# La loi de Zipf est-elle pertinente ? Un examen méthodologique

**Alexandra Schaffar**

IREMIA - Université de la Réunion

LEAD - Université de Toulon – Var

schaffar@univ-reunion.fr

*VERSION PROVISOIRE*

Papier présenté au XLVème colloque de l'ASRDLF

Rimouski, Canada - Août 2008

## Introduction. La loi de Zipf : questions méthodologiques

La loi de Zipf caractérise la distribution de la taille des villes et conditionne, aujourd'hui, toute tentative d'interprétation des mécanismes économiques qui régissent la formation et l'évolution des hiérarchies urbaines d'un pays ou d'une région. Qualifiée jadis de « *mystère urbain* » par Krugman (1996, p.40), la loi de Zipf est, selon Gabaix et Ioannides (2004), « *un des faits les plus frappants en économie et en sciences sociales en général* » (Gabaix et Ioannides, 2004, p.739).

Le travail initial de Zipf (1941) applique la loi de Pareto sur les sciences du langage en ayant, comme support, l'œuvre magistrale de James Joyce, *Ulysse*. Cependant, rapidement, l'auteur étend son travail sur d'autres domaines, tels que les hiérarchies urbaines. En utilisant la base de données de Wickens (1921), Zipf étudie, alors, la distribution rang taille de 256 villes australiennes de plus de 3000 habitants en 1921 (Zipf, 1949, p.137).

Selon Zipf, si $X$ est la variable qui associe à chaque ville sa population, la fonction densité de $X$ suit une loi de Pareto, en épousant la forme :

$$f(x) = Cx^{-a} \qquad (1)$$

où $f(x)dx$ correspond au nombre de villes avec une population comprise entre $x$ et $x + dx$, $C$ est une constante indiquant la taille de la plus grande ville du pays et $a$ un degré de hiérarchisation. La fonction de répartition complémentaire associée à $X$ est :

$$F_{>}(x) = \Pr(X > x) = \int_x^{\infty} f(t)dt = \int_x^{\infty} Ct^{-a}dt$$

$$= \frac{Cx^{-a+1}}{a-1} \quad avec \quad a > 1 \qquad (2)$$

$\Pr(X > x)$ est la probabilité qu'une ville ait une population supérieure à $x$. Si les villes sont rangées, selon leur taille, de la façon suivante : $x_1 > x_2 > x_3 > \ldots > x_i > \ldots > x_n$, avec $r(x)$ le rang de la taille de la ville $x$, on obtient :

$$\Pr(X > x) = \frac{r(x)}{n} \qquad (3)$$

En tenant compte de la relation (2), on peut alors écrire l'égalité suivante :

$$\frac{r(x)}{n} = \frac{C}{a-1} x^{-a+1} \text{ avec } a \succ 1 \qquad (4)$$

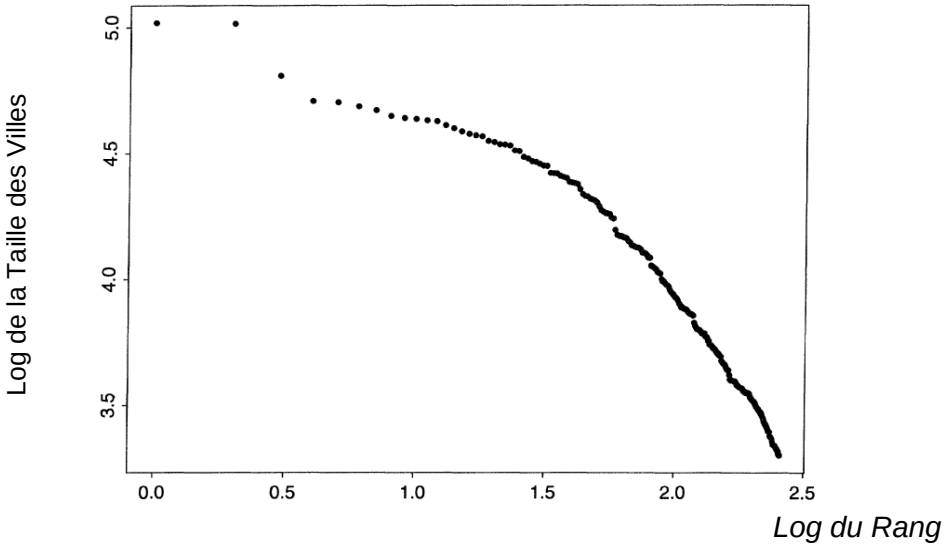
En posant $k = \frac{cn}{a-1}$ et $\beta = a - 1$ (avec $\beta > 0$), on obtient :

$$r(x) = k \cdot x^{-\beta} \qquad (5)$$

avec *k* un paramètre qui dépend de la taille de la plus grande ville et $\beta$ 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 :

$$\ln r(x) = -\beta \ln x + \ln k \qquad (6)$$

Lorsque $\beta$ est égal à 1, on obtient la loi de Zipf, représentée dans la figure 1.




Source : Zipf (1949), schéma reproduit par Feuerverger et Hall (1999)

*Figure 1 : La relation entre le logarithme du rang et le logarithme de taille des villes pour $\beta = 1$ (Loi de Zipf)*

Comme Pareto, Zipf trouve une vague justification de la régularité de cette loi dans la nature humaine qui conditionne les choix de vie et de localisation des individus. Il est intéressant, cependant, de noter que, dans son étude initiale sur l'Australie, Zipf conclut que la distribution des villes australiennes dévie de la distribution de Pareto, notamment à cause de l'allure de la queue de la distribution. Il insiste, d'ailleurs, sur l'intérêt sociologique que représente un tel cas par rapport aux distributions parétiennes.

A travers cette observation, Zipf introduit d'emblée les problèmes de biais liés à l'échantillonnage. Dans le cas australien, Zipf considère qu'en retenant, dans un même échantillon, à la fois les communautés urbaines et rurales, on obtient une distribution composite, issue de la combinaison de deux distributions de Pareto, d'où la déviance de sa queue (Zipf, 1949, p.423, voir aussi la généralisation de la loi de Zipf par Mandelbrot, 1955).

La question de l'échantillonnage et de la borne inférieure (taille de ville minimale) utilisée, va alimenter un débat animé parmi les économistes et les statisticiens sur la véritable nature de la distribution de la taille des villes. En considérant l'asymétrie qui caractérise cette distribution, certains auteurs supposent une déviation des distributions réelles vis-à-vis de la loi de Zipf (Rosen et Resnick, 1980), tandis que d'autres émettent l'hypothèse que cette loi est une construction fictive liée essentiellement à la procédure d'échantillonnage tandis que la distribution des tailles des villes, lorsque l'on considère la totalité de la population urbaine, suit une loi lognormale (Malecki, 1980, Parr, 1985, Eeckout, 2004, Anderson et Ge, 2005) ou une double Pareto (Reed, 2001). Sur un plan méthodologique, la question de l'échantillonnage est la première menace qui pèse sur la loi de Zipf.

La seconde menace est celle de la méthode de calcul du coefficient de Pareto. L'immense littérature sur la distribution rang taille des villes, avec ses comparaisons internationales et diachroniques, s'appuie fortement sur la valeur de ce coefficient pour atteindre un ensemble de conclusions sur la nature des hiérarchies urbaines dans les différents pays et régions. Or, depuis quelques années, un ensemble de chercheurs montre l'extrême volatilité des résultats obtenus, en fonction de la méthode de calcul utilisée. Dans la littérature récente sur la loi de Zipf, certains auteurs (Gabaix et Ioannides, 2004, Nishiyama et Osada, 2004, Soo, 2005) comparent les résultats obtenus par différents estimateurs, afin de déterminer le sens de leur biais et proposer une correction adéquate. Il n'y a pas, cependant, d'étude qui propose une comparaison efficace de l'ensemble des méthodes d'estimation connues.

