L'ESSAIMAGE STATISTIQUE, UNE GÉNÉRALISATION DU BOOTSTRAP

Alain MORINEAU[1], Thi Minh Thao HUYNH[2]
Roland MARION-GALLOIS[3]

[1] *DEENOV 16 villa de Lourcine 75014 Paris* alain.morineau@deenov.com
[2] *Association MODULAD Lourcine Boîte 51, 75014 Paris* thiminhthao@gmail.com
[3] *MEDTRONIC, 1131 Tolochenaz (Switzerland)* roland.marion-gallois@medtronic.com

Résumé. L'essaimage statistique permet de simuler des données de taille quelconque semblant échantillonnées dans la même population qu'un échantillon observé. Ces pseudo-échantillons vont jouer un rôle analogue aux répliques Bootstrap pour l'évaluation des intervalles de confiance. L'essaimage permet d'estimer, à partir d'un jeu de données de taille fixée et sans hypothèses supplémentaires, comment les intervalles de confiance de type Bootstrap varient en fonction de la taille qu'auraient les observations. Les applications sont multiples.

Mots-clés. Essaimage, Échantillonnage, Bootstrap, Analyse des données, Simulation

Abstract. Statistical swarming is a simulation method allowing generating pseudo-samples of whatever size, as if they were directly extracted from the same population as an observed sample. Similar to bootstrap replicates, these pseudo-samples can be used to compute confidence intervals. Using only the data set and without any a priori hypotheses, swarming allows to compute confidence intervals similar to Bootstrap and to assess how the intervals are influenced by sample size. Applications are multiple.

Keywords. Swarming, Sampling, Bootstrap, Data analysis, Simulation

1 La procédure d'essaimage statistique

Par ses objectifs, l'essaimage statistique s'apparente à de la simulation puisqu'il s'agit de fabriquer de nouveaux échantillons à partir d'un échantillon observé. Les méthodes traditionnelles s'appuient sur l'échantillon pour acquérir de l'information sur la population parente et, à partir de la population hypothétique, l'induction statistique permet d'échantillonner de nouvelles données. L'essaimage s'apparente aussi au Bootstrap car il permet de simuler en s'appuyant uniquement sur les données sans passer par le support d'hypothèses a priori (ou hypothèses de commodité) introduites pour les besoins mathématiques de l'induction.

Correction par les distances

$$\bar{x}_{ij}^{*} = \frac{1}{k}\sum_{c=1}^{k}\left\{\frac{\left\{\left(\sum_{c=1}^{k} d_{ic}^{2}\right) - d_{ic}^{2}\right\}}{\sum_{c=1}^{k} d_{ic}^{2}}\right\} z_{ij(c)} \quad [1]$$

Correction par les fréquences

$$\bar{x}_{ij}^{**} = \frac{1}{k}\frac{1}{z_{.j}}\sum_{c=1}^{k}\left\{\frac{\left\{\left(\sum_{c=1}^{k} d_{ic}^{2}\right) - d_{ic}^{2}\right\}}{\sum_{c=1}^{k} d_{ic}^{2}}\right\} z_{ij(c)} \quad [2]$$

Formule de double pondération des k Plus Proches Voisins
pour tenir compte des distances [1] et équilibrer le rôle des fréquences [2]

