Nations Unies ECE/CES/2010/12

Conseil économique et social

Distr. générale
30 mars 2010
Français
Original: anglais

Commission économique pour l'Europe

Conférence des statisticiens européens

Cinquante-huitième réunion plénière
Paris, 8-10 juin 2010
Point 6 de l'ordre du jour provisoire
Statistiques spatiales

Associer variables spatiales et mailles pour améliorer la visualisation des données

Note du Census Bureau des États-Unis

Résumé

La présente note donne un aperçu de l'utilisation de la géographie juridique, statistique et administrative par le Census Bureau des États-Unis. Elle analyse les ressemblances et les différences entre les données spatiales et les statistiques présentées en mailles, et expose les avantages et inconvénients de leur utilisation. Dans certains cas, données spatiales et mailles statistiques sont associées pour aboutir à une solution intégrée. Deux études de cas servent à illustrer différentes manières d'utiliser les statistiques spatiales: l'analyse de la population haïtienne et les statistiques agricoles des États-Unis.

I. Introduction

1. La possibilité d'associer plus facilement données spatiales et statistiques à des niveaux moins élevés de la géographie a suscité ces dernières années un intérêt croissant pour l'utilisation des statistiques spatiales. Les statistiques par zone sont généralement exprimées sous forme de polygones représentés par des unités administratives ou un maillage géométrique. La forme à choisir dépend du but de la présentation des données, de l'analyse, des caractéristiques des données statistiques et spatiales ainsi que de la restitution graphique que présentera l'utilisateur. L'emploi conjugué de la maille pour le volet statistique et de polygones pour le volet administratif présente des avantages et des inconvénients. Dans certains cas, leur association débouchera sur une solution intégrée.

2. La nature des éléments de données spatiales et des caractéristiques qui sous-tendent la géographie du recensement influe sur leur utilisation. Les statistiques spatiales portent surtout sur les configurations et les groupes d'activité. La présente note donne un aperçu de l'utilisation de la géographie juridique, statistique et administrative par le Census Bureau des États-Unis en établissant des comparaisons avec les caractéristiques du maillage géométrique. Elle expose également deux études de cas qui correspondent à une utilisation différente des statistiques spatiales. Elle soulève un certain nombre de questions pour encourager la poursuite du débat et l'approfondissement de ce domaine d'étude.

II. Données spatiales

3. Les données spatiales peuvent se répartir en trois groupes: les données géostatistiques, les données à configuration ponctuelle et les données en réseau (Cressie, 1993). Les données géostatistiques sont des données recueillies sur un domaine spatial continu par référence à la Terre. Elles sont caractérisées par des observations associées à une variation continue dans l'espace, généralement en fonction de la distance (Anselin, 1992). Les échantillons de sol recueillis à travers toute une région en seraient un exemple.

4. Lorsque l'intérêt se porte sur les lieux où se produisent les événements, les données référencées sont désignées sous le nom de données à configuration ponctuelle. Ces données se concentrent sur l'emplacement de chacun des points et tout particulièrement sur la configuration spatiale créée (Cressie, 1993). Les occurrences de données spatiales à configuration ponctuelle sont espacées irrégulièrement. Avec ce type de données spatiales, il n'est pas possible de prédire avec confiance l'emplacement de l'occurrence. L'emplacement des unités d'habitation peut être un exemple de données à configuration ponctuelle.

5. Les données en réseau sont recueillies sur un réseau régulier ou irrégulier assorti d'une certaine structure de voisinage déterminante (Cressie, 1993). La région où les points sont réunis comprend un nombre fini de sites. L'étendue de l'espace est limitée. Des valeurs sont conférées à des points et les emplacements de ces points sont connus. Le nombre de personnes à l'intérieur de chaque comté d'un État en serait un exemple.