Ce papier prétend combler ce besoin, tout en abordant également la première question méthodologique, liée à l'échantillonnage. En proposant une simulation Monte Carlo sur 20000 échantillons, son objectif est double : en premier lieu, mettre en évidence le biais de chaque méthode d'estimation, selon la taille de l'échantillon utilisée ; en second lieu, proposer une comparaison des différentes méthodes afin de sélectionner la plus pertinente d'entre elles.

La première partie du papier propose une revue des principales méthodes d'estimation, ainsi que des corrections qui ont été apportées par différentes contributions théoriques, durant ces dernières années. La seconde partie contient les résultats de la simulation Monte-Carlo effectuée.

# 1. Une revue des différentes méthodes d'estimation du coefficient de hiérarchisation de la loi de Zipf

De façon générale, deux grands types d'estimateurs se distinguent, ceux fournis par la méthode des Moindres Carrés Ordinaires (MCO) et ceux fournis par les méthodes semi-paramétriques dont la plus connue est celle de Hill (1975).

La section 1.1 explore l'efficacité de l'estimation du coefficient de hiérarchisation des MCO, tant pour le modèle de Zipf que celui de Lotka. La section 1.2 propose une correction du biais de l'estimateur pour les échantillons de petite taille, soit par le modèle de Gabaix et Ibragimov (2006) soit par les Moindres Carrés Généralisés (MCG). La section 1.3 aborde les problèmes des valeurs extrêmes et des queues épaisses des différentes distributions et présente les solutions apportées par les méthodes d'estimation semi-paramétrique du coefficient de hiérarchisation, notamment celle de Hill.

## 1.1 Estimation du coefficient de Pareto par la méthode des Moindres Carrés Ordinaires (MCO)

Le plus grand nombre d'études sur la loi de Zipf cherche à calculer le coefficient de hiérarchisation, en utilisant le modèle log-log du rang en fonction de la taille que l'on peut rappeler :

$$\ln R_i = \ln A - \beta \ln T_i \qquad (7)$$

Lorsque le coefficient de hiérarchisation $\beta$ est inférieur à 1, l'effet agglomération est renforcé et les villes de grande taille ont un poids plus important que dans une distribution qui suit la loi de Zipf. A l'inverse, si $\beta$ est supérieur à 1, on est en présence d'un espace polycentrique où le nombre des villes moyennes et plus important que dans une distribution rang taille des villes conforme à la loi de Zipf.

De nombreux chercheurs préfèrent à ce modèle, le modèle correspondant de Lotka qui met en relation la taille en fonction du rang. L'intérêt de cette formulation tient tout simplement sur le fait que, contrairement au modèle précédent, la valeur du coefficient de hiérarchisation de Lotka $\gamma$ augmente lorsque le degré de hiérarchisation (le poids des grandes villes) augmente et vice-versa.

$$\ln T_i = \ln C - \gamma \ln R_i \qquad (8)$$

Lorsque la distribution de la taille des villes suit parfaitement une loi de Pareto, $\gamma$ est l'inverse du coefficient de Pareto $\beta$, tandis que $C = A^{\gamma}$ est un indicateur de la taille de la plus grande ville du système. L'augmentation de $C$, observée durant une grande partie du 20ème siècle traduit un processus généralisé d'urbanisation croissante et de concentration des populations urbaines dans les plus grandes villes. Cependant, depuis le début des années quatre-vingt-dix, dans de nombreux

systèmes urbains, *C* a tendance à stagner, voire baisser, en partie à cause des effets de congestion observés dans les grandes capitales qui conduisent à une migration relative de la population vers des centres urbains secondaires dont le taux de croissance démographique s'accélère (Ades et Glaeser, 1995 ; Dobkins et Ioannides, 2000).

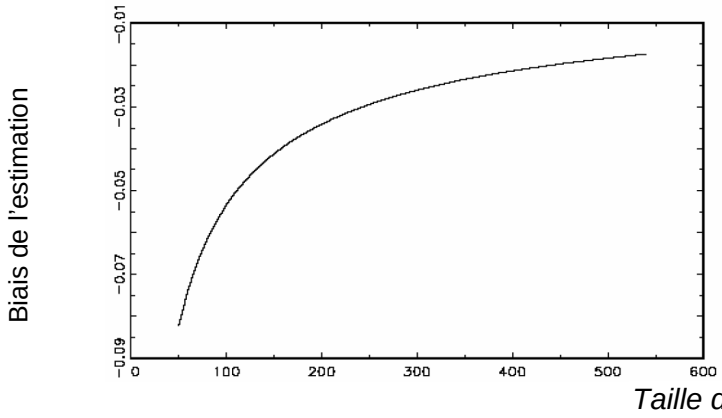
Malgré la grande popularité de ces modèles, plusieurs auteurs accusent le biais des estimateurs du coefficient de hiérarchisation, notamment pour les échantillons de petite taille. Gabaix et Ioannides (2004) construisent une population de villes qui obéit parfaitement à la loi de Zipf, puis, par des simulations Monte Carlo, proposent des estimations pour des échantillons de taille $n$ différente ($n$ = 20, 50, 100,…, 500). Pour les petits échantillons, Gabaix et Ioannides (2004) trouvent un coefficient de hiérarchisation systématiquement inférieur à 1 (qui est la valeur réelle du coefficient de la distribution construite).

La valeur de ce coefficient augmente au fur et à mesure que la taille de l'échantillon augmente (dans l'exemple des deux auteurs, $\hat{\beta} = 0{,}90$ pour $n = 20$, $\hat{\beta} = 0{,}94$ pour $n = 100$, mais $\hat{\beta} = 0{,}98$ pour $n = 500$). La méthode des MCO a ainsi tendance à sous-estimer la valeur du coefficient de hiérarchisation pour les petits échantillons, ce qui conduit à sur-dimensionner la taille de la plus grande ville du système.

Un deuxième souci dans l'estimation du coefficient de hiérarchisation par les MCO réside dans le calcul de sa variance. Gabaix et Ioannides (2004) montrent que la différence des valeurs entre l'écart-type calculé et l'écart-type réel peut être très importante pour les petits échantillons, les premières étant nettement inférieures aux secondes (dans la simulation de Gabaix et Ioannides, pour $n = 100$, l'écart-type calculé $\sigma\left(\hat{\beta}\right)$ est de 0,013, quand l'écart-type réel $\sigma(\beta)$ est de 0,13, ce qui signifie qu'à un risque d'erreur de 5%, l'intervalle de confiance estimé par les MCO [0,974 ; 1,026] est beaucoup plus restreint que l'intervalle réel [0,68 ;1,20]. En sous-estimant les écarts à la moyenne, la méthode des MCO conduit, ainsi, à un rejet exagéré de la loi de Zipf.

