Contributions de la Télédétection et des Systèmes D'information Géographiques pour l'étude de la transmission de la dengue.

Remote-sensing and GIS contributions for the dengue transmission study

Laurent Girdary[1] Enguerran Grandchamp[2]

[1,2]Université des Antilles et de la Guyane, Guadeloupe, France
[1]laurentgirdary@hotmail.fr

[2]LAMIA Laboratory, French West Indies University,
Campus de Fouillole, 97157 Pointe-à-Pitre, Guadeloupe, France
egrandch@univ-ag.fr

Résumé L'objectif de cette étude est de définir les contextes d'habitation en Guadeloupe continentale afin d'analyser les relations avec la transmission de la dengue. Les informations collectées incluent les données environnementales, les caractéristiques physiques des bâtis et les variables démographiques. Une classification dite *raster* en deux étapes a été utilisée afin de classifier les contextes et analyser l'évolution de leur distribution spatiale en 1996, 2000 et 2004. L'approche *raster* consiste à découper le territoire en cellules régulières. Pour cette première étape d'analyse, une classification non supervisée (clustering) a été utilisée afin d'avoir des informations précises pour l'identification et la définition des principaux contextes d'habitation. Pour cela, une classification hiérarchique et une analyse Bayésienne discriminante sont utilisées. Concernant la seconde étape, une classification supervisée est effectuée afin de stabiliser les contextes d'habitation obtenus lors de la première étape. Il est appliqué une procédure d'apprentissage basée sur un échantillonnage des mailles stables obtenues lors de la classification non supervisée. Pour les deux étapes de la classification, huit classes correspondant aux contextes d'habitation sont identifiées et définies. Ces étapes ont permis la classification de l'ensemble du territoire habité et ont montré une répartition spatiale cohérente dans le temps, en tenant compte de l'évolution de l'urbanisation entre 1996 et 2004. C'est la première description de l'application d'une méthode standardisée tenant compte des facteurs environnementaux et démographiques pour la définition des contextes d'habitation. Même si la méthode *raster* présente quelques limites, elle apporte une classification indiscutable des contextes d'habitation à une petite échelle, ce qui permet une analyse de leur relation avec la transmission de la dengue.

Mots clés Classification, apprentissage, contexte d'habitation, maladie vectorielle, dengue

1. Introduction

La dengue est la plus importante maladie virale transmise par les moustiques, affectant les humains à travers le monde entier. Cette maladie est particulièrement présente dans les régions tropicales et subtropicales (Mackenzie 2004) où des dizaines de millions de cas apparaissent chaque année. Le principal vecteur du virus de la dengue est le moustique *Aedes (Stegomyia) aegypti* (Guedes, 2010). Ce moustique est un vecteur hautement efficace, qui se nourrit préférentiellement sur les humains (Hopp, 2001) et qui est abondant dans l'environnement domestique et péridomestique. Les conditions favorables pour le développement du moustique et la transmission de la dengue sont liées à des variables socio-démographiques et physiques. Par exemple, les habitations localisées dans des régions concentrées avec une forte densité de population sont plus exposées aux piqures de moustiques et aux infections par le virus de la dengue.

L'archipel de la Guadeloupe est exposé régulièrement aux épidémies de dengue. Entre 2000 et 2008, trois épidémies ont été relevées dans le département (2001, 2005 et 2007)[1]. Le nombre de cas reportés a progressivement augmenté avec certaines régions de l'île plus affectées. Dans le même temps, l'environnement a évolué en Guadeloupe (nouvelles constructions, développement urbain, etc.). A partir de ces observations, beaucoup de questions peuvent être posées: est ce que ces deux phénomènes sont liés, la distribution spatiale des bâtis contribue-t-elle à l'hétérogénéité de la distribution spatiale des cas dans l'île?

L'objectif de l'étude est de déterminer les contextes d'habitation en Guadeloupe continentale et d'analyser l'évolution de leur distribution spatiale entre 1996 et 2004. Pour cela, deux étapes de classification *raster* seront effectuées. La première étape consiste en une classification non supervisée afin d'identifier les principaux contextes et de définir leurs caractéristiques (Torres Moreno, 2009). La seconde étape de la méthode correspond à une classification supervisée (Kotsiantis, 2007) qui permettra de stabiliser et d'affiner les premières tendances obtenues lors de la première classification.

