LA LOI DE ZIPF SOUS LE PRISME DE L'AUTO-CORRELATION SPATIALE

LES CAS DE LA CHINE ET DE L'INDE

Alexandra SCHAFFAR *

Résumé : *Cet article utilise un modèle rang-taille des villes sur des données de panel. En introduisant la dimension géo-localisée des données, il propose de prendre en compte les effets d'auto-corrélation spatiale dans l'étude de la validité de la loi de Zipf. La méthodologie utilisée pour détecter les effets d'auto-corrélation spatiale est celle initiée par Debarsy et Ertur (2010). Elle permet de choisir la meilleure spécification du modèle rang-taille parmi trois alternatives : le modèle autorégressif spatial (SAR), le modèle à auto-corrélation spatiale des erreurs (SEM) et le modèle spatial général à effets fixes sur des données de panel (SARAR). En Chine, la prise en compte des effets d'auto-corrélation spatiale ne modifie pas fondamentalement les conclusions du modèle a-spatial tout en permettant de l'affiner : la distribution rang-taille des villes suit la loi de Zipf dans le long terme. En Inde, le modèle a-spatial conduit à un rejet de la loi de Zipf et l'introduction de la dimension spatiale conduit à valider la loi de Zipf sur le long terme.*

Mots clés : LOI DE ZIPF, HIÉRARCHIES URBAINES, CHINE, INDE, AUTOCORRÉLATION SPATIALE.

Classification JEL : R10, R12, C31

L'auteur souhaite remercier Julie Le Gallo, Université de Besançon, et Nicolas Debarsy, Université de Louvain, pour les échanges et les remarques fructueuses qui ont permis d'améliorer cet article.

* Université de Toulon, LEAD, EA 3163, 83957 La Garde, France ; schaffar@univ-tln.fr.

1. INTRODUCTION

Depuis plusieurs années, la littérature économique sur la démographie des villes s'interroge sur la nature et la forme des hiérarchies urbaines contemporaines. De nombreux travaux mettent en évidence la validité de la loi de Zipf (1949), selon laquelle la distribution rang-taille des villes d'un pays ou d'une région épouse une distribution de type Pareto avec une pente égale à -1, ce qui confirme les observations empiriques antérieures de Auerbach (1898) et de Lotka (1941). Bien que le « mystère » de la loi de Zipf (Krugman, 1996) fasse l'objet de plusieurs travaux d'ordre méthodologique et/ou empirique, sa dimension spatiale fut étrangement négligée sauf quelques tentatives récentes isolées.

L'objectif de ce papier est de tester l'existence d'effets d'auto-corrélation spatiale sur les distributions des tailles urbaines en Chine et en Inde, entre 1980 et 2005. Durant cette période, les deux pays ont connu des évolutions importantes, tant du point de vue économique que démographique ou institutionnel. En Chine, la libéralisation économique des années quatre-vingt a généré d'importants flux migratoires de main d'œuvre sur un plan intra-régional et interrégional. Ces flux se sont intensifiés après les réformes de 1993 qui ont assoupli le système *Hukou*, en relâchant les multiples contraintes et réglementations qui pesaient sur la mobilité des travailleurs. En Inde, la période 1981-2001 correspond au passage d'un pays autocentré à une économie ouverte à l'échange international. Ce mouvement s'est accéléré après les réformes et les changements politiques intervenus en 1991, tandis que les grandes villes indiennes acquièrent progressivement un statut de métropoles internationales (Schaffar et Dimou, 2012 ; Schaffar, 2010).

De nombreuses études empiriques sur les hiérarchies urbaines emploient des modèles économétriques sur données transversales, permettant d'obtenir une information sur le degré de hiérarchisation d'un système régional ou national de villes pour une seule date. Nous privilégions, ici, une analyse sur des données de panel permettant de déterminer un coefficient de hiérarchisation pour l'ensemble de la période de référence. Comme ces données sont géo-localisées, il est nécessaire d'étudier l'existence d'effets d'auto-corrélation spatiale dans la distribution rang-taille des villes, en s'appuyant sur une batterie d'outils économétriques récents (Elhorst, 2010 ; Lee et Yu, 2010 ; Debarsy et Ertur, 2010).

Dans une deuxième section, nous présentons les développements théoriques récents sur la loi de Zipf pour les villes. La troisième section est axée sur des questions méthodologiques liées à l'utilisation des modèles spatiaux sur données de panel. La quatrième section présente les résultats obtenus pour les distributions chinoise et indienne durant la période 1981-2005. Enfin, la dernière section conclut sur les résultats obtenus et les apports des modèles spatiaux sur données de panel dans l'analyse des hiérarchies urbaines.

2. LA LOI DE ZIPF POUR LES VILLES

Les études empiriques sur les hiérarchies urbaines admettent généralement l'hypothèse que la probabilité $\Pr(S > S_0)$ pour que la taille S d'une ville soit supérieure à une taille minimale S_0 est :

$$\Pr\left(S > S_0\right) = \int_s^{\infty} f(t)\,dt = \int_s^{\infty} Ct^{-a}\,dt = \frac{CS^{-a+1}}{a-1}, a > 1$$

Lorsque les villes sont rangées, selon leur taille $S_1 > S_2 > S_3 > ... > S_i > ... > S_n$, avec *R(s)* le rang de la ville de taille S, on obtient :

$$\Pr(S > S_0) = \frac{R}{n} = \frac{C}{a-1} S^{-a+1}, a > 1$$

Si on pose $k = \dfrac{Cn}{a-1}$ et $\beta = a - 1$ ($\beta > 0$), le modèle rang-taille des villes est donné par la relation :

$$R(s) = k \cdot S^{-\beta}$$

avec *k* un paramètre qui dépend de la taille de la plus grande ville et β un coefficient de hiérarchisation, appelé communément coefficient de Pareto. La relation linéaire entre les logarithmes de rang et de taille correspond à la version la plus connue de la loi rang-taille qui stipule que le rang d'une ville donnée est inversement proportionnel à sa taille :

$$\ln r(s) = -\beta \ln s + \ln k$$

Lorsque β est égal à 1, on obtient la loi de Zipf.

