvoir identifier ces secteurs. C'est le rôle des modèles prédictifs.

La modélisation prédictive : état de l'art

Il est admis que la répartition des sites archéologiques ne relève pas d'un processus aléatoire, mais qu'elle résulte 1) de choix d'implantations liés à des stratégies socio-économiques ou symboliques et 2) de la transmission ou non dans le temps de cette information par préservation-destruction de ces mêmes sites. La première approche a été développée dès les années 1950 afin de déterminer des « localisations préférentielles » de sites. La prise en compte du biais taphonomique est plus récente et n'est clairement exprimée que depuis la fin des années 1980.

La modélisation prédictive s'est imposée comme une composante de l'archéologie spatiale dans les années 1980. Son but est l'évaluation *a priori* du potentiel archéologique par étude de sa relation présence-absence de site/données environnementales. Les paramètres environnementaux sont supposés constituer : 1) soit des facteurs explicatifs de la localisation (modèles inductifs) ; 2) soit de simples indicateurs de la localisation (modèles déductifs).

Les paramètres utilisés sont variés : accès à la ressource en eau (rivière, aquifère, source…), type de formation végétale dans l'environnement immédiat, type de sols, contexte topographique (altitude, pente, exposition au vent, au soleil…), géologie…

Cependant, et c'est là une limite de taille, la relation entre l'information environnementale actuelle et lors de l'occupation n'est pas directe (Kohler, Parker, 1986). Pourtant, il est possible d'établir une liaison statistique entre présence de l'information et présence de sites, sans que ce lien soit causal (Kvamme, Kenneth, 1985). C'est pourquoi nous préférons la notion d'indicateur de présence et d'absence, les hypothèses de comportement n'étant alors pas nécessaires pour asseoir la théorie du modèle.

De très nombreuses méthodes de modélisation prédictive sont utilisées, mais peu d'études en ont comparé les résultats. Cela rend difficile le choix d'une

méthode sur une base objective. De plus, très peu de projets sont conduits jusqu'à l'étape de validation des sorties de modèles. C'est là l'aspect le plus décevant, car il est de ce fait impossible d'apprécier la pertinence des différents résultats.

Les méthodes les plus simples – ce qui ne signifie pas nécessairement les plus fausses – se basent sur des méthodes empiriques (de type expert, par ex. Kvamme, Kenneth, 1985 ; Kohler, Parker, 1986 ; Warren, 1990 ; Krier, 2004) ou statistiques simples ($\chi 2$, par ex. Davtian, 2003). Elles peuvent être utilisées seules ou combinées par exemple avec des méthodes booléennes – addition ou multiplication de couches d'information entre elles –, classiques en analyse SIG (Wheatley, Gilling, 2002). Ces méthodes dites graphiques (Della Bona, 1994) ont été appliquées dans de nombreux projets : Stancic et Kvamme (1999), Della Bona (1994)… Elles présentent toutefois un certain nombre de faiblesses, qui en limitent l'utilisation : nombre de sites par unité spatiale pour effectuer un test de $\chi 2$ significatif, binarisation de l'information, et donc perte pour une grande partie de l'intérêt de la base de données environnementale, difficulté de pondération du poids des différentes couches d'information lors de leur addition booléenne, forte redondance de l'information qui fausse le résultat final… Enfin, le résultat de ces modèles est le plus souvent qualitatif (de « faible » à « forte » probabilité), ce qui rend la validation des sorties de modèles difficile par un retour terrain.

D'autres méthodes plus complexes, comme la régression logistique (Warren, Asch, 2000) ou la logique floue, supposent de posséder une information sur les points où il n'y a pas de site, ce qui est rarement le cas, sauf dans le cadre de prospections systématiques sur de petits territoires.