2. Région d'étude et recueil de données

2.1 Région d'étude

La Guadeloupe est un des trois départements français d'Amérique avec la Martinique et la Guyane. C'est un archipel de l'hémisphère nord situé dans la région caraïbe, localisé entre les latitudes 15°59' et 16°40' nord et longitudes 61°10' et 61°50' ouest. Avec une surface de 1785 km^2, il inclut la Guadeloupe continentale (composées de deux îles adjacentes: Grande Terre et Basse Terre) et trois petites îles: La Désirade, Les Saintes et Marie-Galante[2].

2.2 Recueil de données

La Guadeloupe est composée de 26 communes divisées en plusieurs sections, la commune possède une plus grande surface que la section. L'île est découpée par 129 IRIS qui correspondent à un regroupement de plusieurs sections d'une même commune. Plusieurs variables socio-démographiques et environnementales sont déterminées en fonction des constructions (pourcentage de maisons traditionnelles, en bois ou en dur), de leur confort (pourcentage de maisons sans toilettes et douches, avec climatisation), de leur taille (nombre moyen de pièces). D'autres variables sont prises en compte comme le type de propriétés (principale, secondaire ou occasionnelle) et le nombre de voitures par foyer (zéro, deux et plus). Les variables physiques sont également définies telles que la proportion d'aires agricoles, naturelles et urbaines. Au final, nous avons un total de 19 variables (13 socio-démographiques and 6 physiques). Le choix a été fait de diviser l'ensemble du territoire Guadeloupéen en cellules régulières (carrées de 250×250 mètres) (Jianquan, 2002) afin de pouvoir calculer les différentes variables à l'échelle de ces carrées et par la suite de pouvoir déterminer les contextes d'habitation.

3. Détermination des contextes d'habitation

Il y a deux principales étapes:
-la classification non supervisée (Guis, 2007) afin de définir les principaux contextes d'habitation sans aucune donnée a priori en entrée ;
- la classification supervisée qui a pour but de stabiliser et d'affiner plus précisément chaque contexte définis auparavant.

3.1 Classification non supervisée

Elle comprend tout d'abord une Analyse en Composante Principale (ACP) (Jolliffe, 2002) qui permet de réduire la dimension des données grâce à la corrélation entre certaines variables en entrée pour aboutir à un nombre moindre de variables au final. Ensuite, une classification hiérarchique

[1] http://www.invs.sante.fr/surveillance/dengue/peh_guadeloupe.html
[2] http://www.crguadeloupe.fr/archipel/?ARB_N_ID=731&ARB_N_S=732

ascendante (Husson, 2010) est effectuée ; un dendrogramme (Kaufman, 1990) est obtenu, correspondant à un arbre de décision qui illustre les relations entre les différents contextes. Cette classification hiérarchique permet ainsi de regrouper ensemble les contextes selon leurs similarités. Une analyse Bayésienne discriminante (Srivastava, 2007) permet de construire un modèle de classification à partir des résultats du dendrogramme avec l'utilisation de probabilités postérieures afin de valider les différentes classes correspondant aux différents contextes d'habitation. Après classification du territoire en 1996, 2000 et 2004, les classes sont identifiées sémantiquement selon les valeurs maximales et minimales des différentes variables en entrée. La répartition spatiale des contextes d'habitation est également utilisée pour valider leur identification grâce à des experts en urbanisation.

