Techniques d'enquête, décembre 1997
Vol. 23, n° 2, pp. 107-115
Statistique Canada

Une procédure adaptative d'estimation robuste du taux d'évolution de l'investissement

PHILIPPE RAVALET[1]

RÉSUMÉ

La présence d'observations extrêmes dans les données d'enquête est un problème récurrent de la statistique appliquée auquel l'enquête de l'INSEE sur l'investissement industriel est aussi confrontée. La prévision du taux de croissance des dépenses d'équipement dans l'industrie se ramène, de ce fait, à l'estimation robuste d'un total dans une population finie. Dans une première partie, cet article analyse l'estimateur actuellement utilisé dans l'enquête Investissement. Nous montrons qu'il suit une stratégie de repondération de l'estimateur linéaire. Mais la dichotomie stricte imposée entre les points extrêmes, tous supposés non représentatifs, et les autres points n'est pas entièrement satisfaisante d'un point de vue à la fois théorique et pratique. L'adoption d'une approche modélisée et l'estimation par les GM-estimateurs, appliqués au cas d'une population finie, permet de pallier ces défauts. Nous construisons ensuite une procédure adaptative robuste qui détermine l'estimateur approprié en fonction des résidus observés sur l'échantillon lorsque ceux-ci peuvent être supposés symétriques. Enfin, cette méthode est appliquée aux données de l'enquête Investissement sur la période 1990-1995.

MOTS CLÉS: Enquêtes de conjoncture; valeurs extrêmes; estimation robuste; GM-estimateur; procédure adaptative.

1. INTRODUCTION

Depuis 1952, l'Institut National de la Statistique et des Études Économiques (INSEE) réalise une enquête sur l'investissement qui fournit des estimations prévisionnelles de l'évolution des dépenses d'équipement dans l'industrie, bien avant la publication des Comptes Nationaux et des résultats d'enquêtes exhaustives. L'estimation du taux de croissance de l'investissement s'appuie sur les déclarations d'environ 2 500 chefs d'entreprise concernant leurs dépenses et intentions de commande en biens d'équipement.

La présence quasi systématique de valeurs extrêmes dans ces données constitue une difficulté majeure. Celles-ci peuvent en effet perturber gravement l'estimation du taux de croissance moyen et conduire à des résultats inacceptables. Selon Chambers (1986), on peut distinguer deux types de points extrêmes. Les points non représentatifs correspondent soit à des erreurs de mesure, que l'on s'efforce de corriger lors de la collecte des données, soit à des individus uniques dans la population. A contrario, les points extrêmes représentatifs désignent des individus curieux mais qui ne peuvent être considérés comme exceptionnels. Il en existe certainement de semblables dans la population non interrogée et l'information qu'ils contiennent doit être intégrée dans l'estimation.

Le problème posé ici s'identifie à celui de l'estimation robuste d'un total dans une population finie avec information auxiliaire, problème auquel la théorie n'apporte pas de réponse définitive. Néanmoins diverses techniques, revues dans Lee (1995), peuvent être appliquées. La méthode d'estimation actuellement utilisée dans l'enquête Investissement suit la logique de repondération de l'estimateur linéaire selon Hidiroglou et Srinath (1981). Toutefois, l'identification et le traitement des points extrêmes ne sont pas entièrement satisfaisants. En particulier, tous les points extrêmes sont supposés non représentatifs et la dichotomie entre points «normaux» et points extrêmes rend l'estimation très sensible au choix de ces derniers.

L'introduction d'un modèle linéaire de superpopulation, qui décrit l'évolution individuelle de l'investissement, permet de mieux apprécier le caractère singulier d'une observation et de définir son niveau de représentativité. Son estimation par les GM-estimateurs constitue alors une alternative séduisante à la méthode des moindres carrés dont la propriété d'absence de biais est très coûteuse en termes de variance. Le réglage de la fonction de poids dépend a priori des caractéristiques de la population selon des critères maintenant bien décrits dans la littérature. Ces caractéristiques pouvant changer d'une strate à l'autre, mais aussi au cours du temps, l'intérêt d'une procédure adaptative est évident. A partir d'une première estimation robuste, on détermine l'allure de la distribution des résidus, puis on choisit l'estimateur à utiliser selon une règle prédéfinie. Suivant Hogg, Bril, Han et Yul (1988), on construit une procédure adaptative s'appuyant sur des indicateurs d'épaisseur de queue et de concentration estimés sur l'échantillon, l'asymétrie des résidus n'étant pas envisagée. Cette procédure est appliquée sur les données de l'enquête Investissement pour la période 1990-1995.

2. L'ESTIMATEUR DE L'ENQUÊTE INVESTISSEMENT

2.1 Principe de l'estimation

Dans une population finie $U = \{1, ..., N\}$, correspondant ici à une strate de l'enquête, on tire un échantillon

[1] Philippe Ravalet, Division des enquêtes de conjoncture, INSEE, 15 Bd G. Péri, BP 100, 92244 MALAKOFF CEDEX.

$s = \{1, ..., n\}$ de taille n, et $\bar{s} = \{n + 1, ..., N\}$ désigne la population non interrogée. Chaque entreprise est interrogée sur ses dépenses d'investissement pour deux années consécutives $t - 1$ et t, notées respectivement x et y.

Connaissant le montant total X des investissements de l'année $t - 1$ dans la population, on peut déduire de l'estimation $\hat{Y}$ du total des investissements pour l'année t, le taux d'évolution moyen des dépenses d'équipement entre $t - 1$ et t:

$$\hat{\theta} = \frac{\hat{Y} - X}{X}.$$

Pour simplifier les notations, on définit le paramètre $\Theta = 1 + \theta = Y/X$, estimé par $\hat{\Theta} = \hat{Y}/X$.

