VÉRIFICATION SÉLECTIVE DES DONNÉES ET QUALITÉ DU PRÉDICTEUR UTILISÉ

Philippe Brion[1]

[1] *Consultant, philippe.brion55@gmail.com*

Résumé. Les méthodes de vérification sélective des données (*selective editing*) s'appuient sur des fonctions de score permettant de sélectionner les unités à contrôler de manière approfondie, ces unités ayant une valeur du score au-dessus d'un certain seuil. Généralement, le score est calculé, pour une variable donnée, à partir de la différence entre la valeur observée de la variable et celle d'un « prédicteur » (qui peut être, par exemple, la valeur observée pour la même variable lors d'une enquête précédente). Les scores calculés pour les différentes variables sont ensuite combinés pour obtenir un score global.

L'objet de la communication est d'étudier la manière dont la qualité du prédicteur joue sur l'efficacité des méthodes proposées, à partir d'une formalisation assez simple et en se limitant à une seule variable.

Mots-clés. Vérification sélective des données, fonction de score

1 Introduction

Les méthodes de vérification sélective ont pour objectif de séparer l'ensemble des données (questionnaires d'enquêtes ou données administratives) en deux lots : un à contrôler manuellement, l'autre à traiter de manière automatique, ou pour lequel on se contente de reprendre les données brutes. Elles concernent plutôt la production de statistiques économiques contenant une majorité de variables quantitatives (et de ce fait sont plus adaptées aux enquêtes auprès des entreprises), et s'appuient en général sur des fonctions de score, calculées à partir des scores locaux (c'est-à-dire relatifs à une variable) qui sont ensuite combinés en un score global. Pour une présentation plus complète, on peut se référer entre autres au chapitre 6 de De Waal & al (2011).

Une des fonctions de score couramment utilisée (références) est la fonction $DIFF = w_i(Y_i - Y_i^p)$, où :

- Y_i est la donnée observée (contenant donc potentiellement une erreur d'observation) pour l'unité i ;

- Y_i^p est un « prédicteur » de la vraie valeur (par exemple une donnée obtenue lors d'une enquête passée et pour laquelle il ne reste plus d'erreur d'observation, ou une donnée disponible sans une autre source, par exemple le répertoire d'entreprises, ou encore une donnée calculée sur le fichier de l'enquête, comme par exemple la médiane des observations) ;

- w_i est le poids de sondage affecté à l'unité i .

La pratique montre que ce type de méthode, proposé en général de manière empirique, fonctionne bien pour certaines variables et moins bien pour d'autres.

Ce papier a pour objectif d'étudier si l'on peut, à partir d'une modélisation concernant à la fois l'erreur d'observation et la « qualité » du prédicteur, justifier cette méthode. Il se limite volontairement à deux situations simplifiées par rapport à la pratique des enquêtes statistiques :

- d'une part, on se limite à une seule variable : dans la pratique de la vérification sélective, on doit « agréger » les scores calculés pour chaque variable en un score global servant à sélectionner les unités qui sont contrôlées manuellement (pour plus de détails sur le sujet, voir De Waal & al. 2011) ;

- d'autre part, on se place dans le cas où on réalise un recensement (donc pas de poids de sondage).

Il reprend un certain nombre de points développés par Hesse (2005), mais avec une approche simplifiée.

2 Modélisation proposée

Les notations utilisées sont les suivantes : pour une variable Y et pour chaque unité i, il existe trois valeurs, dont deux sont connues :
- Y_i la donnée observée ;
- Y_i^v la donnée « vraie », donc inconnue ;
- Y_i^p le « prédicteur » de la vraie valeur.

On introduit les deux grandeurs :
- $e_i = Y_i - Y_i^v$, l'erreur – potentielle, car un nombre important d'informations collectées n'en contiennent pas - contenue dans la donnée observée ;
- $u_i = Y_i^p - Y_i^v$, indicateur de la qualité du prédicteur.