C'est pourquoi nous avons recherché un modèle permettant : 1) une utilisation de seules variables de présence de site, 2) de travailler sur un très grand nombre de variables simultanément, 3) de pondérer l'information inter- et intra-couche, 4) de modifier le processus de modélisation le plus simplement possible en pouvant à tout moment retirer ou introduire des données dans le modèle.

Ces critères nous ont conduits à choisir la méthode dite « Weights of Evidence » ou « pondération de l'information probante ».

La méthode « Weights of Evidence »

La méthode « Weights of Evidence » a été développée initialement dans le domaine médical. Elle a ensuite été appliquée au domaine de la prospection minière par Agterberg *et al.* (1989) puis Bonham-Carter (1994). Il s'agit d'une méthode d'optimisation de la probabilité *a priori* basée sur le théorème de Bayes. Cette méthode a été appliquée à la modélisation prédictive, notamment dans les deltas de Californie (Hansen, 2000 ; Kaberline, Schwemm, 2002 ; Hansen, 2004).

La méthode se décompose en 5 étapes ci-après :

1. Estimation de la probabilité *a priori*

Sachant le nombre d'occurrences connues dans une zone délimitée appelée aire d'étude, il est possible de calculer la probabilité d'occurrence par aire unitaire – la « maille » de calcul. Si l'aire d'étude de t km^2 est divisée en cellules

unitaires ayant une aire fixée à u km^2, alors le nombre total de cellules unitaires de l'aire d'étude est $T = t / u$.

S'il y a D cellules unitaires contenant une occurrence, alors la probabilité *a priori* qu'une cellule unitaire tirée au hasard contienne une occurrence est $P(D) = D/T$.

2. Calcul des poids

Soit B_j l'aire en aires unitaires où le paramètre probant j est présent, alors l'aire où le paramètre probant est absent se définit par $\overline{B}_j = T - B_j$. Dans le cas où le thème comprend des zones non renseignées ou « données manquantes », il convient de les déduire également.

La probabilité conditionnelle qu'une cellule tirée au hasard contienne une occurrence et le paramètre j s'écrit $P(D|B_j) = (B_j \cap D)/B_j$.

De même, trois autres probabilités conditionnelles peuvent être définies :

$P(\overline{D}|B_j) = (B_j \cap \overline{D})/B_j$ c'est-à-dire la probabilité de ne pas trouver d'occurrences si le thème probant est présent ;

$P(D|\overline{B}_j) = (\overline{B}_j \cap D)/\overline{B}_j$ c'est-à-dire la probabilité de trouver des occurrences si le thème est absent ;

$P(\overline{D}|\overline{B}_j) = (\overline{B}_j \cap \overline{D})/\overline{B}_j$ c'est-à-dire la probabilité de ne pas trouver d'occurrences si le thème est absent.

Les calculs des poids positif et négatif sont alors les suivantes :

$W_j^+ = ln\,[(P(B_j|D)/P(B_j|\overline{D})]$ *et* $W_j^- = ln[(P(\overline{B}_j|D)/P(\overline{B}_j|\overline{D})]$.

Le contraste $C = W^+ - W^-$ dans le cas d'un thème binaire, ou $C = W_{max} - W_{min}$ dans le cas d'un thème à plusieurs classes, donne une mesure utile de la corrélation entre le thème et les occurrences.

3. Test d'indépendance conditionnelle pour chaque carte probante

La méthode repose sur une hypothèse d'indépendance conditionnelle des couches d'information. Comme le soulève Bonham-Carter (1994), cette indépendance conditionnelle est souvent violée en pratique. Il s'agit de la tester afin d'évaluer la magnitude de la redondance et d'identifier les thèmes qui posent problème ; ceux-ci peuvent alors être rejetés du modèle, reclassifiés ou amalgamés.

4. Calcul de la probabilité *a posteriori* et estimation de l'incertitude

Le calcul de la probabilité *a posteriori* s'effectue pour chaque combinaison unique de thèmes probants.