3.2 Classification supervisée

Afin de définir plus précisément chaque contexte d'habitation, une classification supervisée a été effectuée (Kotsiantis, 2007). Les classes obtenues lors de la première classification sont utilisées pour la phase d'apprentissage (Vapnik, 2000) de l'algorithme. Cette procédure est appliquée en sélectionnant les cellules représentatives et les plus stables issues de la classification non supervisée entre 1996 et 2004 afin de pouvoir classifier les autres cellules pour les trois années d'étude. Deux arbres de décision ont été utilisés avec les algorithmes *Functional Tree* (FT) (Gama 2004 et Landwehr 2005) et *C4.5* (Quinlan 1993). Au final, il y a 6 341 cellules stables sur l'ensemble du territoire. Un échantillon de 458 cellules a été choisi. Le nombre et la distribution de ces cellules dans les classes dépendent de la taille de la classe et sont calculés selon la formule suivante (Cohran, 1977) :

$$\text{Nombre de cellules échantillonnées par classe} = N^2 u^2_{(1-\alpha/2)} / u^2_{(1-\alpha/2)} + 4(N-1)\, d^2$$

où N est le nombre de cellules représentatives dans une même classe, $u(1-\alpha/2)$ est la valeur quantile d'une distribution normale (= 1,96) et d une valeur fixe de pourcentage d'erreur (2, 3, 4 et 10%).

4. Résultats

Un total de 12122, 12607 et 13564 cellules a été analysé respectivement pour les années 1996, 2000 et 2004. L'augmentation du nombre de cellules est le premier résultat observable, correspondant à la construction de nouveaux bâtiments durant la période étudiée.

4.1 Classification non supervisée

Huit contextes d'habitation ont pu être identifiés entre 1996 et 2004, correspondant aux contextes urbains, périphériques, agricoles, naturelles, touristiques, résidentiels, ruraux et intermédiaires. Dans un premier temps, les centres urbains et les zones périphériques sont clairement identifiés. Deux principaux centres urbains sont observés correspondant à la plus petite classe. Ils se situent dans les villes de Pointe-à-Pitre et Basse-Terre où il existe les concentrations de population les plus importantes. Quant aux zones périphériques, elles se situent à proximité immédiate de ces centres urbains et concernent principalement les villes de Gosier, Abymes et Baie-Mahault ou encore Saint-Claude. Le contexte agricole est caractérisé par une part importante de terres agricoles à proximité des habitations. La classe naturelle correspond aux maisons isolées entourées par une végétation importante. Les zones touristiques représentent principalement les hôtels et gîtes où logent majoritairement la population non autochtone. Les zones résidentielles sont caractérisées par la prédominance de maisons pavillonnaires dans des quartiers résidentielles. Les régions intermédiaires sont des zones tampons intercalées entre deux (ou plus) autres classes.
Un certain nombre de contextes (touristique, naturel, agricole et rural en particulier) ont une évolution irrégulière (le nombre de cellules augmente entre 1996 et 2000 puis régresse entre 2000 et 2004 ou inversement). Ces résultats seraient liés à l'aspect non supervisé de la méthode de classification.

La Figure 1 illustre la distribution spatiale des huit contextes d'habitation issus de la classification non supervisée. Un zoom a été fait sur le sud de la Grande-Terre et la région de Pointe-à-Pitre afin de montrer l'évolution de chaque contexte en 1996, 2000 et 2004.