III. Statistiques spatiales

6. Les statistiques spatiales, également connues sous le nom de géostatistiques, sont une forme de statistiques qui analysent des ensembles de données spatiotemporelles. Elles se distinguent des autres formes de statistiques en ce qu'elles se rapportent à l'emplacement des valeurs de données. Toutes les données ont des attributs spatiaux et temporels. La

proximité de ces données est souvent un indicateur de leur similarité. Comme Waldo Tobler l'indiquait dans sa Première loi de la géographie, «tout interagit avec tout, mais deux objets proches ont plus de chances de le faire que deux objets éloignés» (1970). Il y a donc une plus grande probabilité que les données soient similaires lorsqu'elles sont proches les unes des autres, dans l'espace ou dans le temps, que lorsqu'elles sont plus éloignées. Les statistiques spatiales font appel à diverses techniques pour l'étude des données et de leurs attributs topologiques, géométriques et géographiques. Ces statistiques ont pour but de déterminer la quantité de variations spatiales entre des données ponctuelles qui varient dans l'espace et/ou dans le temps. Elles peuvent servir à décrire les caractéristiques spatiales d'un ensemble de données et à établir des valeurs interpolées à partir d'un ensemble déterminé de données dans le cas de zones pour lesquelles les informations sont rares, voire inexistantes. Dans les statistiques spatiales chaque emplacement s'assortit d'une configuration spatiale, que ce soit l'environnement, le climat, la pollution, l'urbanisation ou la santé de l'être humain.

7. L'homme a toujours essayé de trouver des configurations dans le monde qui l'entoure. Il est donc possible de faire remonter l'analyse spatiale au premier temps de la géographie, de la cartographie et de l'arpentage. Toutefois, l'étude en bonne et due forme des statistiques spatiales n'est apparue que vers la fin du XX^{e} siècle. De nos jours, l'analyse spatiale repose sur l'informatique en raison des énormes quantités de données géographiques, de la complexité des programmes d'analyse statistique et géographique et de la sophistication de la modélisation spatiale. L'abondance de données résulte des nouvelles technologies. Il est possible à l'heure actuelle de recueillir des données fournies par des images de télédétection, des systèmes de transport intelligents et des dispositifs mobiles équipés de systèmes de positionnement mondial (GPS) qui peuvent localiser un emplacement pratiquement en temps réel. Grâce à la pléthore de systèmes d'information géographique (SIG), la gestion d'énormes stocks de données devient pratique courante. De ce fait, l'analyse spatiale est devenue un instrument à la portée d'un large public, ce qui a permis à un grand nombre de personnes de devenir des analystes ainsi que de calculer et d'analyser les relations entre les données et les configurations qui s'en dégagent.

8. La surabondance de données disponibles permet de relever de nouveaux défis en matière de stockage, de représentation, de récupération de transmission et surtout de synthétisation des données. L'étude des statistiques spatiales ne peut se faire sans méthode automatisée de synthétisation, de classification et de prévision ou de modélisation. De nombreuses disciplines utilisent les statistiques spatiales mais le point commun, c'est la configuration des données qui est omniprésente. Les données recueillies dans l'espace et dans le temps sont souvent liées les unes aux autres du fait de leur interaction et apparaissent dans les configurations spatiales, ce qui guide l'étude des statistiques spatiales. L'objectif général de ces statistiques est de faire ressortir et permettre d'étudier ces interactions et les configurations qui en découlent, de les classer puis de modéliser les interactions et les configurations en prévision de futures données.

9. Les statistiques spatiales ont pour but d'apporter des réponses à quatre questions fondamentales:

a) Comment les données sont-elles distribuées?

b) Quelle est la configuration créée par les données?

c) Où se trouvent les groupes?

d) Quelles sont les relations entre les ensembles de données ou les valeurs?

