Méthodes de détection des unités atypiques:Cas des enquêtes structurelles ukrainiennes

Michel Grun-Rehomme, Olga Vasechko

Méthodes de détection des unités atypiques: Cas des enquêtes structurelles ukrainiennes

Olga A. Vasechko
Research Institute of Statistics, 3 Shota Rustaveli str., 01023 Kyiv, Ukraine
E-Mail : O.Vasechko@ukrstat.gov.ua

Michel Grun-Réhomme
Université Paris 2, ERMES-EA4441-CNRS, 12 place du Panthéon, 75005 Paris, France
E-Mail: grun@u-paris2.fr

Résumé
De nombreuses méthodes algébriques, graphiques ou probabilistes existent pour détecter les valeurs extrêmes (« outliers ») sur une variable. Dans cette présentation, deux nouvelles méthodes non paramétriques simples de détection des unités atypiques dans le cadre univarié sont proposées. La première dans le domaine de la V-robustesse permet de tenir compte de l'éloignement de l'observation par rapport au centre de la distribution et de la forme de la fin de la distribution. Et la seconde, basée sur une statistique du coefficient de variation, permet d'accentuer le caractère atypique des unités et de mesurer l'effet d'une unité atypique sur la qualité de l'information. De plus, une troisième méthode, obtenue comme combinaison convexe de deux techniques de la théorie des valeurs extrêmes, en minimisant la variance, est également proposée.
Ces différentes méthodes sont comparées sur des données des enquêtes structurelles ukrainiennes. Il est nécessaire d'utiliser une technique facile à mettre en production.

Mots clés
Statistiques d'entreprise, outliers, valeurs extrêmes, robustesse.

Abstract
Many algebraic methods, graphs and probabilistic exist to detect extreme values ("outliers"). In this presentation, two new non-parametric methods for detecting single atypical units in the univariate framework are proposed. The first in the area of V-robustness takes into account the distance of observation from the center of the distribution and shape of the end of the distribution. And the second, based on a statistical coefficient of variation, can increase the atypical units and measure the effect of an unusual unity on the quality of information. In addition, a third method, obtained as a convex combination of two techniques of extreme value theory, minimizing the variance is also proposed.
These different methods are compared on data from structural surveys Ukrainian. It is necessary to use an easy technique to put into production.

Key words
Business statistics, outlier detection, coefficient of variation.

1. Introduction

Le problème de la détection d'unités atypiques ou de valeurs extrêmes est général et très ancien. Il se pose à tous ceux qui ont à analyser des données réelles.
Les enquêtes structurelles produites par les offices statistiques sont en général réalisées par sondage stratifié auprès des petites entreprises. Les strates sont le plus souvent construites à partir d'un croisement de deux ou trois variables : le secteur d'activité, la taille de l'entreprise (effectif salarié) et éventuellement un critère régional. Le chiffre d'affaires X (ou les ventes ou la production) est considéré comme variable principale. L'objectif est d'estimer le total (ou la moyenne) du chiffre d'affaires selon les secteurs d'activité. La présence d'unités atypiques (petites entreprises dont le chiffre d'affaires est très important, et donc de ce point de vue, elles se comportent comme de grandes entreprises) dans l'échantillon, met en défaut l'hypothèse d'homogénéité des strates, augmente la

variance des estimations et peut introduire des biais en surévaluant les résultats. La détection d'entreprises atypiques permet aussi, pour la prochaine enquête, de les mettre dans une même strate particulière qui sera enquêtée exhaustivement.
Les techniques usuelles sont rappelées dans la section 2. Deux nouvelles variantes non paramétriques simples de détection des unités atypiques seront exposées dans la section 3. La section 4 propose une autre méthode issue de la théorie des valeurs extrêmes. Ces différentes méthodes (usuelles et nouvelles) sont alors comparées sur des données issues des enquêtes structurelles ukrainiennes. Une conclusion termine ce papier.