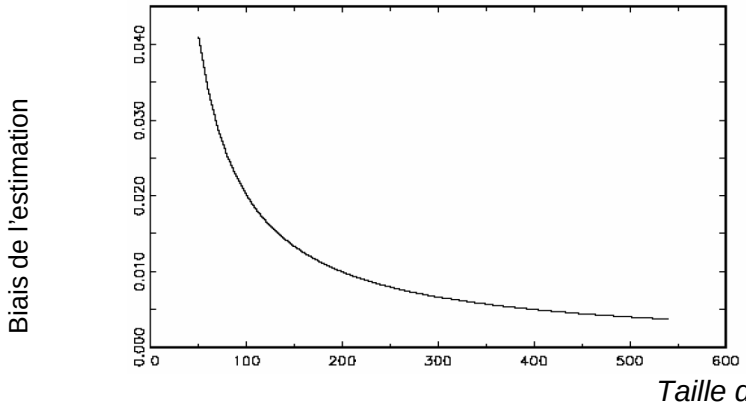
Selon Nishiyama, Osada et Morimune (2004), le biais dans l'estimation par les MCO est lié au processus de classement des villes par rang, c'est-à-dire au fait que l'on construit la variable expliquée $R$ , en ordonnant la variable explicative $T$ , ce qui est à l'origine de l'apparition de corrélations positives entre les résidus supposés indépendants. Par les résultats de leurs simulations Monte Carlo, Gan, Li et Song (2006) émettent même l'hypothèse que la loi de Zipf n'est qu'un faux phénomène statistique lié à cette procédure de classement.

Les figures 2 et 3 montrent le degré du biais de l'estimation du coefficient de hiérarchisation et de sa variance par la méthode des MCO, en fonction de la taille de l'échantillon, pour une distribution qui suit une loi de Pareto (Nishiyama, Osada et Morimune, 2004).




Source : Nishiyama, Osada et Morimune (2004)

*Figure 2 : Le biais de l'estimateur du coefficient de Pareto par la méthode des MCO*




Source : Nishiyama, Osada et Morimune (2004)

*Figure 3 : Le biais de la variance du coefficient estimée par les MCO*

Ces graphiques montrent, de façon claire, que le biais de l'estimation diminue au fur et à mesure que la taille de l'échantillon augmente. Gabaix et Ioannides (2004) soutiennent que l'estimation du coefficient de hiérarchisation dans le modèle utilisé par Zipf présente un biais (vers le bas) légèrement inférieur que celui (vers le haut) obtenu pas la méthode de Lotka.

### 1.2 La correction du biais de l'estimation par les Moindres Carrés Ordinaires

De nombreux auteurs ont essayé d'apporter quelques corrections au biais de l'estimation du coefficient de Pareto par la méthode des MCO. En s'appuyant sur la généralisation du modèle de Zipf par Mandelbrot, Gabaix et Ibragimov (2006) considèrent la relation suivante :

$$\ln(R_i - \xi) = \ln A - \beta \ln T_i \qquad (9)$$

avec $0 \leq \xi < 1$ et où $\xi = 0$ correspond au modèle de Zipf. En s'appuyant sur les démonstrations de Kratz et Resnick (1996) et de Csörgö et Viharos (1997), Gabaix et Ibragimov (2006) montrent que l'espérance d'estimation du coefficient de hiérarchisation $\hat{\beta}$, compte tenu de la valeur du coefficient réel $\beta$, obéit à :

$$\begin{aligned} &E\left((\hat{\beta}/\beta) - 1\right) = \frac{(2\gamma - 1)\ln^2 n}{4n} + u\left(\frac{(\ln n)^2}{n}\right) \\ &\Rightarrow \hat{\beta}/\beta = 1 + \sqrt{\frac{2}{n}} N(0,1) + \frac{(\ln n)^2 (2\xi - 1)}{4n} + u\left(\frac{(\ln n)^2}{n}\right) \end{aligned} \qquad (10)$$

Il est, alors, aisé de montrer que la meilleure estimation de $\beta$ est fournie lorsque $\xi = 1/2$. La relation rang taille des villes, pour laquelle le biais de l'estimation du coefficient pour des petits échantillons disparaît, prend alors la forme :

$$\ln\left(R_i - \frac{1}{2}\right) = \ln C - \hat{\beta} \ln T_i \qquad (11)$$

tandis que son écart-type peut être estimé par :

$$\sigma(\hat{\beta}) = \sqrt{\frac{2}{n}}\hat{\beta} \qquad (12)$$

En s'appuyant sur des simulations Monte-Carlo, Gabaix et Ibragimov (2006) montrent que le modèle Rang $(-1/2)$ fournit la meilleur approximation du coefficient de hiérarchisation $\beta$, y compris pour des échantillons de petite taille, lorsque l'on utilise la méthode des MCO.

De leur côté, en travaillant sur le modèle de Lotka, Nishiyama et Osada (2004) proposent une correction de l'estimation du coefficient de hiérarchisation de Lotka, en le multipliant par une constante qui ne dépend que du nombre de villes *n* retenues dans l'échantillon :

$$\bar{\gamma} = \frac{n\sum (\ln i)^2 - (\sum \ln i)^2}{n\sum_{i=1}^{n} \ln i\left(\frac{1}{n} + \ldots \frac{1}{i} - 1\right)}\hat{\gamma} \qquad (13)$$

Comme la constante multiplicative est inférieure à 1 en valeur absolue, cette correction réduit fortement le biais de l'estimation du coefficient par les MCO.

Par la suite, Nishiyama, Osada et Morimune (2004) proposent une amélioration de l'estimation du coefficient d'hiérarchisation, dans le modèle de Lotka, en utilisant la méthode des Moindres Carrés Généralisés (MCG). Nishiyama et Osada (2004) et Nishiyama, Osada et Sato (2006) fournissent une présentation rapide de la méthode des MCG appliquée à la distribution rang taille des villes. Si :

$$Y' = [\ \ln T_1 \quad \ln T_2 \quad \cdots \quad \ln T_n\ ] \qquad (14)$$

$$X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \ln 1 & \ln 2 & \cdots & \ln n \end{bmatrix} \qquad (15)$$

et

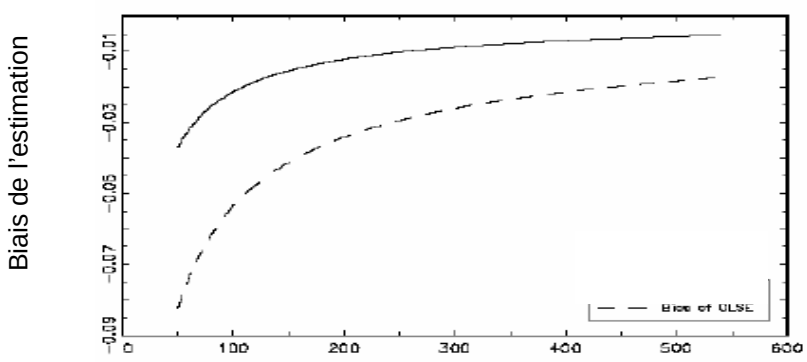
$$\Omega = Var(Y) \qquad (16)$$

Les estimations de $\gamma$ et A ainsi que leurs variances sont égales à :

$$\begin{bmatrix} \widehat{A} \\ \widehat{\gamma} \end{bmatrix} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y \qquad (17)$$

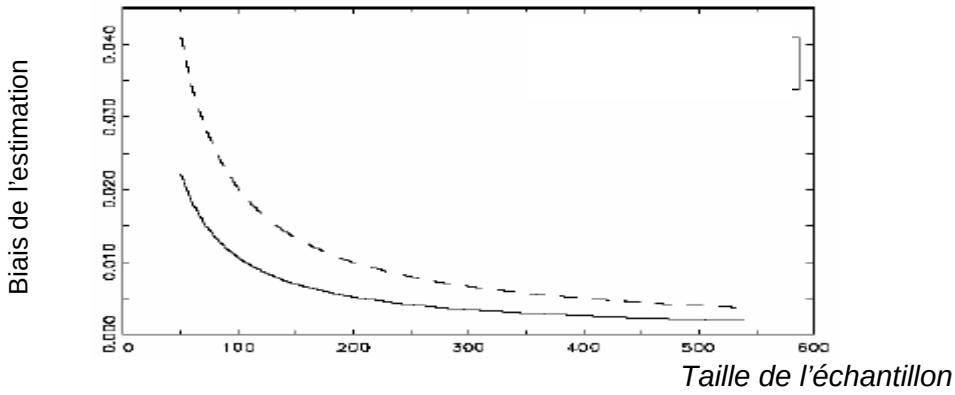
$$Var\begin{bmatrix} \widehat{A} \\ \widehat{\gamma} \end{bmatrix} = (X'\Omega^{-1}X)^{-1} \qquad (18)$$

L'estimation du coefficient de hiérarchisation (de Lotka) par la méthode des MCG réduit considérablement le biais observé dans l'estimation par les MCO, comme on peut le constater dans les figures 4 et 5.




