99-09

L'ESTIMATION DE L'EMPLOI À PARTIR DE LA DISTRIBUTION DES ÉTABLISSEMENTS PAR CLASSE DE TAILLE

André LEMELIN

INRS-Urbanisation
3465, rue Durocher
Montréal, Québec
H2X 2C6

Juin 1999

L'ESTIMATION DE L'EMPLOI À PARTIR DE LA DISTRIBUTION DES ÉTABLISSEMENTS PAR CLASSE DE TAILLE

André LEMELIN

Juin 1999

RÉSUMÉ

La rareté des données est un problème endémique dans les recherches économiques appliquées à l'échelle régionale, *a fortiori* à l'échelle infra-métropolitaine. Cette note discute trois méthodes pragmatiques d'estimation de l'emploi à partir de la distribution des établissements par classe de taille : la méthode du point milieu de l'intervalle (méthode PM), le lissage lognormal et le lissage log-logistique. On compare les méthodes au moyen de deux ensembles de données : celles du Recensement des établissements et de l'emploi de Montréal de 1996 (Banque de Données et d'Information urbaine, INRS-Urbanisation et Ville de Montréal) et celles de Statistique Canada sur les industries manufacturières à deux chiffres de la CTI au Québec en 1995. Les résultats montrent que les modèles lognormal et logistique sont nettement supérieurs à la méthode PM, notamment à cause du biais à la hausse inhérent à cette méthode.

SUMMARY

The scarcity of data is a widespread problem in applied regional and metropolitan economic research. This note discusses three pragmatic methods to estimate employment from the distribution of establishments by size-class : the class-interval mid-point method (MP method), lognormal smoothing, and log-logistic smoothing. The methods are compared using data from two sources : the 1996 *Recensement des établissements et de l'emploi de Montréal* (Banque de Données et d'Information urbaine, INRS-Urbanisation and Ville de Montréal), and Statistics Canada two-digit SIC manufacturing industry data for Québec in 1995. Results show that the models are clearly superior to the simple MP method, partly because of that method's inherent upward bias.

TABLE DES MATIÈRES

LES MAINS SALES...

Cet article [1] terre-à-terre s'adresse à ceux qui se salissent les mains dans les données. Il est issu d'une expérience de l'auteur qui a été appelé, il y a quelques années, à réaliser une recherche appliquée pour le compte d'une société de gestion immobilière qui voulait connaître, à l'échelle infra-métropolitaine, l'environnement économique des zones où elle possédait des actifs. Pour examiner la structure économique de ces zones, il fallait connaître, au moins approximativement, l'emploi par branche d'activité. À cette échelle spatiale, les données de Statistique Canada sont évidemment confidentielles. Quant au répertoire Scott [2], bien connu, il contenait pour le Québec des données sur le secteur manufacturier seulement. En outre, les données du Scott comportent des lacunes importantes, particulièrement quant à la localisation des établissements d'entreprises qui ont plusieurs places d'affaires (par exemple, la Division Canadair de Bombardier ne figure pas parmi les établissements ayant une adresse à St-Laurent, dans la région métropolitaine de Montréal) [3]. Il a donc fallu extraire les données nécessaires d'autres répertoires [4]. Or, pour certaines des zones à l'étude, la seule indication relative au chiffre de l'emploi était la classe de taille à laquelle appartenait chaque établissement... De telles situations sont courantes en recherche économique appliquée à l'échelle locale et régionale : ceux qui s'adonnent à ce type d'activité le savent.

Cette note aborde donc, d'un point de vue pragmatique, le problème de l'estimation de l'emploi à partir de la distribution des établissements par classe de taille. Car cette information est généralement plus facile d'accès, ou moins coûteuse à obtenir, que des données exactes sur l'emploi. D'une part, en effet, la classe de taille est souvent considérée comme moins confidentielle que le nombre d'emplois, tant par les agences qui possèdent des données que par les entreprises elles-mêmes [5]. Et d'autre part, s'il s'agit de recueillir des données par enquête, cette information est plus rapide et moins coûteuse à obtenir, aussi bien pour l'enquêteur que pour le répondant.

Spécifiquement, la présente note compare trois méthodes d'estimation de l'emploi à partir de la distribution des établissements par classe de taille. Les sections qui suivent présentent successivement : la méthode du point milieu de l'intervalle (méthode

1 Cette recherche a bénéficié de l'appui du Conseil de recherches en sciences humaines du Canada. L'auteur remercie William Coffey, du Département de géographie de l'Université de Montréal, et Mario Polèse, de l'INRS-Urbanisation, qui ont bien voulu commenter une version préliminaire de ce texte.

2 *Scott's Directories Selectory Prospector*, Southam Magazine and Information Group, Don Mills, ON, Canada, 1996.

3 Pour donner une idée de l'ordre de grandeur du travail requis pour préparer ces données aux fins d'analyse, disons qu'il en a coûté environ 40 000 $ à une équipe de l'INRS-Urbanisation pour réaliser un travail de validation et de correction des données du Scott de 1991 pour la région métropolitaine de Montréal.

4 Remerciements à Madame Myriam Tétrault, du Service de développement économique de la Ville de Saint-Laurent et à Monsieur Sylvain Laurendeau, du Centre d'emploi Montréal-Olympique, Développement des ressources humaines Canada.