Les modèles rang-taille ont généré une littérature empirique importante. Rosen et Resnick (1980) entament une comparaison des distributions de villes de plus de 100 000 habitants dans 44 pays. Ils obtiennent des coefficients de hiérarchisation dans un intervalle entre 0,81 et 1,96, avec 75% des pays affichant une valeur absolue de l'exposant supérieure à 1, ce qui les conduit à conclure que la distribution rang-taille des villes est plus égalitaire que celle supposée par la loi Zipf.

Dans sa revue de la littérature empirique sur la loi de Zipf, entre 1913 et 1980, Carroll (1982) montre que les distributions dont le coefficient de Pareto est égal à 1 ne représentent qu'un cas parmi d'autres hiérarchies urbaines existantes. Plutôt que de tester la validité de la loi de Zipf, il suggère de comparer l'organisation des différents systèmes urbains dans un *continuum* entre deux configurations extrêmes, la primatie absolue d'une part, l'égalité parfaite des tailles des villes, d'autre part. Gabaix et Ioannides (2004) se focalisent également sur le degré d'adéquation des différentes distributions avec la loi de Zipf plutôt que sur sa validité au sens strict du terme.

Moriconi-Ebrard (1993) propose une analyse de la distribution des villes de plus de 10 000 habitants dans 78 pays, entre 1950 et 1980. Il trouve un indice de hiérarchisation global égal à 1,05, avec un écart-type relativement faible, ce qui semble valider la loi de Zipf à l'échelle mondiale. Un véritable clivage apparaît entre l'Ancien (Europe, Asie, Afrique) et le Nouveau Monde (Amérique, Océanie), ce qui conduit Moriconi-Ebrard à expliquer les différentes structurations des systèmes urbains par des facteurs historiques. De son côté, dans son étude sur les distributions des villes de 73 pays, Soo (2005) rejette la validité empirique de la loi de Zipf dans 40% des cas. Il trouve un coefficient de hiérarchisation plus fort pour les pays à revenu élevé que pour les pays à revenu faible ou intermédiaire, en concluant que le niveau du développement économique a une influence déterminante sur les hiérarchies urbaines.

Enfin, dans sa méta-analyse sur 515 études, Nitsch (2005) propose une comparaison des différentes estimations du coefficient de hiérarchisation. L'ensemble des estimations du coefficient varie dans un intervalle de valeurs entre 0,49 et 1,96, avec une moyenne de 1,09. Nitsch (2005) montre, dans un premier temps, que la définition de la ville joue un rôle fondamental dans la valeur estimée du coefficient de hiérarchisation. Les échantillons de villes définies de façon administrative présentent des distributions plus égalitaires que les échantillons des agglomérations, bien que ce résultat puisse être lié au fait que le nombre des premières est systématiquement inférieur à celui des secondes. Dans un deuxième temps, il montre que la moyenne des estimations baisse de façon conséquente jusqu'au milieu du 20ème siècle puis se stabilise, voire remonte légèrement par la suite, ce qui serait un signe de déconcentration urbaine, dans certains pays, à partir des années 1960 et 70.

Bien qu'une littérature conséquente semble confirmer la loi de Zipf dans sa version plus ou moins stricte, plusieurs travaux mettent en cause sa construction statistique parfois artificielle (Black et Henderson, 1999).

En premier lieu, certains chercheurs considèrent que si l'on tient compte de toute la distribution des tailles urbaines et non seulement de sa partie haute, celle-ci épouse davantage la forme d'une distribution log-normale que d'une distribution de Pareto (Eeckhout, 2004). La contestation d'une borne inférieure arbitraire pour l'échantillonnage devient cependant problématique pour la définition de la ville qui, dans certains cas, n'est plus qu'un simple bourg voire un habitat isolé. Or, dans de nombreux pays, la majorité de la population vit dans les centres urbains et les agglomérations, ce qui permet de considérer que la distribution tronquée est une bonne approximation des hiérarchies urbaines (Levy, 2009 ; Eeckhout, 2009 ; Ioannides et Scouras, 2009).

En second lieu, la taille de l'échantillon et/ou la méthode d'estimation du coefficient de hiérarchisation influent fortement sur sa valeur. Gabaix et Ioannides (2004) ou Soo (2005) montrent que pour de nombreuses distributions, la loi de Zipf est validée ou pas, selon la méthode d'estimation utilisée, tandis que Gabaix et Ibragimov (2011) élaborent une correction pour les petits échantillons. Schaffar (2009) propose une comparaison des biais du coefficient selon les différentes méthodes d'estimation utilisées et leurs corrections.

Enfin, en dernier lieu, les distributions rang-taille des villes traitent des données géo-localisées. Il semble, alors, logique que les tailles urbaines puissent être géographiquement corrélées : par exemple, les régions rurales sont peuplées de petites villes, tandis que certaines régions marquées par la présence d'industries traditionnelles sont caractérisées par un tissu de villes moyennes spécialisées ; parfois, l'ombre d'agglomération d'une grande métropole empêche le développement d'autres grandes et ou moyennes villes à proximité (Krugman, 1996). Peu d'études ont cherché à prendre en compte la dimension spatiale des distributions rang-taille.