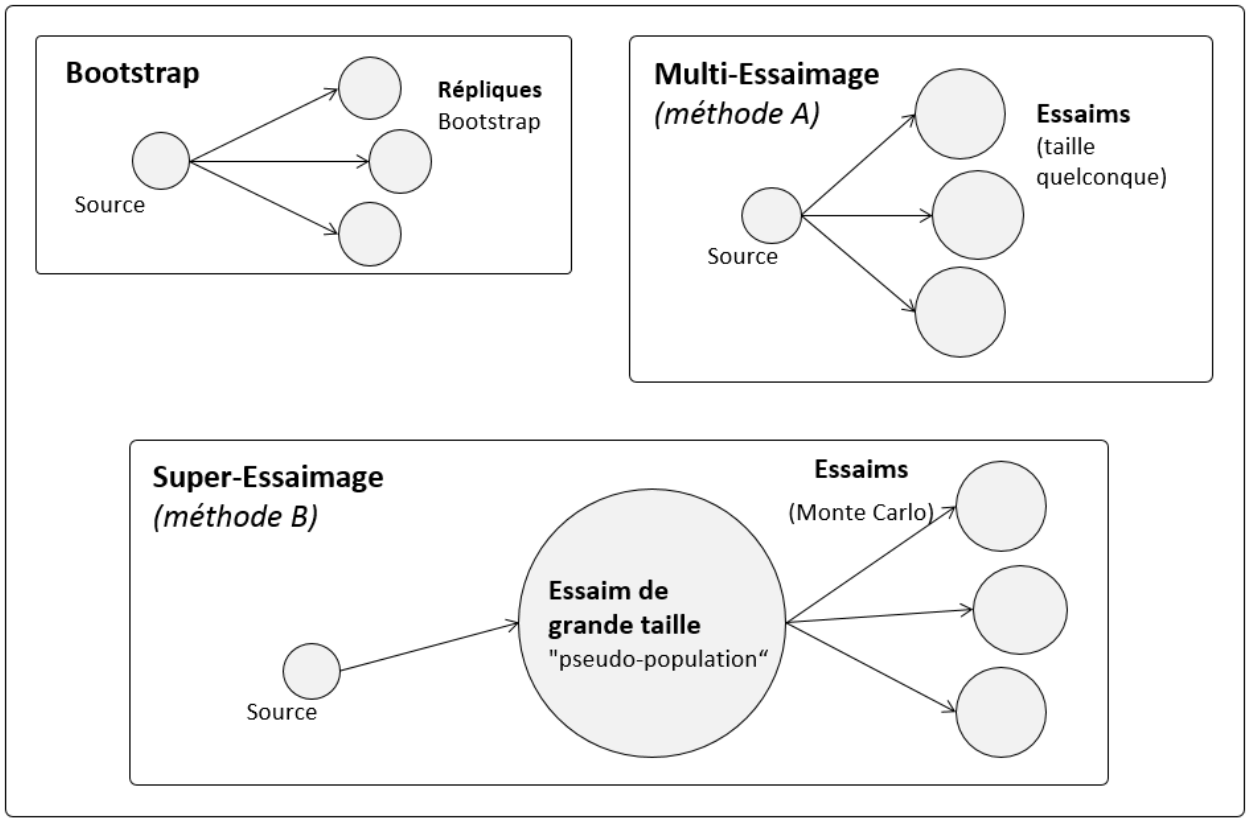
La procédure d'essaimage (explicitée dans [réf 1]) peut se résumer ainsi. On dispose d'un échantillon multivarié de taille n. On réalise une analyse factorielle des données et on extrait un

nombre suffisant d'axes factoriels pour capturer les liaisons structurelles entre les variables (réf [3]). Les distributions des coordonnées des données sur ces axes servent à simuler de nouveaux points en nombre N. Chacun de ces points est ensuite remplacé par le point moyen de ses k plus proches voisins dans le nuage des n observations dans l'espace factoriels, le voisinage permettant de consolider les liaisons entre les variables dans le nuage simulé. La moyenne utilisée obéit à une double pondération qui permet d'une part de remplir l'espace au-delà de l'enveloppe convexe des observations, d'autre part d'équilibrer les rôles des variables en fonction de leur poids.

2 Deux approches pour estimer les intervalles de confiance

Le Bootstrap utilisé pour évaluer les variabilités s'appuie sur la construction de *répliques Bootstrap* qui jouent le rôle de nouvaux échantillons de même taille que la source. La variabilité d'une statistique est approchée par la variabilité de ses valeurs dans l'ensemble des répliques. On l'évalue souvent par l'intervalle de confiance dit *percentile*.

La méthode d'essaimage a été validée empiriquement de façon approfondie dans la publication [réf 2] (ERCIM-2014) en utilisant le contexte suivant. On se place dans le cas où la population mère est parfaitement connue. On s'intéresse à une statistique dont la valeur est connue de façon exacte dans la population et on cherche à l'estimer à partir d'un échantillon extrait de la population. On montre comment l'essaimage fournit des intervalles de confiance de type percentile, avec la possibilité de vérifier le taux de recouvrement de la vraie valeur.



Graphique 1 : Schéma de fonctionnement du Bootstrap
et des deux méthodes d'essaimage : *multi-essaimage* et *super-essaimage*.

On peut évaluer la variabilité d'une statistique par essaimage en adoptant une des deux démarches suivantes. La *méthode A* consiste à fabriquer un grand nombre d'essaims de même taille. Elle mimique directement le Bootstrap mais le généralise puisque la taille des essaims est quelconque (et non celle de la source comme dans le Bootstrap). Cette méthode est lourde en calculs puisqu'elle nécessite de répéter complètement la procédure d'essaimage pour chaque nouvel essaim.