2. Méthodes empiriques usuelles

On envisage cette détection des unités atypiques sur une seule variable d'intérêt X . On peut distinguer trois grandes catégories de méthodes, utilisées par les offices nationaux de statistique, pour détecter les unités atypiques : algébriques, graphiques, probabilistes.

Méthodes algébriques

On trouve de nombreuses démarchent où la détection se fait le plus souvent à partir de la distance relative d'une unité x_i au centre de la distribution X , sans supposer de loi a priori sur X. Plus précisément, si m (moyenne, médiane,...) désigne un paramètre de tendance centrale et s (variance, Median Absolute Deviation, intervalle interquartile,...) un paramètre d'échelle, la distance relative d'une unité i (ou x_i) au centre est définie par :

$$d_i = \frac{|x_i - m|}{s}$$

Une unité i sera déclarée comme atypique si d_i est supérieure à une certaine valeur déterminée empiriquement.
Hidiroglou et Berthelot, (1986), proposent pour les enquêtes périodiques auprès des entreprises, de détecter les unités atypiques à l'extérieur d'un intervalle de la forme $[M - k\,Q_1, M + k\,Q_3]$, où M désigne la médiane et k un paramètre déterminé de façon empirique par l'utilisateur.
Ces méthodes algébriques, nombreuses et relativement proches, présentent à nos yeux plusieurs inconvénients : La détermination du seuil, à partir duquel l'entreprise est déclarée atypique, ne repose pas sur la forme de la fin de la distribution de la variable X et cette détection ne tient pas compte de l'effectif de la strate.

Méthodes graphiques

Tukey (1975, 1977) définit la région des outliers à partir du boxplot comme étant $\{x \in X \,/\, x < LF \text{ ou } x > LU\}$, où $LF = Q_1 - k_l (Q_3 - Q_1)$ et $LU = Q_3 + k_u (Q_3 - Q_1)$.

Les paramètres k_l et k_u peuvent être déterminés empiriquement ou à partir de la connaissance de la fonction de répartition de X et d'un seuil fixé dépendant du nombre d'observations dans l'échantillon.

Méthodes probabilistes

La connaissance a priori de la loi de probabilité de X peut être utilisée pour détecter les unités typiques en fixant un seuil, mais ceci n'est pas souvent le cas en pratique. La distribution X étant fortement asymétrique, sa transformation par une fonction concave comme $Log\ X$ ou $\sqrt{X}$ diminue la distance entre une valeur extrême et la moyenne de la distribution et donc la valeur de ses moments. On obtient ainsi obtenir une distribution transformée qui peut se rapprocher d'une loi normale.

3. Nouvelles variantes non paramétriques

Nous proposons deux nouvelles méthodes (ou variantes) de détection des unités atypiques, l'une basée sur la variance, l'autre sur le coefficient de variation. Ces deux techniques se placent dans le cadre de la V-robustesse (on privilégie une diminution de la variance).

Méthode basée sur la variance

Pour chaque unité *i*, on calcule un indicateur de contribution relative à la variance de la distribution de la façon suivante :

$$I_n(i) = \frac{V(X) - (1 - w_i)V(X - i)}{V(X)}$$

où $V(X - i)$ désigne la variance de l'ensemble des observations privé de l'unité i , $V(X)$ la variance totale de X et w_i le poids de sondage. Cette approche peut être vue comme une version discrète de la fonction « Change of Variance » introduite par Hampel (1974) et redécouverte par Rousseeuw en 1981 (Hampel et *al.*, 1981). Cet indicateur vérifie les propriétés suivantes :

(1) $0 \le I_n(i) \le 1$ pour tout *i* et *n*, et il peut atteindre ses bornes.

(2) $I_n(i)$ est croissant sur l'ensemble des valeurs supérieures à la moyenne de X .

(3) $I_n(i)$ peut admettre un point d'inflexion sur l'ensemble des valeurs supérieures à la moyenne de X .

Plus l'indicateur $I_n(i)$ est proche de 1, plus l'observation i est éloignée du centre de la distribution.