IV. Mailles statistiques

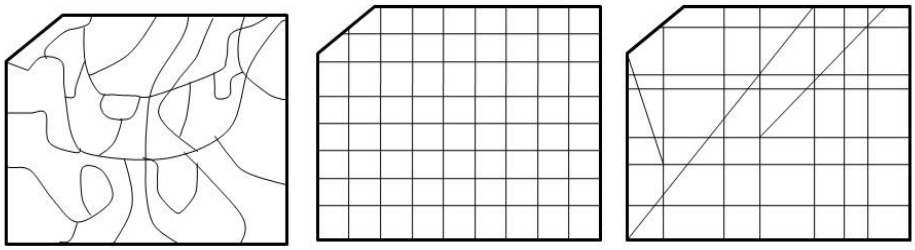
10. Les mailles statistiques sont des rectangles contenant des données spatiales; ces rectangles sont généralement de même dimension et ont une taille correspondant à un usage déterminé. On crée une surface maillée en commençant par établir des espacements réguliers à partir de l'origine le long des axes X et Y (horizontal et vertical). Les mailles donnent un éclairage relationnel d'une cellule par rapport à l'autre sur toute la surface maillée. En raison de leur configuration géométrique, il est possible de modifier l'échelle des mailles pour obtenir une résolution plus ou moins fine des données. Une maille est un paramètre fictif, un espace pour stocker les occurrences de données. L'espace de la maille n'a en lui-même aucune définition ni signification.

11. Les données statistiques sont reportées dans la cellule de la maille. Les données ponctuelles sont dispersées et forment une configuration régulière ou une configuration aléatoire irrégulière. L'un des objectifs est de faire en sorte que plusieurs données ponctuelles n'occupent pas le même emplacement. Pour qu'il existe une véritable association, les données classées sont reportées sur toute la surface de la cellule de la maille et reliées aux cellules environnantes, soit de la même classe soit de classes différentes. Les similitudes font apparaître des groupes et des configurations tandis que les différences dénotent des événements uniques ou atypiques.

12. En superposant une maille sur un réseau spatial irrégulier (des configurations de mobilité, par exemple), l'utilisateur peut voir les relations entre les deux références géographiques. De manière générale, on obtient de meilleures comparaisons des données en utilisant des mailles relativement grandes et des caractéristiques spatiales à relativement petite échelle (plus grande zone). Des caractéristiques résultant d'activités humaines, des réseaux de transport (routes et voies ferrées) par exemple, ont des formes irrégulières et amorphes. Des caractéristiques propres à la nature telles que des chaînes de montagne et des rivières ont en commun des propriétés analogues liées à une orientation irrégulière.

13. Les limites des circonscriptions administratives sont souvent irrégulières. Dans certains cas, des circonscriptions administratives rectangulaires, par exemple celles établies par le Système par lignes et rangées (township and range system) aux États-Unis, peuvent coïncider avec une maille statistique (fig. 1). La probabilité d'une coïncidence dépend de nombreux facteurs tels que l'origine, la taille et la raison d'être de la maille.

Figure 1
Comparaison de diverses aires géographiques



14. Les mailles ouvrent une fenêtre mobile et de taille variable sur les données. Elles constituent également un mécanisme permettant d'intégrer des données en provenance d'autres sources. L'intégration des données pose actuellement un problème complexe lorsque les données spatiales sont irrégulières. Les mailles présentent des propriétés bien définies mais n'expriment pas le caractère irrégulier des phénomènes du monde réel comme le font les données spatiales.

V. Similitudes et différences entre les données spatiales et les mailles statistiques

15. Les données spatiales expriment des phénomènes du monde réel. Certaines données concernent la nature et sont presque toujours imprévisibles. Les données résultant d'activités humaines sont généralement imprévisibles; toutefois certaines caractéristiques découlent de configurations qui suivent des plans, des spécifications par exemple. Des arbres plantés dans un verger, un espace libre entre des directions opposées sur une route à accès limité et les bornes kilométriques situées à intervalles réguliers le long d'une route en sont des exemples.

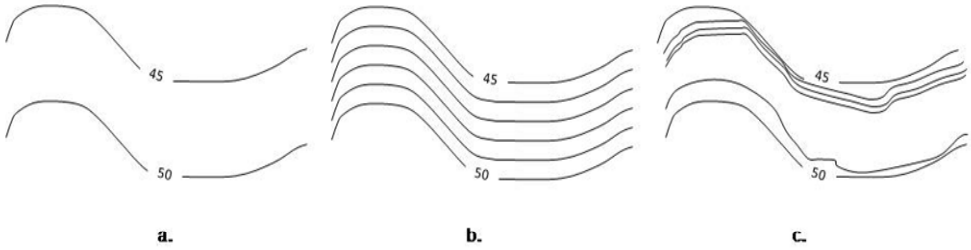
16. Les données spatiales varient dans l'espace et dans le temps. D'après Griffith et Paelinck, les données correspondant au monde réel sont bruitées, imprécises et imparfaites (2007). Les données sont bruitées parce qu'incertaines (ou imprévisibles). La situation au sol pourrait révéler un paysage relativement aride dans lequel apparaîtrait inopinément une étendue d'eau alimentée par des sources naturelles.