La *méthode B* consiste à créer un ou plusieurs *super-essaims*, c'est-à-dire des essaims de très grande taille qui vont jouer le rôle de *pseudo-populations*. Pour étudier la variabilité d'une statistique dans un échantillon de taille n quelconque, on extrait par tirage aléatoire simple (Mont Carlo) un grand nombre d'échantillons de taille n dans chacune des pseudo-populations. En bref, le Bootstrap d'une source est généralisé et remplacé par l'échantillonnage direct dans des super-essaims de la source.

Les simulations menées sur divers exemples montrent que les variabilités estimées par les méthodes A et B sont très semblables. Pour la suite, on se restreint à la méthode B de super-essaimage pour son avantage déterminant en termes de poids des calculs. Cet algorithme mixte constitue la méthode d'essaimage.

3 Expérience de comparaison des intervalles par Bootstrap et par Essaimage

On observe 17 variables nominales totalisant 50 modalités. On dispose d'une population connue de taille 9108 dans laquelle on va tirer des échantillons de taille 100 appelés sources. Pour chaque source on évalue sa *qualité* pour représenter la population par la valeur du RCME. Le RCME est la racine du carré moyen des écarts entre les fréquences dans la population et dans la source, en éliminant pour chaque variable la dernière modalité (dont la valeur se déduit des autres). Après 100 000 tirages de sources de taille 100, on dispose d'une évaluation par simulation de la distribution *théorique* du RCME correspondant à ces données (voir la Figure 1).

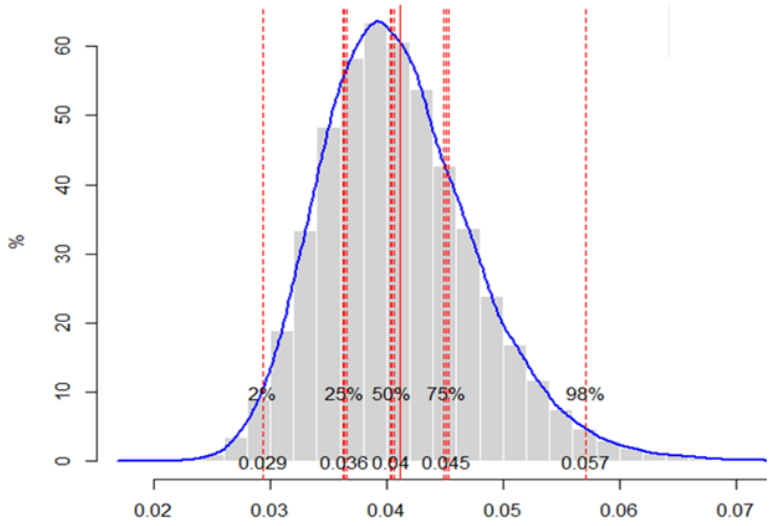


Figure 1 : Distribution théorique du RCME approchée à l'aide de 100 000 simulations avec définition des zônes de qualité des sources (2%, médiane, 98%)

On s'intéresse à la fréquence d'une modalité d'une variable particulière, cette fréquence étant 13,04% dans la population. On calcule les intervalles de confiance en faisant varier la qualité de la source en tirant dans les 3 zones de qualité : 2%, médiane (50%) et 98%. La zone appelée 2% correspond à des sources de très bonne qualité, dont le RCME est inférieur à 2% ; la zone dite médiane correspond à des sources dont le RCME se situe dans un petit intervalle autour de la valeur médiane ; enfin la zone 98% correspond aux sources de mauvaise qualité, c'est-à-dire dont le RMCE est parmi les 2% les plus grands (Figure 1). On effectue d'une part un essaimage de taille 100, ce qui permet la comparaison avec les intervalles Bootstrap d'une source de taille 100 ; et d'autre part pour un essaimage de taille 500 à partir de la même source de taille 100.

Les résultats sont sur les trois lignes du Tableau 1, associées aux 3 zones de qualité des sources (2%, med et 98%) repérées par dans la colonne de gauche.