Deux possibilités sont envisagées pour détecter les unités atypiques à l'aide de cet indicateur :

- A partir du point d'inflexion de $I_n(i)$. Ce qui revient à $d_i = \frac{|x_i - m|}{s} \ge \frac{\sqrt{n}}{2}$. Cette approche nous permet donc de proposer une valeur, dépendante de l'effectif *n* pour détecter les unités atypiques.
- A partir de l'histogramme de $I_n(i)$.

Méthode basée sur le coefficient de variation

Le coefficient de variation a l'avantage de s'exprimer sans dimension, il est indépendant des unités choisies et permet de comparer la variabilité de plusieurs variables au sein d'un même échantillon et de comparer la variabilité de plusieurs échantillons sur une même variable.

Soit X une distribution (le chiffre d'affaires) qui admet, pour un échantillon de taille *n*, la représentation : $x_1, x_2, ...x_i, ..., x_n$. On suppose les valeurs ordonnées, $x_i \le x_{i+1}$ et toutes positives. On calcule alors le coefficient de variation (CV) de la distribution cumulée, à savoir, pour tout *i* (*i=1* à *n*) :

$$CV(i) = CV\{x_k / k \le i\}$$

Puis on calcule la différence entre deux coefficients successifs :

$$Diff(i) = CV(i) - CV(i-1)$$

On peut alors aisément déterminer le nombre d'unités atypiques à retenir en examinant la suite des dernières valeurs de ces différences ou l'histogramme de ces valeurs. Une rupture dans la suite des valeurs ou dans le changement de pente de l'histogramme indique le rang au-delà duquel une unité sera déclarée comme atypique. Lorsque des valeurs des x_i sont égales ou très proches, la croissance de $CV(i)$ est très faible et donc $Diff(i)$ proche de zéro. De plus, cette approche permet de mesurer les effets des valeurs extrêmes sur le coefficient de variation de la distribution et donc sur la qualité de l'information.

4. Méthode mixte de la théorie des valeurs extrêmes

Nous supposons que la théorie des valeurs extrêmes est connue. Nous proposons cette nouvelle méthode pour déterminer un seuil, à partir duquel une unité sera déclarée atypique en minimisant la variance d'une combinaison convexe des seuils obtenus par la fonction moyenne des excès et la loi de Pareto généralisée (on a estimé un quantile extrême avec une probabilité de 99,9% d'être une valeur extrême pour la distribution du chiffre d'affaires avec un niveau de confiance de 95%).

Soient U_1 le seuil, au-delà duquel une unité est déclarée comme extrême, obtenu par la fonction moyenne des excès et U_2 le seuil, au-delà duquel une unité est déclarée comme extrême, obtenu par la fonction GPD. Soit $U = \alpha U_1 + (1 - \alpha)U_2$ et $0 < \alpha < 1$, α minimise la variance de *U*.

On obtient $\alpha = \dfrac{V_2 - \mathrm{cov}(V_1, V_2)}{V_1 + V_2 - 2\mathrm{cov}(V_1, V_2)}$, où V_i correspond à la variance de U_i.

Les variances et covariances sont estimées par une technique de bootsrap. Ce résultat se généralise au cas *p* variables aléatoires U_i.

5. Comparaison des méthodes sur des données structurelles ukrainiennes

Les données utilisées concernent les chiffres d'affaire annuels (en 2007) des petites entreprises ukrainiennes qui appartiennent à l'une des divisions suivantes (de la nomenclature européenne Nace, NAF Rév. 1) : La division 28 (Fabrication de produits métalliques ou division 25 dans la NAF Rév. 2 de 2008), la division 45 (construction), la division 52 (commerce de détail) et la division 72 (activités informatiques).