Source : Nishiyama, Osada et Morimune (2004)

*Figure 4 : Comparaison du biais de l'estimation du coefficient de hiérarchisation $(\gamma)$ par la méthode des MCO et des MCG*




Source : Nishiyama, Osada et Morimune (2004)

*Figure 5 : Comparaison du biais de l'estimation de la variance du coefficient de hiérarchisation par la méthode des MCO et des MCG*

L'estimation par les MCG est identique à celle obtenue par la méthode du Maximum de Vraisemblance, lorsque les erreurs sont indépendantes et leur distribution suit une Loi Normale (Charnes, Frome et Yu, 1976).

### 1.3 L'estimateur de Hill

Contrairement à d'autres types de distribution (binomiale, Normale, Poisson), une distribution qui suit une loi de Zipf se caractérise par une probabilité moins faible d'apparition de valeurs extrêmes, ce qui conduit à la formation d'une « queue épaisse » qui peut avoir un comportement spécifique par rapport au reste de la distribution. C'est le cas de la distribution rang taille des villes où, selon la valeur minimale de la taille de ville considérée (borne inférieure), on peut se retrouver avec

un échantillon qui affiche des valeurs extrêmes (par exemple, le ratio entre New York, la plus grande ville des Etats-Unis avec 8 millions d'habitants et Duffield, Virginia, avec 52 habitants est de $1/150000$).

Malgré le fait qu'il n'y a pas de consensus établi, parmi les statisticiens, quant à la définition *stricto sensu* d'une distribution à queue épaisse (Werner et Upper, 2002), on peut considérer que si $X$ est une variable aléatoire, $\mu_X$ et $\sigma_X$ respectivement sa moyenne et son écart-type, la distribution de $X$ est une distribution à queue épaisse si :

$$E\left[\frac{(X - \mu_X)^4}{\sigma_X}\right] > 3 \qquad (19)$$

Malakarne et al. (2001) identifient le problème des queues épaisses à celui du « bornage » de la distribution rang taille des villes : en premier lieu, lorsque l'on procède à des comparaisons internationales, les bornes inférieures ne peuvent pas être les mêmes, pour chaque pays. A titre d'exemple, les villes avec une population supérieure à 100 000 habitants représentent 15% des villes américaines, mais seulement 4% des villes brésiliennes. En second lieu, lorsque l'on augmente considérablement la taille de l'échantillon, en incluant les plus petites villes, une déviation peut apparaître vis-à-vis de la loi de Pareto.

La combinaison de ces deux problèmes fait que la loi de Zipf est vérifiée, le plus souvent, pour des échantillons d'une taille donnée et dès lors que l'on s'éloigne de cette taille « idéale » (en augmentant ou en réduisant le nombre de villes), on s'éloigne de la loi de Zipf. C'est aussi la raison pour laquelle la loi de Zipf a tendance à se confirmer plus facilement pour des pays de grande taille qui possèdent de nombreuses grandes villes, par exemple les Etats-Unis, que les pays à taille plus réduite, par exemple de nombreuses nations européennes (Candeal et al, 2001).

La théorie des valeurs extrêmes essaye d'apporter des réponses au problème d'estimation du coefficient de distributions avec des queues épaisses et cherche à comprendre le comportement des maxima ou minima d'une variable aléatoire sur une période de temps fixe (pour une revue de littérature complète sur ces travaux, voir Embrechts, Kluppelberg et Mikosch, 1997, et, plus récemment, Stoev, Michailidis et Taqqu, 2008).

Afin d'améliorer le calcul de la pente des distributions avec une queue épaisse, de nombreux chercheurs font appel à l'utilisation de méthodes semi et non-paramétriques. Dans ce contexte, Hill (1975) a proposé une méthode semi-paramétrique largement utilisée, aujourd'hui, dans l'estimation du coefficient de hiérarchisation de la distribution des tailles des villes. L'estimateur de Hill est celui de la méthode du maximum de vraisemblance, lorsque la distribution étudiée épouse parfaitement une loi de Pareto.

On suit, ici, la présentation et la formulation de cet estimateur proposées par Gabaix et Ioannides (2004), car elles sont les plus populaires, parmi les chercheurs qui travaillent sur la loi rang taille des villes.

Pour un échantillon de *n* villes avec des tailles $T_1 > ...T_j > ...T_n$, l'estimateur de Hill est égal à :

$$\hat{\beta} = \frac{n-1}{\sum_{j=1}^{n-1}(\ln T_j - \ln T_n)} \qquad (20)$$

tandis que écart-type pour $\frac{1}{\hat{\beta}}$ est donné par l'équation

$$\sigma_n(\frac{1}{\hat{\beta}}) = \left( \frac{\sum_{j=1}^{n-1} j(\ln T_j - \ln T_{j+1})^2}{n-1} - \frac{1}{\hat{\beta}^2} \right)^{\frac{1}{2}} (n-1)^{-\frac{1}{2}} \qquad (21)$$

Si $\frac{1}{\hat{\beta}} > \sigma_n(\frac{1}{\hat{\beta}})$, l' écart-type de l'estimation $\hat{\beta}$ est égal à :

$$\sigma_n(\hat{\beta}) = \hat{\beta}^2 \left( \frac{\sum_{j=1}^{n-1} j(\ln T_j - \ln T_{j+1})^2}{n-1} - \frac{1}{\hat{\beta}^2} \right)^{\frac{1}{2}} (n-1)^{-\frac{1}{2}} \qquad (22)$$

Dans sa comparaison internationale des distributions rang taille des villes, Soo (2005) montre que l'utilisation de l'estimateur de Hill permet de baisser sensiblement le rejet de la loi de Zipf, par rapport à la méthode des MCO. Néanmoins, McCullogh, 1996, Embrechts et al., 1997, Weron, 2001 mettent l'accent sur le biais que comporte cet estimateur très sensible à la sélection de la borne inférieure et à la taille de l'échantillon, qui peut conduire à une sous-estimation systématique de l'écart-type réel de $\beta$.

Si *k* est le nombre des plus grandes villes (celles qui forment la queue épaisse) d'un échantillon donné de taille *n*, Huisman et al. (2001) et Durfour et Kurz-Kim (2005) montrent que le biais de l'estimateur de Hill augmente avec *k*, tandis que la variance de l'estimateur est proportionnelle à $1/k$. Ce biais peut être approximé de la façon suivante :

$$\hat{\beta} = \beta + \lambda k + u \text{ avec } k = 1, 2, \cdots, m \qquad (23)$$

Lorsque $k$ tend vers 0, le biais de l'estimateur $\lambda k$ diminue, mais sa variance augment, ce qui conduit à un choix « délicat » entre la réduction du biais et celle de la pertinence de l'estimation (Huisman et al, 2001). Beirlant et al. (1999) proposent diverses corrections du biais de l'estimation de Hill.