D'après le théorème de Bayes, on obtient que $P(D|B_j) = [P(B_j|D)\,P(D)/P(B_j)]$ et que $P(D|\overline{B}_j) = [P(\overline{B}_j|D)\,P(D)/P(\overline{B}_j)]$.

D'après les formules des poids, la probabilité *a posteriori* peut alors s'écrire sous sa forme logarithmique : $ln(D|B_j) = W_j^+ + ln\,(D)$ et $ln\,(D|\overline{B}_j) = W_j^- + ln(D)$.

L'avantage de cette écriture réside dans la possibilité d'additionner les poids. Ainsi, si tous les thèmes probants sont conditionnellement indépendants, le logarithme de la probabilité *a posteriori* globale, en additionnant tous les poids est :

$ln\,(D|B^k{}_1 \cap B^k{}_2 \cap B^k{}_3 \ldots B^k{}_n) = \Sigma^n{}_{j=1}\, W^k{}_j + ln\,(D)$.

5. L'incertitude de la probabilité *a posteriori*

Elle est évaluée à travers deux composantes : l'incertitude due à la valeur des poids et celle due au fait qu'un ou plusieurs thèmes probants soient incomplets, c'est-à-dire présentant des données manquantes. La formule donnée pour une unique condition de recouvrement est : $\_^2$ *(totale)* = $\_^2$ *(poids)* + $\sum^n_{j=1}$ $\_^2_j$ *(manquant)*.

L'incertitude totale ainsi estimée par des tests de Student peut être représentée sous la forme d'une carte de confiance.

Après le calcul final de la probabilité *a posteriori*, un dernier test permet de vérifier l'indépendance conditionnelle générale.

Résultats et discussion

La base de données utilisée pour la modélisation est une modification de Patriarche (2003). Elle comprend plus de 1000 sites, toutes périodes confondues. Seuls les sites post-mésolithiques ont été pris en compte.

La base environnementale se compose d'un Modèle numérique de terrain à résolution spatiale de 92 m (SRTM), des Scan 25 de l'IGN, utilisés comme support pour la digitalisation des cartes géologique et pédologique au 1:25 000^e, de la carte géologique du Roussillon, dérivée des cartes géologiques au 1:50 000^e et 1:100 000^e du BRGM modifiées, de la carte des sols, issue de la digitalisation de la carte au 1:50 000^e de Servat et Callot (1966) et de l'occupation du sol issue de la classification Corine Land Cover (base Géozoum).

Ces données constituent la base primaire. Elles permettent de générer des couches secondaires par généralisation ou par calcul. Plusieurs essais de modélisation ont été effectués.

Modèle 1 : échantillon aléatoire de sites et ensemble de la base (fig. 1)

Ce modèle a été réalisé en utilisant un échantillon de 253 sites tirés aléatoirement de l'ensemble de la base de données environnementales. Pour ce modèle, la probabilité *a priori* est de 0,003.

Les thèmes les plus probants avant généralisation sont la pédologie, la géologie et l'exposition. Ce sont de bons indicateurs de la présence de sites. La pente, notamment les valeurs comprises entre 6 et 7°, se révèle un assez bon indicateur de l'absence. Quant à l'occupation du sol actuelle, c'est un déterminant médiocre.

L'intégration de l'ensemble de ces paramètres permet d'obtenir la carte suivante (fig. 1). La probabilité *a posteriori* est nettement améliorée et s'élève jusqu'à 0,27, soit une chance sur 3 d'obtenir une tranchée positive. Les secteurs à fort potentiel sont les secteurs urbains (*sic* !) et les collines sableuses du Roussillon central. Les secteurs de basse plaine, les bourrelets alluviaux des cours d'eau et la plaine d'inondation historique, montrent de très faible probabilité, ce qui est cohérent.