L'estimateur actuellement utilisé dans l'enquête de l'INSEE s'inspire de la méthode du ratio, avec pour information auxiliaire l'investissement réalisé en $t - 1$:

$$\hat{Y}_{\text{ratio}} = \frac{X}{\sum_s x_i} \sum_s y_i.$$

Cet estimateur peut s'écrire comme un estimateur linéaire pondéré:

$$\hat{Y}_{\text{ratio}} = \sum_s w_i z_i. \qquad (1)$$

Dans cette expression, $w_i = Xx_i/\sum_s x_j$ est le poids de l'individu i et $z_i = y_i/x_i$ l'évolution annuelle de son investissement. Un tel estimateur sera sensible à la présence de points extrêmes à la fois sur z et w. *Un point atypique* présentera une évolution z très différente de celle des autres, tandis qu'un *point influent* aura un poids w suffisamment important pour attirer, par effet levier, le taux d'évolution moyen de la strate vers son propre taux d'évolution. Le critère décisif pour qualifier une observation de point extrême étant que le produit wz soit assez grand pour perturber l'estimation $\hat{Y}_{\text{ratio}}$, la distinction entre points atypiques et points influents est bien entendu arbitraire. Le terme générique *grands investisseurs* (ou GI en abrégé) désignera l'ensemble de ces points extrêmes tandis que le terme *extrapolables* fera référence aux autres individus de l'échantillon.

Ayant réalisé une partition a posteriori de l'échantillon $s = \{\text{GI}\} \cup \{\textit{extrapolables}\}$, on estime le total des investissements du reste de la population $\bar{s}$ à partir du comportement des seuls individus extrapolables selon la méthode du ratio:

$$\hat{Y}_{\text{GI}} = \sum_s y_i + \left(\sum_{\bar{s}} x_i\right) \frac{\sum_{\{\text{extra}\}} y_i}{\sum_{\{\text{extra}\}} x_i}. \qquad (2)$$

Dans (2), le poids des extrapolables $1 + \sum_{\bar{s}} x_i / \sum_{\{\text{extra}\}} x_i$ est bien strictement plus grand que celui des grands investisseurs qui vaut 1.

2.2 Sélection des Grands Investisseurs

Les grands investisseurs sont choisis, au niveau de chaque strate, en fonction de leur influence sur l'estimation de Θ selon une procédure itérative. Pour commencer, les individus sont tous supposés extrapolables et on calcule pour chacun d'eux un indice de non prise en compte, mesurant l'impact sur $\hat{\Theta}$ de son exclusion de l'échantillon, $\text{NPEC} = (\hat{Y}^i_{\text{GI}} - \hat{Y}_{\text{GI}})/X$ où $\hat{Y}^i_{\text{GI}}$ est le total estimé sans l'individu i.

L'entreprise ayant le plus grand indice NPEC en valeur absolue est déclarée grand investisseur. On réestime alors $\hat{Y}_{\text{GI}}$ avec cette nouvelle partition de U, puis on identifie le grand investisseur suivant. La sélection s'interrompt dès que tous les individus extrapolables ont une influence sur l'estimation inférieure à un seuil donné. Cette condition est d'autant plus facilement vérifiée que le nombre et la masse des observations sont importants. Inversement, elle se révélera impossible à réaliser si le nombre d'individus est trop faible; dans ce cas, le gestionnaire d'enquête veille simplement à ce qu'aucun individu n'ait une influence beaucoup plus grande que les autres, introduisant ainsi une dose de subjectivité dans la procédure.

Par ce mécanisme itératif, les phases habituelles de détection et de traitement des points extrêmes sont réalisées de façon simultanée. La principale difficulté tient dans le fait que le statut d'un individu n'est pas une qualité intrinsèque, mais dépend de la composition de l'échantillon. Celui-ci peut changer d'une enquête à l'autre. En outre, cette procédure peut conduire dans certains cas de figure (Ravalet 1996) à exclure inutilement certains individus car, à aucun moment, le statut de grand investisseur n'est remis en question.

2.3 La stratégie de repondération de l'estimateur linéaire

L'estimateur GI suit en fait de la stratégie de repondération de l'estimateur linéaire (1) présentée par Hidiroglou et Srinath (1981) sur l'exemple de l'estimation d'un total sans information auxiliaire. Ayant réalisé a priori une partition $s = s_1 \cup s_2$ de l'échantillon distinguant les points extrêmes s_1 (en nombre n_1) des autres observations s_2, les auteurs proposent de réduire, dans $\hat{Y} = (N/n)\sum_s y_i$, le poids N/n des points extrêmes à une valeur plus faible λ en posant

$$\hat{Y}_\lambda = \lambda \sum_{s_1} y_i + \frac{N - \lambda n_1}{n - n_1} \sum_{s_2} y_i$$

soit

$$\hat{Y}_\lambda = \sum_s y_i + \frac{N - n}{n - n_1} \sum_{s_2} y_i + n_1(\lambda - 1)\left[\frac{1}{n_1}\sum_{s_1} y_i - \frac{1}{n - n_1}\sum_{s_2} y_i\right].$$

La valeur optimale de λ qui minimise l'écart quadratique moyen de cet estimateur, conditionnellement ou non au nombre de valeurs extrêmes dans l'échantillon, est fonction de plusieurs paramètres de la population. Sans information a priori, le choix de λ est délicat.