5 Bien que les entreprises soient généralement peu réticentes à divulguer leur nombre d'emplois.

PM), le modèle lognormal et le modèle log-logistique. Les résultats de l'application des trois méthodes sont ensuite comparés au moyen des données du Recensement des établissements et de l'emploi de Montréal de 1996 (Banque de Données et d'Information urbaine, INRS-Urbanisation et Ville de Montréal).

LA MÉTHODE DU POINT MILIEU DE L'INTERVALLE

Le problème à résoudre est formulé mathématiquement comme suit. Il y a P classes de taille, définies par leurs bornes supérieures τ_i, i = 1, ..., P (en général, la dernière classe est ouverte : $\tau_P = \infty$; nous adoptons en outre la convention que $\tau_0 = 0$). On connaît le nombre d'établissements par classe de taille :

$$N_i = \sum_{t=\tau_{i-1}+1}^{\tau_i} n_t \text{ , pour } i = 1, ..., P$$

On veut trouver :

$$\sum_t n_t \, t$$

où n_t est le nombre d'établissements de t employés.

On pourrait estimer l'emploi de chaque classe de taille en attribuant aux établissements de cette classe le nombre d'emplois correspondant au point milieu de la classe. Dans la suite de ce texte, nous désignerons cette méthode par l'expression «méthode PM». Pour les établissements de la classe i, qui correspond à l'intervalle $[\tau_{i-1}+1;\ \tau_i]$, l'emploi moyen selon la méthode PM est donné par :

$$\frac{\tau_i + \tau_{i-1} + 1}{2}$$

Cette méthode repose toutefois sur une hypothèse implicite quant à la distribution des établissements au sein de la classe i. Par exemple, la méthode PM est compatible avec la distribution uniforme, qui suppose qu'il y a un nombre égal d'établissements de chacune des tailles qui sont comprises dans la classe i. Mais il se trouve que les hypothèses compatibles avec la méthode PM sont toutes contraires à ce qu'ont révélé les études sur la distribution des établissements par classe de taille (voir Sutton, 1997). Ces études montrent en effet que la distribution des établissements par taille est fortement asymétrique, la fréquence des établissements diminuant avec la taille (sauf peut-être, dans certains cas, pour les tout petits établissements, de moins de 5 employés), de sorte que la taille moyenne est normalement inférieure au point milieu de l'intervalle.

Dans ces conditions, la méthode PM conduit inévitablement à surestimer le chiffre de l'emploi. Cloutier (1995) a d'ailleurs montré dans un autre contexte (l'étude de

la répartition du revenu) à quel point les résultats obtenus dépendent de l'hypothèse que l'on fait sur la forme de la distribution des individus à l'intérieur des classes.

En plus de ce défaut majeur, la méthode PM a celui de n'offrir aucune indication quant à l'emploi moyen de la dernière classe, ouverte.

LE MODÈLE DE LA DISTRIBUTION LOGNORMALE

La méthode que nous proposons s'appuie sur le modèle de la distribution lognormale. Une variable dont la distribution est lognormale est une variable dont le logarithme a une distribution normale. Il s'agit donc d'une distribution asymétrique, dont le domaine de variation est limité inférieurement à zéro. Les propriétés de la distribution lognormale sont exposées en détail dans l'ouvrage classique d'Aitchison et Brown (1957). Parmi ces propriétés, mentionnons celle-ci : sous certaines conditions, la distribution lognormale peut être considérée comme l'aboutissement d'un processus de croissance des individus (ici, des établissements) où le taux de croissance de chaque individu est aléatoire, indépendant des autres taux de croissance et indépendant de la taille de l'individu (c'est la «Loi de l'effet proportionnel» de Gibrat) [6].

Lorsque la distribution lognormale représente assez bien, au moins en première approximation, une distribution observée, cette dernière propriété fournit une hypothèse attrayante quant à la genèse de la distribution. Aussi n'est-il pas étonnant que la distribution lognormale ait été abondamment utilisée en sciences sociales : elle a été appliquée notamment à la distribution par tailles des villes (De Cola, 1985; Parr et Suzuki, 1973), à la répartition du revenu (Cloutier, 1995) et à la distribution par tailles des entreprises ou des établissements (voir la liste des auteurs cités par Sutton, 1997).

Pour ce qui est de la distribution des établissements, les études récentes citées par Sutton (1997) concluent que, si la distribution lognormale est acceptable en première approximation, le modèle de sa genèse repose sur des hypothèses irréalistes, qui ne rendent pas compte, par exemple, de l'apparition et de la disparition d'établissements. Pour nos fins cependant, il suffit que la distribution lognormale soit une approximation acceptable.

On fait donc l'hypothèse que la distribution des établissements par taille est lognormale. Il s'ensuit que la probabilité qu'un établissement tiré au hasard compte X employés ou moins est donnée par

$$\text{Prob}\left[t \leq X\right] = \text{Prob}\left[\ln t \leq \ln X\right] = \Phi(z)$$

où

[6] Nous passons sous silence les multiples modèles qui ont été développés pour tenir compte notamment de la création et de la disparition d'individus. Voir Sutton (1997).

$$z = \frac{\ln X - \mu}{\sigma},$$

μ et σ sont les paramètres inconnus de la distribution, et

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp\left(\frac{-x^2}{2}\right) dx$$

est la fonction de probabilité cumulative normale standard [7].

Les observations et la validation du modèle lognormal

Les données fournissent les points de repère suivants :

$$s_{\tau_i} = \frac{\sum_{j=1}^{i} N_j}{\sum_{j=1}^{P} N_j}$$

s_{τ_i} est la fraction des établissements de taille inférieure ou égale à τ_i.