17. Des données imprécises comportent différentes incohérences qui peuvent consister en données incomplètes ou en éléments atypiques et anomalies. Il n'est pas possible, s'agissant d'ensembles de données au niveau national, de maintenir des données spatiales complètes et exactes en temps réel étant donné que des faits se produisent et/ou que la situation évolue. Très peu de données spatiales sont rassemblées en temps réel, même avec des capteurs fonctionnant en permanence. Certaines données sont incomplètes en raison de facteurs tels que les limites des sources, la qualité des données et les ressources disponibles. Les éléments atypiques sont les extrêmes. Par exemple, des données de température à l'une ou l'autre des extrémités de l'échelle de température qui ont été recueillies lors de situations microclimatiques locales et inhabituelles peuvent indiquer les extrêmes. Les anomalies sont les exceptions. Le terme «comté» (county) a une définition commune correspondant au deuxième niveau de l'administration publique. En Louisiane, le terme «paroisse» (parish) équivaut à «comté» et constitue une exception par rapport au terme classique.

18. Les données imparfaites comportent généralement des points faibles au niveau de l'observation. Les maisons individuelles sont identifiables au sol et généralement à partir d'une image satellite. Cependant, un garage converti en unité d'habitation ne peut normalement être identifié sans un complément d'information ou une vérification plus poussée.

19. L'espace blanc sur une carte comporte des caractéristiques spatiales qui ne sont pas indiquées. Cet état de fait prend de l'ampleur à mesure que l'échelle est agrandie et que la carte comporte plus d'espace blanc. L'interpolation est difficile en raison de la nature des données spatiales. Le calcul de données spatiales par imputation pour combler des lacunes ou le remplacement des données manquantes sur une carte peut s'apparenter davantage à une estimation empirique. Soit l'exemple classique de l'interpolation des courbes. L'équidistance des courbes de niveau et la cote le long d'une courbe de niveau sont déterminées par des procédés photogrammétriques et géodésiques ou d'arpentage relativement précis. Les cotes entre une courbe maîtresse et la suivante ne sont pas connues (fig. 2 a)). On peut interpoler l'orientation de la courbe dont la cote change progressivement et de façon régulière (fig. 2 b)), mais il est de fait que l'on ne peut connaître la pente ou l'élévation relative stable sans un calcul précis (fig. 2 c)).

Figure 2
Comparaison de diverses aires géographiques



20. Dans le monde des données spatiales, les métadonnées sont la seule nouveauté qui permet d'améliorer l'utilisation des données et surtout qui facilite les efforts d'intégration des données géospatiales. Les règles applicables aux métadonnées indiquent une marche à suivre pour étayer des informations telles que la qualité des données, l'année de leur collecte, la source et autres informations nécessaires à chaque occurrence de la caractéristique. Les métadonnées sont parfois volumineuses mais elles permettent de prendre des décisions justes et en connaissance de cause concernant l'utilisation et l'intégration de données discrètes.

21. Les coordonnées géographiques ne suffisent pas à elles seules pour définir des données spatiales. Les attributs leur donnent plus de sens et indiquent implicitement la raison d'être des données géoréférencées. Des exemples d'attributs tels qu'un système de classification et un nom géographique ainsi qu'une multitude d'autres descripteurs concourent à l'établissement d'une définition complète des caractéristiques et donnent plus de valeur au phénomène géographique.

22. La gestion des données géospatiales est une opération complexe. Les risques d'introduction d'une erreur dans les données spatiales sont légion. Comme les données géographiques sont référencées par rapport à la Terre, l'un des premiers risques d'erreur concerne l'inexactitude de l'emplacement des données. D'autres complications peuvent surgir du fait de la multitude d'attributs qui définit l'occurrence de la caractéristique. De surcroît, les opérations géospatiales sont propices à la propagation d'erreurs.

