Traitement des regroupements de réponses dans les enquêtes auprès des entreprises

Henri Bodet et Théo Leroy

Insee

Colloque francophone sur les sondages - 26 octobre 2018

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?
2. Quelques enquêtes auprès des entreprises sur l'environnement

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?
2. Quelques enquêtes auprès des entreprises sur l'environnement
3. Estimation de l'impact des regroupements sur la charge de réponse

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?
2. Quelques enquêtes auprès des entreprises sur l'environnement
3. Estimation de l'impact des regroupements sur la charge de réponse
4. Comment traite-t-on les regroupements de réponses ?

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?
2. Quelques enquêtes auprès des entreprises sur l'environnement
3. Estimation de l'impact des regroupements sur la charge de réponse
4. Comment traite-t-on les regroupements de réponses ?
5. Évaluation par simulation de l'erreur due aux regroupements

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?
2. Quelques enquêtes auprès des entreprises sur l'environnement
3. Estimation de l'impact des regroupements sur la charge de réponse
4. Comment traite-t-on les regroupements de réponses ?
5. Évaluation par simulation de l'erreur due aux regroupements
6. Expression générale de l'erreur induite par les regroupements

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?
2. Quelques enquêtes auprès des entreprises sur l'environnement
3. Estimation de l'impact des regroupements sur la charge de réponse
4. Comment traite-t-on les regroupements de réponses ?
5. Évaluation par simulation de l'erreur due aux regroupements
6. Expression générale de l'erreur induite par les regroupements
7. Utilisation de l'expression générale pour repérer les regroupements à risque

Plan

1. Qu'entend-on ici par "regroupement de réponses" ?
2. Quelques enquêtes auprès des entreprises sur l'environnement
3. Estimation de l'impact des regroupements sur la charge de réponse
4. Comment traite-t-on les regroupements de réponses ?
5. Évaluation par simulation de l'erreur due aux regroupements
6. Expression générale de l'erreur induite par les regroupements
7. Utilisation de l'expression générale pour repérer les regroupements à risque
8. L'application de la méthode généralisée de partage des poids

Regroupements de réponses - définition

Dans le cadre de ce travail, on appelle regroupement de réponses la situation où : une unité interrogée répond pour un ensemble d'autres en agrégeant les réponses

-et non pas en fournissant toutes les réponses.

Il ne s'agit donc pas d'une situation de sondage par grappes ou de sondage indirect mais d'une cas où on a une information dégradée : le total uniquement - alors qu'on avait prévu de recueillir les réponses individuelles.

Quelques enquêtes portant sur l'environnement

EACEI : Enquête Annuelle sur les Consommations d'Énergie dans l'Industrie

Porte sur les types d'énergies consommées, l'autoproduction d'énergie, la facture énergétique de l'Industrie, l'utilisation des sources d'énergie