Parmi celles-ci, il convient de citer le travail de Chasco et Le Gallo (2008) qui utilisent un modèle spatial SUR pour tester l'auto-corrélation spatiale de la distribution rang-taille des villes espagnoles durant le vingtième siècle. Xu et Harriss (2010), de leur côté, introduisent l'auto-corrélation spatiale dans un modèle de croissance urbaine appliqué aux villes du Texas entre 1850 et 2000, tandis que dans leur travail sur la distribution rang-taille des grandes villes américaines, Rey et Ye (2007) cherchent à spatialiser la loi de Zipf en identifiant les Etats d'appartenance des villes les plus dynamiques, dont le rang a fortement augmenté entre 1960 et 2000. Enfin Schaffar (2010), propose un test de cointégration afin d'identifier les éventuels effets d'appartenance régionale sur la croissance démographique des villes indiennes durant les trente dernières années du vingtième siècle.

3. L'INTRODUCTION DES EFFETS D'AUTO-CORRÉLATION SPATIALE : ÉLÉMENTS MÉTHODOLOGIQUES

Nous cherchons à prendre en compte les effets d'auto-corrélation spatiale dans l'examen de la validité de la loi de Zipf en Chine et en Inde, à la fin du vingtième siècle. Ce travail s'appuie sur la méthodologie économétrique développée par Lee et Yu (2010) et Debarsy et Ertur (2010).

La base de données pour les villes chinoises est extraite *des Chinese Urban Statistical Yearbooks*. Les villes sont définies selon les critères statistiques de 1984 corrigés en 2000. L'échantillon comprend les 225 villes de plus de 100 000 habitants en 1984, ce qui correspond à 59% de la population urbaine du pays. Les séries statistiques vont de 1984 à 2004 et couvrent toute la période des mutations économiques sociales, politiques et institutionnelles profondes de ce pays. Les villes de Hong Kong et du Tibet sont exclues de la base statistique.

Pour l'Inde, les données sont issues des *Census* de 1981, 1991 et 2001. Nous considérons, dans ce travail, les 222 villes avec plus de 100 000 habitants en 1981. Pour établir des séries de tailles urbaines annuelles entre 1981 et 2001, deux sources statistiques complémentaires ont été utilisées : celles de la croissance démographique naturelle, c'est-à-dire le solde des naissances et des décès, publiées annuellement par le *Registrar General of India* pour chaque ville et celles des migrations rurales-urbaines, disponibles pour chaque ville, dans les *Migration Tables* annuelles. Ces informations ont été fournies par *Datanet* qui est une émanation du *Department of Statistics, Planning and Public Grievances* du gouvernement indien. Les villes du Jammu et Kashmir et de Himachal Pradesh ne sont pas inclues.

Afin d'étudier la distribution rang-taille des villes des deux pays, nous utilisons la correction de Gabaix et Ibragimov (2011) pour les échantillons de petite taille à travers le modèle Rang – ½. Lorsque l'on dispose des données en panel, ce modèle est donné par l'équation :

$$\ln(R_{it} - 1/2) = -\beta \ln S_{it} + \alpha_i + u_{it}$$

où α_i représente les effets fixes individuels et u_{it} le terme d'erreur qui est i.i.d. avec une moyenne nulle est une variance σ^2.

Nous testons l'hypothèse de la dépendance spatiale dans la distribution rang-taille des villes qui sous-entend une asymétrie spatiale dans la distribution des tailles urbaines en Chine et en Inde. Nous suivons la démarche initiée par Anselin (1998 et 2008) qui introduit des effets d'interaction spatiale dans des modèles sur données de panel, en distinguant le modèle SAR (modèle autorégressif spatial) et le modèle SEM (modèle à auto-corrélation spatiale des erreurs). Le premier incorpore une variable endogène décalée parmi les régresseurs du modèle linéaire standard, tandis que le second admet l'hypothèse d'une relation autorégressive spatiale pour le terme d'erreur u_{it}. Lee et Yu (2010) proposent une combinaison du modèle autorégressif spatial et du modèle à auto-corrélation spatiale des erreurs, à travers un modèle spatial général (SARAR) à effets fixes sur des données de panel.

Afin de prendre en compte l'hétérogénéité spatiale dans la distribution rang-taille des villes, nous suivons la démarche initiée par Debarsy et Ertur (2010) pour spécifier les modèles SAR, SEM et SARAR.

Le modèle autorégressif spatial (SAR) sur des données de panel avec effets fixes est exprimé par l'équation :

$$Y_{nt} = \rho W_n Y_{nt} + X_{nt}\beta + \mu_n + U_{nt}$$

où $Y_{nt} = (y_{1t}, y_{2t}, ..., y_{nt})$ est le vecteur $n \times 1$ de la variable dépendante, $X_{nt} = (x_{1t}, x_{2t}, ..., x_{nt})$ le vecteur $n \times 1$ de la variable exogène, W_n la matrice de poids $n \times n$, μ_n le vecteur $n \times 1$ permettant de prendre en considération les effets fixes individuels. Le terme β désigne le coefficient de hiérarchisation et ρ le paramètre spatial autorégressif.

Le modèle à auto-corrélation spatiale des erreurs (SEM) sur données de panel peut être présenté de la manière suivante :

$$Y_{nt} = X_{nt}\beta + \mu_n + U_{nt}$$
$$U_{nt} = \lambda M_n + V_{nt}$$

avec λ le paramètre spatial autorégressif des erreurs, M_n la matrice de poids $n \times n$ et $V_{nt} = (v_{1t}, v_{2t}, ..., v_{nt})$ le vecteur des termes d'erreur. Le terme v_{it} est *i.i.d.* avec une moyenne nulle et une variance σ^2.

Enfin, les deux types d'auto-corrélation spatiale sont pris en considération dans le modèle SARAR :

$$Y_{nt} = \rho W_n Y_{nt} + X_{nt}\beta + \mu_n + U_{nt}$$
$$U_{nt} = \lambda M_n U_{nt} + V_{nt}$$

où W_n et M_n sont deux matrices spatiales $n \times n$ et λ et ρ des coefficients d'auto-corrélation spatiale à estimer.