Si les observations étaient parfaitement conformes au modèle, on aurait

$$s_{\tau_i} = \text{Prob}\left[t \leq \tau_i\right] = \Phi(z_{\tau_i})$$

où

$$z_{\tau_i} = \frac{\ln \tau_i - \mu}{\sigma}$$

On aurait aussi, par définition,

$$\Phi^{-1}(s_{\tau_i}) = z_{\tau_i} = \frac{\ln \tau_i - \mu}{\sigma}$$

Donc, si les observations se conformaient à une distribution lognormale, on observerait une relation linéaire entre $\Phi^{-1}(s_{\tau_i})$ et ln τ_i. En pratique, évidemment, les observations ne se conforment pas parfaitement à une distribution lognormale, de sorte que la relation n'est qu'approximativement linéaire. On pourrait bien sûr estimer les paramètres μ et σ en faisant la régression de $\Phi^{-1}(s_{\tau_i})$ sur ln τ_i. On pourrait ensuite chercher à tester formellement l'hypothèse d'une distribution lognormale, selon les règles de l'induction statistique. Le nombre d'observations dont on dispose normalement pour ce test est cependant trop petit pour que l'on puisse espérer des résultats

[7] Du point de vue opératoire, mentionnons que la fonction *NORMSDIST* du logiciel *Excel* renvoie la valeur de la probabilité cumulative normale standard et que la fonction *NORMSINV* en est la fonction inverse.

significatifs [8]. De manière tout à fait pragmatique, on se contentera donc en général d'un examen visuel, complété par des mesures descriptives du degré de linéarité de la relation : le coefficient de détermination multiple R^2 de la relation entre $\Phi^{-1}(s_{\tau_i})$ et $\ln \tau_i$; l'indicateur de dissociation [9] entre, d'une part, les fréquences relatives observées et, d'autre part, les fréquences relatives obtenues par interpolation linéaire de la distribution entre la borne supérieure de la première classe de taille et la borne supérieure de la dernière classe de taille bornée (les paramètres de cette interpolation sont les mêmes que ceux de l'extrapolation définie ci-après pour les classes extrêmes).

Lissage par interpolation

Supposons que les observations se conforment exactement à une distribution lognormale. On généralise la notation en posant

$$s_t = \frac{\sum_{j=1}^{t} n_j}{\sum_{j=1}^{P} N_j}$$

Alors la relation

$$\Phi^{-1}(s_{\tau_i}) = z_{\tau_i} = \frac{\ln \tau_i - \mu}{\sigma}$$

qui se vérifie pour tous les τ_i bornes supérieures des classes de tailles, peut servir à calculer par interpolation linéaire les valeurs de $\Phi^{-1}(s_t)$ pour toutes les tailles d'établissement t. Pour chaque taille t de l'intervalle $[\tau_k;\ \tau_{k+1}]$, on a la formule d'interpolation linéaire

$$s_t = \Phi\left(\Phi^{-1}(s_{\tau_k}) + \alpha(t)\, \Delta_{k+1}\Phi^{-1}\right)$$

où

$$\alpha(t) = \frac{\ln t - \ln \tau_k}{\ln \tau_{k+1} - \ln \tau_k}$$

varie de zéro à un et où

$$\Delta_{k+1}\Phi^{-1} = \Phi^{-1}(s_{\tau_{k+1}}) - \Phi^{-1}(s_{\tau_k})$$

---

8 Le nombre d'observations n'est pas le nombre d'établissements, mais plutôt le nombre de classes de taille bornées supérieurement, c'est-à-dire, pour la classification de Statistique Canada, 7 au maximum.

9 Égal à la moitié de la somme des valeurs absolues des écarts.

On peut ensuite calculer

$$n_t = (s_t - s_{t-1})\, N$$

avec

$$N = \sum_{j=1}^{P} N_j$$

le nombre total d'établissements.

En réalité, les observations sont rarement parfaitement conformes à la distribution lognormale. Il est à noter d'ailleurs que les valeurs calculées des n_t ne seront pas en général entières. Si toutefois l'on juge que la lognormale est une approximation acceptable, on peut appliquer la formule d'interpolation linéaire entre les points de repère que constituent les limites supérieures des classes de taille. La distribution obtenue correspond alors à une relation linéaire par segments entre $\Phi^{-1}(s_t)$ et ln t, où les points de raccord entre les segments de droite sont donnés par les valeurs observées des $\Phi^{-1}(s_{\tau_i})$ et ln τ_i. Si l'on trace la fonction de distribution cumulative, on obtient une courbe d'apparence lognormale qui relie entre eux les points (ln τ_i; s_{τ_i}), mais dont la forme est définie par des paramètres μ et σ légèrement différents d'un segment à l'autre.

Extrapolation pour les classes extrêmes

La première classe de taille est définie par l'intervalle [1 ; τ_1] et, naturellement, $s_0 = 0$. Il est donc impossible d'appliquer telle quelle la formule d'interpolation donnée précédemment, puisque $\Phi^{-1}(s_0) = \Phi^{-1}(0) = -\infty$. De même, la dernière classe de taille est le plus souvent ouverte : $\tau_P = \infty$ et $s_P = 1$. Là encore, il est impossible d'appliquer telle quelle la formule d'interpolation puisque $\Phi^{-1}(s_P) = \Phi^{-1}(1) = \infty$.