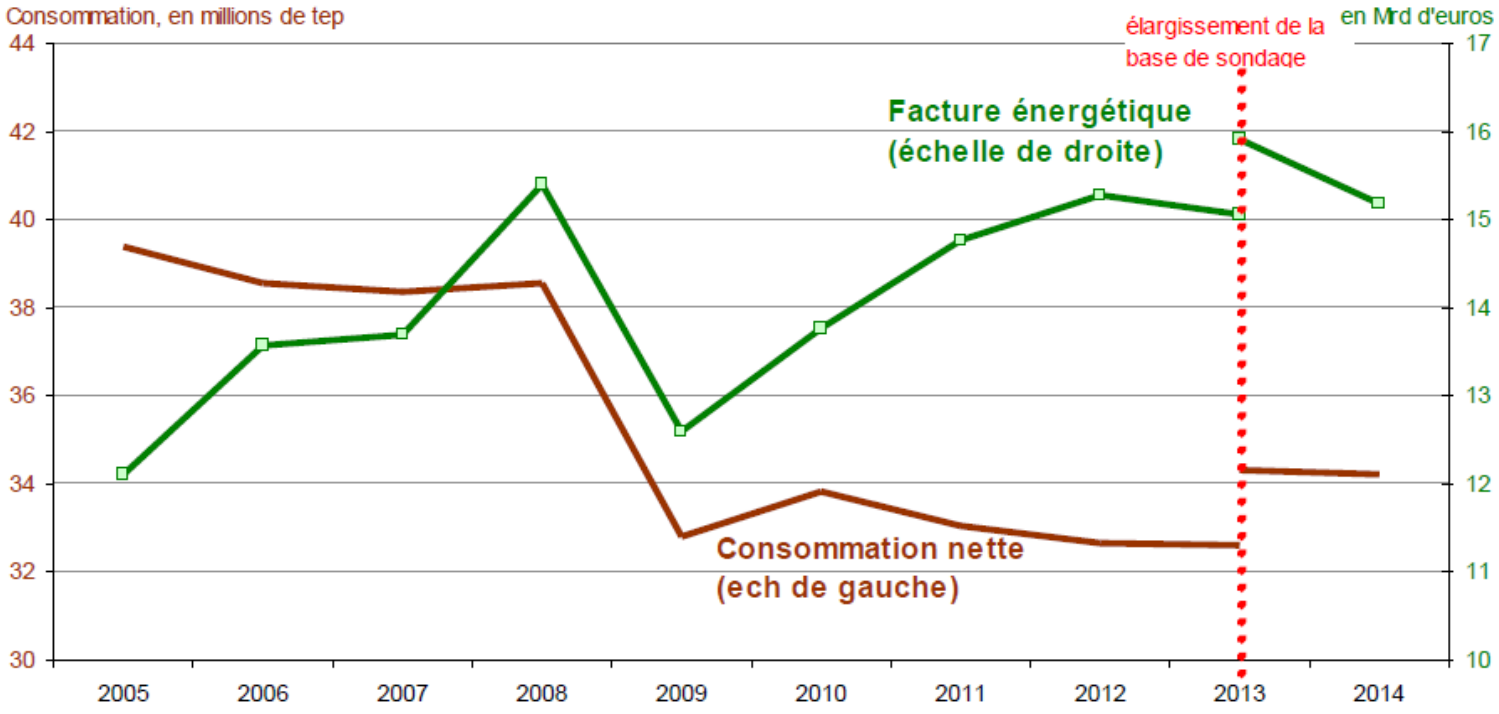
Antipol : Dépenses des entreprises pour protéger l'environnement

Porte sur les dépenses (études, dépenses courantes, investissements) réalisées par les établissements industriels pour protéger l'environnement

Déchets : Déchets non-dangereux du commerce et de l'industrie

Porte sur les types de déchets produits par les établissements, leur quantité et leur gestion

Exemples de résultats obtenus avec l'EACEI



source : Marlène Bahu, présentation aux Rencontres Économiques de la Défense,2016

Quelques enquêtes portant sur l'environnement

Dans les enquêtes auprès des entreprises, il y a plusieurs niveaux d'interrogation possibles :

- l'entreprise statistique (ou le groupe)
- l'unité légale
- l'établissement : l'implantation locale

Quelques enquêtes portant sur l'environnement

Dans les enquêtes auprès des entreprises, il y a plusieurs niveaux d'interrogation possibles :

- l'entreprise statistique (ou le groupe)
- l'unité légale
- l'établissement : l'implantation locale

Les enquêtes précédentes portent toutes auprès des établissements.

Toutefois, dans certains cas, l'information n'est disponible que pour un ensemble d'établissements.