Appliqué au cas de l'estimateur du ratio avec variable auxiliaire x, cela s'écrit:

$$\hat{Y}_{\text{ratio}\,\lambda} = \sum_s y_i + \sum_{\bar{s}} x_i \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} + (\lambda - 1)\left(\frac{\sum_{s_1} y_i}{\sum_{s_1} x_i} - \frac{\sum_{s_2} y_i}{\sum_{s_2} x_i} \right) \sum_{s_1} x_i \quad (3)$$

Les deux premiers termes du second membre de (3) forment une estimation du total Y, sous l'hypothèse implicite que tous les points extrêmes sont dans l'échantillon, et le troisième est une correction tenant compte de la présence éventuelle de tels points dans la population non interrogée. Cette correction est fonction du λ retenu et de la différence des comportements moyens entre les deux types d'individus estimés sur l'échantillon.

En rapprochant (2) et (3), on s'aperçoit que l'estimateur GI est formellement équivalent au cas $\lambda = 1$. L'utilisation de $\hat{Y}_{\text{GI}}$ suppose donc implicitement que les points extrêmes ont été correctement identifiés et sont tous non représentatifs. Dans Ravalet (1996), on a montré que ces deux hypothèses étaient malheureusement rarement vérifiées dans le contexte de l'enquête Investissement.

La procédure d'identification étant manuelle et le critère retenu relativement *ad hoc* en l'absence de toute hypothèse sur la population, il n'est pas exclu que certains points extrêmes échappent à la sélection. L'utilisation du ratio sur les extrapolables pose alors le problème de la robustesse de l'estimation vis à vis du choix des grands investisseurs. En outre, tous ces points ne sont vraisemblablement pas uniques. Les points atypiques, particulièrement nombreux chez les petites et moyennes entreprises, devraient plutôt être considérés comme représentatifs. Toutefois, choisir $\lambda > 1$ introduirait immanquablement la question de la robustesse du troisième terme de (3).

Des modifications de l'estimateur $\hat{Y}_{\text{GI}}$ sont envisageables pour tenter de pallier ces défauts. La moyenne sur les extrapolables peut être par exemple remplacée par un estimateur plus robuste et seuls les points non représentatifs sont déclarés grands investisseurs. Cette technique s'inscrit dans le cadre plus général des M-estimateurs où la donnée d'un modèle facilite à la fois le repérage et le traitement des points extrêmes (Lee 1995). Il ne s'agit plus alors de procéder à une dichotomie stricte entre points extrêmes et autres points mais de définir des zones de plus ou moins grande représentativité.

3. ESTIMATION ROBUSTE PAR LES GM-ESTIMATEURS

3.1 Le modèle linéaire et les GM-estimateurs

On suppose l'existence d'un modèle linéaire ξ reliant pour l'ensemble de la population U les investissements x et y aux dates $t-1$ et t.

$$\xi: y_i = \beta x_i + \epsilon_i$$

avec

$$E(\epsilon_i) = 0$$
$$E(\epsilon_i \epsilon_j) = 0 \qquad \forall\, i \neq j.$$
$$V(\epsilon_i) = \sigma^2 \eta(x_i)$$

La pente β de la droite de régression passant par l'origine dans le modèle de superpopulation s'interprète comme le taux d'évolution Θ dans la population. La variance de y est supposée fonction croissante de x et η est en général une fonction puissance: $\eta(x_i) = x_i^{\gamma}$.

Sous le modèle, le meilleur estimateur linéaire sans biais (Brewer 1963 et Royall 1970) du total est $\hat{Y}_{mc} = \sum_s y_i + \hat{\beta}_{mc} \sum_{\bar{s}} x_i$, où $\hat{\beta}_{mc} = (\sum_s x_i y_i / \eta(x_i)) / (\sum_s x_i^2 / \eta(x_i))^{-1}$ est l'estimateur des moindres carrés.

Dans le cas particulier $\eta(x) = x$, cette expression se réduit à $\hat{\beta}_{mc} = \sum_s y_i / \sum_s x_i$, estimateur du ratio. Cet estimateur sans biais n'est efficace que sous l'hypothèse de normalité des résidus et se montre peu robuste.

Les M-estimateurs (Huber 1981) permettent de définir une version robuste des moindres carrés en substituant à la fonction carré, dans le programme de minimisation, une fonction ρ croissant moins rapidement:

$$\text{Min} \sum_s \rho\left(\frac{y_i - \beta_R x_i}{\sigma \sqrt{\eta(x_i)}} \right).$$

Le M-estimateur $\hat{\beta}_R$ est la solution de l'équation implicite:

$$\sum_s \psi\left(\frac{y_i - \hat{\beta}_R x_i}{\sigma \sqrt{\eta(x_i)}} \right) \frac{x_i}{\sqrt{\eta(x_i)}} = 0$$

où

$$\psi(t) = \frac{\partial \rho(t)}{\partial t}.$$

La fonction ψ, comme la fonction de Huber $\psi(t) = \text{Max}(-c, \text{Min}(t, c))$, dépend d'une (ou plusieurs) constante de réglage c contrôlant la part des observations qui doivent être considérées comme points extrêmes. Cet estimateur sera encore sensible à la présence de valeurs extrêmes sur la variable explicative x. On définit alors une classe plus générale d'estimateurs appelés GM-estimateurs (Hampel, Ronchetti, Rousseeuw et Stahel 1986) par l'équation implicite:

$$\sum_s w\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right)\psi\left(\frac{r_i}{\sigma}v\left(\frac{x_i}{\sigma\sqrt{\eta(x_i)}}\right)\right)\frac{x_i}{\sqrt{\eta(x_i)}}=0$$

avec

$$r_i=\frac{y_i-\hat{\beta}_R x_i}{\sqrt{\eta(x_i)}}.$$

Un choix habituellement retenu est la forme de Mallows: $v(t) = 1$ et $w(t) = 1/t$. Un estimateur robuste $\hat{\beta}_R$ vérifiera donc l'équation implicite

$$\sum_s \psi\left(\frac{y_i-\hat{\beta}_R x_i}{\sigma\sqrt{\eta(x_i)}}\right)=0. \quad (4)$$

