Intérêt de l'intégration de la dimension spatiale dans l'analyse démographique

Août 2016

1 Position du problème

Au début du XXI^e siècle, la population africaine représentait près de 7% de la population mondiale. Ce chiffre est passé à 9,1% en 1950 et à 15,7% en 2014. En 1950, cette population s'établissait à 228 827 000 habitants et en 2014, elle est estimée à 1 138 229 000 habitants soit un taux de croissance annuel moyen de 2,5%. L'Afrique de l'Est se révèle être la plus peuplée avec 33,6% de la population africaine, suivie par l'Afrique de l'Ouest (29,8%), l'Afrique du Nord (18,9%), l'Afrique Centrale (12,2%) et l'Afrique Australe (5,4%). La population africaine est caractérisée par un taux de fécondité estimé en 2013 à 4,7 enfants par femme contre une moyenne mondiale de 2,5 enfants par femme et une espérance de vie à la naissance de 57 ans tandis que la moyenne mondiale est de 69 ans.

Ce rythme de croissance de la population africaine, qui atteint des niveaux jusque là inconnus, et sa part de plus en plus importante dans la population mondiale, interpelle naturellement les démographes et les responsables politiques sur les causes de cette augmentation mais également sur ses conséquences dans l'évolution du bien-être des habitants de ce continent.

L'étude de l'évolution de la population africaine passe par l'analyse de ses caractéristiques que sont la mortalité, la fécondité, les migrations internationales, la structure par âge, la transition démographique etc. Ces variables sont elles-même déterminées par un certain nombre de facteurs d'ordre socio-économique, culturel, politique et écologique. Ainsi, d'une part, pour comprendre la dynamique de la population africaine, il faut correctement expliquer ses caractéristiques par leurs déterminants et, d'autre part, pour saisir les effets de la population sur les indicateurs de développement, il faut les intégrer dans un modèle adéquat et robuste.

Pour ce faire, le chercheur dispose d'un certain nombre d'outils statistiques et de méthodes économétriques tels que la statistique descriptive, l'analyse des données, l'économétrie du modèle linéaire, l'économétrie des variables qualitatives, des séries temporelles etc. En l'absence d'une dimension spatiale, ces méthodes

sont tout à fait fiables et conduisent à des estimations robustes.

Toutefois, l'hypothèse de l'absence d'une dimension spatiale dans les caractéristiques de la population ou dans leurs déterminants n'est pas toujours pertinente. S'agissant de la fécondité par exemple, on sait qu'elle est causée, en Afrique, par un certain nombre de variables intermédiaires telles que la profession du mari, l'activité économique et le niveau d'instruction de la femme... Or, en Afrique il existe une corrélation (positive ou négative) entre ces variables et la localisation géographique de l'individu. En effet, ces variables intermédiaires sont elles-même expliquées par des facteurs socio-économiques et culturels tels que l'ethnie et la religion qui concentrent les individus en des lieux déterminés.

S'il est établi que la répartition des variables démographiques sont le fait d'une concentration régionale conduisant à une autocorrélation spatiale, la non prise en compte de la dimension spatiale dans l'analyse démographique peut entraîner une perte d'information, une mauvaise spécification des erreurs, une absence de convergence et une inefficacité des estimateurs qui aboutissent, de fait, à une mauvaise connaissance des caractéristiques de la population africaine et à un biais dans l'évaluation des effets de la population sur les indicateurs de niveau de vie (Thomas-Agnan, 2012).

2 L'analyse statistique des données spatiales : définition des concepts

L'application de l'analyse spatiale nécessite au préalable la compréhension d'un certain nombre de concepts tels que les données spatiales, les effets spatiaux, les matrices de voisinage, l'indice et le test de Moran, le diagramme de Moran et les modèles économétriques spatiales.

2.1 Les données spatiales

Ce sont des données comportant une dimension spatiale, c'est à dire pour lesquelles une information géographique est attachée à chaque unité statistique. Elles peuvent être des données géo-référencées (des observations localisées dans l'espace) ou des objets spatiaux. Les données géo-référencées apportent des informations sur la valeur de la variable en plus de l'information sur la localisation tandis que les objets spatiaux peuvent être des coordonnées géographiques dé-

terminant la localisation d'objets ou d'évènements, des lignes ou arcs. Ces types de données sont le plus souvent insérés dans des "shapefiles".