23. Il est souhaitable que les données spatiales soient exactes. Il est de la plus haute importance de garantir des relations géographiques correctes dans le domaine statistique. L'application de concepts topologiques dans les opérations géospatiales garantit le respect de l'obligation de maintenir des relations correctes entre les primitives géospatiales (points, lignes et surfaces). Une irrégularité dans la forme ou l'état d'une aire géographique ne pose pas de problème lorsqu'il s'agit de garantir des relations géographiques correctes. Par exemple, les principes topologiques garantissent la relation entre un point représentant une unité d'habitation et son îlot de recensement, quelle que soit la forme ou l'étendue géographique de cet îlot. L'unité d'habitation se trouve dans l'îlot correct et elle est relativement exacte même si les limites de l'îlot de recensement ne sont pas positionnées correctement.

24. Il existe plusieurs différences entre l'application de statistiques spatiales et l'utilisation des mailles statistiques. Visualiser ces deux démarches permet de mieux en comprendre les caractéristiques, les différences et les ressemblances. Les instruments qui ont été mis au point permettent de déterminer plus facilement la meilleure démarche selon l'objet et l'utilisation des données.

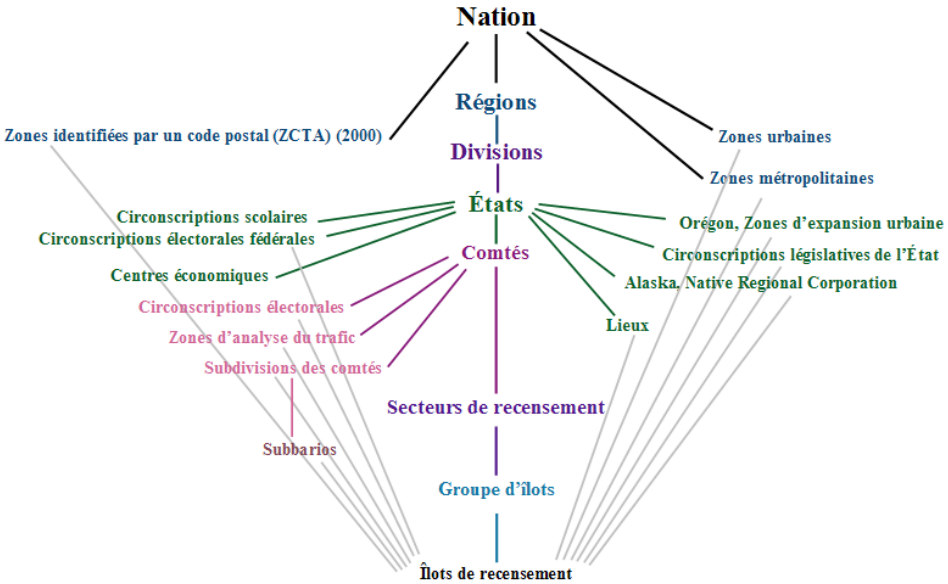
VI. Cadre géographique utilisé par le Census Bureau des États-Unis

25. Le Census Bureau des États-Unis utilise une géographie par îlots de recensement. L'îlot est la plus petite aire géographique au niveau de laquelle il rassemble les données et les présente en tableaux dans le cadre du recensement décennal. Les routes et autres caractéristiques visibles forment les limites des polygones correspondant aux îlots de recensement. D'autres caractéristiques, telles que les limites d'une ville (qui peuvent être un périmètre arpenté invisible sur le sol), sont également utilisées comme limites des îlots de recensement. Les États-Unis et leurs Territoires comptent des millions d'îlots de recensement. Toutes les aires sont affectées à un îlot de recensement.