- Une source d'énergie peut être achetée par un établissement pour plusieurs autres
- Parfois, l'information doit être reconstruite : il est préférable qu'une personne au siège s'en occupe pour l'ensemble des établissements

Impact sur la charge statistique

Le fait d'admettre des réponses regroupées devrait diminuer la charge statistique pesant sur les entreprises - puisque moins d'entreprises répondent. Toutefois :

- Répondre pour un ensemble d'établissements peut être plus lourd que répondre que pour son seul établissement
- Le regroupement déborde souvent l'échantillon : répondre pour des établissements qui ne sont pas interrogés n'est pas un gain de temps pour eux

Impact sur la charge statistique

Le fait d'admettre des réponses regroupées devrait diminuer la charge statistique pesant sur les entreprises - puisque moins d'entreprises répondent. Toutefois :

- Répondre pour un ensemble d'établissements peut être plus lourd que répondre que pour son seul établissement
- Le regroupement déborde souvent l'échantillon : répondre pour des établissements qui ne sont pas interrogés n'est pas un gain de temps pour eux

Nous avons entrepris d'estimer le solde de ces effets.

Impact sur la charge statistique

Pour chaque enquête auprès des entreprises, l'Insee recueille le temps de réponse.

Si l'entreprise répondante omet de le mentionner, on lui impute un temps de réponse. On dispose donc :

- d'un temps de réponse
- d'une procédure d'imputation

Impact sur la charge statistique

Nous avons travaillé avec l'EACEI 2016.

Le temps de réponse à cette enquête vérifie *grosso modo* :

- temps moyen 45 minutes
- 20 minutes supplémentaires si l'unité regroupe des réponses

Impact sur la charge statistique

Nous avons construit un contrefactuel - pour estimer le temps de réponse qu'il y aurait eu sans regroupements.

Nous avons imputé des temps de réponses dans une situation fictive :

- pour les regroupants : s'ils n'avaient pas regroupé de réponses
- pour les regroupés : s'ils avaient répondu directement

La statistique qui nous intéresse est la somme des temps de réponse sur tout l'échantillon.

On obtient les résultats suivants :

Situation	Temps de réponse cumulé (minutes)
avec regroupements	437 900
sans regroupements	439 400

–

- Globalement, pas d'effet sur la charge statistique

On obtient les résultats suivants :

Situation	Temps de réponse cumulé (minutes)
avec regroupements	437 900
sans regroupements	439 400

- Globalement, pas d'effet sur la charge statistique

Toutefois, ce bilan n'inclut pas :

- Le gain apporté par les réponses qu'on n'aurait pas eues autrement
- La charge que cette pratique fait peser sur le service enquêteur

Le traitement des regroupements

Les regroupements sont acceptés pour les raisons suivantes :

- Cela permet d'obtenir des réponses que l'on perdrait autrement
- Même si on perd le détail d'une donnée, on a une information exacte sur le total

Le traitement des regroupements

Ils sont traités ainsi :

- On ventile les réponses au prorata des effectifs des établissements composant le regroupement.
- On fait comme si chaque unité regroupée appartenant à l'échantillon avait répondu à l'enquête.

De la sorte, même si les réponses individuelles sont fausses, le total semble bien conservé.

On peut penser qu'on ne perd en précision qu'à un niveau détaillé.

De plus, si la variable d'intérêt est à peu près proportionnelle aux effectifs salariés, l'erreur sera moindre.

Le traitement des regroupements

On peut exprimer l'erreur que l'on commet en traitant ainsi un regroupement : $E_g = \sum_{i \in s \cap g} w_i (y_i - \frac{x_i}{X_g} Y_g)$

Il y a deux façons pour que cette erreur soit nulle :

Première condition : que la procédure de dégroupement soit "exacte" (c'est-à-dire que $y_i = \frac{x_i}{X_g} Y_g$ pour toutes les unités.)

ou

Deuxième condition : que toutes les unités du regroupement soient dans l'échantillon et aient le même poids w_g.

Évaluation par simulation

En pratique, l'erreur n'est pas nulle : peut-on l'estimer par simulation ?

En s'appuyant sur l'EACEI 2015 qui a connu une extension exceptionnelle : 14 000 établissements interrogés sur 23 000.

Parmi les répondants à cette enquête se trouvent beaucoup de réponses d'établissements appartenant à la même unité légale.