2.2 Les effets spatiaux

Ces effets sont l'autocorrélation spatiale et l'hétérogénéité spatiale

2.2.1 L'autocorrélation spatiale

L'autocorrélation spatiale indique une relation fonctionnelle entre ce qui se passe entre une unité spatiale et ses voisines. Elle fournit une information supplémentaire par rapport aux statistiques traditionnelles, invariantes par rapport à la configuration spatiale des données. L'autocorrélation spatiale est différente de l'autocorrélation temporelle dans le sens qu'elle est multidimensionnelle. Par conséquent, les techniques valables pour les séries temporelles ne sont pas directement transposables au cas spatial (Ertur, 2012).

Les sources de l'autocorrélation spatiale sont multiples : interactions spatiales, contagion, externalités spatiales, effets de débordement géographique, effets d'imitation ou encore une mauvaise spécification du modèle économétrique causée par l'omission de variables elles-mêmes spatialement autocorrélées. De manière transposée, les sources d'absence d'autocorrélation sont entre autres une répartition spatiale des valeurs de la variable aléatoire, une invariance par permutation aléatoire des individus (aléas spatial), des valeurs observées en une localisation qui ne dépendent pas des valeurs observées en des localisations voisines, une égalité de degré de vraisemblance entre le schéma spatial observé et tout autre schéma spatial et une modification de la localisation des individus sans conséquence sur le contenu informationnel des données.

Il existe deux types d'autocorrélation : l'autocorrélation positive et l'autocorrélation négative. L'autocorrélation positive apparait lorsque des valeurs similaires tendent à se concentrer dans l'espace. Elle est présente dans les processus de diffusion et contagion bien qu'il faille signaler que la diffusion tend à produire une autocorrélation spatiale positive mais que la réciproque n'est pas nécessairement vraie. Quant à l'autocorrélation négative, elle intervient lorsque les valeurs sur lesquelles porte la concentration sont dissimilaires.

2.2.2 L'hétérogénéité spatiale

L'hétérogénéité spatiale est liée à la différenciation des comportements dans l'espace. Il s'agit de l'instabilité structurelle due à une instabilité spatiale des coefficients de la régression et à la non-linéarité des formes fonctionnelles ou de l'hétéroscédasticité provenant de variables omises ou de toute autre forme de mauvaise spécification (Ertur, 2012).

Ce phénomène se retrouve à plusieurs échelles : les comportements et les phénomènes économiques ne sont pas les mêmes dans le centre d'une ville et dans sa périphérie ou dans une région urbaine et dans une région rurale (Le Gallo, 2004).

L'instabilité structurelle provient de l'absence de stabilité dans l'espace des comportements ou d'autres relations étudiées : les formes fonctionnelles et les paramètres varient selon leurs localisations et ne sont donc pas homogènes. Il est donc nécessaire de mobiliser des modélisations prenant en compte les caractéristiques particulières de chaque localisation de l'échantillon.

S'agissant de l'hétérogénéité spatiale, dans les modèles économétriques, elle peut venir de variables manquantes ou de toute autre forme de mauvaise spécification. Par exemple, les unités spatiales elles-mêmes ne sont généralement ni de formes régulières, ni homogènes : des régions peuvent avoir des formes et des aires différentes, des niveaux de développement technologique variables, des populations plus ou moins importantes, etc.

Enfin, le traitement de l'hétérogénéité spatiale est effectué en prenant en compte l'usage de variables explicatives pour modéliser la tendance. Certaines de ces variables peuvent être spatiales de nature comme, par exemple, la distance. Cependant, il faut noter qu'il n'est pas suffisant de prendre en compte ces variables dans la moyenne pour évacuer totalement la structure spatiale du problème qui peut rester présente à l'ordre 2 (Thomas-Agnan, 2012).

2.3 Les matrices de voisinage

La mesure du voisinage d'un point peut être abordée de plusieurs manières. Elle est souvent effectuée avec une matrice de voisinage qui est la version spatiale de l'opérateur retard en séries temporelles. Cette matrice de voisinage est également appelée matrice de poids.

Pour N sites, la matrice de voisinage W (qui peut être autre que géographique) est définie comme la matrice de dimension $N \times N$ telle que chacun de ses éléments

w_{ij} indique l'intensité de la proximité de la zone i par rapport à la zone j. La matrice W n'est pas nécessairement symétrique. Elle peut cependant être symétrisée avec la transformation $W_s = (W + W')/2$.

La matrice de voisinage est dite normalisée lorsqu'on impose la contrainte $\sum_{j=1}^{N} w_{ij} = 1$. L'intérêt de cette normalisation est qu'elle permet de rendre les paramètres spatiaux comparables entre divers modèles. On impose également, en général, que la diagonale soit nulle c'est à dire $\forall i, w_{ii} = 0$.