Selon de nombreux chercheurs (Baumont et al, 2006), la définition de la matrice de poids joue un rôle fondamental dans la détection des effets d'auto-corrélation spatiale. L'utilisation de données de panel implique que le poids de la distance géographique entre les villes reste le même durant toute la période de référence.

Dans ce travail, deux matrices de poids ont été utilisées : la première est une matrice de contiguïté où la valeur des « poids spatiaux » est égale à 1 lorsque deux villes sont « voisines » et à 0, à défaut. Nous admettons que deux villes sont définies comme voisines si elles appartiennent à la même province (Chine) ou au même Etat (Inde). Ceci permet de prendre en compte les aspects institutionnels dans l'organisation hiérarchique urbaine des deux pays comme, par exemple, le fait que les migrations interprovinciales restent interdites et/ou fortement contrôlées en Chine, avant les réformes de 1993.

La deuxième est une matrice de distance définie par $w_{ij} = d_{ij}^{-1}$ avec $i \neq j$, où d_{ij} est la distance entre les villes *i* et *j*. Cette matrice est standardisée en ligne, ce qui signifie que chaque élément de la ligne *i* a été divisé par la somme totale de ses *j* éléments : $w_{ij} = w_{ij} \big/ \sum_j w_{ij}, \forall i$. Les poids spatiaux sont alors compris entre 0 et 1. Comme le soulignent Le Gallo et al. (2005), cette démarche modifie l'interprétation des poids spatiaux et met davantage l'accent sur la distance relative que sur la distance absolue entre les villes.

En suivant Debarsy et Ertur (2010), nous opérons cinq différents tests de spécification, le test joint, le test d'omission d'une auto-corrélation spatiale des erreurs, le test d'omission d'une variable spatialement décalée et deux tests conditionnels :

i. $H_0 : \rho = \lambda = 0$. Sous l'hypothèse nulle, la dimension spatiale ne joue aucun rôle dans le modèle rang-taille des villes.

ii. $H_0 : \lambda = 0$, en supposant que $\rho = 0$. Sous l'hypothèse alternative, on est en présence d'un modèle SEM.

iii. $H_0 : \rho = 0$, en supposant que $\lambda = 0$. Sous l'hypothèse alternative, on est en présence d'un modèle SAR.

iv. $H_0 : \rho = 0$, avec $\lambda \neq 0$. Sous l'hypothèse nulle, on obtient un modèle SEM, tandis que sous l'hypothèse alternative, un modèle SARAR.

v. $H_0 : \lambda = 0$, avec $\rho \neq 0$. Sous l'hypothèse nulle, on obtient un modèle SAR et sous l'alternative, un modèle SARAR.

Pour réaliser les différents tests, nous utilisons le multiplicateur de Lagrange (LM) dont l'intérêt réside dans le fait qu'il nécessite uniquement l'estimation du modèle sous hypothèse nulle, ($LM_{\lambda\rho}, LM_{\lambda}, LM_{\rho}, LM_{\rho/\lambda}, LM_{\lambda/\rho}$), ce qui facilite considérablement l'inférence statistique.

4. LA LOI DE ZIPF EN CHINE ET EN INDE SOUS LE PRISME DE L'AUTO-CORRÉLATION SPATIALE

Dans un premier temps, nous estimons le coefficient de Pareto pour les distributions rang-taille des villes chinoises et indiennes sans prendre en considération la dimension spatiale des données de panel. Le tableau 1 fournit les résultats des estimations pour chaque pays, d'une part pour l'ensemble de la période de référence (1984-2004 en Chine ; 1981-2001 en Inde), d'autre part par sous-période : avant et après les réformes de 1993 en Chine et celles de 1991 en Inde.

Le coefficient de Pareto obtenu avec le modèle rang-taille a-spatial n'est pas significativement différent de 1 pour la Chine (0,913) mais significativement inférieur à 1 pour l'Inde (0,816), lorsque l'on considère l'ensemble de la période de référence. Ceci signifie que pour la période 1984-2004, la distribution des tailles des villes chinoises valide la loi de Zipf, tandis que le système urbain indien apparait plus hiérarchisé, marqué par le poids prépondérant de quelques très grandes métropoles.

Lorsque l'on étudie les distributions des tailles urbaines par sous-période, ces conclusions doivent être amendées. En Chine, le coefficient de Pareto baisse très sensiblement durant la deuxième sous-période, ce qui traduit une hiérarchisation croissante marquée du système urbain de ce pays, suite aux réformes de 1993 et la libéralisation des migrations interprovinciales des ménages. En Inde, à l'inverse, le coefficient reste étonnamment stable durant les deux sous-périodes, ce qui impliquerait que les changements radicaux dans les orientations des politiques macroéconomiques qui s'opèrent à partir de 1991 n'ont pas de conséquences fondamentales sur les hiérarchies urbaines.

Nous introduisons, par la suite, l'hypothèse d'une auto-corrélation spatiale dans les séries des tailles urbaines chinoises et indiennes, en utilisant deux matrices de poids, une matrice de contiguïté et une matrice de distance. Les tableaux 2 et 3 résument les résultats des différents tests d'auto-corrélation spatiale qui permettent de spécifier les modèles rang-taille respectivement pour les villes chinoises et indiennes. Dans tous les cas de figure, le test joint $LM_{\lambda\rho}$ conduit à un rejet de l'hypothèse nulle et confirme la présence d'effets d'auto-corrélation spatiale dans les deux panels de données.