La simulation a consisté à considérer qu'il s'agissait de regroupements potentiels et à en sélectionner un certain nombre au hasard.

Ceci permet d'obtenir un éventail des estimateurs possibles.

En outre, on dispose d'une référence : l'estimateur sans regroupements.

Évaluation par simulation

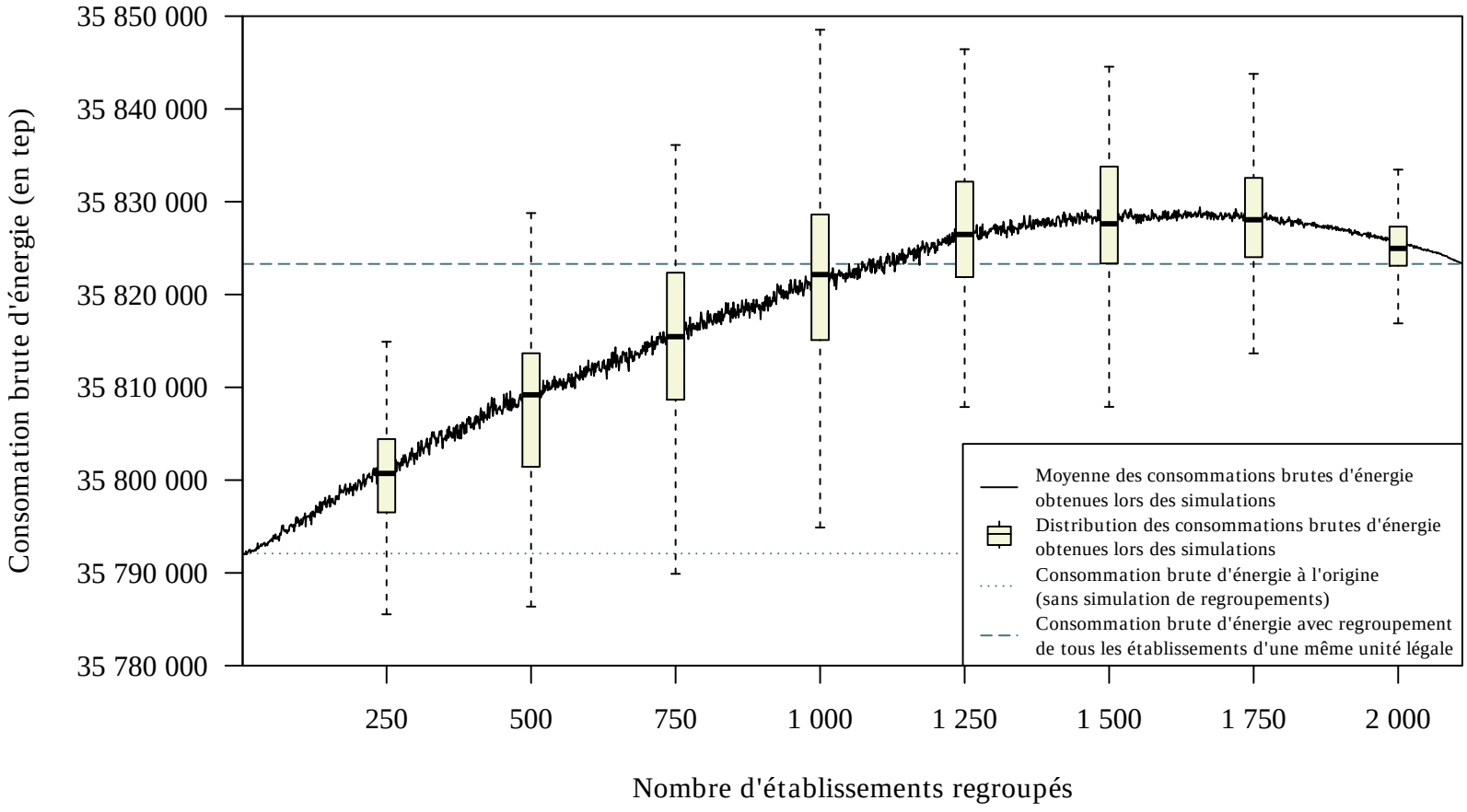


FIGURE 1 – Erreur induite sur toute la population - consommation brute d'énergie

amplitude maximale : 0,8 % du vrai total