Tirage des sources	Source n=100	BOOTSTRAP - n=100 - Intervalle des erreurs (10 000 répliques)			
zone	RCME	Moy RCME	borne inf	borne sup	Delta
2%	12,99	12,99	**12,57**	**13,39**	0,82
med	12,81	12,81	**12,39**	**13,31**	0,92
98%	12,68	12,68	**12,24**	**13,11**	0,87
Moy	12,83	*12,83*	*12,40*	*13,27*	***0,87***

Tirage des sources	Source n=100	ESSAIMAGE - n=100 - Intervalle des erreurs (1000 essaims dans 10 super-essaims)			
zone	RCME	Moy RCME	borne inf	borne sup	Delta
2%	12,99	12,86	**12,45**	**13,27**	0,82
med	12,81	12,94	**12,55**	**13,32**	0,77
98%	12,68	12,61	**12,14**	**13,06**	0,92
Moy	12,83	*12,89*	*12,49*	*13,28*	***0,79***

Tirage des sources	Source n=100	ESSAIMAGE - n=500 - Intervalle des erreurs (1000 essaims dans 10 super-essaims)			
zone	RCME	Moy RCME	borne inf	borne sup	Delta
2%	12,99	12,85	**12,67**	**13,04**	0,37
med	12,81	12,94	**12,76**	**13,11**	0,35
98%	12,68	12,60	**12,40**	**12,81**	0,41
Moy	12,83	*12,89*	*12,71*	*13,07*	***0,36***

Tableau 1 : Comparaison des intervalles de confiance par Bootstrap et par Essaimage

La table du haut concerne le Bootstrap sur des sources de taille 100 ; au milieu se trouve la table correspondant à un essaimage de même taille sur les mêmes sources ; dans la table du bas se trouve un essaimage de taille 500 sur les mêmes sources. On peut comparer les valeurs moyennes du critère de qualité RCME, les bornes inférieures et supérieures des intervalles percentiles à 95% et enfin, l'amplitude de l'intervalle dans la dernière colonne.

Si on compare l'essaimage de taille 100 au Bootstrap concurrent (taille 100 nécessairement), on constate des résultats assez semblables avec cependant une moyenne des amplitudes de l'intervalle de l'essaimage (0.79) inférieure à celle du Bootstrap (0.87). Quand on compare l'essaimage de taille 100 et l'essaimage de taille 500 (les deux tables en dessous du Bootstrap), on constate une diminution très nette de l'amplitude de l'intervalle percentile à 95% : on passe de 0.79 à 0.36. Les remarques faites sur ce cas particulier se retrouvent dans l'ensemble des expériences que nous avons pu faire, sans que ceci constitue évidemment une démonstration. Sur la Figure 2, on peut comparer l'amplitude de l'intervalle percentile pour le Bootstrap et pour l'essaimage concurrent, puis voir la décroissance de cette amplitude quand la taille de l'essaim augmente de 100 à 1000.

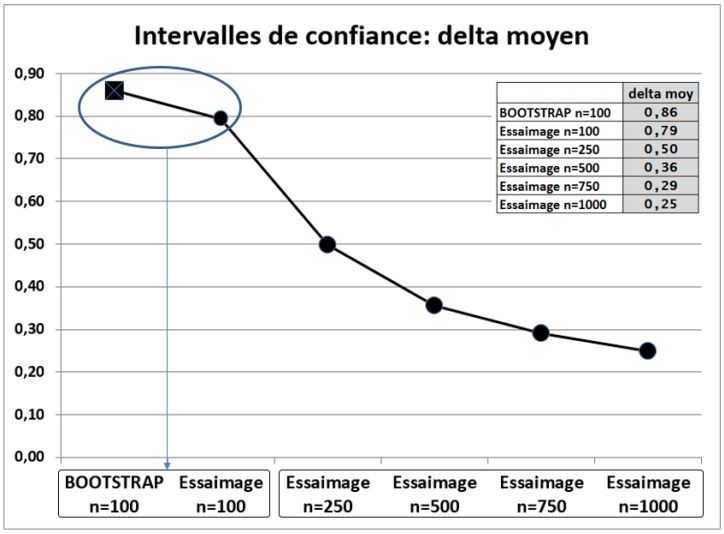


Figure 2 : Delta moyen des intervalles pour le Bootstrap et pour des essaims de taille croissante

Pour affiner les comparaisons, on travaille ici sur 5 zones de qualité et 5 tailles d'essaims.

La figure 3 montre l'effet conjugué de la qualité de la source et de la taille de l'essaim sur l'amplitude de l'intervalle percentile à 95%. Quelle que soit la taille de l'essaim, l'amplitude de l'intervalle de confiance diminue quand la qualité de la source augmente ; rappelons que la qualité de la source (critère RCME) évalue dans quelle mesure la source est « représentative » de la population dont elle est extraite. On voit sur la même figure comment l'amplitude de l'intervalle diminue quand la taille de l'essaim augmente.