# 2. Le choix de la méthode d'estimation. Une simulation Monte Carlo

Afin de comparer les performances des différentes méthodes d'estimation du coefficient de hiérarchisation, on a effectué une simulation Monte Carlo. La section 2.1 présente la démarche suivie et les résultats obtenus lorsque l'on teste les performances de chaque méthode selon la taille de l'échantillon. La section 2.2 affiche un certain nombre de résultats complémentaires, permettant de désigner la méthode la plus pertinente pour le calcul du coefficient de Pareto.

## 2.1 Méthodes d'estimation et taille de l'échantillon

Dans la simulation Monte Carlo réalisée, on a considéré cinq tailles différentes : 20, 50, 100, 200 et 500 villes et, pour chacune de ces tailles, on a construit 20000 échantillons qui suivent une loi de Zipf (avec un coefficient de Pareto $\beta$ égal à 1). Dans ce cas, la fonction de répartition de la loi suivie par ces échantillons est la suivante :

$$F(x) = 1 - C \quad x^{-1} \qquad (24)$$

Pour construire ces échantillons, on a utilisé la méthode d'inversion de la fonction de répartition. Cette méthode stipule que si $X$ est une variable aléatoire de fonction de répartition $F$ inversible, d'inverse $F^{-1}$, et $U$ suit une loi uniforme sur $[0;1]$, alors $F^{-1}(U)$ suit la même loi que $X$. En appliquant l'inverse de la fonction de répartition, on obtient :

$$F^{-1}(u) = \frac{C}{1-u} \qquad (25)$$

Pour chacune des tailles considérées, on a produit 20000 estimations permettant d'étudier le comportement des différentes méthodes : le calcul du coefficient de Pareto par les MCO (*7*), par les MCO corrigés par Gabaix et Ibragimov (*11*), par la méthode de Hill (*20*) et le calcul du coefficient de hiérarchisation de Lotka par les MCO (*8*), en lui appliquant la correction de Nishiyama et Osada (*13*) et, enfin, en utilisant la méthode des MCG (*17*).

| Taille échantillon | 20 | 50 | 100 | 200 | 500 |
|---|---|---|---|---|---|
| $\beta$ (MCO) | | | | | |
| Moyenne | 0,9018 | 0,9225 | 0,9437 | 0,9613 | 0,9782 |
| Ecart-type réel | 0,2804 | 0,1813 | 0,1324 | 0,0962 | 0,0624 |
| Moyenne des e-types calculés | 0,0485 | 0,0235 | 0,0137 | 0,0078 | 0,0036 |
| IC à 5% | [0,52 ; 1,42] | [0,65 ; 1,24] | [0,74 ; 1,17] | [0,81 ; 1,13] | [0,88 ; 1,08] |
| $\beta_{cor}$(MCO corr) | | | | | |
| Moyenne | 1,0504 | 1,0101 | 1,0016 | 0,9986 | 0,9984 |
| Ecart-type réel | 0,3203 | 0,1944 | 0,1378 | 0,0983 | 0,0630 |
| Moyenne des e-types calculés | 0,0588 | 0,0263 | 0,0147 | 0,0081 | 0,0038 |
| Ecart type estimé | 0,3162 | 0,2000 | 0,1414 | 0,1001 | 0,0632 |
| IC à 5% | [0,61 ; 1,65] | [0,72 ; 1,35] | [0,79 ; 1,24] | [0,84 ; 1,16] | [0,90 ; 1,10] |
| $\beta$ (Hill) | | | | | |
| Moyenne | 1,0586 | 1,0219 | 1,0112 | 1,0054 | 1,0021 |
| Ecart-type réel | 0,2570 | 0,1482 | 0,1019 | 0,0708 | 0,0449 |
| Moyenne des e-types calculés | 0,2324 | 0,1433 | 0,1008 | 0,0710 | 0,0448 |
| IC à 5% | [0,71 ; 1,53] | [0,80 ; 1,29] | [0,85 ; 1,19] | [0,89 ; 1,13] | [0,93 ; 1,08] |
| $\gamma$ (MCO) | | | | | |
| Moyenne | 1,1395 | 1,0823 | 1,0529 | 1,0339 | 1,0180 |
| Ecart-type réel | 0,3321 | 0,2013 | 0,1403 | 0,0991 | 0,0629 |
| Moyenne des e-types calculés | 0,0646 | 0,0291 | 0,0159 | 0,0086 | 0,0038 |
| IC à 5% | [0,67 ; 1,75] | [0,78 ; 1,44] | [0,84 ; 1,30] | [0,88 ; 1,21] | [0,92 ; 1,13] |
| $\gamma_{cor}$ ($\gamma$ corrigé) | | | | | |
| Moyenne | 0,9987 | 1,0001 | 0,9995 | 0,9998 | 0,9997 |
| Ecart-type réel | 0,2910 | 0,1860 | 0,1332 | 0,0958 | 0,0618 |
| Moyenne des e-types calculés | 0,0566 | 0,0269 | 0,0151 | 0,0083 | 0,0037 |
| IC à 5% | [0,58 ; 1,53] | [0,72 ; 1,33] | [0,80 ; 1,23] | [0,85 ; 1,17] | [0,90 ; 1,11] |
| $\gamma$ (MCG) | | | | | |
| Moyenne | 1,1025 | 1,0523 | 1,0298 | 1,0169 | 1,0078 |
| Ecart-type réel | 0,2657 | 0,1543 | 0,1049 | 0,0723 | 0,0453 |
| Moyenne des e-types calculés | 0,3655 | 0,2708 | 0,2104 | 0,1609 | 0,1109 |
| IC à 5% | [0,71 ; 1,58] | [0,81 ; 1,32] | [0,86 ; 1,21] | [0,90 ; 1,14] | [0,93 ; 1,08] |

*Tableau 1 : Estimations du coefficient de Pareto –Résultats de la simulation Monte Carlo (20000 estimations)*

Le tableau 1 fournit les résultats obtenus : la moyenne, l'écart-type réel, la moyenne des écart-types calculés pour chaque coefficient de la régression, ainsi que l'intervalle de confiance à 5% du coefficient pour chaque série d'estimations. Pour les estimations avec la méthode des MCO corrigée (*11*), on calcule également l'estimation de l'écart-type avec la formule proposée (*12*) par Gabaix et Ioannides (2004) et Gabaix et Ibragimov (2006).

Comme on peut le constater, toutes les estimations s'améliorent au fur et à mesure que la taille de l'échantillon augmente : les moyennes s'approchent de 1, les écart-types diminuent et les intervalles de confiance se réduisent. Le biais est globalement positif pour le calcul du coefficient de hiérarchisation avec le modèle de Zipf et négatif pour le calcul du coefficient de hiérarchisation avec le modèle de Lotka.

Force est, cependant, de noter que pour toutes les méthodes, sauf pour celle de Hill, et pour toutes les tailles, la différence entre l'écart-type réel et l'écart type calculé est conséquente. Notons, à cet égard, que la formule de calcul de l'écart-type proposée par Gabaix et Ioannides (2004) permet une estimation très proche de l'écart-type réel.

Les histogrammes issus des 20000 estimations pour les échantillons de taille de 20, 100 et 500 villes sont présentés dans la figure 6. Ils complètent les informations obtenues dans le tableau 1, car permettent d'examiner les effets d'asymétrie qui caractérisent notamment les petits échantillons. Afin d'obtenir une meilleure visualisation de ces effets, l'échelle de l'axe des abscisses est différente, ici, pour chaque taille d'échantillon. La réduction de l'échelle des abscisses, toute méthode d'estimation confondue, est synonyme de réduction de l'intervalle de confiance de l'estimateur, au fur et à mesure que la taille de l'échantillon augmente.