Évaluation par simulation

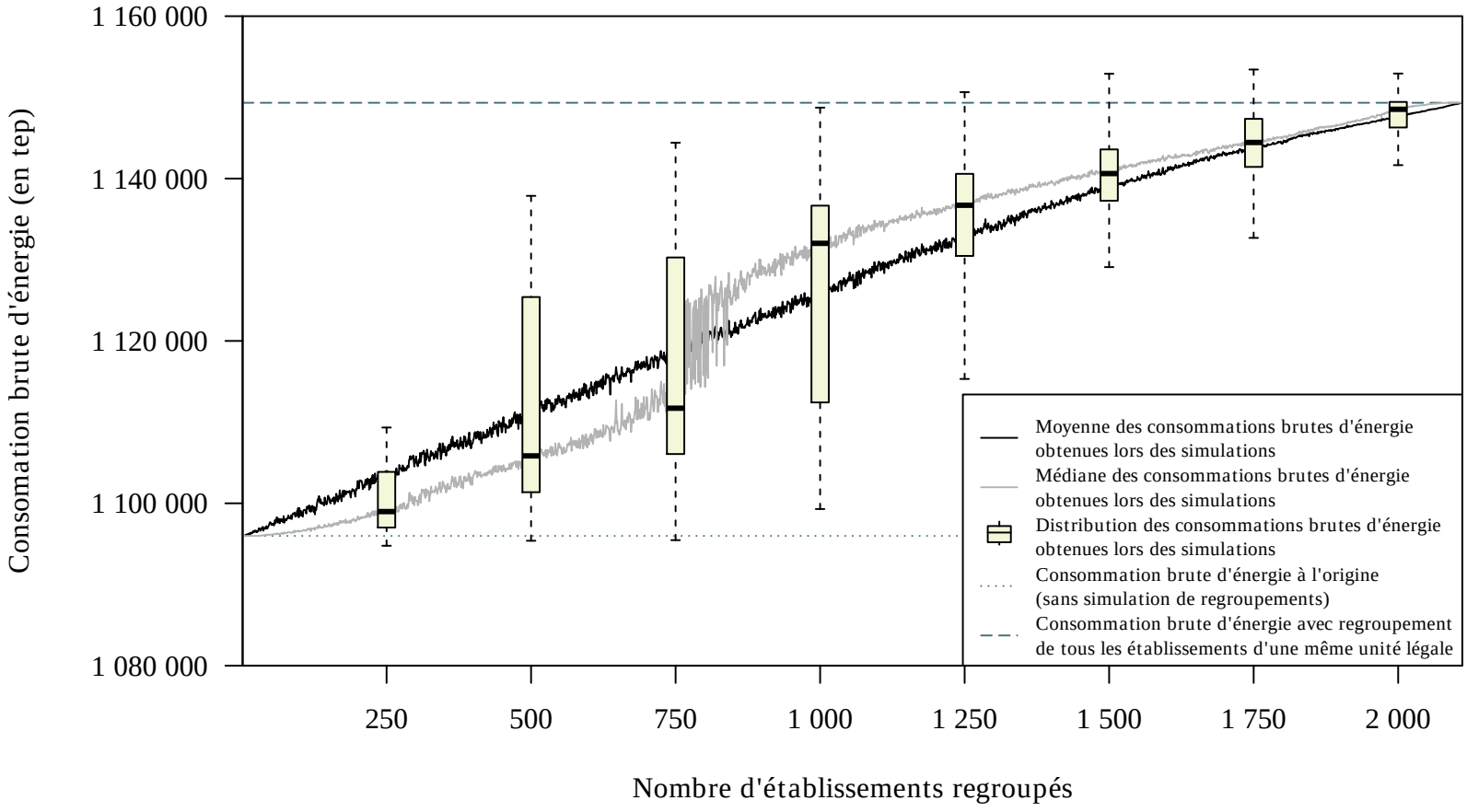


FIGURE 2 – Erreur induite sur l'île-de-France- consommation brute d'énergie

amplitude maximale : 4,5 % du vrai total

Expression de l'erreur induite par un regroupement

Cas où on regroupe deux unités

α_i : le poids de l'unité i dans la variable auxiliaire x au sein du regroupement β_i : poids de cette unité dans le total de la variable y

$$E_{1+2} = (y_1 + y_2)(w_2 - w_1)(\beta_1 - \alpha_1) \tag{1}$$