On applique donc aux classes extrêmes une formule d'extrapolation linéaire. L'une des solutions possibles consiste à supposer que la pente sur l'intervalle [0; τ_1] et la pente au-delà de τ_{P-1} sont toutes deux égales à la pente sur l'intervalle [τ_1; τ_{P-1}]. Pour $t \leq \tau_1$ ou $t > \tau_{P-1}$, on a donc :

$$s_t = \Phi\left(\Phi^{-1}(s_{\tau_1}) + \alpha'(t)\, \Delta'\Phi^{-1}\right)$$

où

$$\Delta'\Phi^{-1} = \Phi^{-1}(s_{\tau_{P-1}}) - \Phi^{-1}(s_{\tau_1})$$

et où

$$\alpha'(t) = \frac{\ln t - \ln \tau_1}{\ln \tau_{P-1} - \ln \tau_1}, \text{ pour } t \leq \tau_1$$

$$\alpha'(t) = \frac{\ln t - \ln \tau_{P-1}}{\ln \tau_{P-1} - \ln \tau_1}, \text{ pour } t > \tau_{P-1}$$

Il est à noter que $\alpha'(t) < 0$ lorsque $t < \tau_1$. Par ailleurs, pour $t > \tau_{P-1}$, on devrait théoriquement poursuivre les calculs à l'infini. En pratique, il suffit de calculer jusqu'à ce que l'on ait déterminé la taille de la presque totalité des établissements, c'est-à-dire jusqu'à ce que $(1 - s_t)$ devienne insignifiant. Ou encore, s'agissant de calculer le nombre d'emplois, on peut choisir d'interrompre les calculs au moment où l'accroissement de l'emploi

$$t\, n_t = t\, N\, (s_t - s_{t-1})$$

est suffisamment petit.

De façon plus concrète encore, le nombre d'établissements de grande taille étant relativement petit, il est toujours hasardeux d'appliquer une méthode mécanique pour estimer le nombre de leurs employés. C'est pourquoi, lorsque cela est possible, la méthode exposée ici devrait être complétée par une enquête auprès des plus grands établissements [10].

UN MODÈLE CONCURRENT : LA DISTRIBUTION LOGISTIQUE

On le sait, la distribution logistique est assez semblable à la normale, tout en étant plus dense aux extrémités. Au lieu de poser l'hypothèse que la distribution des établissements par taille est lognormale, on pourrait supposer qu'elle est «log-logistique».

Sur le plan théorique, on pourrait justifier le recours au modèle logistique de la façon suivante. Chaque directeur d'établissement choisit le nombre d'employés qu'il embauche dans son établissement. Chacun choisit évidemment la taille d'établissement (le nombre d'employés) dont il attend la meilleure rentabilité. Pour une taille d'établissement donnée, supposons que la rentabilité soit constituée de deux éléments : une composante systématique, dont la valeur pourrait être calculée (moyennant connaissance des paramètres), et une composante aléatoire, qui reflète l'ensemble des variables manquantes qui influencent la rentabilité attendue. Si l'on fait l'hypothèse que les termes aléatoires ont des distributions Gumbel identiques et indépendantes, et qu'en outre la composante systématique de la rentabilité attendue est une fonction linéaire du logarithme de la taille, il en découle que la distribution des établissements par taille est log-logistique. Si cette fable n'est pas convaincante, on peut justifier le recours à la logistique de la même façon que l'on justifie le recours à la lognormale, comme approximation empirique.

[10] Dans le cadre d'une étude menée pour SITQ Immobilier, nous avons effectué un sondage téléphonique auprès de tous les établissements de 500 employés et plus. Ces établissements n'étaient qu'une quarantaine sur les quelque 5000 recensés par nos sources.

Si on fait l'hypothèse que la distribution des établissements par taille est log-logistique, la probabilité qu'un établissement tiré au hasard compte X employés ou moins est donnée par

$$\text{Prob}\left[t \le X\right] = \text{Prob}\left[\ln t \le \ln X\right] = \Phi(z)$$

où

$$z = a + b \ln X$$

a et b sont les paramètres inconnus de la distribution, et

$$\Phi(z) = \frac{1}{1+e^{-z}}$$

est maintenant la fonction de probabilité cumulative logistique

Si les observations étaient parfaitement conformes au modèle logistique [11], on aurait donc

$$s_{\tau_i} = \text{Prob}\left[t < \tau_i\right] = \frac{1}{1+e^{-(a+b\ln\tau_i)}}$$

et la transformation logistique de s_{τ_i} serait une fonction linéaire de ln τ_i :

$$\text{LOGIT}(s_{\tau_i}) = \ln\left(\frac{s_{\tau_i}}{1-s_{\tau_i}}\right) = a + b\ln(\tau_i)$$

Comme pour le modèle de distribution lognormale, on peut alors voir dans quelle mesure les données se conforment au modèle logistique en examinant l'hypothèse que la relation entre LOGIT(s_{τ_i}) et ln τ_i est linéaire. On peut aussi appliquer au modèle logistique la même méthode de lissage par interpolation, avec extrapolation pour les classes extrêmes, que nous avons proposée pour le modèle lognormal.

L'ÉPREUVE DE LA RÉALITÉ

Application au Recensement des établissements et de l'emploi de Montréal de 1996 (RÉEM)

Si l'hypothèse de distribution uniforme sous-jacente à la méthode PM est clairement discréditée, les arguments énoncés en faveur du lissage lognormal ou logistique ne garantissent pas d'emblée que ces méthodes donnent à coup sûr de meilleurs résultats. Pour examiner et comparer la performance des trois manières d'estimer l'emploi à partir de la distribution des établissements par classe de taille, nous

[11] C'est pour alléger le texte que nous laissons tomber l'expression plus exacte «log-logistique».

avons utilisé les données du Recensement des établissements et de l'emploi de Montréal de 1996 (RÉEM).