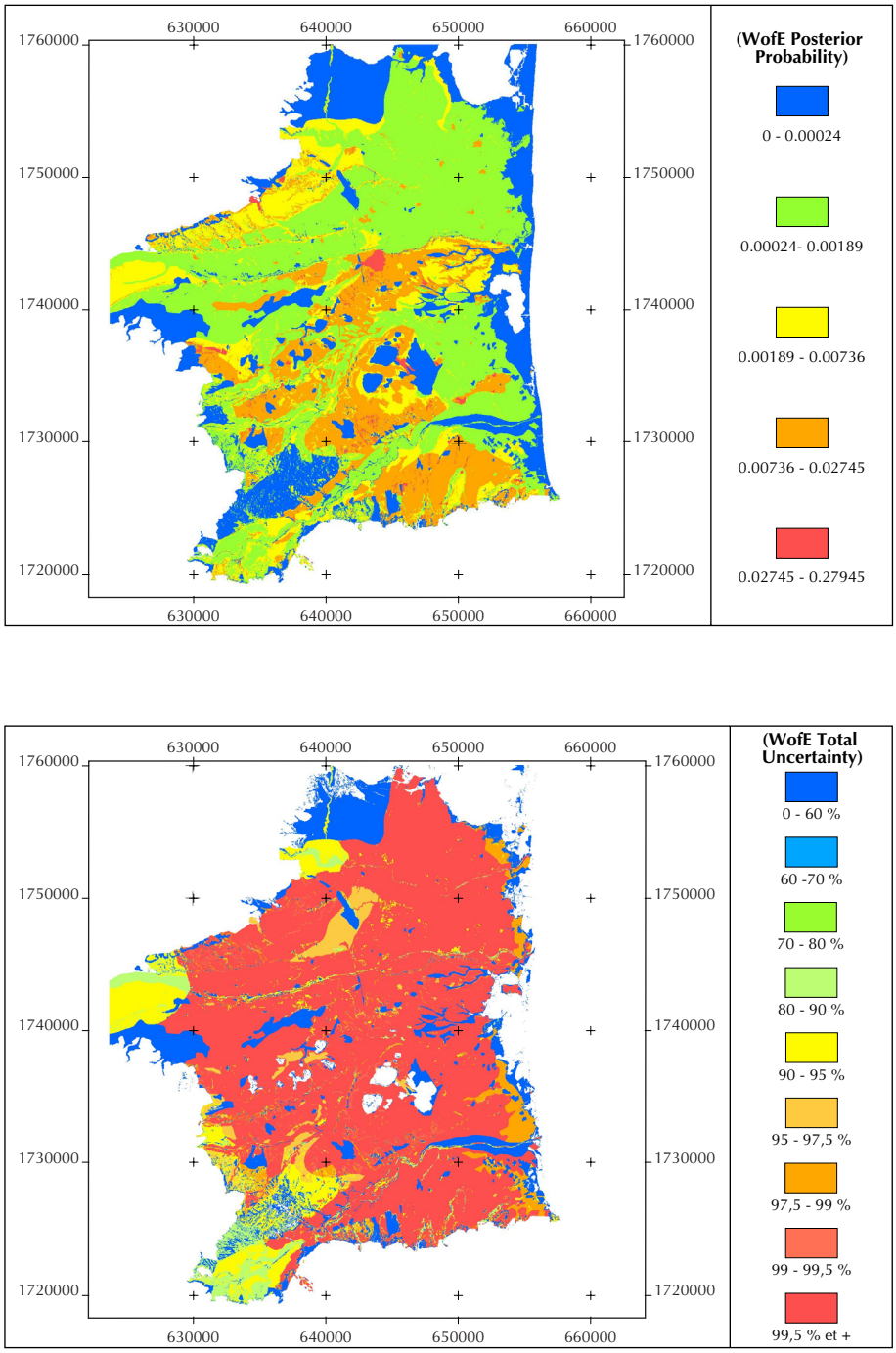


Fig. 1. *Cartographie prédictive – Modèle 1. En haut, carte de probabilité* a posteriori. *En bas, carte de confiance.*

La carte de confiance est plutôt satisfaisante. Seuls les secteurs où l'information occupation du sol est manquante dans le calcul voient leur confiance s'abaisser sous 0,90.