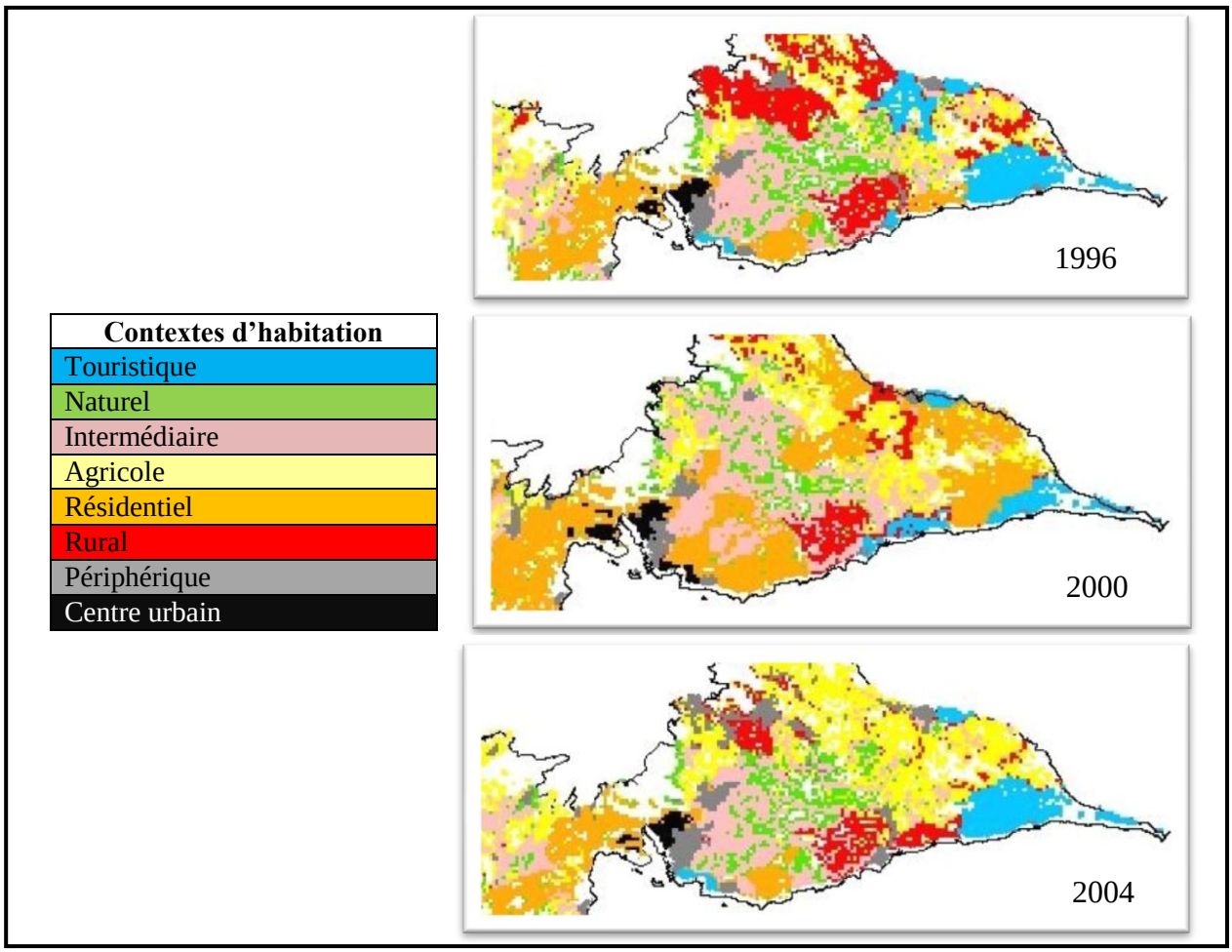

Figure 1 – Distribution spatiale des contextes d'habitation par la méthode non supervisée

4.2 Classification supervisée

Deux algorithmes *C4.5* et *FT* ont été testés afin de stabiliser et d'affiner les premières tendances de la répartition spatiale des différents contextes issus de la classification non supervisée. De ce point de vue, l'algorithme *FT* donne de meilleurs résultats que l'algorithme *C4.5* avec une évolution plus stable des contextes durant l'ensemble de la période étudiée. En effet, il y a une évolution irrégulière des contextes intermédiaires, résidentiels, ruraux et périphériques avec l'algorithme *C4.5*. Lorsqu'on observe la distribution spatiale des contextes (Figure 2), les deux approches montrent une évolution globale cohérente même si l'algorithme *FT* confirme une meilleure stabilité des contextes.