Tableau 1. Le modèle rang-taille pour les villes chinoises et indiennes

Chine	1984-1993	1994-2004	1984-2004
Modèle	a-spatial	a-spatial	a-spatial
β	-1.108	-0.742	-0.913
(t-stat)	(-122.779)	(-70.721)	(-130.225)
Inde	1981-1991	1991-2001	1981-2001
Modèle	a-spatial	a-spatial	a-spatial
β	-0.809	-0.835	-0,816
(t-stat)	(-122.779)	(-70.721)	(-130.225)

Pour la Chine, le test d'omission d'une auto-corrélation spatiale des erreurs LM_{λ} et le test d'omission d'une variable spatialement décalée LM_{ρ} conduisent à rejeter l'hypothèse nulle H_0 pour $\lambda = 0$ et pour $\rho = 0$ (tableau 2). Lorsque l'on utilise la matrice de distance, on obtient des valeurs plus élevées avec le test LM_{ρ} qu'avec le test LM_{λ}, ce qui permet de supposer que le modèle SEM est plus approprié pour décrire la relation rang-taille des villes chinoises.

Les tests conditionnels $LM_{\lambda/\rho}$ et $LM_{\rho/\lambda}$ permettent de spécifier la nature de l'auto-corrélation spatiale et de choisir le modèle le plus adéquat parmi les différents modèles alternatifs (SAR, SEM et SARAR). Lorsque l'on utilise la matrice de contiguïté, l'hypothèse nulle est rejetée avec les deux tests pour l'ensemble de la période de référence, ainsi que pour la période 1984–1993. Dans ce cas, le modèle SARAR, c'est-à-dire le modèle spatial autorégressif avec des erreurs spatialement auto-corrélées, apparaît comme le plus pertinent pour décrire la relation rang-taille des villes chinoises. A l'inverse, le modèle SEM semble plus adéquat pour la deuxième sous-période (1994-2004).

Lorsqu'on utilise la matrice de distance, le test $LM_{\rho/\lambda}$ n'est pas significatif. Dans ce cas, le modèle rang-taille des villes chinoises est un modèle SEM, à la fois pour l'ensemble de la période de référence et pour les deux sous-périodes avant et après les réformes de 1993.

Pour l'Inde, lorsqu'on utilise la matrice de distance, le test d'omission d'une variable spatialement décalée LM_{ρ} n'est pas significatif (l'hypothèse nulle ne peut pas être rejetée) pour la première décennie, ce qui implique que le modèle le plus adéquat est le modèle à auto-corrélation spatiale des erreurs. Les résultats des tests conditionnels $LM_{\lambda/\rho}$ et $LM_{\rho/\lambda}$ opérés pour la spécification du modèle durant la seconde décennie et pour l'ensemble de la période 1981-2001, conduisent également au choix d'un modèle SEM.

Tableau 2. Tests d'auto-corrélation spatiale dans le modèle rang-taille des villes chinoises

	W1 : matrice de distance			W2 : matrice de contiguïté		
	1984-1993	1993-2003	1984-2004	1984-1993	1993-2003	1984-2004
$LM_{\rho\lambda}$	1466.312 (0.000)	26.318 (0.000)	3442.665 (0.000)	47.015 (0.000)	2156.932 (0.000)	4423.473 (0.000)
LM_{λ}	1156.424 (0.000)	4.423 (0.035)	31.022 (0.000)	39.038 (0.000)	19.877 (0.000)	4313.099 (0.000)
LM_{ρ}	1994.91 (0.000)	26.313 (0.000)	54.522 (0.000)	20.586 (0.000)	21.542 (0.000)	63.118 (0.000)
$LM_{\rho/\lambda}$	0.174 (0.676)	1.981 (0.159)	0.507 (0.476)	14.154 (0.000)	3.362 (0.067)	21.075 (0.000)
$LM_{\lambda/\rho}$	1166.955 (0.000)	23.953 (0.000)	2395.108 (0.000)	13.400 (0.000)	1726.515 (0.000)	3491.924 (0.000)
Model	SEM	SEM	SEM	SARAR	SEM	SARAR

Avec la matrice de contiguïté, aucun test ne permet de réfuter l'hypothèse nulle H_0. Pour chaque sous-période, comme pour l'ensemble de la période de référence le modèle le plus pertinent est le modèle SARAR.

Le tableau 4 réunit les résultats des estimations des paramètres spatiaux autorégressifs λ et ρ ainsi que du coefficient de hiérarchisation β pour la Chine, calculés à travers la méthode du maximum de vraisemblance. Ces résultats montrent clairement que la prise en compte des effets d'auto-corrélation spatiale ne modifie pas, de façon fondamentale, les interprétations de l'évolution du coefficient de Pareto avant et après les réformes de 1993, obtenues par le modèle a-spatial.

En premier lieu, la distribution rang-taille des villes chinoises suit, dans le long terme, la loi de Zipf, peu importe le modèle utilisé (a-spatial, avec utilisation de la matrice de distance, avec utilisation de la matrice de contiguïté).

En second lieu, la dynamique de la distribution rang-taille des villes suit une tendance contradictoire avant et après les réformes de 1993 : le coefficient β est élevé durant la période 1984-1993, indiquant une distribution de la population urbaine entre les villes plus équitable que celle préconisée par la loi de Zipf. Ceci est la conséquence d'une politique volontariste centralisée de contrôle des migrations et de canalisation de la croissance démographique urbaine. Après les réformes de 1993 et la levée de certaines restrictions du système *Hukou*, l'accélération des flux migratoires des travailleurs vers les villes les plus dynamiques induisent des changements profonds dans la structuration urbaine du pays et un renforcement des hiérarchies urbaines. La croissance des villes de grande ou de moyenne taille (à l'exception des trois plus grandes métropoles, toujours sous le contrôle du système *Hukou*) est supérieure à celle des petites villes ce qui traduit la baisse significative du coefficient β : 0,799 (0,749) entre 1994 et 2004, contre 1,024 (1,086) durant la décennie précédente, lorsque l'on utilise la matrice de distance (ou de contiguïté).