Cependant, le calcul de l'indépendance conditionnelle indique une forte redondance, susceptible de nuire à la fiabilité des résultats. Les couches les plus redondantes ont été éliminées du second modèle.

Modèle 2 : échantillon aléatoire et pédologie, exposition et occupation du sol (fig. 2).

Dans ce second modèle, les résultats obtenus montrent une probabilité *a posteriori* qui s'élève jusqu'à 0,38. La carte de confiance est de meilleure qualité, seules les zones où l'information est manquante étant inférieures à 0,90. Les secteurs à fort potentiel sont stables par rapport au modèle précédent et sont représentés par les zones urbaines, les collines pliocènes ainsi que les cônes alluviaux pléistocènes du piémont des Aspres.

L'indépendance conditionnelle entre les thèmes est de 0,93, ce qui est excellent.

Modèle 3 : échantillon aléatoire de sites gallo-romains et ensemble de la base environnementale (fig. 3)

L'échantillon est représenté par 101 sites gallo-romains, donnant une probabilité *a priori* de 0,0012. Le faible nombre de sites induit une forte indétermination pour beaucoup de classes qui ont été éliminées lors de l'étape de généralisation. La pédologie reste la variable la plus déterminante de la présence des sites. En particulier, les vieux sols fersiallitiques des terrasses et des cônes anciens semblent très attractifs. À l'opposé, les sols jeunes et hydromorphes de la plaine sont de bons déterminants en termes d'absence. Un comportement similaire peut être déduit des informations géologiques. L'occupation du sol actuelle se révèle peu probante. Cependant, les cultures permanentes et les vergers se révèlent de mauvais déterminants du facteur présence, ce qui est probablement lié à la difficulté de leur prospection au sol. En revanche, les secteurs urbanisés n'apparaissent plus comme de bons déterminants de la présence.

Les résultats obtenus montrent une probabilité *a posteriori* qui s'élève jusqu'à 0,15. Cette valeur est la conséquence du faible nombre de sites pris en compte. De larges plages ont une probabilité très faible mais également une confiance faible pour la même raison. Les probabilités les plus fortes sont obtenues dans le secteur nord de la plaine, ce qui est conforme aux connaissances archéologiques (Kotarba, comm. pers.). Enfin, il est intéressant de noter la disparition de la trame d'occupation urbaine actuelle, liée au caractère non probant de la classe urbaine du thème d'occupation du sol. Il confirme que la trame urbaine antique différait probablement de la trame postérieure, d'origine médiévale.

Les résultats obtenus dans ce travail sont encourageants. Ils montrent la robustesse de la méthode « Weights of Evidence » pour une application archéologique. L'amélioration significative de la prédiction entre sites réels et points fictifs aléatoires en constitue une preuve. Cependant, la validation d'un tel modèle ne peut se faire que par le terrain. C'est pourquoi, nous avons engagé