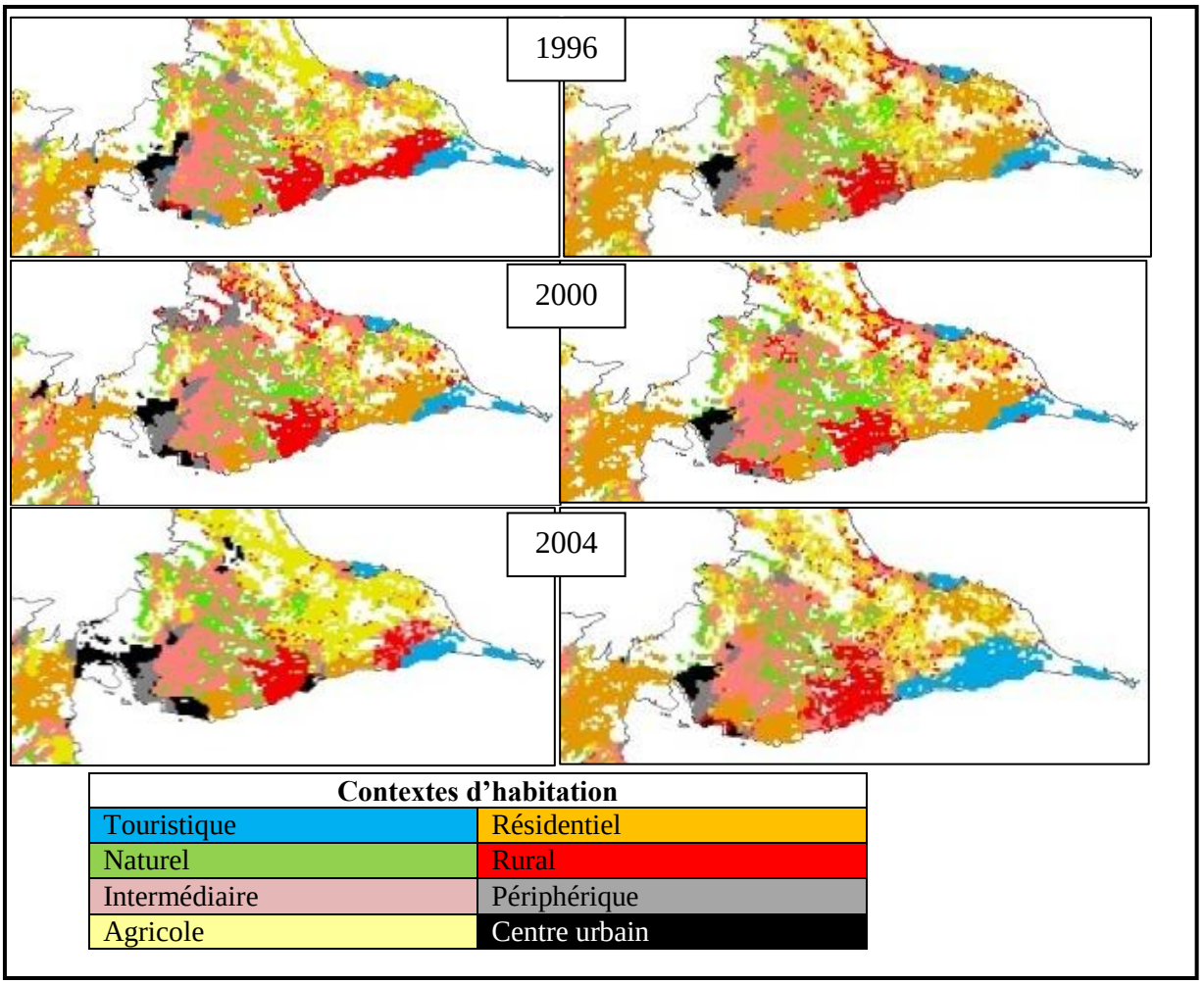

Figure 2 – Répartition spatiale de la classification supervisée selon les algorithmes *C4.5* à gauche et *FT* à droite

5. Discussion

Toutes les cellules couvrant les régions habitées de la Guadeloupe continentale ont été classifiées par la méthode non supervisée dans un premier temps. Les contextes urbains, périphériques et résidentiels sont reconnus sans ambigüité et la précision de leurs limites géographiques est confirmée par les cartes administratives et les experts locaux dans l'aménagement du territoire et les architectes. En plus de ces trois contextes, cinq autres contextes d'habitation sont définis: touristiques, naturels, intermédiaires, agricoles et ruraux. La forte corrélation entre les variables démographiques et les caractéristiques physiques des bâtis ont rendu aisé l'identification de ces contextes. Néanmoins, la classification de certaines cellules était instable avec des changements de classification entre 1996 et 2004.