Tableau 3. Tests d'auto-corrélation spatiale dans le modèle rang-taille des villes indiennes

	W1 : matrice de distance			W2 : matrice de contiguïté		
	1981-1991	1992-2001	1981-2001	1981-1991	1992-2001	1981-2001
$LM_{\rho\lambda}$	281.112 (0.000)	387.841 (0.000)	781.975 (0.000)	53.975 (0.000)	381.932 (0.000)	402.263 (0.000)
LM_{λ}	54.93 (0.000)	1040.287 (0.000)	1204.22 (0.000)	53.475 (0.000)	37.527 (0.000)	40.239 (0.000)
LM_{ρ}	0.454 (0.502)	133.171 (0.000)	197.16 (0.000)	7.186 (0.007)	362.832 (0.000)	6.238 (0.013)
$LM_{\rho/\lambda}$	4.671 (0.031)	2.032 (0.296)	0.027 (0.870)	6.471 (0.011)	754.965 (0.000)	6.145 (0.013)
$LM_{\lambda/\rho}$		178.145 (0.000)	2208.679 (0.000)	42.240 (0.000)	342.35 (0.000)	33.99 (0.000)
Modèle	SEM	SEM	SEM	SARAR	SARAR	SARAR

Tableau 4. Paramètres du modèle rang-taille (Chine)

	W1 : matrice de distance			W2 : matrice de contiguïté		
	1984-1993	1993-2003	1984-2004	1984-1993	1993-2003	1984-2004
Modèle	SEM	SEM	SEM	SARAR	SEM	SARAR
β	-1.024	-0.799	-0.944	-1.086	-0.749	-0.919
σ_{β}	0.098	0.076	0.090	0.101	0.071	0.086
(t-stat)	(-88.401)	(-72.842)	(-132.284)	(-97.051)	(-75.061)	(-131.825)
ρ	—	—	—	0.087	—	0.049
(t-stat)	—	—	—	(2.308)	—	(-1.106)
λ	0.773	0.981	0.816	0.292	0.800	0.790
(t-stat)	(18.077)	(25.939)	(42.066)	(3.276)	(24.081)	(29.971)

Comme on peut le constater dans le tableau 4, les effets de voisinage sont renforcés dans la croissance urbaine chinoise, après les réformes de 1993.

Dans le modèle SEM, valable dans le cas chinois lorsque l'on utilise la matrice de distance, la dépendance spatiale est liée uniquement aux effets de diffusion spatiale. Un choc exogène qui affecte la taille démographique d'une ville affecte également la taille démographique de toutes les autres villes, mais de façon inversement proportionnelle à la distance qui les sépare de celle-ci. Les effets d'un tel choc exogène diminuent, donc, au fur et à mesure que l'on s'éloigne de la ville où il s'est produit. Le tableau 4 montre que lorsque l'on utilise le modèle SEM, l'intensité de l'auto-corrélation spatiale des résidus passe de 0,773 à 0,981 entre la première et la seconde sous-période, ce qui signifie que les effets de diffusion spatiale des chocs exogènes localisés sur la démographie urbaine se renforcent sensiblement après 1993.

Tableau 5. Paramètres du modèle rang-taille (Inde)

	W1 : matrice de distance			W2 : matrice de contiguïté		
	1981-1991	1991-2001	1981-2001	1981-1991	1991-2001	1981-2001
Modèle	SEM	SEM	SEM	SARAR	SARAR	SARAR
β	-0.934	-1.113	-1.070	-0.932	-1.160	-1.067
σ_β	0.094	0.107	0.101	0.089	0.110	0.096
(t-stat)	(-105.29)	(-504.76)	(-95.12)	(-112.57)	(-468.45)	(-97.62)
ρ	—	—	—	0.0527	0.171	0.041
(t-stat)	—	—	—	(2.77)	(12.24)	(3.39)
λ	0.975	0.443	0.551	0.281	1.001	0.414
(t-stat)	(287.12)	(25.93)	(12.89)	(4. 38)	(56.78)	(5.414)

Dans le modèle SARAR, obtenu quand on privilégie la matrice de contiguïté, l'interprétation de la diffusion spatiale des effets des chocs exogènes aléatoires sur la démographie des villes est plus complexe, car elle doit prendre en compte l'omission de variables déterminantes, en particulier lorsque celles-ci sont spatialement corrélées (Fingleton, 1999 ; Ertur et Thiaw, 2005). Il convient de signaler que les valeurs de β obtenues avec le modèle SARAR sont très proches de celles obtenues avec le modèle a-spatial.

De façon générale, la détection des effets d'auto-corrélation spatiale dans la distribution rang-taille des villes chinoises illustre l'existence de clusters de villes, géographiquement proches à un niveau intra ou interprovincial, dont la démographie évolue de façon parallèle, surtout après 1993 et la libéralisation des mouvements migratoires. Les corrections apportées à l'estimation du coefficient de Pareto lorsqu'on tient compte de ces effets permettent une meilleure adéquation entre la distribution rang-taille des villes et la loi de Zipf, du moins avec l'utilisation d'une matrice de distance.

L'introduction d'effets d'auto-corrélation spatiale dans le modèle rang-taille des villes indiennes permet également de corriger certaines interprétations obtenues à partir du modèle a-spatial. Le tableau 5 fournit les résultats des estimations des paramètres spatiaux autorégressifs λ et ρ et du coefficient de hiérarchisation β, calculés à travers la méthode du maximum de vraisemblance. Il convient, néanmoins de signaler qu'on ne distingue pas, ici, les effets directs liés aux effets de voisinage de premier ordre et les effets indirects (Lesage et Pace, 2009), ce qui devrait introduire une certaine prudence dans les interprétations formulées.