Le Recensement des établissements et de l'emploi de Montréal de 1996 (REM96P) fait partie de la Banque de Données et d'Information Urbaine, conçue et développée conjointement par l'INRS-Urbanisation et la Ville de Montréal. C'est un recensement des places d'affaires et des emplois situés sur le territoire de Montréal en 1996. Le fichier contient l'adresse de chaque place d'affaires recensée, son domaine d'activité (code CTI), le nombre d'emplois réguliers (30 heures ou plus par semaine) et le nombre total d'emplois qu'on y trouve.

Ces données permettent de comparer les erreurs d'estimation des trois méthodes, non seulement au niveau de l'ensemble d'une branche d'activité, mais au niveau de chaque classe de taille. C'est un avantage considérable par rapport, notamment, aux données publiées par Statistique Canada sur les industries manufacturières (Cat. 31-203-XPB), avec lesquelles des comparaisons aussi fines sont impossibles.

Nous avons donc appliqué les trois méthodes à 64 branches d'activité [12] au niveau de deux chiffres de la CTI. Pour chaque branche d'activité, les établissements ont été regroupés par classe de taille, selon la nomenclature de Statistique Canada :

1-4 employés
5-9 employés
10-19 employés
20-49 employés
50-99 employés
100-199 employés
200-499 employés
500-999 employés
1000 employés et plus

Nos calculs ont produit une masse considérable de chiffres. Aussi, afin d'aller droit au but, le tableau 1 présente plusieurs indicateurs de performance qui permettent de comparer les trois méthodes. Les indicateurs de performance retenus sont :

Indicateurs de biais

- la moyenne non pondérée des erreurs en pourcentage
- la moyenne pondérée des erreurs en pourcentage, les poids étant proportionnels aux nombres d'emplois
- le rapport du nombre de cas de surestimation sur le nombre de cas de sous-estimation

[12] Nous avons écarté de l'analyse les branches d'activité de moins de dix établissements, ainsi que la branche 11 – *Boissons*, qui compte 14 établissements, parce qu'elle ne compte aucun établissement de la classe 1-4 employés, ni aucun des classes de 100 à 999 employés.

Indicateurs de dispersion des erreurs

- la moyenne non pondérée des valeurs absolues des erreurs en pourcentage
- la moyenne pondérée des valeurs absolues des erreurs en pourcentage
- l'écart type non pondéré
- l'écart type pondéré

Indicateurs de performance globale

- les «cotes» des méthodes l'une contre l'autre, c'est-à-dire le rapport du nombre de cas où l'erreur d'estimation de la première méthode est plus petite que celle de la seconde, sur le nombre de cas où l'erreur de la seconde est petite que celle de la première.

Comme on s'y attendait, la méthode PM montre une nette tendance à la surestimation pour toutes les classes de taille. Avec le modèle lognormal et le modèle logistique, par contre, on observe :

- une nette tendance à la sous-estimation pour la classe de taille de 1 à 4 emplois;
- une légère tendance à la surestimation pour les classes de 5 à 199 emplois;
- une légère tendance à la sous-estimation pour les classes de 200 à 999 emplois;
- une surestimation grossière pour la classe de 1000 emplois et plus, surtout avec le modèle logistique (rappelons toutefois que la méthode PM en fournit pas la moindre indication quant au nombre d'emplois dans les établissements de 1000 emplois et plus).

Au total, pour les établissements de moins de 1000 emplois, on constate une certaine compensation des erreurs, de sorte que le biais devient assez faible, tant avec le modèle lognormal qu'avec le modèle logistique. Lorsqu'on inclut les établissements de 1000 emplois et plus, la forte tendance à surestimer les emplois de cette dernière classe se répercute évidemment sur l'ensemble.

Pour ce qui est de la dispersion des erreurs, telle que mesurée par l'écart type, elle est assez semblable entre les trois méthodes. Par contre, lorsque l'on compare les moyennes des valeurs absolues des erreurs, on voit que le modèle lognormal et le modèle logistique donnent généralement de meilleures estimations que la méthode PM. Mieux encore, la performance des deux modèles proposés est meilleure quand les erreurs sont pondérées.

Enfin, les cotes de chacun des modèles contre la méthode PM confirment leur supériorité. Pour l'ensemble des établissements de moins de 1000 emplois, ces deux modèles dominent la méthode PM à 15 contre 1 (c'est-à-dire qu'ils donnent une meilleure estimation pour 60 branches d'activité contre seulement quatre pour la méthode PM).

Il n'a guère été question jusqu'ici des différences entre le modèle lognormal et le modèle logistique. D'ailleurs, les indicateurs de performance sont assez voisins, sauf

pour ce qui est des établissements de 1000 emplois et plus, pour lesquels le modèle logistique donne des estimations qui sont bien pires que celles du modèle lognormal.

Mais l'examen de la cote du modèle lognormal contre le modèle logistique révèle un paradoxe : pour toutes les classes de taille sauf celle de 1000 emplois et plus, le modèle lognormal donne un meilleur estimé dans moins du tiers des cas (cote inférieure à 0,5, ou un contre deux); néanmoins, lorsque l'on considère l'ensemble des établissements de moins de 1000 emplois, le modèle lognormal est meilleur que le modèle logistique dans un peu plus de la moitié des cas (33 branches d'activité sur 64). À première vue, c'est comme si une équipe sportive perdait chacune des parties, mais gagnait le championnat ! Ce qui se passe, c'est que jeu de la compensation des erreurs favorise davantage le modèle lognormal [13].