La deuxième étape de l'analyse, la classification supervisée est plus efficace dans l'amélioration de la stabilité des cellules durant l'ensemble de la période d'étude. Les meilleurs résultats sont obtenus avec l'utilisation de l'algorithme *FT* après une phase d'apprentissage sur un échantillonnage de cellules stables de chaque contexte. Cette procédure a permis d'obtenir une classification stable d'une majorité de cellules et une évolution cohérente des contextes d'habitation entre 1996 et 2004, notamment l'extension de la zone touristique et résidentielle. Le contexte touristique peut jouer un rôle important dans la transmission de la maladie. En effet, les populations vivant dans les hôtels et gîtes n'ont peu ou jamais été en contact immunologiquement avec les différents sérotypes du virus dans la région Caraïbe, elles ne sont donc pas protégées de l'infection par le virus de la dengue. Les zones résidentielles offrent un environnement favorable avec de la végétation et des réserves d'eau près des

bâtis (gouttières, regards, etc.) qui contribue à l'augmentation de la population de moustiques. Au final, ces deux contextes présentent des conditions favorables pour le développement d'*Aedes aegypti*, et une part importante des populations qui y vivent susceptibles d'être infectées. Les cartes obtenues durant cette procédure pourront être utiles dans l'étude de la corrélation entre les contextes d'habitations et la transmission de la dengue. Elles pourront apporter aux autorités de santé publique des informations précises afin d'aider ces derniers dans la planification de leur intervention et identifier les mesures de contrôle du vecteur le plus efficace (Kolivras, 2006). Plus généralement, l'augmentation du nombre de cellules entre 1996 et 2004 soit 1 442 nouvelles cellules (12%) traduit une pression de l'urbanisation sur les espaces naturels et agricoles (Thinon, 2007).

Même si les résultats préliminaires sont très prometteurs, quelques possibilités existent afin d'améliorer la classification et la détermination des contextes. En effet, la rastérisation de l'information, basée sur la décomposition systématique et arbitraire de l'espace en carrées de dimensions réguliers pourrait être complétée et affinée par une représentation vecteur des données. Les données *raster* ne sont pas toujours adaptées au traitement pour les systèmes d'information géographique (Benz, 2004; Lewinski, 2004), induisant des effets de bord pour les carrées localisées aux limites administrative (IRIS par exemple). De plus, des variables définies à une échelle donnée ne sont fréquemment pas interprétables à une autre échelle (Liebhold, Rossi and Kemp, 1993). En utilisant une représentation vecteur avec un espace découpé selon l'occupation du terrain et des limites administratives, nous pourrions utiliser plus efficacement les informations inclues dans les variables socio-démographiques et physiques.

6. Conclusion

La dengue est un problème majeur de santé publique. Les contextes d'habitation peuvent avoir un impact sur la transmission de la dengue à travers différents facteurs (population, environnement, conditions sociales, etc.). Afin de réduire la propagation du virus, il est nécessaire d'analyser la relation entre l'habitation et la transmission de la dengue.

Même si les huit contextes d'habitation choisis sont bien détectés par les deux types de classification (supervisée et non supervisée), plusieurs différences existent entre les deux méthodes. La contribution de la classification supervisée est meilleure et plus particulièrement l'algorithme *FT*. Celui-ci conduit ainsi à une classification et une évolution stables des contextes entre 1996 et 2004, avec une extension des zones touristiques et résidentielles entre autres. L'évolution des espaces urbains, naturels et agricoles dans le temps pourrait également avoir un impact important dans les populations de moustique et dans la variation du niveau de transmission de la dengue (Vanwambeke, 2007). L'identification de hauts risques de transmission de la dengue dans certains contextes pourrait aider les autorités à cibler leurs actions et augmenter le taux d'avantages/coûts des mesures de contrôle.

Références

BENZ U. and Al., 2004, Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information, Journal of Photogrammetry and Remote Sensing, Vol. 58, Issue: 3-4, pp 239– 258

COHRAN W., 1977, Sampling techniques. Wiley, Harvard University, paperback: 428 pages, ISBN: 978-0-471- 16240-7.