En général, le paramètre σ est inconnu et doit être remplacé dans cette expression par une estimation robuste $\hat{\sigma}$ de la dispersion des résidus

$$\sum_s \psi\left(\frac{y_i-\hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right)=\sum_i \psi\left(\frac{r_i}{\hat{\sigma}}\right)=0.$$

L'estimateur du total sera finalement:

$$\hat{Y}_{\beta R}=\sum_s y_i+\hat{\beta}_R\sum_{\bar{s}} x_i. \quad (5)$$

Cet estimateur est étudié par Gwet et Rivest (1992). En général, il n'est pas sans biais par rapport au plan de sondage. Chambers (1986) propose de corriger ce biais en introduisant dans (5) un troisième terme qui l'estime de façon robuste:

$$\hat{Y}_{\text{Chambers}}=\sum_{i\in s} y_i+\hat{\beta}_R\sum_{i\in\bar{s}} x_i+$$

$$\left(\sum_{i\in s}\frac{x_i/\hat{\sigma}\sqrt{\eta(x_i)}}{\sum_{j\in s} x_j^2/\hat{\sigma}^2\eta(x_j)}\psi_E\left(\frac{y_i-\hat{\beta}_R x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right)\right)\sum_{i\in\bar{s}} x_i$$

Choisir une fonction ψ_E bornée semble un bon compromis entre biais et variance de l'estimateur. Par exemple, Welsh et Ronchetti (1994) optent pour une fonction de Huber avec une constante de réglage grande $c = 15$. Mais le réglage de ψ_E, sans information préalable sur la densité des points extrêmes, est toujours délicat.

3.2 Choix de l'estimateur

Les propriétés souhaitables des fonctions ψ sont désormais bien connues par référence au problème de l'estimation d'une tendance centrale. Celles-ci doivent être bornées, continues, équivalentes à l'identité au voisinage de zéro. On distingue habituellement les fonctions strictement monotones (Huber) des fonctions redescendantes comme la fonction bicarrée de Tukey, le sinus d'Andrew et la fonction de Hampel ou de Cauchy. Parce que leur fonction d'influence tend vers zéro, ces estimateurs seront moins sensibles à la présence de points extrêmes que la fonction de Huber. La vitesse de convergence vers zéro est une caractéristique essentielle des fonctions redescendantes. Celles, nulles à distance finie (Hampel, Tukey ou Andrew) excluent les points extrêmes de l'estimation de β alors que les autres leur accordent une faible représentativité.

Le choix et le réglage de la fonction ψ sont délicats. Ils dépendent beaucoup de la nature des données et plus précisément de la distribution des résidus (Hoaglin, Mosteller et Tukey 1983, chap. 11). Une idée, ne serait-ce qu'approximative, de l'allure de la distribution des résidus devrait permettre de mieux cibler le choix et le réglage de l'estimateur, donc de rendre l'estimation plus efficace. Cette remarque intuitive est à l'origine des procédures adaptatives, présentées notamment par Hogg (1974) et (1982). L'idée est d'apprécier la nature de la distribution des résidus, calculés à partir d'une première estimation robuste (du type norme L_1 par exemple), à l'aide d'indicateurs robustes bien choisis (épaisseur de queue, asymétrie, concentration *etc.*). La donnée de ces indicateurs permet alors de choisir, selon une règle de décision prédéfinie, l'estimateur adapté à cette situation et on résout l'équation implicite (4) en prenant comme valeur initiale la première estimation robuste de β.

Le principe d'une procédure adaptative apparaît d'autant plus séduisant qu'il systématise l'étude préalable nécessaire au choix et au réglage d'un estimateur. Celle-ci peut en effet s'avérer extrêmement coûteuse si elle doit être réalisée manuellement pour chaque strate de l'échantillon et renouvelée à chaque enquête.

4. CONSTRUCTION D'UNE PROCÉDURE ADAPTATIVE

On décrit ici la construction d'une procédure adaptative pour le calcul du taux d'évolution moyen de l'investissement à partir des données de l'enquête de conjoncture. Aussi certains choix ont-ils été effectués sachant la nature et les caractéristiques propres de ces données et ne sont pas nécessairement transposables à d'autres modèles de régression. En particulier, on a retenu, après vérification sur les données, l'hypothèse de symétrie de la distribution des résidus et exclu le cas de distributions à queue fine.

La construction d'une procédure adaptative, qui s'inspire des travaux de Moberg, Ramberg et Randles (1980), s'effectue en plusieurs étapes. On choisit la fonction (ou famille de fonctions) ψ à utiliser, puis on sélectionne l'ensemble des critères servant à qualifier la distribution des résidus. La donnée de ces critères permet la construction d'une règle de classification. Enfin, à chaque classe est associé le réglage de l'estimateur à utiliser.

4.1 Choix de la fonction ψ

Les fonctions monotones du type Huber n'assurant pas une protection suffisante contre les points extrêmes, seules les fonctions redescendantes ont été prises en considération. Parmi celles-ci, on a retenu les fonctions de Cauchy généralisée (utilisées notamment par Moberg et coll. 1980 pour approximer les fonctions lambda généralisées) et bicarrée de Tukey:

$$\psi_c(r) = \frac{cr}{(b + r)^2 + c}, \quad \forall r$$

et

$$\psi_T(r) = \frac{r}{c}\left(1 - \frac{r^2}{c^2}\right)^2, \quad \forall |r| \le c.$$

Ces deux estimateurs se différentient nettement dans le traitement des points extrêmes (voir figure 1). La fonction bicarrée suit l'identité plus longtemps que la fonction de Cauchy mais présente en revanche un point de rejet fini: les résidus au-delà de $c^*\sigma$ n'interviennent pas dans l'estimation alors que la fonction de Cauchy leur accorde une certaine représentativité. Le paramètre b permet, en principe, de contrôler l'asymétrie de ψ en fonction de celle des résidus.