26. Comparée à la forme régulière des mailles statistiques, celle des îlots de recensement est irrégulière. Le choix des caractéristiques servant à délimiter ces îlots répond à des critères spécifiés. Même si les polygones correspondant aux îlots ont des formes irrégulières, en particulier hors de la configuration des rues au cœur des villes, leur taille satisfait à une tolérance générale. Il est difficile de mener à bien des opérations sur le terrain lorsque les îlots sont trop grands. Tous les relevés spatiaux comprennent des îlots «à problème» qui présentent des anomalies. Par exemple, une route et/ou une voie ferrée peut suivre le bord d'une rivière qui serpente le long d'une vallée. Les polygones étroits correspondant au caractère continu de ces caractéristiques se traduisent souvent par des îlots de recensement allongés.

27. D'autres aires géographiques utilisées pour les recensements se composent d'îlots. C'est là une caractéristique commune à tous les niveaux de la géographie du recensement. De nombreux autres niveaux géographiques s'emboîtent jusqu'à un certain point l'un dans l'autre, c'est-à-dire qu'un niveau plus élevé se définit en partie en intégrant des niveaux moins élevés qui s'emboîtent à l'intérieur de ses délimitations. Le comté en offre un exemple courant en ce sens qu'il comprend des secteurs de recensement, puis des groupes d'îlots à l'intérieur de ces secteurs et enfin des îlots de recensement tels qu'indiqués dans la figure 3 de la hiérarchie type des entités géographiques utilisées pour les recensements.

Figure 3
Hiérarchie type des entités géographiques utilisées pour les recensements



28. Il existe trois types de zones de recensement: les zones juridiques, les zones statistiques et les zones administratives. Les premières sont déterminées à d'autres niveaux de l'administration publique. Par exemple, la délimitation d'une ville dotée de la personnalité morale est approuvée par les élus de la ville. Les zones statistiques sont déterminées par le Census Bureau des États-Unis, souvent en concertation avec des partenaires d'organismes de planification ou de groupements analogues afin de délimiter des zones à bon escient pour la présentation en tableaux des données. Les zones administratives sont par exemple les circonscriptions scolaires dans une ville.

29. Le niveau de précision obtenu pour une aire géographique donnée dépend de divers facteurs. Pour délimiter des zones juridiques, des villes par exemple, on réalise une enquête (a boundary and annexation survey) pour déterminer les modifications territoriales et les nouvelles annexions. La qualité de l'information dépend des sources consultées et de la qualité de cette consultation. Les zones statistiques sont délimitées sur la base de critères. Comme leur interprétation et les intérêts locaux varient, il en va de même des résultats.

30. Du fait de leur taille réduite et de leur caractère utilitaire, les îlots de recensement sont des unités qui conviennent tout-à-fait pour l'acquisition, la gestion et l'utilisation des statistiques spatiales. Leur remplacement par des mailles à ce niveau inextensible de la géographie offre probablement moins de possibilités et présente une plus grande complexité. De façon générale, on peut d'autant plus s'attendre à une plus grande précision que l'on se situe à un niveau moins élevé de la géographie du recensement. L'utilisation de l'îlot du recensement a suscité de nouveaux espoirs quant à la disponibilité et la qualité des données.

VII. Amélioration de la visualisation de l'analyse − études de cas

31. Dans un article qui paraîtra sous peu, la Division de la population du Census Bureau des États-Unis a procédé à une analyse de la population haïtienne (Azar *et al.*, à paraître). Il s'agissait de cartographier la population avec des mailles de 100 mètres, en analysant les données de recensement et images satellites. La population dénombrée a été répartie entre les cellules du maillage en fonction des surfaces imperméables créées par l'homme (constructions telles que bâtiments et routes) correspondant à chaque cellule. La carte quadrillée a ensuite été améliorée au moyen d'outils de cartographie en ligne (Census Bureau des États-Unis, 2010), ce qui permet d'avoir une vision des données selon les besoins et de créer une base commune pour procéder à des analyses dans tout le pays.

32. Une publication précédente, l'Agricultural Atlas of the United States de 1992, avait associé des mailles à la géographie administrative, ce qui avait permis d'afficher avec plus de précision les emplacements des occurrences dans une série de cartes de répartition par points (USDA, 2010). La publication contenait environ 190 de ces cartes ainsi que plus de 120 cartes choroplèthes. Les cartes par points des États-Unis ont été établies à l'aide de données recueillies au niveau des comtés.