CROSS V.V., 2001, Fuzzy extensions for relationships in a generalized object model, International Journal on Intelligent Systems, Vol.16, pp. 843–861.

GAMA J., 2004, Functional Trees, Machine learning, Vol. 55, pp.219-250.

GUEDES D.R.D., CORDEIRO M.T., MELO-SANTOS M.A.V., MAGALHAES T., MARQUES E., REGIS L., FURTADO A.F. and AYRES C.F.J., 2010, Patient-based dengue virus surveillance in *Aedes aegypti* from Recife, Brazil, Journal Vector Borne Disease: Vol. 47 pp. 67–75.

GUIS H., 2007, Géomatique et épidémiologie: Caractérisation des paysages favorables à Culicoides imicola, vecteur de la fièvre cattarhale ovine en Corse, thesis (PhD)- Franche Comté University.

HUSSON F., JOSSE J. and Pages J.,2010, Principal component methods - hierarchical clustering-partitioning clustering: why would we need to choose for visualizing data, Technical Report – Agrocampus Applied Mathematics Department. (http://www.agrocampus-ouest.fr/math/).

JOLLIFFE I. T., 2002, Principal component analysis Springer; 2nd edition paperback: 502 pages ISBN-13: 978-0387954424.

KAUFMAN L. and ROUSSEEUW P.J., 1990, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

KOLIVRAS K. N., 2006, Mosquito Habitat and Dengue Risk Potential in Hawaii: A Conceptual Framework and GIS Application, The Professional Geographer, Vol.58, No. 2, Issue 2, pp 139–154.

LANDWEHR N., HALL M., EIBE F., 2005, Logistic Model Trees. Machine learning, Kluwer, Vol. 59, Issue 1-2, pp. 161-205.

LEWINSKI S., ZAREMSKI K., 2004, Examples of Object Oriented Classification Performed On High Resolution Satellite Images. Miscellanea Geographica. Vol. 11.

HOPP M. J. and FOLEY J. A., 2001, Global-scale relationships between climate and the dengue fever vector, *Aedes aegypti*, Climatic Change: Vol. 48, pp. 441–463.

LIEBHOLD A.M., ROSSI R.E. and KEMP W.P., 1993, Geostatistics and geographic information systems in applied insect ecology, annual review of entomology, Vol. 38, pp. 303-327.

MACKENZIE J.S., GUBLER D.J. and PETERSEN L.R., 2004, Emerging flaviviruses: the spread and resurgence of Japanese encephalitis, West Nile and dengue viruses, Nature medicine Vol. 10, N° 12 pp. 98-109.

QUINLAN R., 1993, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, paperback: 302 Pages.

SRIVASTAVE S., GUPTA M. R. and FRIGYIK B.A., 2007, Bayesian Quadratic Discriminant Analysis. Journal of Machine Learning Research, pp. 1277-1305.

TACOLI C., 1998, Rural-urban interactions: a guide to the literature, Environment and Urbanization, Vol. 10, N. 1, pp.147-166.

THINON P., MARTIGNAC C., METZGER P. and CHEYLAN J.-P., 2007, Analyse géographique et modélisation des dynamiques d'urbanisation à La Réunion, Cybergeo : Revue européenne de géographie, N°389 (http://www.cybergeo.eu/index8692.html).

TORRES-MORENO J.M., BOUGRAIN L. and ALEXANDRE F., 2009, Combining Supervised and Unsupervised Learning for GIS Classification, arXiv: 09052347.

VANWAMBEKE S.O., LAMBIN E.F., EICHHORN M.P. and Al., 2007, Impact of land-use change on dengue and malaria in Northern Thailand. EcoHealth, Vol. 4, No. 1, pp. 37–51.

VANWAMBEKE S.O., BENNETT S.N. and DURRELL K.D., 2011, Spatially disaggregated disease transmission risk: land cover land use and risk of dengue transmission on the island of Oahu, Tropical Medicine and International Health Vol. 16, Issue 2, pp 174–185.

VAPNIK V. N., 2000, The Nature of Statistical Learning Theory (2nd Edition), Springer-Verlag.