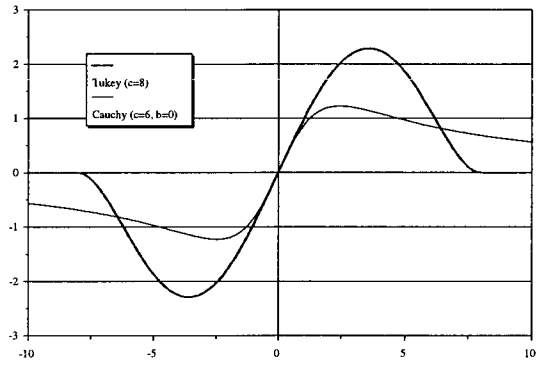

Figure 1. Fonction de Cauchy et de Tukey

4.2 Paramètre d'échelle, algorithme de calcul et critères de sélection

De façon générale un estimateur $\hat{\sigma}$ de dispersion est défini par une équation implicite $\sum\chi(r_i/\hat{\sigma}) = 0$, où χ est une fonction paire. Il s'agit donc de résoudre le système d'équations non linéaires en $(\hat{\beta}, \hat{\sigma})$ suivant:

$$\begin{cases} \sum_i \psi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0 \\ \sum_i \chi\left(\dfrac{y_i - \hat{\beta}x_i}{\hat{\sigma}\sqrt{\eta(x_i)}}\right) = 0. \end{cases} \tag{6}$$

Rivest (1989) montre sur quelques exemples que la résolution du système (6) peut poser des difficultés en raison d'une éventuelle multiplicité des solutions, même dans le cas d'une fonction ψ monotone. Suivant ses recommandations, nous procédons en deux étapes. Dans un premier temps, le paramètre de dispersion σ est estimé à l'aide de la médiane des valeurs absolues (MAD) des résidus définis à partir de la médiane des taux d'évolution individuels. Ensuite β est calculé par (4) en utilisant la valeur de σ trouvée précédemment.

Pour la résolution de (4), nous avons préféré l'algorithme de repondération à l'algorithme de Newton-Raphson, car il semble converger plus facilement, notamment lorsque la constante de réglage est petite.

L'efficacité d'une procédure adaptative reposant sur celle du processus décisionnel, la plus grande attention doit être portée sur la nature, la qualité et la robustesse des informations commandant le choix de l'estimateur.

L'épaisseur de queue est un indicateur indispensable car elle renseigne sur l'importance relative des points extrêmes dans l'échantillon, donc dans la population (voir Hoaglin et coll. 1983, chap. 10). On a retenu comme indicateur d'épaisseur de queue la proposition de Hogg (1974):

$$\tau(p) = \frac{\bar{U}(p) - \bar{L}(p)}{\bar{U}(0{,}5) - \bar{L}(0{,}5)}$$

$\bar{U}(p)$ (resp. $\bar{L}(p)$) est la moyenne des np plus grandes (resp plus petites) statistiques d'ordre, en utilisant une interpolation linéaire lorsque np n'est pas entier. On a choisi $p = 0{,}05$; pour la loi normale $\tau(,05)$ vaut 2,59.

De plus, il nous a semblé important, comme Hogg et coll. (1988), de tester la présence éventuelle d'une distribution du type double exponentielle, en mesurant la concentration des résidus par l'indicateur pk suivant:

$$pk = \frac{\bar{X}(1 - \beta, 1 - \alpha) - \bar{X}(\alpha, \beta)}{\bar{X}(0{,}5,\ 1 - \beta) - \bar{X}(\beta, 0{,}5)}$$

où $\bar{X}(a, b)$ est la moyenne des statistiques d'ordre entre la na-ième et la nb-ième, avec des grandeurs interpolées si na ou nb ne sont pas entiers. On a retenu $\alpha = 0{,}05$ et $\beta = 0{,}15$, soit $pk = 2{,}7$ pour une distribution normale.

Enfin, des études (Moberg et coll. 1980, Hogg et coll. 1988) ont souligné l'importance de la dissymétrie des distributions. En effet, en présence de résidus asymétriques, le biais des estimateurs robustes peut être important, rendant ainsi leur utilisation délicate (Chambers et Kokic 1993). Dans l'enquête Investissement de l'INSEE, les résidus sont théoriquement asymétriques puisque minorés ($r = y - \beta x \ge -\beta x$). Toutefois, nous avons constaté empiriquement que cette asymétrie était très légère et qu'elle pouvait être négligée sans dommages. L'échec de la correction d'un éventuel biais par la fonction ψ_E dans

l'estimateur de Chambers, de plus, confirme cette observation. Seul le cas symétrique est considéré ici; le biais des estimateurs définis par (5) est nul par conséquent.

4.3 Classification des distributions et réglage de l'estimateur

La définition de la règle de décision s'est appuyée sur l'étude de huit distributions symétriques particulières illustrant diverses situations d'épaisseur de queue et de concentration (voir tableau 1). La famille des distributions contaminées $CN(\alpha, K)$, de fonction de répartition $F(x) = (1 - \alpha)\Phi(x) + \alpha\Phi(x/K)$ où Φ est la fonction cumulative de la loi $N(0,1)$, nous a paru intéressante car ces lois donnent une bonne représentation de données réelles (Hoaglin et coll. 1983 chap. 10) et notamment celles de l'enquête investissement (Ravalet 1996). Gaussienne en leur milieu, elles contiennent néanmoins plus d'observations extrêmes que la loi normale $N(0,1)$.