33. Avec l'évolution des questions touchant à l'agriculture, il était clair que, sur le territoire de nombreux comtés, un placement aléatoire des points indiquerait une activité agricole là où elle était impossible ou très peu probable, par exemple des pâturages dans des zones urbaines ou des cultures dans la toundra. Les sources comprenaient des fichiers généralisés distincts des limites des comtés, des traits de côte et des zones urbaines. Un fichier au format raster a été utilisé pour les données relatives à l'utilisation des terres/la couverture du sol. À de plus petites échelles, ces données ont été regroupées par thème sous forme de fichiers polygonaux et intégrées dans les frontières administratives des comtés.

34. Les données agricoles ont été réparties entre les cinq grandes catégories suivantes: exploitations agricoles polyvalentes; cultures; terres de parcours et pâturages; bétail;

vergers. Il a été assigné à chaque rubrique de l'utilisation des terres (par exemple forêts et pâturages boisés) une valeur correspondant à la probabilité d'une occurrence pour chacune des catégories agricoles (élevée pour les terres de parcours et pâturages). Un certain nombre de points, calculé en pourcentage en fonction du classement ainsi établi, a été inscrit dans la partie du comté correspondant à l'utilisation spécifiée des terres. Chaque point d'une carte donnée représentait un nombre spécifié d'occurrences (valeur du point).

35. L'association de deux types disparates de données a fourni une représentation plus fidèle de données cartographiées. La configuration des données à la disposition des analystes correspondait à l'impact de l'utilisation des terres sur diverses activités agricoles. De même, l'utilisation combinée de données maillées et de la géographie administrative peut produire des résultats plus favorables.

VIII. Conclusions

36. La convergence relativement récente des données, de la technologie, des logiciels et de la puissance des ordinateurs a ouvert de nouvelles perspectives aux analystes désireux d'étudier les données statistiques en fonction de leur emplacement. Des fonctions essentielles se sont dégagées des systèmes d'information géographique utilisés précédemment, mais l'on cherche de plus en plus à évaluer les configurations et les groupes d'occurrences ainsi qu'à prévoir les activités futures.

37. Lorsque l'on peut utiliser des aires géographiques correspondant à un niveau relativement bas, les îlots de recensement par exemple, il y a tout lieu de considérer ce type de données comme une alternative à un maillage statistique étant donné que la nature des données spatiales influe sur les occurrences de données statistiques. Il existe des possibilités de fusionner les unités administratives et le maillage statistique, ce qui pourrait offrir aux analystes de nouvelles utilisations de données intégrées. Les efforts déployés au cours de ces trois dernières décennies se sont concentrés sur la construction d'ensembles de données spatiales et l'utilisation des caractéristiques selon des méthodes conventionnelles. La possibilité de visualiser l'effet des statistiques spatiales sous diverses formes offre aux analystes différents axes de recherche à approfondir et ouvre la voie à une analyse plus large.

IX. Bibliographie

Anselin, L. (1992), Spatial Data Analysis with GIS: An Introduction to Application in the Social Sciences, Technical Report 92-10, National Center for Geographic Information and Analysis

Azar D., *et al.* (Forthcoming), Spatial refinement of census population distribution using remotely sensed estimates of impervious surfaces in Haiti, *International Journal of Remote Sensing*.

Cressie, N.A.C., (1993), *Statistics for Spatial Data*, Wiley: New York.

Griffith, D.A. and Paelinck, J.H.P. (2007), An equation by any other name is still the same: Spatial econometrics and spatial statistics, *Annals of Regional Science*, Vol. 41.

Tobler W. (1970), A computer movie simulating urban growth in the Detroit region, Economic Geography, 46(2).

United States Census Bureau. 2010. Haiti Earthquake: United States Census Bureau Population Data. Online: https//www.geoint-online.net/community/haitiearthquake/default.aspx.

USDA. 2010. Agricultural Atlas of the United States. Online: http://www.agcensus.usda.gov/Publications/1992/Agricultural_Atlas/textfile/introduc.asc.