Le tableau 1, ainsi que la figure 6, permettent de constater l'efficacité des améliorations proposées par Gabaix et Ibragimov (2006) et par Nishiyama et Osada (2004) pour le calcul des coefficients de Pareto et de Lotka respectivement. Avec la correction de Gabaix et Ibragimov (2006), le biais de l'estimateur diminue de 0,10 à 0,05 pour les échantillons de taille 20 et de 0,022 à 0,002 pour les échantillons de taille 500. La correction de Nishiyama et Osada (2004) est encore plus performante : le biais diminue de 0,14 à 0,01 pour les échantillons de taille 20 et de 0,018 à 0,0003 pour les échantillons de taille 500.

Cependant, la relation entre la réduction du biais et l'augmentation de la taille de l'échantillon est plus asymptotique, lorsque l'on applique ces deux corrections, sauf pour les échantillons de plus petite taille ($T = 20$). Il semblerait ainsi qu'il y ait une taille optimale de l'échantillon (entre 100 et 200), pour laquelle la correction de Gabaix et Ibragimov procure une estimation quasi-parfaite ($\beta$ égal à 1) –ce qui confirmerait la critique de Eeckhout (2004)-, tandis que le coefficient de Lotka, corrigé par Nishiyama et Osada est moins sensible à la modification de la taille de l'échantillon, au-delà d'une taille 50.

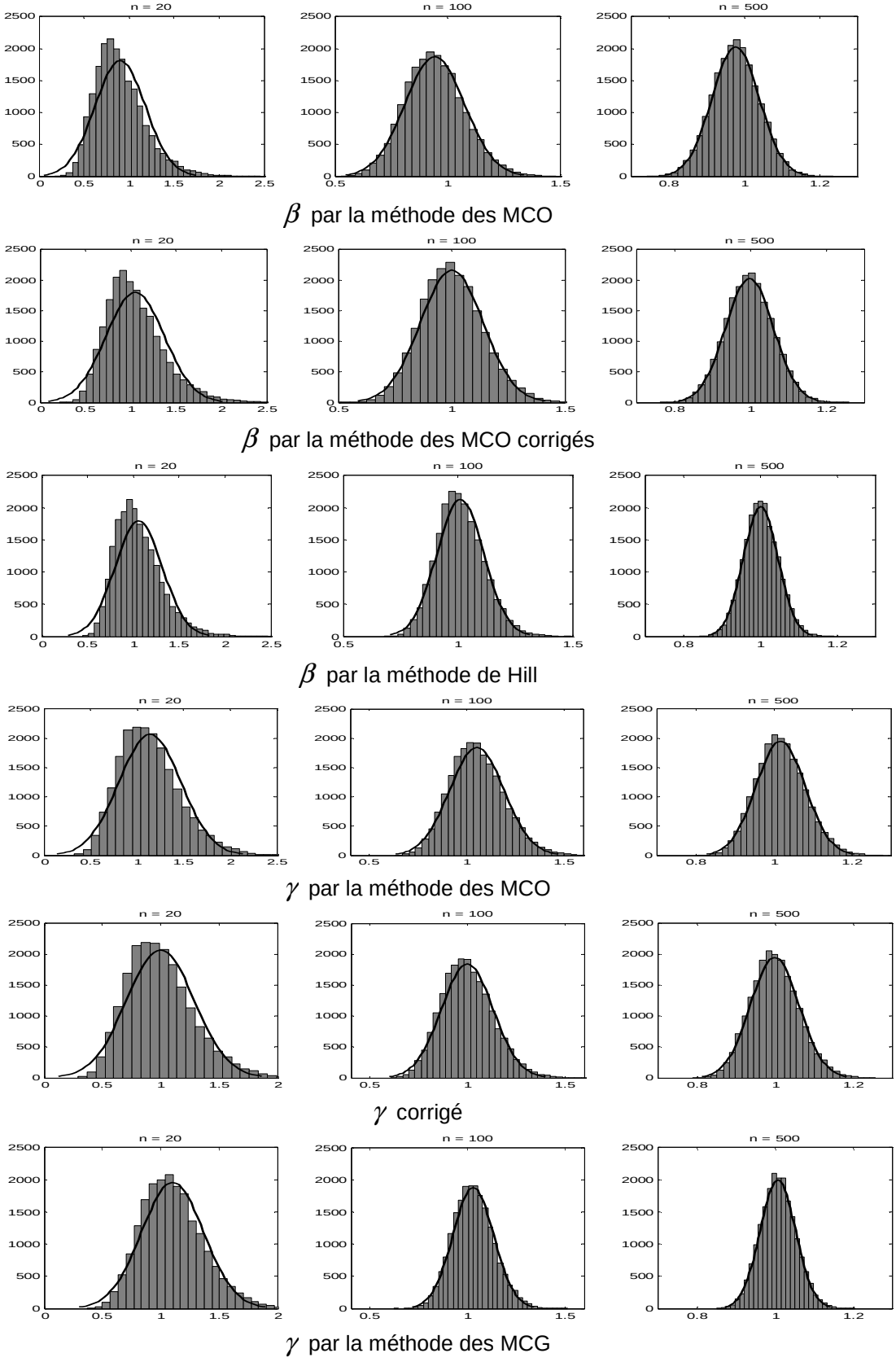



*Figure 6 : Histogrammes des estimateurs du coefficient de hiérarchisation obtenus par 20 000 simulations Monte Carlo*

L'ensemble des résultats obtenus montre que le calcul du coefficient de hiérarchisation par la méthode des MCO, pour le modèle de Zipf ou de Lotka, sans les corrections de Gabaix et Ibragimov (2006) ou de Nishiyama et Osada (2004), est nettement moins performant que celui proposé par les autres méthodes d'estimation. Or, la très grande majorité d'études sur les distributions rang taille des villes emploient cette méthode.

Dans le travail le plus complet de comparaison des différentes estimations du coefficient de hiérarchisation, par le biais d'une méta-analyse portant sur 29 études et 515 estimations, Nitsch (2005) montre que l'ensemble des estimations du coefficient varie dans un intervalle [0,49 ; 1,96], avec une moyenne de 1,09. Cependant, 90% de ces études s'appuient sur un calcul du coefficient de hiérarchisation par la méthode des MCO sur des échantillons de taille inférieure à 100 villes, d'où l'important doute que l'on peut formuler quant à la validité de ces résultats.

### 2.2 Une comparaison des différentes méthodes d'estimation

Les résultats précédents sont synthétisés dans la figure 7 pour les échantillons de taille 500, permettant de comparer les moyennes et les écart- types des différents estimateurs. Ces observations sont sensiblement les mêmes pour les différentes tailles d'échantillons.