Les figures 1 et 2 résument les résultats. On y compare les erreurs d'estimations, en pourcentage, de l'emploi total des établissements de moins de 1000 emplois pour chacune des 64 branches [14]. Dans la figure 1, c'est à l'intérieur des deux cônes horizontaux formés par les droites à 45° que les modèles proposés donnent des résultats supérieurs à ceux de la méthode PM : il est clair que, pour la très grande majorité des 64 branches d'activité, les modèles donnent un meilleur estimé que la méthode PM. Ensuite, il saute aux yeux que les deux modèles produisent des estimés moindres que ceux de la méthode PM : tous les points de la figure 1 sont situés en-dessous de la droite de 45°. Il ressort aussi très clairement que la méthode PM tend à surestimer le nombre d'emplois, comme on s'y attendrait : il n'y a que deux paires de points (sur 64) qui soient situés à gauche de l'axe vertical.

La figure 2 compare plus directement le modèle lognormal et le modèle logistique. On constate que les deux modèles donnent des résultats assez semblables. Le modèle logistique tend à produire des estimés plus élevés, ce qui lui est favorable lorsque le modèle lognormal sous-estime et défavorable dans le cas contraire.

Application aux données de Statistique Canada sur les industries manufacturières au Québec en 1995

Mais ces résultats sont-ils particuliers à la Ville de Montréal en 1996 ? Peuvent-ils se généraliser ? Seule une multiplication des études de cas permettrait de répondre de façon parfaitement convaincante à ces questions. Néanmoins, pour voir si l'on

[13] Plus exactement, les erreurs de sous-estimation de la méthode lognormale pour les classes 1-4, 200-499 et 500-999 sont plus marquées que celles du modèle logistique, et elles compensent donc une plus grande fraction des erreurs de surestimation dans les autres classes de taille, qui sont similaires pour les deux modèles.

[14] Il n'eût pas été utile d'inclure la classe des établissements de 1000 emplois et plus : d'une part, la méthode du point milieu ne fournit pas d'estimé des emplois de cette classe; d'autre part, les deux modèles, lognormal et logistique, conduisent *dans tous les cas* à une surestimation, le modèle logistique étant *toujours* pire que l'autre.

pouvait entretenir un «préjugé favorable» à l'égard des résultats obtenus à partir des données du Recensement des établissements et de l'emploi de Montréal de 1996 (RÉEM), nous avons mené une étude complémentaire au moyen des données publiées par Statistique Canada sur les industries manufacturières au Québec pour 1995 (Cat. 31-203-XPB).

Plus précisément, nous avons appliqué les trois méthodes aux industries manufacturières au niveau de deux chiffres de la CTI. Les industries 12 *Tabac* et 39 *Autres* ont été écartées : le chiffre de l'emploi étant confidentiel dans ces deux cas, il est impossible d'évaluer la performance des méthodes.

Le tableau 2 présente, pour le secteur manufacturier au Québec en 1995, les mêmes indicateurs que le tableau 1. Puisque Statistique Canada ne publie pas le chiffre de l'emploi pour chaque classe de taille, seuls peuvent être examinés les estimés de l'emploi total des branches. De plus, rappelons que la méthode PM ne fournit aucun repère quant à l'emploi de la dernière classe de taille, ouverte; par conséquent, aux fins de la comparaison entre chaque modèle et la méthode PM, on complété les résultats de cette dernière en attribuant à la dernière classe de taille le chiffre d'emploi estimé à l'aide du modèle qui fait l'objet de la comparaison : pour les comparaisons avec le modèle lognormal, il faut donc considérer les résultats étiquetés «PM1» et pour les comparaisons avec le modèle logistique, les résultats étiquetés «PM2».

On constate à la lecture du tableau 2 que les indicateurs de biais (erreur moyenne et rapport du nombre de cas de surestimation sur le nombre de cas de sous-estimation) confirment la tendance de la méthode PM à la surestimation. Par ailleurs, s'agissant des modèles lognormal et logistique, on note un très net contraste entre les résultats obtenus pour les branches avec, et sans établissements de 1000 emplois ou plus : dans le premier cas, une tendance à la surestimation (quoique mitigée avec le modèle lognormal) et dans le second, une tendance assez claire à la sous-estimation.

Par ailleurs, la moyenne des valeurs absolues des erreurs est toujours moindre avec les modèles qu'avec la méthode PM, la différence étant naturellement plus marquée pour les branches sans établissement de 1000 emplois ou plus. Enfin, les cotes confirment la domination du modèle lognormal sur le modèle logistique, ainsi que la domination des deux modèles sur la méthode PM.

Dans l'ensemble, donc, cette seconde série de résultats est assez conforme à la première. La seule différence digne de mention serait un biais plus net des modèles dans le cas des branches sans établissement de 1000 emplois ou plus : tous deux ont tendance à sous-estimer l'emploi.

CONCLUSION

Nous avons comparé les modèles lognormal et logistique avec la méthode PM à l'aide de deux ensembles de données : celles du Recensement des établissements et de l'emploi de Montréal de 1996 (RÉEM) et celles de Statistique Canada sur les industries manufacturières à deux chiffres de la CTI au Québec en 1995 (Cat. 31-203-XPB). Pour pouvoir tirer en toute confiance des conclusions générales, il aurait fallu répéter avec plusieurs autres ensembles de données les calculs que nous avons faits avec celles-là. À ce stade-ci, nos conclusions sont donc provisoires.