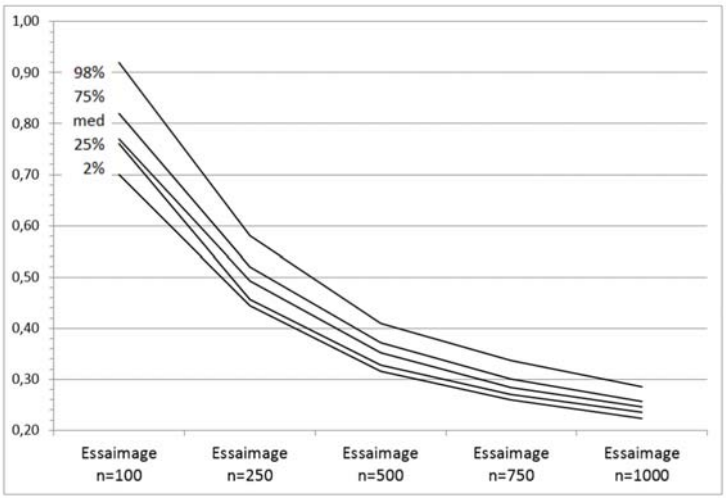


Figure 3 : Variation du delta moyen des intervalles percentiles à 95%
en fonction de la qualité de la source et de la taille de l'essaim

4 Essaimage *vs* Bootstrap

On résumera la comparaison entre les deux procédures avec les points suivants :

- L'essaimage, comme le Bootstrap, permet d'évaluer les variabilités en s'appuyant uniquement sur la source (les données observées).
- En ce qui concerne la partie calcul, le Bootstrap ne nécessite que des tirages avec remise. L'essaimage doit enchainer une analyse factorielle de la source, suivie de tirages (sans remise) et d'une procédure de retour aux variables d'origine.
- Les variabilités estimées par Bootstrap ne s'entendent que pour des échantillons ayant la taille de la source. L'essaimage permet de faire des évaluations de variabilité pour des échantillons de taille quelconque qui seraient issus de la même population que la source.
- L'essaimage, passant par une étape de simulation de type Monte Carlo, jouit de toutes les applications possibles des procédures de simulation, ce qui évoque un vaste champ d'applications supplémentaires.
- Avec le Bootstrap, on se trouve dans la situation rêvée de posséder un nombre illimité d'échantillons de taille fixée (ce sont les répliques Bootstrap). Avec l'essaimage, c'est autant d'échantillons de taille quelconque (ce sont les essaims).
- Le Bootstrap est une procédure utilisée depuis longtemps, étudiée de façon approfondie et ayant fait ses preuves dans de multiples contextes. L'essaimage est une procédure toute nouvelle, dont les mérites nécessitent consolidation et maturation.

5 Perspectives

La technique d'essaimage a été introduite pour répondre à des questions considérées techniquement difficiles, en particulier l'étude de la spécificité et la sensibilité d'un test statistique réalisé sur les données, en fonction de la taille que pourrait avoir l'échantillon observé ; ou encore l'étude de la variation d'amplitude de l'intervalle de confiance d'un coefficient de régression logistique si on doublait la taille de l'échantillon.

Mais l'essaimage ouvre aussi la voie à des applications d'un type nouveau, rendues possibles par cette approche purement calculatoire. Citons par exemple le *clonage* de données : l'échantillon observé est remplacé par autant d'individus ayant les mêmes propriétés que les observations (*anonymisation automatique* de tout fichier de données).

L'essaimage pourrait être aussi une voie d'exploration multidimensionnelle des variabilités au sens suivant. On peut évaluer par Bootstrap la variabilité de l'écart-type d'une variable isolée, mais si la variable est accompagnée d'autres variables, l'estimation Bootstrap n'en tient pas compte. L'essaimage, travaillant dans l'espace factoriel des observations, prendra en compte la dépendance entre toutes les variables pour estimer l'écart-type de cette variable particulière (sans passer par une modélisation).

Bibliographie

[1] Morineau, A., Huynh, T.M.T., Marion-Gallois, R. (2013), *Swarming sampling, a method of increasing small samples*, http://papersjds13.sfds.asso.fr/submission_39.pdf Actes des Journées de Statistique, Toulouse.

[2] Morineau, A., Huynh, T.M.T., Marion-Gallois, R. (2014), *The statistical swarming method and its validation*, Conference E993, Computational and Methodological Statistics ERCIM-2014, Pisa http://www.cmstatistics.org/ERCIM2014/docs/BoA%20CFE-ERCIM%202014.pdf

[3] Lebart, L., Morineau, A., Warwick K.M. (1984), *Multivariate descriptive statistical analysis*, J. Wiley (New-York).