Il existe plusieurs de types de matrices de voisinage :

- la matrice de contiguïté ;
- la matrice basée sur la distance entre centroïdes ;
- la matrice basée sur les plus proches voisins ; et
- la matrice de triangulation de Delauney.

2.4 L'indice et le test de Moran

2.4.1 L'indice de Moran

Soit w_{ij} les éléments d'une matrice de voisinage W vérifiant $w_{ii} = 0$ et une variable $x_i = x_{s_i}, i = 1, ...n$, l'indice de Moran est défini comme suit :

$$\mathbf{I} = \frac{N}{S_0} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum (x_i - \bar{x})^2} \tag{1}$$

avec $S_0 = \sum_i \sum_j w_{ij}$.
En posant $z_i = x_i - \bar{x}$, $z_j = x_j - \bar{x}$ et $z = (z_1, z_2, ..., z_N)'$, il vient la forme matricielle suivante :

$$\mathbf{I} = \frac{N}{S_0} \frac{z' W z}{z' z} \tag{2}$$

Notons que pour une matrice W standardisée en lignes, $S_0 = N$.
L'indice de Moran définit le rapport d'une sorte de covariance entre unités contiguës à la valeur du champ. C'est dans une certaine mesure un coefficient d'autocorrélation. Par ailleurs, **I** est invariant à un changement d'unité de mesure et à une symétrisation de la matrice W. Les valeurs positives et fortes de **I** indiquent une autocorrélation spatiale positive, les valeurs négatives et fortes de **I** une autocorrélation spatiale négative et les valeurs proches de 0 une absence d'autocorrélation.

2.4.2 Le test de Moran

Le test de Moran est un test d'absence d'autocorrélation spatiale. Il en existe deux types : le test de Moran pour variable surfacique continue (qui est celui qui nous intéresse et que nous allons développer) et le test de Moran pour variable surfacique qualitative.

Les hypothèses du test pour variable surfacique continue sont les suivantes :

$$\begin{cases} H_0 & \text{absence d'autocorrélation spatiale} \\ H_1 & \text{présence d'autocorrélation spatiale} \end{cases} \tag{3}$$

2.5 Le diagramme de Moran

Le diagramme de Moran est un nuage de points présentant la variable d'intérêt x en abscisse et la variable spatialement décalée Wx en ordonnée. La variable x est centrée en abscisse et par conséquent la variable spatialement décalée Wx en ordonnée est également centrée lorsque W est normalisée.

Les points du quadrant $x \geq 0$, $y \geq 0$ sont autocorrélés positivement et localement par rapport à la variable d'intérêt de même que les points du quadrant $x \leq 0$, $y \leq 0$. Les points des quadrants $x \leq 0$, $y \geq 0$ et $x \geq 0$, $y \leq 0$ sont quant-à eux autocorrélés négativement et localement par rapport toujours à la variable d'intérêt. Une non linéarité du nuage traduit plusieurs régimes d'association spatiale.

2.6 Les modèles de régression spatiale

Soit le modèle suivant :

$$Y = X\beta + \epsilon \tag{4}$$

$\mathbb{E}(\epsilon) = 0$, $Var(\epsilon) = \sigma^2\Omega$ où Ω est la matrice de variance-covariance.
Dans le cas où $\Omega = I_n$, on procède à une estimation par moindres carrés ordinaires (MCO) et dans le cas contraire, il faut utiliser les moindres carrés pondérés (MCP) ou les moindres carrés généralisés (MCG).

Étant donné une matrice de voisinage normalisée W et une variable Z, il vient automatiquement que WZ présente une autocorrélation spatiale avec Z. La famille des modèles spatiaux simultanés consiste à introduire une telle variable dans le modèle non spatial (MCO, MCP ou MCG) à divers endroits (Thomas-Agnan, 2012) :

- introduire WX en explicative dans le modèle non spatial conduit au modèle *SLX* (Spatially lagged-X) ;

- introduire WY dans le membre de droite du modèle non spatial conduit au modèle *SAR* (Spatial-Autoregressive Model) encore appelé *LAG* ;
- introduire WX dans le modèle *SAR* conduit au modèle *SDM* (Spatial Durbin Model) ;
- utiliser le modèle *SAR* pour le terme d'erreur conduit au modèle *SEM* (Spatial Error Model) ;
- combiner les modèles *SAR* et *SEM* conduit au modèle général *SAC* ;
- introduire $W\epsilon$ dans le modèle non spatial conduit au modèle *MA* (Moving Average) ; et
- combiner les modèles *SAR* et *MA* conduit au modèle *SARMA*.