Tableau 1
Huit distributions particulières

		$\tau(,05)$	pk
1	loi normale	2,59	2,76
2	loi contaminée $CN(.05,3)$	2,94	2,83
3	loi double exponentielle	3,28	3,41
4	loi contaminée $CN(.05,10)$	4,47	2,85
5	loi contaminée $CN(.10,10)$	5,42	3,05
6	loi contaminée $CN(.20,10)$	5,64	4,44
7	loi Slash	7,65	4,19
8	loi de Cauchy	7,82	4,78

Les deux indicateurs $\tau(0,5)$ et pk, ont été simulés sur ces huit lois, et ce, pour plusieurs tailles d'échantillon. Le graphe de $(\tau(0,5), pk)$ permet de distinguer quatre groupes de distributions: les distributions à queue peu épaisse et peu concentrée du type loi normale ou $CN(,05, 3)$, les distributions à queue épaisse du type $CN(,05,10)$, $CN(,10,10)$, et $CN(,20,10)$, puis les distributions à queue très épaisse du type Slash et Cauchy, et enfin les distributions concentrées comme la loi double exponentielle. Ces quatre classes sont définies (voir figure 2) par les frontières d'équation:

Classe I: $\tau(0,5) \le 3,6 - \frac{14}{n}$ et $pk \le 3,20$

Classe II: $3,6 - \frac{14}{n} < \tau(0,5) \le 5,8 - \frac{35}{n}$

Classe III: $5,8 - \frac{35}{n} < \tau(0,5)$

Classe IV: $\tau(0,5) \le 3,6 - \frac{14}{n}$ et $pk > 3,20$.

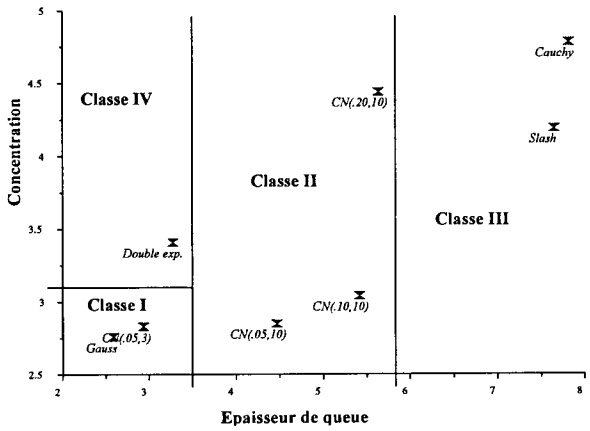

Figure 2. Quatre classes de distributions

L'ultime étape consiste à fixer le réglage des deux estimateurs dans chaque classe. Puisque l'on ne s'intéresse qu'au cas symétrique, le paramètre b de la fonction de Cauchy est nul. Par simulations, on a déterminé pour les huit lois de référence les constantes c optimales des fonctions de Tukey et de Cauchy (*i.e.*, minimisant la variance de ces estimateurs ou, ce qui revient au même ici, leur écart quadratique moyen). Celles-ci diminuent bien avec l'épaisseur de queue, si l'on excepte naturellement le cas de la loi double exponentielle qui requiert un réglage voisin de ceux utilisés pour les lois Slash et Cauchy.

L'estimateur de Tukey est plus efficace sur les lois normale ou contaminées, mais il nécessite en général un réglage plus fin. La figure 3 montre l'exemple de la loi contaminée $CN(,10, 10)$. Enfin, si le choix de la constante apparaît relativement critique pour les lois à queue épaisse ou concentrées, une large bande de valeur est envisageable pour les lois proches de la normale.

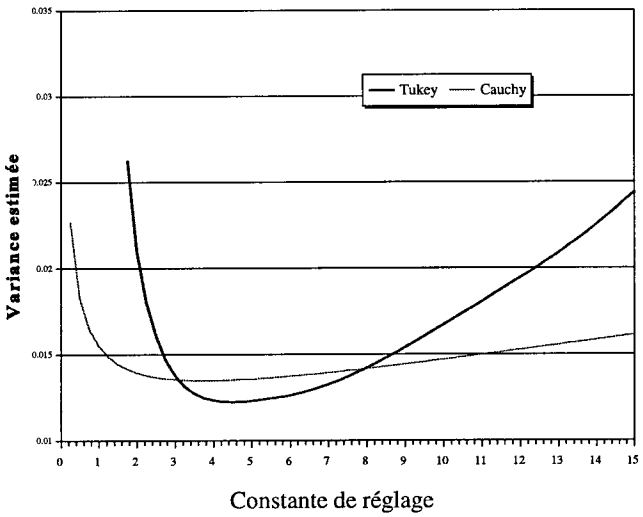

Figure 3. Variance des estimateurs de Tukey et de Cauchy pour la loi $CN(,10, 10)$ $(n = 100)$

La synthèse de ces résultats permet de définir les réglages à employer sur chaque classe de distribution. Ces réglages, établis pour des échantillons de taille 100 (tableau 2), restent tout à fait acceptables pour des échantillons dont la taille est comprise entre 50 et 150.

Tableau 2
Réglage des estimateurs selon la classe des distributions des résidus ($n = 100$)

Classe	Tukey	Cauchy
I	7	7
II	4,5	4
III	3	1
IV	3	1

5. APPLICATION À L'ENQUÊTE INVESTISSEMENT

5.1 Le problème de la stratification

Les strates utilisées pour l'estimateur GI sont définies par le croisement d'une activité (18 secteurs manufacturiers) et d'une tranche de taille d'entreprise (petites, moyennes et grandes). Parmi ces 54 strates, une vingtaine environ ne regroupent jamais plus de vingt observations. Cette stratification est donc trop fine pour l'utilisation correcte de la procédure adaptative qui suppose un nombre minimal d'observations.