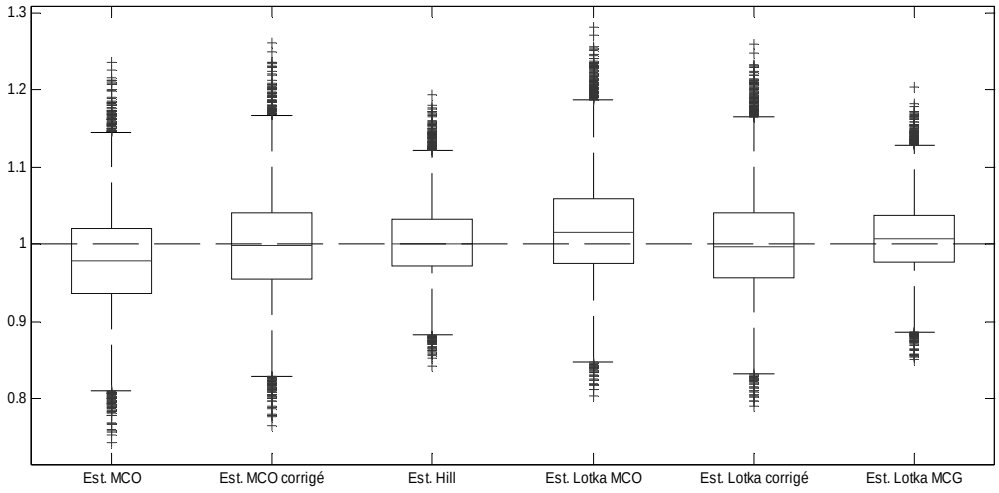



*Figure 7 : Comportement des différents estimateurs pour les échantillons de taille 500 (20000 simulations Monte Carlo)*

Cette figure montre assez clairement que la méthode des MCO (pour le calcul du coefficient de Pareto ou de Lotka) est la moins performante. Les corrections proposées par Gabaix et Ibragimov (2006) et Nishiyama et Osada (2004) permettent de la rendre aussi efficace que les méthodes de Hill ou des MCG, mais le calcul de l'écart-type, dans ces dernières, est systématiquement inférieur à l'écart-type réel. La

méthode des MCG permet d'améliorer les résultats obtenus, tandis que la méthode Hill semble globalement la plus performante.

Ceci est également confirmé par le tableau 2 qui procure, pour chaque taille $T$, la part (en %) des (20000) échantillons pour laquelle une méthode donne une meilleure estimation par rapport à une autre. Par exemple, dans le premier sous-tableau, la première ligne montre que la méthode des MCO procure une meilleure estimation que la méthode des MCO corrigés dans 40,86% des 20 000 échantillons, une meilleure estimation que la méthode Hill dans 36,82% des échantillons, etc.

Comme on peut le constater, la méthode Hill procure systématiquement des meilleurs résultats, lorsqu'elle est comparée à une autre méthode d'estimation, indépendamment de la taille de l'échantillon. Seule la méthode des MCG (pour le modèle de Lotka) permet d'obtenir des résultas assez proches de ceux de la méthode de Hill, et ceci pour les échantillons de plus grande taille (pour $T = 200$ et $T = 500$, la méthode de Hill fournit une meilleure estimation que celle de MCG pour 53,15% et 52,63% des échantillons respectivement.

Notons, enfin, que contrairement à l'idée communément admise, selon laquelle au dessus d'une certaine taille de l'échantillon les différentes méthodes pourraient permuter, la corrélation ($R^2$) entre l'estimateur de Hill et les autres estimateurs baisse systématiquement au fur et à mesure que la taille de l'échantillon augmente. Ce résultat étonnant, lié à la baisse conjointe des écarts-types des différents estimateurs, montre que la spécificité de chaque estimateur se confirme au fur et à mesure que le nombre d'informations augmente.

Taille 20

| | MCO | MCO corr | Hill | Lotka MCO | Lotka corr | MCG |
|---|---|---|---|---|---|---|
| MCO | | 40,86 | 36,825 | 56,96 | 45,05 | 43,285 |
| MCO corr | 59,14 | | 38,435 | 55,975 | 46,225 | 48,695 |
| Hill | 63,175 | 61,565 | | 60,085 | 56,415 | 54,74 |
| Lotka MCO | 43,04 | 44,025 | 39,915 | | 45,335 | 35,105 |
| Lotka corr | 54,95 | 53,775 | 43,585 | 54,665 | | 44,435 |
| MCG | 56,715 | 51,305 | 45,26 | 64,895 | 55,565 | |

Taille 50

| | MCO | MCO corr | Hill | Lotka MCO | Lotka corr | MCG |
|---|---|---|---|---|---|---|
| MCO | | 39,585 | 35,205 | 52,19 | 43,04 | 36,295 |
| MCO corr | 60,415 | | 37,51 | 56,38 | 45,41 | 42,35 |
| Hill | 64,795 | 62,49 | | 61,925 | 59,91 | 54,675 |
| Lotka MCO | 47,81 | 43,62 | 38,075 | | 44,335 | 34,695 |
| Lotka corr | 56,96 | 54,59 | 40,09 | 55,665 | | 40,27 |
| MCG | 63,705 | 57,65 | 45,325 | 65,305 | 59,73 | |

Taille 100

| | MCO | MCO corr | Hill | Lotka MCO | Lotka corr | MCG |
|---|---|---|---|---|---|---|
| MCO | | 40,32 | 34,545 | 49,385 | 42,96 | 34,1 |
| MCO corr | 59,68 | | 36,53 | 55,595 | 45,65 | 39,325 |
| Hill | 65,455 | 63,47 | | 62,945 | 61,58 | 54,12 |
| Lotka MCO | 50,615 | 44,405 | 37,055 | | 44,855 | 34,375 |
| Lotka corr | 57,04 | 54,35 | 38,42 | 55,145 | | 38,7 |
| MCG | 65,9 | 60,675 | 45,88 | 65,625 | 61,3 | |

Echantillons 200 :

| | MCO | MCO corr | Hill | Lotka MCO | Lotka corr | MCG |
|---|---|---|---|---|---|---|
| MCO | | 40,94 | 34,3 | 46,875 | 43,24 | 33,535 |
| MCO corr | 59,06 | | 35,55 | 54,745 | 45,715 | 36,855 |
| Hill | 65,7 | 64,45 | | 63,84 | 63,2 | 53,15 |
| Lotka MCO | 53,125 | 45,255 | 36,16 | | 45,27 | 34,42 |
| Lotka corr | 56,76 | 54,285 | 36,8 | 54,73 | | 37,115 |
| MCG | 66,465 | 63,145 | 46,85 | 65,58 | 62,885 | |

Echantillons 500 :

| | MCO | MCO corr | Hill | Lotka MCO | Lotka corr | MCG |
|---|---|---|---|---|---|---|
| MCO | | 42,155 | 34,185 | 45,18 | 43,845 | 33,325 |
| MCO corr | 57,845 | | 35,78 | 53,865 | 46,535 | 35,675 |
| Hill | 65,815 | 64,22 | | 64,335 | 63,695 | 52,63 |
| Lotka MCO | 54,82 | 46,135 | 35,665 | | 45,975 | 34,685 |
| Lotka corr | 56,155 | 53,465 | 36,305 | 54,025 | | 36,06 |
| MCG | 66,675 | 64,325 | 47,37 | 65,315 | 63,94 | |

*Tableau 2 : Performances comparées des différentes méthodes d'estimation du coefficient de hiérarchisation (20000 échantillons)*

## Conclusion. La loi de Zipf est elle un indicateur pertinent des hiérarchies urbaines ?

La littérature sur la loi de Zipf et la distribution rang taille des villes, connaît un développement spectaculaire, depuis une dizaine d'années, avec la prolifération d'un ensemble de modèles dynamiques cherchant à déterminer les mécanismes économiques qui sont à l'origine d'un processus de croissance urbaine, conduisant à une distribution rang taille des villes qui obéit à la loi de Zipf (Fujita et al, 1999, Gabaix et Ioannides, 2004, Brakman et al, 2004, Duranton, 2006, Bosker et al., 2006). Par ailleurs, une nombre conséquent d'études empiriques examinent et comparent l'évolution des hiérarchies urbaines des différents pays et régions, en s'appuyant sur les informations fournies par les coefficients de hiérarchisation qui caractérisent la relation rang taille des villes.