3 Exemple d'application de l'analyse spatiale en démographie : l'accès au soin de santé des ménages du Sénégal

Prenons comme exemple illustratif (très réducteur) l'accès au soin de santé des ménages au Sénégal. Cette variable est liée à la santé des membres du ménage en général et à celle de la mère en particulier, toute chose qui explique le taux de fécondité et le taux de mortalité.

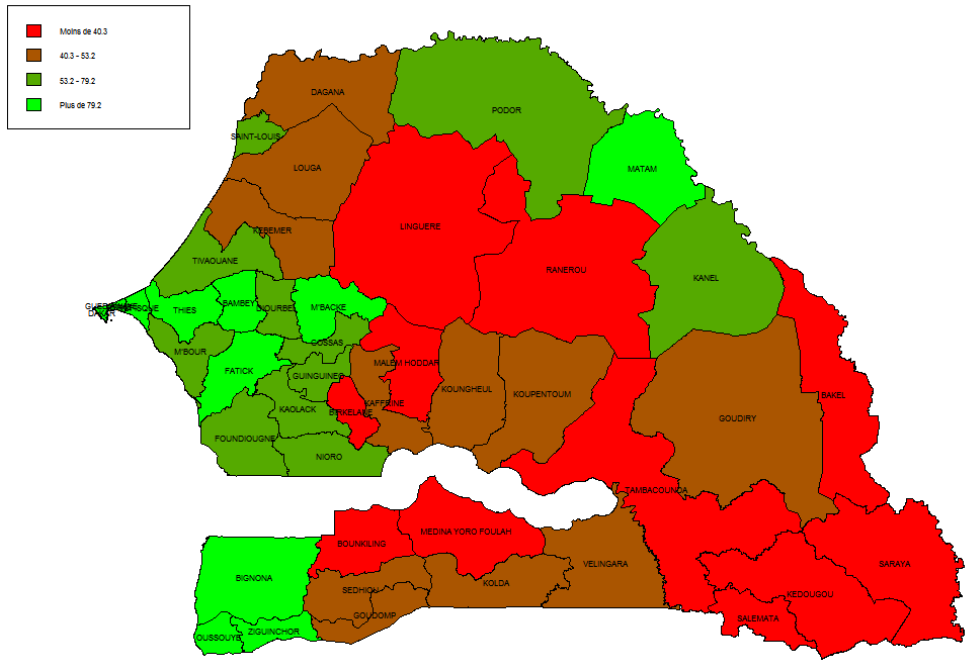
3.1 Données

Les données sont tirées de l'Enquête Village de 2009 sur l'accès au services sociaux de base conçue pour suppléer au manque d'informations permettant le ciblage des zones rurales jugées prioritaires dans la mise en place de programmes de réduction de la pauvreté. Un ménage est considéré comme ayant accès aux services de santé s'il se trouve à moins de 5 kilomètres du poste de santé le plus proche.

Pour prendre en compte la dimension spatiale, un "shapefile" du Sénégal en découpage départemental a été utilisé et la matrice de contiguité a été retenue comme matrice de voisinage.

La figure suivante permet de s'apercevoir de la concentration spatiale des ménages ayant accès aux services de soins de santé.