Supposons néanmoins que nos résultats soient représentatifs. Que conclure alors ? D'abord, le biais à la hausse de la méthode PM semble très réel; pour la recherche appliquée, cela signifie qu'on peut considérer les estimés obtenus par cette méthode comme une limite supérieure de la fourchette d'estimation. Ensuite, les modèles lognormal et logistique donnent plus souvent qu'autrement des estimés qui sont meilleurs que ceux de la méthode PM.

Toutefois, les deux modèles, tel que nous les avons appliqués, conduisent à une surestimation systématique de l'emploi des établissements de la classe de 1000 emplois ou plus (rappelons cependant que la méthode PM ne fournit aucun point de repère pour cette classe d'établissements). En outre, le biais est plus accentué dans le cas du modèle logistique. Cette tendance à la surestimation pourrait inciter à raffiner les modèles. Mais à nos yeux, elle confirme surtout l'utilité de combiner l'application des modèles avec la cueillette de données auprès des établissements de grande taille.

RÉFÉRENCES

Aitchison, J. et Brown, J. A. C. (1957) *The lognormal distribution (with special reference to its uses in economics),* Cambridge University Press, Cambridge.

Cloutier, Norman R. (1995) «Lognormal extrapolation and income estimation for poor Black families», *Journal of Regional Science*, 35(1), 165-171.

De Cola, L. (1985) «Lognormal estimates of macroregional city-size distributions, 1950-1970», *Environment and Planning A*, 17, 1637-1652.

Parr, J. et Suzuki, K. (1973) «Settlement populations and the lognormal distribution», *Urban Studies*, 10, 335-352.

Statistique Canada (1997) *Industries manufacturières du Canada : niveaux national et provincial, 1995*, Cat. 31-203-XPB.

Sutton, John (1997) «Gibrat's legacy», *Journal of Economic Literature*, 35(1), 40-59.

Tableau 1 – Indicateurs de performance des méthodes
Données du RÉEM de 1996

		Classes de taille										
		1-4	**5-9**	**10-19**	**20-49**	**50-99**	**100-199**	**200-499**	**500-999**	**1-999**	**1000 et +**	**Total**
Erreur moyenne												
	PM	6%	8%	12%	19%	15%	14%	24%	14%	15%		
non-pondér.	Logno.	-9%	2%	6%	3%	3%	0%	-1%	-2%	-1%	22%	5%
	Logist.	-3%	2%	5%	2%	3%	1%	0%	-4%	-1%	227%	16%
	PM	14%	9%	12%	18%	12%	13%	19%	12%	14%		
pondér.	Logno.	-14%	2%	5%	4%	3%	4%	-1%	-1%	1%	40%	7%
	Logist.	-1%	2%	4%	3%	3%	3%	-1%	-1%	1%	153%	26%
Rapport du nombre de cas de sur/sous-estimation												
	PM	1,9	6,0	20,0	11,6	13,8	5,0	9,0	7,0	31,0		
	Logno.	0,3	2,6	5,3	2,3	2,7	1,3	0,5	0,4	0,7	17/0	1,0
	Logist.	0,6	2,0	5,3	2,0	2,3	1,5	0,7	0,4	1,1	17/0	1,6
Moyenne des valeurs absolues des erreurs												
	PM	11%	10%	13%	19%	16%	16%	26%	15%	15%		
non-pondér.	Logno.	13%	6%	7%	8%	8%	10%	17%	13%	5%	81%	11%
	Logist.	9%	6%	7%	8%	7%	10%	16%	10%	5%	227%	21%
	PM	15%	9%	12%	18%	12%	14%	19%	13%	14%		
pondér.	Logno.	15%	3%	5%	6%	4%	7%	8%	8%	3%	40%	9%
	Logist.	7%	2%	5%	4%	4%	6%	7%	8%	3%	153%	27%
Écart-type												
	PM	13%	9%	9%	12%	12%	14%	21%	11%	7%		
non-pondér.	Logno.	12%	8%	8%	11%	10%	15%	22%	15%	7%	106%	22%
	Logist.	12%	8%	8%	10%	10%	14%	21%	12%	7%	120%	45%
	PM	11%	3%	4%	7%	6%	8%	12%	9%	4%		
pondér.	Logno.	8%	3%	3%	5%	5%	9%	11%	11%	4%	51%	14%
	Logist.	9%	3%	3%	5%	5%	8%	11%	10%	4%	86%	32%
Cotes												
Lognorm./PM		0,6	4,3	9,5	4,7	6,4	3,5	2,6	1,7	15,0		
Logist./PM		1,1	4,3	9,5	4,7	5,6	3,5	3,0	2,0	15,0		
Logno./Logist.		0,4	0,5	0,2	0,3	0,4	0,5	0,5	0,3	1,1	17/0	2,2
Nombre de cas		64	64	63	63	59	54	40	24	64	17	64

Tableau 2 – Indicateurs de performance des méthodes
Données de 1995 sur le secteur manufacturier au Québec