En premier lieu, avec l'introduction de la dimension spatiale des données, le coefficient de Pareto emprunte des valeurs beaucoup plus proches de l'unité, peu importe la matrice de poids utilisée. Ceci signifie que la prise en considération des effets d'auto-corrélation spatiale permet de valider la loi de Zipf pour les villes, contrairement aux conclusions établies avec le modèle a-spatial.

En second lieu, comme avec le modèle a-spatial, le coefficient β augmente après 1991, ce qui traduit une diffusion de la croissance urbaine indienne et une distribution plus égalitaire des tailles urbaines. Bien que les différences entre les deux sous-périodes soient plus marquées avec les deux modèles SEM et SARAR qu'avec le modèle a-spatial, force est de constater que l'évolution du coefficient est assez faible durant la période de référence, ce qui permet de conclure en faveur d'une certaine stabilité des hiérarchies urbaines entre 1981 et 2001. Les réformes d'ordre macro-économique de 1991 n'ont pas d'incident sur la structuration du système urbain indien.

Enfin, en troisième lieu, lorsque l'on utilise la matrice de distance, les effets d'auto-corrélation spatiale semblent progressivement s'amenuiser, avec le paramètre spatial autorégressif λ qui diminue fortement entre les deux sous-périodes. Ceci signifie que la dimension spatiale des données joue un rôle de moins en moins important dans la structuration des hiérarchies urbaines indiennes, ce qui pourrait être expliqué par une dissociation progressive des performances de croissance démographique de chaque ville de celles de son Etat d'appartenance (Schaffar, 2010).

CONCLUSION

Cet article utilise un modèle rang-taille des villes sur des données de panel. En introduisant la dimension géo-localisée des données, il propose de prendre en compte les effets d'auto-corrélation spatiale dans l'étude de la validité de la loi de Zipf. La méthodologie utilisée pour détecter les effets d'auto-corrélation spatiale est celle initiée par Debarsy et Ertur (2010). Elle permet de choisir la meilleure spécification du modèle rang-taille parmi trois alternatives : le modèle autorégressif spatial (SAR), le modèle à auto-corrélation spatiale des erreurs (SEM) et le modèle spatial général à effets fixes sur des données de panel (SARAR).

Ce travail est appliqué sur les distributions des tailles des villes chinoises et indiennes entre 1981 et 2005. Durant cette période, les deux pays ont connu des mutations démographiques, économiques et institutionnelles importantes. En Chine, en 1993, des réformes importantes alègent le contrôle des déplacements et des migrations internes et libéralisent les processus de croissance démographique urbaine. En Inde, l'année 1991 marque un virage important dans l'orientation des choix macro-économiques du pays, avec une libéralisation des marchés des capitaux et du travail et une ouverture à l'internationalisation. Dans cet article, nous examinons les distributions rang-taille des villes pour l'ensemble de la période de référence, mais aussi par sous-période : avant et après 1993 pour la Chine, avant et après 1991 en Inde, afin d'étudier les impacts de ces changements et réformes sur leurs hiérarchies urbaines. L'introduction de la dimension spatiale se fait à travers l'utilisation de deux matrices de poids : une matrice de contiguïté et une matrice de distance.

De façon globale, trois séries de conclusions peuvent être tirées :

En premier lieu, sur un plan méthodologique, l'utilisation d'un modèle rang-taille avec des données de panel fournit des résultats plus robustes quant à l'interprétation des hiérarchies urbaines que les modèles rang-taille classiques qui portent sur une seule série de données annuelles. L'introduction de la dimension spatiale des données permet d'affiner ce modèle et fournir une estimation plus pertinente du coefficient de hiérarchisation.

En second lieu, en Chine, la prise en compte des effets d'auto-corrélation spatiale ne modifie pas fondamentalement les conclusions du modèle a-spatial tout en permettant de l'affiner : la distribution rang-taille des villes chinoises suit la loi de Zipf dans le long terme, tandis que lorsque l'on raisonne par sous-période, le coefficient de Pareto baisse sensiblement après 1993, ce qui traduit une hiérarchisation croissance du système urbain chinois. En Inde, à l'inverse, l'introduction de la dimension spatiale des données conduit à valider la loi de Zipf sur le long terme, contrairement aux conclusions du modèle a-spatial. Néanmoins, dans les deux modèles, le coefficient de hiérarchisation entre les deux sous-périodes reste relativement stable, ce qui signifie que les changements obtenus en 1991 n'interviennent pas de façon significative dans la détermination des hiérarchies urbaines.

Enfin, en troisième lieu, l'intensité des effets d'auto-corrélation spatiale évolue différemment dans chaque pays : en Chine, ces impacts se renforcent après 1993 et conduisent à l'hypothèse d'une formation de clusters spatiaux de villes qui évoluent de façon parallèle du point de vue démographique. En Inde, à l'inverse, ces impacts s'amenuisent, ce qui peut être interprété comme une désindexation des processus de croissance urbaine de chaque ville et des évolutions démographiques de son Etat d'appartenance. Ceci pourrait traduire l'émergence d'un processus long de substitution des migrations interétatiques aux migrations intra-étatiques qui prévalaient jusqu'à la fin du vingtième siècle dans ce pays.

Ce travail s'inscrit dans une démarche originale d'introduction de la dimension spatiale des données dans l'étude des modèles rang-taille des villes. Il élimine, de ce fait, un biais important dans l'examen de la loi de Zipf, lié à l'auto-corrélation spatiale des tailles urbaines, que les nombreuses études sur ce sujet ont curieusement omis d'étudier. L'étape suivante de cette démarche exploratoire passe par la prise en compte de la dimension spatiale dans la croissance urbaine, permettant, par là, d'étudier les effets dynamiques de l'auto-corrélation spatiale sur la trajectoire démographique des villes.