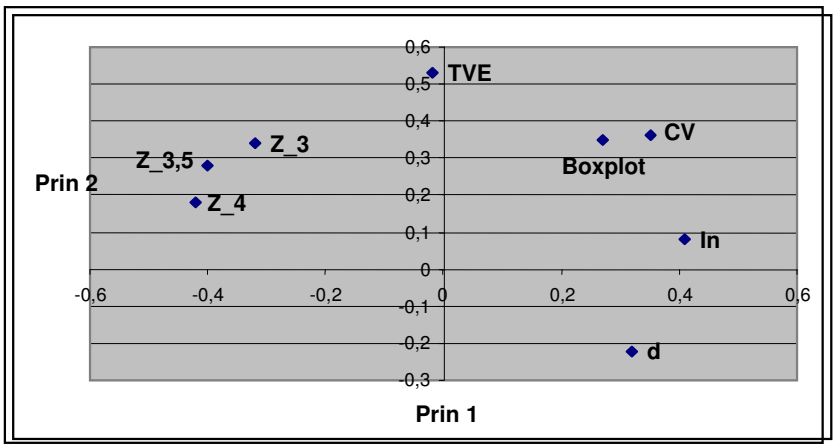
Pour chacun de ces secteurs d'activité, cinq méthodes de détection des entreprises atypiques (unités atypiques ou valeurs extrêmes) sont envisagées:

- Le graphique boxplot.
- La méthode algébrique classique en utilisant le logarithme du chiffre d'affaires avec différents seuils Z_0 (trois cas).
- L'approche basée sur la variance, en utilisant l'histogramme de l'indicateur ou le point d'inflexion (deux cas).
- La dernière approche basée sur le coefficient de variation.
- Le critère mixte de détermination d'un seuil en utilisant la théorie des valeurs extrêmes.

Tableau : Nombre de valeurs extrêmes selon la méthode utilisée

Méthodes	**Div. 28**	**Div. 45**	**Div. 52**	**Div. 72**
Boxplot	1	6	3	6
$I_n(i)$	4	3	3	6
Diff(i)	2	3	3	6
$d_i \geq \dfrac{\sqrt{n}}{2}$	1	0	0	1
$Log\ X\ , Z_0 = 3$	1	0	23	6
$Log\ X\ ,\ Z_0 = 3{,}5$	1	0	5	0
$Log\ X\ ,\ Z_0 = 4$	0	0	1	0
Mixte, TVE	2	10	12	11

Une analyse en composantes principales sur l'ensemble des méthodes paramétriques et non paramétriques présentées dans ce texte donne la représentation suivante :



Le premier axe principal explique plus de 56% de l'inertie totale du nuage et il correspondant à un facteur de taille. Le second axe (26% de l'inertie expliquée) oppose les critères qui divergent sur le nombre d'outliers pour les divisions 28 et 52.

Conclusions sur les données ukrainiennes

Sans faire de distinction entre les secteurs, l'existence de chiffres d'affaires extrêmes peut être liée à la création de petites entreprises en vue de transfère ou de blanchiment d'argent. On peut vérifier que les mois précédents des élections politiques majeures, le nombre d'entreprises atypiques augmente considérablement.
Pour mieux comprendre l'existence de ces valeurs extrêmes on peut envisager plusieurs hypothèses qui méritent une véritable discussion avec les spécialistes.

Hypothése1 *: L'existence des valeurs extrêmes peut être expliquée par une mauvaise classification des entreprises selon leurs tailles.*

En effet pour le cas Ukrainien, une entreprise est de petite taille si son chiffre d'affaires annuel est inférieur à 500000 Euro ou si le nombre moyen de salariés est inférieur à 20. Dans le cas idéal, le capital de l'entreprise serait un bon critère de classification de l'entreprise, mais on n'arrive pas à suivre le capital d'une entreprise qui peut varier d'une année à l'autre, et suivre l'évolution du chiffre d'affaires s'avère plus facile. Ainsi la valeur du capital n'est intéressante que l'année de création de l'entreprise.

Cette spécificité de l'enquête structurelle, qui utilise deux critères pour définir les petites entreprises (conformément à la législation en vigueur), conduit à avoir de mauvaises classifications quand le chiffre d'affaires de l'entreprise dépasse pour un des deux critères le seuil législatif.

Hypothése2 *: L'observation est considérée comme atypique.*
Il peut s'agir dans ce cas d'un report d'une activité à une autre permettant à l'entreprise de petite taille une augmentation considérable de son chiffre d'affaires, une activité annexe peut aussi générer un excèdent de chiffre d'affaires.