		Branches avec des établissements de 1000 emplois ou plus	Branches sans établissement de 1000 emplois ou plus
Erreur moyenne			
Non-pondér.	PM1	13,5%	7,3%
	Logno.	6,5%	-6,9%
	PM2	40,6%	7,3%
	Logist.	32,9%	-7,0%
pondér.	PM1	8,1%	9,7%
	Logno.	0,4%	-3,6%
	PM2	27,6%	9,7%
	Logist.	19,2%	-3,7%
Rapport du nombre de cas de sur/sous-estimation			
	PM1	9,0	9,0
	Logno.	0,7	0,1
	PM2	10/0	9,0
	Logist.	4,0	0,1
Moyenne des valeurs absolues des erreurs			
Non-pondér.	PM1	15,1%	10,1%
	Logno.	13,8%	7,1%
	PM2	40,6%	10,1%
	Logist.	34,7%	7,2%
pondér.	PM1	10,4%	10,0%
	Logno.	10,2%	3,7%
	PM2	27,6%	10,0%
	Logist.	21,9%	3,8%
Écart type			
Non-pondér.	PM1	17,6%	7,4%
	Logno.	18,8%	10,0%
	PM2	37,8%	7,4%
	Logist.	40,0%	9,9%
pondér.	PM1	2,1%	0,1%
	Logno.	2,4%	0,2%
	PM2	8,6%	0,1%
	Logist.	9,8%	0,2%
Cotes			
Lognorm./PM1		1,5	2,3
Logist./PM2		9,0	2,3
Logno./Logist.		4,0	1,5
Nombre de cas		10	10

Figure 1 – Erreurs d'estimation comparées, 64 branches, établissements < 1000 emplois

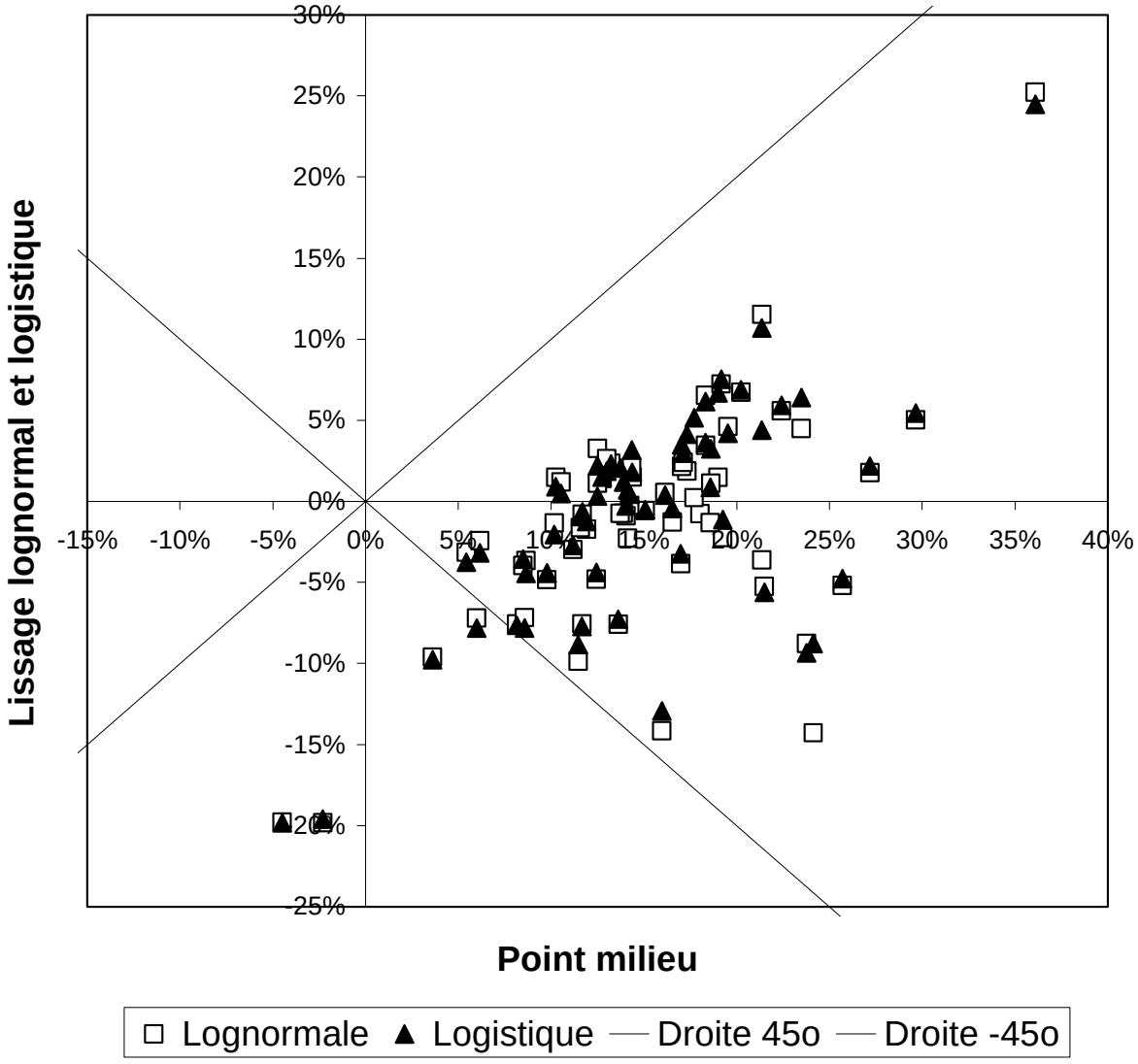

Figure 2 – Erreurs d'estimation comparées, 64 branches, établissements < 1000 emplois