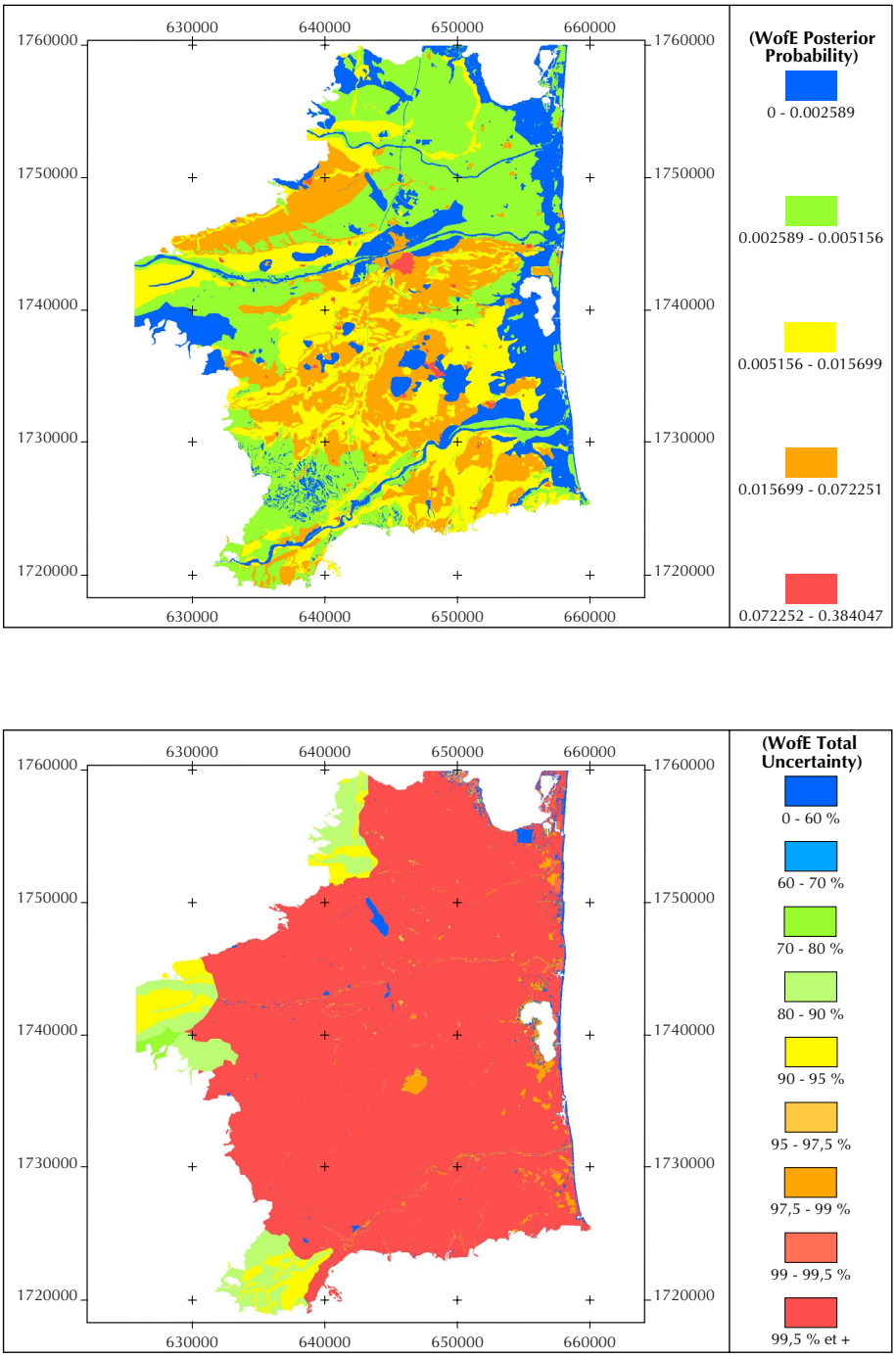
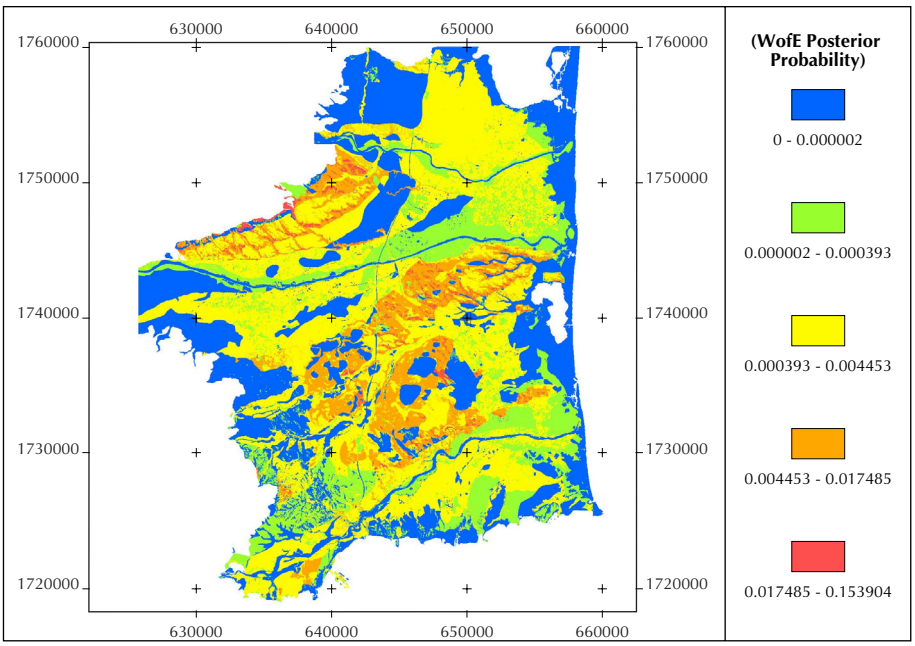