Comme les petites entreprises se distinguent assez nettement des moyennes et des grandes, en termes de dispersion et d'épaisseur de queue des résidus, on conserve la différentiation par taille. Des secteurs doivent donc être regroupés. La méthode, utilisée par Sohre (1995), qui consiste à regrouper après la collecte des données les secteurs ayant des paramètres (ici l'évolution moyenne de l'investissement) les plus proches, n'a pas été retenue. La proximité est en effet impossible à apprécier sur de petites strates et les regroupements obtenus sont susceptibles de changer d'une enquête à l'autre, rendant les comparaisons difficiles. Nous avons préféré redéfinir 15 nouvelles strates à partir d'un niveau de nomenclature supérieur distinguant quatre secteurs seulement: biens intermédiaires, biens d'équipement professionnel, automobile et biens de consommation.

5.2 Caractéristiques des strates

L'hypothèse d'une variance des résidus indépendante de x dans le modèle ξ ne peut être acceptée. Le choix de γ dans la fonction η s'effectue de façon à ce que la courbe des résidus (en valeur absolue) en fonction du régresseur, lissée par la méthode du LOESS, ne présente pas de tendance (Cleveland 1979). Pour la strate – biens intermédiaires, taille moyenne – à l'enquête d'avril 1995 (voir figure 4), $\gamma = 1{,}3$ est un compromis acceptable entre l'apparition d'une tendance à la baisse pour les x petits et l'annulation de la tendance à la hausse pour les plus grandes valeurs de x. Un examen similaire sur les autres strates a confirmé ce choix pour l'ensemble de l'industrie manufacturière.

Dans chaque strate, la distribution des résidus apparaît systématiquement à queue plus épaisse que la loi normale, sans être à queue très épaisse. Dans un même secteur d'activité, l'indice d'épaisseur de queue décroît avec la taille des entreprises. La grande majorité des strates représentant les petites et moyennes entreprises ont été affectées dans la classe 2. Les grandes entreprises présentent plus souvent des distributions de résidus à queue peu épaisse, proches soit de la loi normale (classe 1), soit de la loi double exponentielle (classe 4). La classe 2 est largement majoritaire et représente 75 % des cas. Seulement 20 % des distributions sont reconnues à queue peu épaisse et affectées en proportions égales dans les classes 1 et 4. En revanche, les distributions à queue très épaisse (classe 3) sont exceptionnelles (moins de 5 % des cas). S'il semble exister une certaine rémanence de la classification, celle-ci n'est pas parfaite. Et les changements sont bien réels puisqu'ils résistent à une légère modification des frontières entre les classes. Ceci justifie donc parfaitement l'utilisation d'une procédure adaptative.

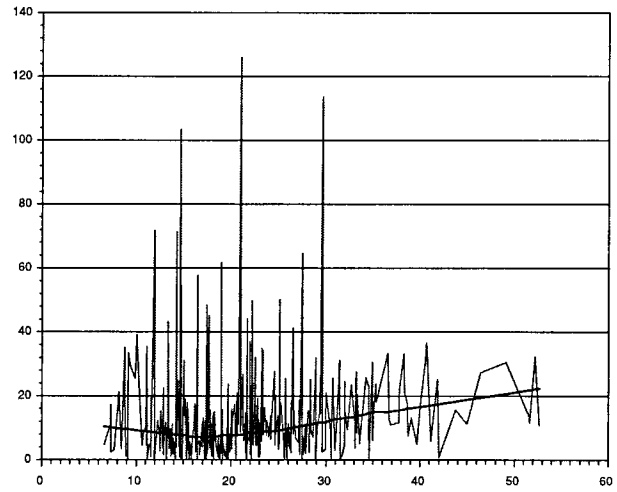

Figure 4. Valeur absolue des résidus ($\gamma = 1{,}3$, biens intermédiaires taille 2, Avril 95)

5.3 Les estimations réalisées

La procédure d'estimation basée sur (5), appliquée aux six enquêtes couvrant la période 1990-1995, a donné les résultats portés sur la figure 5. On y trouvera aussi les estimations de la Comptabilité Nationale, celles obtenues par l'estimateur GI ainsi que celles calculées issues de l'Enquête Annuelle d'Entreprise (E.A.E) qui est exhaustive.

Sur l'ensemble de l'industrie manufacturière, les résultats de la procédure adaptative sont comparables à ceux de l'estimateur GI. La fonction bicarrée conduit à des estimations toujours inférieures à celles obtenues avec la fonction de Cauchy. Avec un point de rejet fini, la fonction de Tukey est en effet moins influencée par la légère

asymétrie vers la droite de la distribution des résidus. Par ailleurs, ces nouvelles estimations se rapprochent plus de celles de l'E.A.E que des Comptes Nationaux. Ceci n'est guère très surprenant vu l'excellente corrélation entre les données individuelles de l'E.A.E et les réponses obtenues à l'enquête. Les écarts en 1991 et 1994 par rapport aux comptes demeurent pour l'instant inexplicables. En dehors de l'année 1994, les estimations obtenues avec la fonction de Cauchy sont tout à fait acceptables dans les secteurs des biens intermédiaires, de l'automobile, et dans une moindre mesure des biens d'équipement professionnel. En revanche, dans les biens de consommation, les résultats sont assez éloignés des Comptes Nationaux. On se heurte ici vraisemblablement à un problème de qualité de l'échantillon. Ce secteur est très hétérogène et quelques activités comme l'imprimerie sont mal couvertes par l'enquête.