FIGURE 1 – Répartition spatiale de l'accès aux services de soins de santé



Source : Données de l'enquête village 2009, ANSD, notre construction

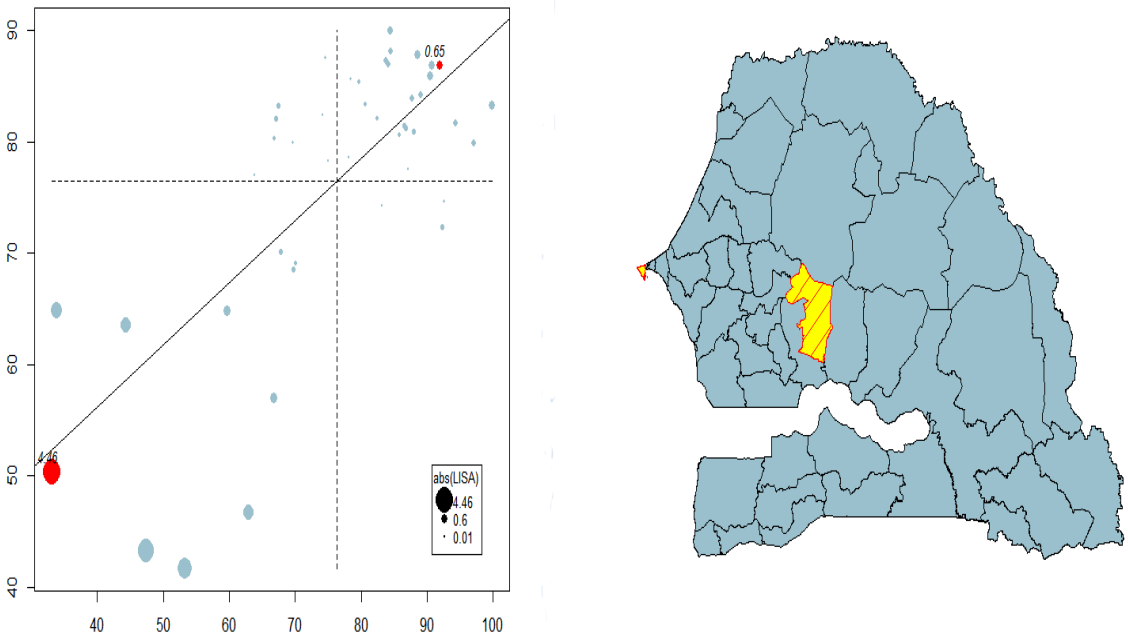
Les ménages se trouvant dans les départements de l'Ouest du Sénégal ont le plus accès aux services de soins de santé tandis que ceux de l'Est en ont un accès moindre.

3.2 Test et diagramme de Moran

Le test de Moran fournit une statistique de test de 0,56 avec une p-value inférieure à 0,0001. Ceci conduit à accepter l'hypothèse de l'existence d'une autocorrélation positive.

Le diagramme de Moran vient confirmer l'existence de cette autocorrélation :

FIGURE 2 – Diagramme de Moran



Source : Données de l'enquête village 2009, ANSD, notre construction

La largeur des points indique les intensités de l'autocorrélation. Les points du quadrant $x \geq 0$, $y \geq 0$ et ceux du quadrant $x \leq 0$, $y \leq 0$ représentent les départements où l'autocorrélation spatiale est positive avec cependant une importante différence : les départements représentés dans le quadrant $x \leq 0$, $y \leq 0$ ont relativement accès aux services de soins de santé tandis que ceux du quadrant $x \geq 0$, $y \geq 0$ n'y ont pas accès.

Ainsi, la variable accès aux services de soins de santé présente une autocorrélation spatiale. De fait, si l'on expliquait la fécondité par exemple à partir d'un certain nombre de variables explicatives dont l'accès aux services de soins de santé, le modèle économétrique standard de la forme (FEC :fécondité, X : des variables explicatives de la fécondité et $ASANT$: l'accès aux services de soins de santé) :

$$FEC = X\alpha + ASANT\beta + \epsilon$$

n'est pas toujours approprié. Il faudrait alors impérativement passer à des modèles économétriques spatiaux tels que les modèles SAR, $SARMA$ ou SEM.

Références bibliographiques

Anselin L., 1980, Estimation Methods for Spatial Autoregressive Structures, Cornell University, Regional Science Dissertation and Monograph Series no 8, Ithaca, NY

Anselin L., Kelejian H., 1997, Testing for spatial error autocorrelation in the presence of endogenous regressors, International Regional Science Review, 20, 153-182

Aten B., 1997, Does space matter ? International comparisons of the prices of tradables and nontradables, International Regional Science Review, 20, 35-52.

Benirschka A., Binkley J., 1996, Land price volatility in a geographically dispersed market, American Journal of Agricultural Economics, 76, 185-195.

Cliff A.D., Ord J.K., 1972, Testing for spatial autocorrelation among regression-residuals, Geographical Analysis , 4, 267-284

Cliff A.D., Ord J.K., 1973, Spatial Autocorrelation, Pion, Londres.

Cliff A.D., Ord J.K., 1981, Spatial Processes : Models and Applications, Pion, Londres.

Le Gallo J, (2000), Econométrie spatiale, Université de Bourgogne

Le Sage J.P, R. K. Pace (2009), Introduction to spatial econometrics, CRC Press.

Pace R. K., Gilley O.W., 1997, Using the spatial configuration of the data to improve estimation, Journal of Real Estate Finance and Economics, 14, 333-340.

Paelinck J.H.P., Klaassen L.H., 1979, Spatial econometrics, Saxon House, Farnborough.

Thomas-Agnan C, (2012), Analyse statistique des données spatiales, Eustat.