Fig. 2. *Cartographie prédictive – Modèle 2. En haut, carte de probabilité* a posteriori. *En bas, carte de confiance.*



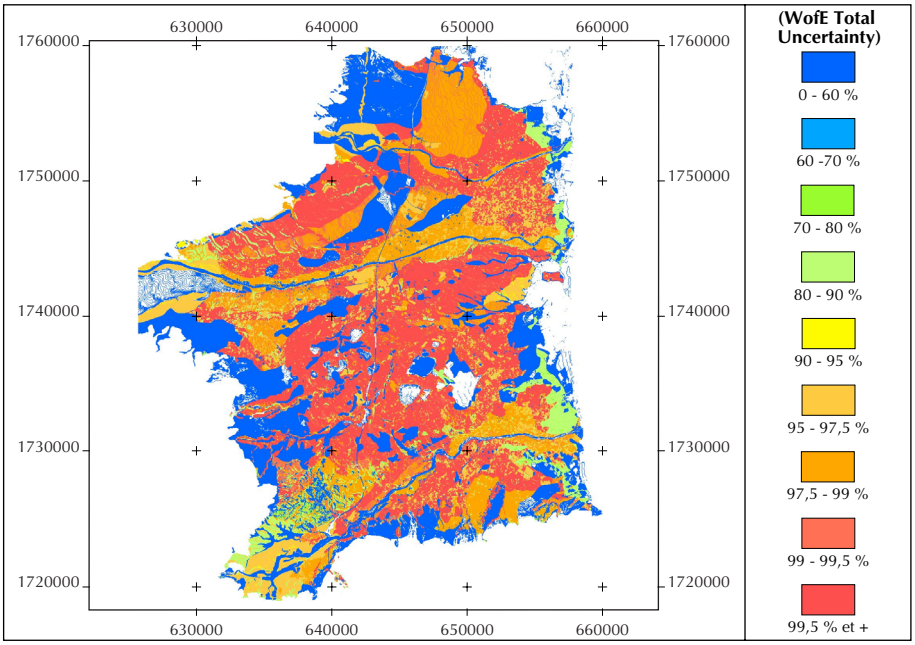


Fig. 3. *Cartographie prédictive – Modèle 3. En haut, carte de probabilité* a posteriori. *En bas, carte de confiance.*

une procédure de suivi sur le tracé de la future ligne LGV Barcelone-Perpignan. À l'issue du diagnostic, les nouvelles données seront comparées aux sorties du modèle prédictif afin de tester, précisément, sa prédictivité. Par ailleurs, les données recueillies, notamment géomorphologiques, seront réinjectées dans le modèle et serviront de base pour le calcul d'un nouveau modèle pour l'ensemble de la plaine.