La classification des entreprises en se basant sur l'effectif peut être remise en cause, en effet, à coté des salariés déclarés on trouve des intermittents ou des non salariés et donc une capacité de production plus importante que celle attendue. Il existe aussi des petites entreprises qui ont une activité irrégulière conjoncturelle au niveau économique et donc de l'emploi.

Dans le secteur des services (l'informatique dans notre exemple), l'effectif n'est pas significatif pour dire que telle entreprise est de petite ou de grande taille, par exemple les entreprises de création de logiciel n'ont pas besoin d'avoir un effectif important pour fonctionner, en revanche, ces entreprises peuvent générer des flux colossaux.
Cette hypothèse peut être vérifiée pour certains secteurs dont une activité annexe correspond à une capacité de générer des flux supérieurs à ceux générés par l'activité principale.

La conjoncture économique en Ukraine fait que ces résultats peuvent ne pas être considérés comme des anomalies mais comme reflétant une réalité économique. L'avis des experts en entreprenariat est donc d'une grande importance sur le plan de l'analyse économique et sur le plan empirique, pour la détermination du seuil et l'estimation des différents paramètres d'une loi.

En conclusion le chiffre d'affaires n'est peut être pas un bon indicateur de classification des entreprises selon leurs tailles, et pour mieux expliquer l'existence de ces valeurs extrêmes, il serait recommander de faire une analyse en terme d'accroissement du chiffre d'affaires selon les secteurs d'activité, ou faire varier les critères (de définition des petites entreprises) selon les secteurs d'activité.

6. Conclusions

Ces différentes méthodes paramétriques ou non paramétriques ne permettent pas de trancher quant au nombre d'unités atypiques, mais elles permettent de déterminer très rapidement, et pour toutes les strates, une stratégie haute (très peu d'entreprises atypiques) et une stratégie basse de détection des unités atypiques.

Les critères qui comportent un aspect « subjectif » (observation d'un graphique ou point de vue de l'expert), donnent des résultats assez proches, comme le graphique boxplot et $Diff(i)$. Mais il faut noter qu'une différence d'une unité sur le nombre d'entreprises atypiques a un impact non négligeable sur l'estimation d'un total. Le point de vue de l'expert peut être utile dans les cas litigieux, mais il est trop coûteux dans sa généralisation au niveau de la production des données.

La théorie des valeurs extrêmes classique ou basée sur la loi de Pareto généralisée ne résout pas ces difficultés d'un coup. Elle est intéressante dans une optique de prévision, mais dans une économie instable en voie de transition, il parait difficile de l'utiliser dans la pratique. Il en est de même de notre technique, basée sur une diminution de la variance, même si elle semble être un bon compromis empirique de qualité entre la méthode de la fonction moyenne des excès et celle utilisant la loi de Pareto généralisée.

Le graphique boxplot ou notre technique basée sur le coefficient de variation nous semblent bien appropriées au cadre des enquêtes structurelles.

L'approche doit être ouverte et multiforme et en ce sens, il n'y a pas une méthode pour un problème. Les méthodes doivent être utilisées de façon souple, en restant pragmatique.

Bibliographie

Hampel F.R. (1974), *The Influence Curve and its Role in Robust Estimation*, Journal of the American Statistical Association, 69, 383-393

Hampel F.R., Rousseeuw P.J., Ronchetti E.M. (1981), *The Change-of-Variance Curve and Optimal Redescending M-Estimators*, Journal of the American Statistical Association, 76, 643-648

Hidiroglou M.A., Berthelot J-M. (1986), *Statistical Editing and Imputation for Periodic Business Surveys*, Survey Methodologie, 12, pp.73-84

Tukey J. W. (1975), *Mathematics and the picturing of data*, In Proc. Int. Congr. Mathematics, Vancouver, vol.2, 523-531

Tukey J.W. (1977), *Exploratory Data Analysis*, Ed. Addison-Wesley