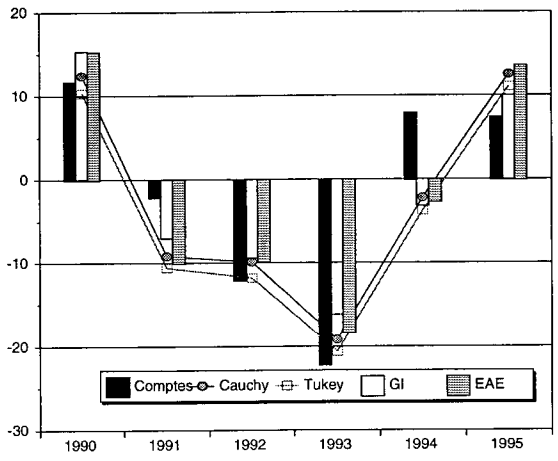

Figure 5. Taux de croissance de l'investissement en valeur dans l'industrie manufacturière

6. CONCLUSIONS

Cet article présente une justification théorique de la procédure actuellement utilisée pour dépouiller l'enquête Investissement, et notamment du principe d'exclusion des points extrêmes ou grands investisseurs. Toutefois la stratégie de repondération de l'estimateur linéaire à la Hidiroglou et Srinath (1981) présente ici des insuffisances, liées pour l'essentiel à l'identification et au traitement des points extrêmes représentatifs. La dichotomie entre individus extrapolables et grands investisseurs apparaît trop radicale et conduit à un manque de robustesse, puisque la courbe d'influence de cet estimateur n'est pas continue.

En revanche, l'hypothèse d'un modèle linéaire de superpopulation et son estimation par les GM-estimateurs nous ont semblé être d'un grand intérêt méthodologique et pratique. L'insertion de ces techniques au sein d'une procédure adaptative permet, de plus, de disposer d'un estimateur robuste pour un ensemble varié de situations. Suivant les principes décrits dans la littérature, la procédure proposée ici utilise des indicateurs d'épaisseur de queue et de concentration des résidus du modèle linéaire calculés sur l'échantillon, pour décider du réglage de la fonction de poids à utiliser, les résidus étant supposés par ailleurs symétriques. Les estimations réalisées avec la fonction de Cauchy ont donné des résultats satisfaisants sur l'industrie manufacturière et valident largement celles déjà publiées. Les avantages de cette méthode par rapport à celle utilisée actuellement s'expriment pour l'essentiel en termes de coûts de mise en oeuvre et d'une plus grande maîtrise de la méthodologie employée.

La procédure adaptative a été construite indépendamment de l'enquête. Aussi l'optimalité de la classification par rapport au contenu des strates n'est pas garantie. Par ailleurs, nous n'avons pas étudié la robustesse de la règle d'affectation à une classe. Cette question est importante lorsque l'on effectue plusieurs mesures successives et l'on désire en interpréter les révisions. A l'évidence, d'autres recherches sur ces méthodes de classification sont nécessaires, pour intégrer, par exemple, l'information livrée par les estimations précédentes ou les enquêtes exhaustives sur la population étudiée.

REMERCIEMENTS

L'auteur tient à remercier Michel Hidiroglou et Dominique Ladiray pour leurs commentaires et suggestions lors de l'élaboration de cet article.

BIBLIOGRAPHIE

BREWER, K.R. (1963). Ratio estimation and finite population: some results deducible from the assumption of an underlying stochastic process. *The Australian Journal of Statistics,* 5, 93-105.

CHAMBERS, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association,* 81, 1063-1069.

CHAMBERS, R.L., et KOKIC P.N. (1993). Outlier robust sample survey inference. *Bulletin de l'Institut International de Statistique, actes de la 49*ième *session,* livraison 2, 55-72.

CLEVELAND, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association,* 74, 829-836.

GWET, J.P., et RIVEST, L.P. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association,* 87, 1174-1182.

HAMPEL, F.R., RONCHETTI, E., ROUSSEEUW, P.J., et STAHEL, W.E. (1986). *Robust Statistics: The Approach Based on Influence Function.* New York: John Wiley.

HIDIROGLOU, M.A., et SRINATH, K.P. (1981). Some estimators of the population total from simple random samples containing large units. *Journal of the American Statistical Association,* 76, 690-695.

HOAGLIN, D.C., MOSTELLER, F., et TUKEY, J.W. (1983). *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley.

HOGG, R.V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of The American Statistical Association,* 69, 909-923.

HOGG, R.V. (1982). On adaptive statistical inferences. *Communication in Statistics*, 11, 2531-2542.

HOGG, R.V., BRIL, G.K., HAN, S.M., et YUL, L. (1988). An argument for adaptive robust estimation. *Probability and Statistics: Essays in Honor of Franklin A. Graybill.* Amsterdam: North-Holland/Elsevier, 135-148.

HUBER, P.J. (1981). *Robust Statistics*. New York: John Wiley.

LEE, H. (1995). Outliers in business surveys. Dans *Business Survey Methods*. New York: John Wiley.

MOBERG, T.F., RAMBERG, J.S., et RANDLES, R.H. (1980). An adaptive multiple regression procedure based on M-estimators. *Technometrics*, 22, 213-224.

RAVALET, P. (1996). L'estimation du taux d'évolution de l'investissement dans l'enquête de conjoncture: analyse et voie d'amélioration. Document de travail de l'INSEE Méthodologie Statistique, 9604.

RIVEST, L.P. (1989). De l'unicité des estimateurs robustes en régression lorsque le paramètre d'échelle et le paramètre de régression sont estimés simultanément. *La Revue Canadienne de Statistique*, 17, 141-153.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57, 377-387.

SOHRE, P. (1995). The Adaptive KOF Procedure for the Estimation of Industry Investment. 22nd CIRET Conference, Singapore.

WELSH, A.H., et RONCHETTI, E. (1994). Bias-Calibrated Estimations of Totals and Quantiles from Sample Surveys Containing Outliers. Rapport Technique, Department of Econometrics, University of Geneva, Switzerland.