Concernant l'erreur, on postule un modèle de loi contaminée, à l'instar de Di Zio et Guarnera (2013), mais à la différence de la modélisation proposée dans leur article, l'erreur n'a pas nécessairement une moyenne égale à zéro, et on ne postule pas de loi sur la variable elle-même. L'erreur e_i a donc une probabilité π_i *a priori*, et suit, dans ce cas, une loi normale de moyenne e_{i0} :

$$f(e_i) = (1-\pi_i)\delta(0) + \pi_i N(e_i, e_{i0,} \sigma_{ei}^2)$$

Concernant la grandeur u_i, qui peut être considérée comme le résidu de la régression de Y_i^p sur Y_i^v, on postule que celui-ci suit une loi normale de moyenne zéro :

$$f(u_i) = N(u_i, 0, \sigma_{ui}^2)$$

A priori, on peut penser que les trois grandeurs inconnues e_{i0}, σ_{ei} et σ_{ui} ne sont pas égales pour toutes les unités, mais plutôt proportionnelles à leur taille. En revanche, la probabilité de l'erreur π_i peut être supposée constante.

3 Détermination de la loi *a posteriori* de l'erreur d'observation

Les grandeurs e_i et u_i sont inconnues, mais la différence entre le prédicteur et la valeur brute fournit la valeur de $u_i - e_i$. On utilise donc la formule de Bayes pour estimer la loi de e_i sachant cette valeur connue :

$$f(e_i / u_i - e_i) = \frac{f(u_i - e_i / e_i) f(e_i)}{f(u_i - e_i)}$$

Les différentes lois apparaissant dans la formule sont (pour la suite, on oublie l'indice i) :

$$f(u-e/e)=N(u-e,-e,\sigma_u^2)$$
$$f(u-e)=(1-\pi)N(u-e,0,\sigma_u^2)+\pi N(u-e,-e_{0,}\sigma_u^2+\sigma_e^2)$$
$$f(e)=(1-\pi)\delta(0)+\pi N(e,e_{0,}\sigma_e^2)$$

On en déduit, après calculs, que :

$$f(e/u-e)=(1-\widetilde{\pi})\delta(0)+\widetilde{\pi} N(e,\frac{e_0\sigma_u^2-(u-e)\sigma_e^2}{\sigma_e^2+\sigma_u^2},\frac{\sigma_u^2\sigma_e^2}{\sigma_u^2+\sigma_e^2})$$

avec $$\widetilde{\pi}=\frac{\pi\frac{\sigma_u}{\sqrt{\sigma_u^2+\sigma_e^2}}\exp(\frac{-2\sigma_u^2(u-e)e_0-\sigma_u^2(e_0^2)+\sigma_e^2(u-e)^2}{2\sigma_u^2(\sigma_e^2+\sigma_u^2)})}{1-\pi+\pi\frac{\sigma_u}{\sqrt{\sigma_u^2+\sigma_e^2}}\exp(\frac{-2\sigma_u^2(u-e)e_0-\sigma_u^2(e_0^2)+\sigma_e^2(u-e)^2}{2\sigma_u^2(\sigma_e^2+\sigma_u^2)})}$$

4 Quelle utilisation pour la vérification sélective ?

On rappelle qu'on se place ici dans le cadre d'un recensement. Si l'on utilise, pour produire l'estimation d'un total, les données brutes, on a une erreur quadratique moyenne qui vaut $EQM=E(\sum_U e_i)^2$, l'espérance étant calculée relativement à la loi de l'erreur d'observation.

Si l'on applique l'inégalité triangulaire, $[E(\sum_U e_i)^2]^{(1/2)}\leqslant\sum_U[E(e_i^2)]^{(1/2)}$.