REFERENCES

Anselin L., 1998, *Spatial econometrics: methods and models*. Kluwer Academic, The Netherlands.

Anselin L., Le Gallo, J, Jayet, J., 2008. *Spatial panel econometrics. The econometrics of panel data: fundamentals and recent developments in theory and practice*, Springer, Berlin Heidelberg.

Auerbach, F., 1913, Das Gesetz der Bevölkerungskonzentration, *Regional Science and Urban Economics* 31, 601-615.

Baumont C., Ertur C., Le Gallo J., 2006, Clubs de convergence et effets de débordements géographiques : une analyse spatiale sur données régionales européennes, 1980-1995, *Economie et prévision*, 173(2), 111-134.

Black D., Henderson J.V., 1999, A Theory of Urban Growth, *Journal of Political Economy*, 107, 252-284.

Chasco C., Le Gallo J., 2008, Spatial analysis of urban growth in Spain, 1900-2001, *Empirical Economics*, Vol. 34(1), 59-80.

Chinese Urban Statistical Yearbooks, Chinese Statistical Office, Beijing.

Debarsy N., Ertur C., 2010, Testing for spatial autocorrelation in a fixed effects panel data model, *Regional Science and Urban Economics*, 40, 453-470.

Eeckhout J., 2004, Gibrat's Law for (all) Cities, *The American Economic Review*, 94(5), 1429-1451.

Eeckhout J., 2009, Reply to Levy M. on Gibrat's Law for (all) Cities, *American Economic Review*, 99, 1676-1683.

Ertur C., Thiaw K., 2005, Growth and Spatial Dependence - The Mankiw, Romer and Weil model revisited, communication au colloque de la European Regional Science Association.

Elhorst J., 2010, Spatial panel data models, In Fischer M.M., Getis A. (Eds.), *Handbook of Applied spatial analysis*, Springer, 377-407.

Fingleton, B., 1999, Estimates of time to economic convergence: an analysis of regions of the European Union, *International Regional Science Review*, 22, 5-34.

Gabaix X., 1999, Zipf's Law for Cities: an Explanation, *Quarterly Journal of Economics*, 114, 739-767.

Gabaix, X., Ibragimov, R., 2011, RANK-1/2: A simple way to improve the OLS estimation of tail exponents, *Journal of Business & Economic Statistic*, 29(1), 24-31.

Gabaix, X., Ioannides, Y., 2004, The evolution of city sizes' distribution, in Henderson J.V et Thisse J-F. (eds) *Handbook of regional and urban economics*, vol.4, Elsevier Science B.B, Amsterdam, 2341-2376.

Ioannides Y., Scouras S., 2009, Gibrat's law for all cities: a rejoinder, working paper, Tufts University.

Krugman P., 1996, Confronting the Mystery of Urban Hierarchy, *Journal of the Japanese and the International Economies*, 10, 399-418.

Lee L., Yu J., 2010, Some recent developments in spatial panel data models, *Regional Science and Urban Economics*, 40, 255-271.

Le Gallo J., Baumont C., Dall'Erba S. Ertur C., 2005, On the property of diffusion in the spatial error model, *Applied Economics Letters*, vol. 12(9), 533-536.

Lesage J., Pace R.K., 2009, *Introduction to spatial econometrics*, CRC Press, Taylor and Francis, 331 p.

Levy M., 2009, Comment on Eeckhout J., 2004, Gibrat's Law for (all) Cities, *American Economic Review*, 99, 1672-1675

Moriconi-Ebrard F., 1993, *L'urbanisation du monde depuis 1950*, Paris, Anthropos.

Nitsch V., 2005, Zipf zipped, *Journal of Urban Economics*, n°57, 86-100.

Parr J., 1985, A note on the size distribution of cities over time, *Journal of Urban Economics*, vol.18, 199–212.

Rey S., Ye X., 2007, Spatial Dynamics in City Size Distribution, The 2nd ICA Workshop on Geospatial Analysis and Modeling, online proceedings.

Rosen, K., Resnick, M., 1980, The size distribution of cities: an examination of the Pareto low primacy, *Journal of Urban Economics*, 8, 165-186.

Schaffar A., 2010, Quelle est la nature de la croissance urbaine indienne ?, *Revue d'Economie du Développement*, 2, 101-120.

Schaffar A., Dimou M., 2012, Rank size dynamics in India and in China: 1981-2004, *Regional Studies*, 46(6), 707-721.

Soo K.T., 2005, Zipf's Law for cities: a cross-country investigation, *Regional Science and Urban Economics*, vol.35, 239-263.

Xu Z, Harriss R, 2010 A Spatial and Temporal Autocorrelated Growth Model for City Rank--Size distribution, *Urban Studies*, 47(2), 321-335

Zipf G.K., 1949, *Human Behavior and the Principle of Least Effort*, Addison-Welsey, Cambridge, MA.

ZIPF'S LAW UNDER SPATIAL AUTOCORRELATION. THE CASE OF INDIA AND CHINA

Abstract: *This paper applies a rank-size model on panel data. By introducing the spatial dimension of data, it aims in taking into account spatial autocorrelation effects on the Zipf law. We apply Debarsy and Ertur (2010) methodology in order to choose the best specification between three alternative models: the spatial autoregressive model (SAR), the spatial error model (SEM) and the general spatial model with fixed effects on panel data (SARAR). Taking into account spatial autocorrelation effects does not modify the conclusions from the a-spatial model for China – the Zipf's law holds but do modify conclusions for India: in this country the a-spatial model doesn't confirm the Zipf's law on the long run, while this changes when we introduce spatial autcorrelation.*

Key-words: ZIPF'S LAW, URBAN HIERARCHIES, SPATIAL AUTOCORRELATION, CHINA, INDIA.