Trois facteurs interviennent :

- $y_1 + y_2$: la valeur regroupée
- $w_2 - w_1$: l'écart entre les poids
- $\beta_1 - \alpha_1$: la "non proportionnalité" de la variable regroupée et de la variable auxiliaire.

Expression de l'erreur induite par un regroupement

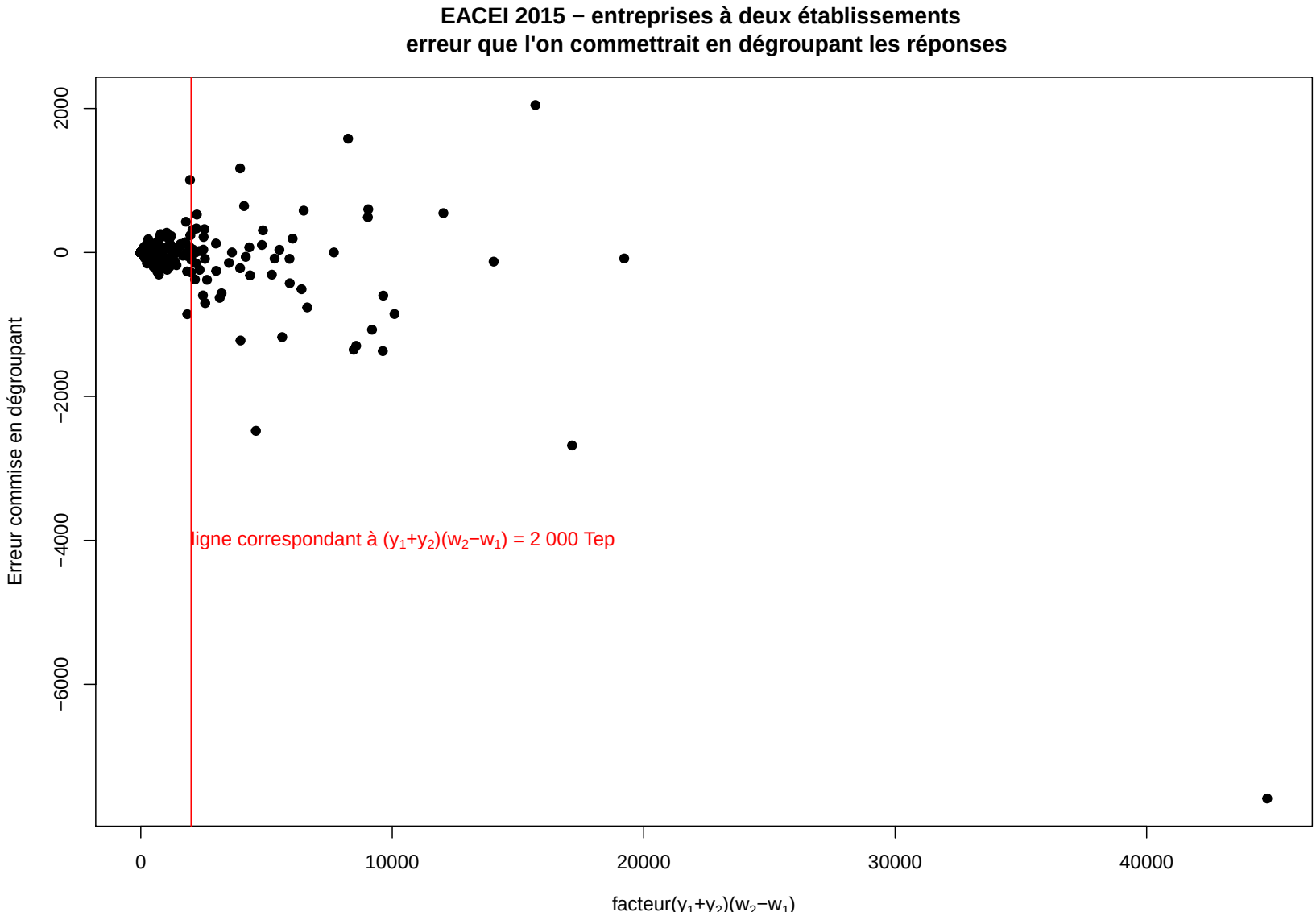
Cette expression permet d'indiquer une majoration de l'erreur potentielle induite par un regroupement par un facteur de risque dont les composantes sont connues.