Bibliographie

AGTENBERG F. P., BONHAM-CARTER G. F., WRIGHT D. F., 1989.– Weights of evidence modelling : a new approach to mapping mineral potential, *in* : F. P. Agtenberg, G. F. Bonham-Carter (eds), *Statistical applications in the Earth Sciences. Geological Survey of Canada*, Paper 89-9.

BONHAM-CARTER G. F., 1994.– *Geographic Information System for Geoscientists : modelling with GIS*, Pergamon Press, Taaytown, 398 p.

DAVTIAN G., 2003.– *La modélisation prédictive*, École thématique CNRS-ISA de Tours, 13-18 mars 2003.

DELLA BONA L., 1994.– *Archeological Predictive Modelling in Ontario's Forests, « vol. 3 : Methodological Considerations », A Report Prepared for the Ontario Ministry of Natural Resources*, Lakehead University, Center for Archaeological Resource Prediction, Thunder Bay, Ontario.

HANSEN D. T., 2000.– *Describing GIS Applications : Spatial Statistics and Weights of Evidence Extension to ArcView in the Analysis of the Distribution of Archaeology Sites in the Landscape*, Presented at the 2000 ESRI User Conference, San Diego.

HANSEN D. T., 2004.– *Describing GIS Applications : Spatial Statistics and Weights of Evidence Extension to ArcView in the Analysis of the Distribution of Archaeology Sites in the Landscape*, Technical Paper, GIScafe Paper 054.

KABERLINE M., SCHWEMM C., 2002.– *Weights of Evidence Analysis for Cultural Resource Site Prediction and Risk Assessment*, PWR GIS 2002, PMIS #79905

KHOLER T. A, PARKER S. C., 1986.– Predictive Models for Archeological Resource Location, *in* : *Advances in Archeological Model and Theory*, vol. 9, Academic Press, New York, p. 397-452.

KRIER V., 2004.– *La plaine alluviale de l'Oise. Milieu et système fluvial*, Service départemental d'archéologie, CDRom.

KVAMME K. L., KENNETH P., 1985.– Determining Empirical Relationships Between the Natural Environment and Prehistoric site Locations : A Hunter-Gatherer Example, *in* : C. Carr (ed.), *For Concordance in Archaeological Analysis*, Westport Publishers, University of Arkansas, p. 208-237.

SERVAT E., CALLOT G., 1966.– *Notice explicative de la carte des sols du Roussillon*, INRA, Service d'étude des sols, Montpellier, 68 p.

STANCIC Z., KVAMME, K. L., 1999.– Settlement patterns modelling throught Boolean overlays of social and environmental variables, *in* : J. A. Bercelo, I. Briz, A. Vila (eds), *New Techniques for Old Times*, CAA98, BAR International Serie 757, Oxford, p. 231-237.

VERHAGEN P., 1995.– De archeologische poteniekaart in Nederland : een methodologie voor het voorspellen van archeologische waarden op basis van archeologishe en landshappeliijke gegenvens, *Westerheem*, 44, 5, p. 177-187.

WARREN R. E. 1990.– Predictive Modelling of Archaeological Site Location : A Case Study in the Midwest, *in* : K. Allen, S. Green, E. Zubrow (eds), *Interpreting Space : GIS and Archaeology*, Taylor and Francis, London, p. 201-215.

WARREN R. E., ASCH D. L., 2000.– A predictive model of archeological site location in the Eastern Prairie Peninsula, *in* : K. L. Westcott, R. J. Brandon (eds), *Practical Applications of GIS for Archaeologists : A Predictive Modeling Kit*, Taylor and Francis, London and New York, p. 5-25.

WHEATLEY D., GILLINGS M., 2002.– *Spatial Technology and Archeology*, The archeological applications of GIS, Taylor and Francis, London, 270 p.