Étant donné que l'on connaît la valeur de u_i-e_i , on peut affiner, et estimer $E(e_i^2/u_i-e_i)$:

$$E(e_i^2/u_i-e_i)=E(e_i/u_i-e_i)^2+V(e_i/u_i-e_i)=\widetilde{\pi}_i^2\frac{(e_{io}\sigma_{ui}^2-(u_i-e_i)\sigma_{ei}^2)^2}{(\sigma_{ei}^2+\sigma_{ui}^2)^2}+\widetilde{\pi}_i^2\frac{\sigma_{ui}^2\sigma_{ei}^2}{\sigma_{ui}^2+\sigma_{ei}^2}$$

Cette formule peut aider à déterminer les unités à contrôler de façon manuelle en priorité (par exemple en fonction d'un budget donné), en prenant celles qui ont les plus grandes valeurs : si l'on dispose d'ordres de grandeur pour les quantités π_i , e_{i0} , σ_{ei} et σ_{ui} , par exemple à partir d'enquêtes passées, on peut obtenir un ordre de grandeur de $E(e_i^2/u_i-e_i)$, ceci afin de conduire les choix destinés à minimiser l'erreur quadratique moyenne due aux unités non contrôlées.

Mais l'intérêt de cette formule est également d'étudier comment la valeur $E(e_i^2/u_i-e_i)$ évolue en fonction des différents paramètres, en particulier σ_{ui} , et d'opérer un retour sur la fonction DIFF. On a vu à la section 3 que la loi de l'erreur sachant u_i-e_i reste une loi contaminée, avec dans le cas où il y a erreur une loi normale dont la moyenne est décalée en fonction de u_i-e_i , et dont l'écart-type est inférieur à la fois à σ_{ei} et σ_{ui} . Si la valeur σ_{ui} tend vers 0 (prédicteur quasi-parfait), alors on voit que $\widetilde{\pi}_i$ tend vers 1 (et la loi normale concernant l'erreur tend elle-même vers une distribution de Dirac), et l'utilisation de la fonction DIFF est justifiée. A l'inverse, quand σ_{ui} tend vers l'infini, on retrouve pour $\widetilde{\pi}_i$ la valeur π_i ; et pour la loi normale de l'erreur on retrouve la loi initiale, ce qui signifie que le prédicteur ne fournit aucune valeur ajoutée.

On voit donc que la prise en compte de la seule valeur absolue de u_i-e_i , dans la fonction DIFF, ne suffit pas, pour beaucoup de variables, à garantir que l'on sélectionne les unités contribuant le plus à « neutraliser » une partie importante de l'erreur quadratique moyenne.

Des études quantifiées menées sur des données passées devraient permettre de retrouver les jugements empiriques opérés par les praticiens sur les variables observées dans les statistiques d'entreprises, jugements qui indiquent par exemple que la vérification sélective est beaucoup plus efficace pour une variable comme le chiffre d'affaires que pour les investissements (dont la valeur affiche beaucoup moins de continuité d'une année sur l'autre).

De telles simulations sur données existantes devraient également permettre de confirmer, ou non, les hypothèses faites au départ pour la modélisation du problème, en particulier concernant la loi de l'erreur d'observation.

Enfin, reste la question soulevée par Hesse (2005) sur la part relative, dans la formule donnant la valeur de $E(e_i^2/u_i - e_i)$, du biais et de la variance. Hesse indique que, en cas d'utilisation d'une méthode de redressement automatique pour les données jugées suspectes mais non sélectionnées comme devant être expertisées à la main[1], la part de la variance devient prépondérante, ce qui peut conduire à revoir le critère utilisé par la méthode de vérification sélective. Des travaux complémentaires devraient être menés sur ce sujet.

Bibliographie

[1] De Waal, T., Pannekoek et Scholtus, S. (2011), Handbook of Statistical Data Editing and Imputation, *John Wiley*.
[2] Di Zio, M. et Guarnera, U. (2013), A Contamination Model for Selective Editing, *Journal of Official Statistics*, Vol. 29, N°. 4, pp 539-555.
[3] Hesse, C. (2005), Vérification sélective de données qualitatives, *Document de travail E2005/04*, *Insee*, Paris.

1Alors que dans ce papier, on se place dans le cas où l'on se contente d'utiliser les données brutes pour les unités non contrôlées à la main.