En s'appuyant sur une simulation Monte Carlo, l'objectif de ce papier était de montrer que la valeur affichée par le coefficient de hiérarchisation dans le modèle de Zipf (ou de Lotka) peut être fortement contestable selon la méthode d'estimation utilisée, ainsi que la taille de l'échantillon considérée. Le papier montre que la méthode semi-paramétrique de Hill paraît comme la méthode la plus pertinente et devrait être privilégiée aux dépens des autres méthodes, peu importe la taille de l'échantillon.

## Bibliographie


Ades A., Glaeser L., 1995, Trade and Circuses: Explaining Urban Giants, *Quarterly Journal of Economics, n°*110(1), pp.195-227.

Anderson G., Ge Y., 2005, The size distribution of Chinese cities, *Regional Science and Urban Economics* 35, pp 756-776.

Beirlant P. and alii, 1999, Tail index estimation and an exponential regression model, *Extremes*, 2177, pp.177-200.

Bosker E., Brakman D., Marrewijk C.., Van de Berg M, 1999, The return of Zipf: towards a further understanding of the Rank-Size rule, *Journal of Regional Science,* 39, pp.183-213.

Brakman D., Garretsen H., Schramm M., 2004, The strategic bombing of German cities during World War II and its impact on city growth, *Journal of Economic Geography*, 4, pp.201-218.

Candeal et al, 2001, *Spurious Zipf's law*, Working paper, University of Zaragosa, Spain.

Charnes A., Frome E, Yu P., 1976, The equivalence of generalised least squares and maximum likelihood estimates n the exponential family, *Journal of the American Statistical Association*, Vol.71, n°353, pp.169-171.

Csörgo, S., Viharos, L., 1997, Asymptotic normality of least squares estimators of tails indices, *Bernoulli*, 3, pp.351-370.

Dobkins, L.H., Ioannides,Y.M., 2000, Dynamic evolution of U.S. cities, *In :* Huriot, J., Thisse, J (Eds.), *The Economics of Cities, Theoretical Perspectives.* Cambridge University Press, Cambridge, pp 217-260.

Duranton G., 2006, Some foundations for Zipf's law: product proliferation and local spillovers, *Regional Science and Urban Economics*, n°36, pp.542-563.

Durfour J-M., Kurz-Kim J-R., 2005 Exact inference and optimal invariant estimation for the tail coefficient of symmetric unstable distributions', *Working paper, Université de Montréal.*

Eeckhout J.,2003, Gibrat's Law for (all) Cities, *American Economic Review,* 94, pp.1429-1451.

Embrechts, P., Kluppelberg, C, Mikosch, T., 1997, *Modelling Extremal Events for Insurance and Finance*, Springer, New York.

Fujita M., Krugman P., Venables A., 1999, *The Spatial Economy*, MIT Press, Cambridge, MA.

Gabaix, X., Ibragimov, R., 2006, *Log(Rank – ½) : a simple way to improve the OLS estimation of tail exponents*, Discussion paper 2106, Harvard Institute of Economic Research, Harvard University.

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, pp.2341-2376.

Gan L., Li D., Song S., 2006, Is Zipf's law spurious in explaining city-size distributions, *Economic Letters*, 92, pp.256-262.

Hill, B.M., 1975, A simple approach to inference about the tail of a distribution, *Annals of Statistics* 3, pp 1163-1174.

Huisman et al., 2001, Tail-index estimates in small samples, *Journal of Business and Economic Statistics, n°*19, pp.208-216.

Kratz M., Resnick S., 1996, The QQ-estimator and heavy tails, *Communications in Statistics, Stochastic models*, 12, pp.699-724.

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

Lotka A., 1941, The law of urban concentration, *Science*, n°94, pp.164.

Malakarne L, et al., 2001, q-exponential distribution in urban agglomeration, *Physical review*, Vol.65, n°017106, pp.65-68.

Malecki E., 1980, Growth and change in the analysis of rank–size distributions: empirical findings, *Environment and Planning A*, n°12, pp.41–52.

Mandelbrot B., 1960, The Pareto-Levy Law and the distribution of income, *International Economic Review*, Vol.1, n°2, pp.79-106.

McCulloch J., 1997, Measuring tail thickness to estimate the stable index alpha: A critique', *Journal of Business and Economic Statistics, n°*15, pp.74-81.

Nishiyama Y., Osada S., 2005, *Statistical theory of rank-size rule regression under Pareto distribution*, Working paper, Kyoto Institute of economic research.

Nishiyama Y., Osada S., Mirumune M., 2004, *Estimation and testing for rank-size rule regression under Pareto distribution*, Working paper, Kyoto Institute of economic research.

Nishiyama Y., Osada S., Sato Y., 2006, OLS-t-test revisited in rank-size rule regression, *DEE Discussion paper 06-3*, Kyoto Institute of economic research.

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

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

Reed J., 2001, The Pareto, Zipf and other power laws', *Economics Letters,* n°74, 15-19.

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

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

Stoev S., Michailidis G., Taqqu M., 2008, *Estimating heavy-tail exponents through max self-similarity*, Research Paper, University of Michigan, Ann Arbor.

Weron R., 2001, Levy-stable distributions revisited: tail index >2 does not exclude the Levy stable regime, International Journal of Modern Physics, vol.12, n°2, pp.110-1115

Werner T., Upper C., 2002, Information Flows in Times of Stress. The Case of Bund Futures and Bonds during the 1998 Turbulences, *BIS Paper, N°.12* .

Wickens, C., 1921, *The Commonwealth of Australia 1921*, H. J. Green, Government Printer, Melbourne.

Zipf G.K., 1941, *National Unity and Disunity: the Nation as a Bio-Social Organism*, Principia Press, Bloomington, IN.

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

## Résumé

La loi de Zipf caractérise la distribution de la taille des villes et conditionne, aujourd'hui, toute tentative d'interprétation des mécanismes économiques qui régissent la formation et l'évolution des hiérarchies urbaines d'un pays ou d'une région. L'immense littérature sur la distribution rang taille des villes, avec ses comparaisons internationales et diachroniques, s'appuie fortement sur la valeur du coefficient de hiérarchisation du modèle logarithmique de Zipf, pour atteindre un ensemble de conclusions sur la nature des hiérarchies urbaines dans les différents pays et régions. Or, depuis quelques années, un ensemble de chercheurs montre l'extrême volatilité des résultats obtenus, en fonction de la méthode de calcul utilisée. Il n'y a pas, cependant, d'étude qui propose une comparaison efficace de l'ensemble des méthodes d'estimation connues. Par le biais d'une simulation Monte Carlo, ce papier prétend combler ce besoin.

**Mots clés :** loi de Zipf, distribution rang taille, villes, Monte-Carlo.

**Title:** How well does Zipf's law hold ? Some methodological issues

## Abstract

Zipf's law, characterizing city size distribution, is one of the most astonishing regularities in urban economics. There is a awesome literature on that subject, resulting in more specific analysis over different countries and regions' urban hierarchies. However, some recent studies bring evidence of important methodological problems that Zipf's model carries, leading, somehow, to wrong conclusions. The aim of this paper is to compare the main methods of estimating the Pareto exponent, by using a Monte Carlo simulation over 20000 samples, in order to design the most adequate method of estimating. From this point of view, the paper is essentially methodological.

**Key words:** Zipf's law, rank size distribution, cities, Monte-Carlo analysis.

**Codes JEL :** R12, C13, C16.