$$\text{Risque}_{1,2} = (y_1 + y_2)(w_2 - w_1) \tag{2}$$

Ce facteur peut permettre de repérer des regroupements à risque.

Expression de l'erreur induite par un regroupement

Sur l'EACEI 2015, nous avons calculé l'erreur que l'on "aurait eu" et le facteur de risque.



Expression de l'erreur induite par un regroupement

Cas d'un regroupement$g$ de n_g unités

Il faut en fixer une - mettons celle qui a le numéro 1 - pour obtenir une expression comparable :

$$E_g = Y_g \sum_{i \in g, i \neq 1} (w_i - w_1)(\alpha_i - \beta_i) \qquad (3)$$

Ce qui conduit à la majoration $| E_g | \leq Y_g (n_g - 1)(w_{\max} - w_{\min})$

Expression de l'erreur induite par un regroupement

On obtient donc un facteur de risque qui peut être utilisé pour repérer les regroupements les plus problématiques :

$$\text{Risque}_g = Y_g(n_g - 1)(w_{\max} - w_{\min})$$

Application de la MGPP

Nous avons vu que la condition suffisante la plus simple à vérifier pour que le regroupement n'induise pas d'erreur serait que :

- Toutes les unités regroupées soient dans l'échantillon
- Elles aient toutes le même poids

La méthode généralisée de partage des poids (MGPP) permet de réunir ces conditions *a posteriori*.

Application de la MGPP

Pour se placer dans le cadre d'application de la MGPP, il faut faire les hypohèses de comportement suivantes :

- Les regroupement existent "avant l'enquête"
- Si une unité du regroupement est interrogée, on reçoit forcément la réponse regroupée

Application de la MGPP

Avec ces hypothèse, l'application de la méthode se résume ainsi :

- inclure toutes les unités regroupées dans l'échantillon
- leur affecter le même poids w_g^*

$$w_g^* = \frac{\sum_{i \in \cap g} w_i}{n_g}$$

On peut ensuite répartir les réponses au prorata d'une variable auxilaire - cette méthode assure que le total sera conservé.

Conclusion

Sur l'impact des regroupements

- La pratique du regroupement n'allège pas la charge pesant sur les entreprises
- Elle se justifie par le fait qu'elle permet à des unités qui ne le feraient pas autrement de répondre
- Les simulations montrent que l'erreur induite est certainement faible par rapport à la variance due au sondage
- Le fait de ventiler les réponses au prorata ne garantit pas la conservation du total
- Nous avons un moyen de repérer les regroupements qui peuvent engendrer l'erreur la plus grande

Conclusion

Sur la façon de traiter les regroupements

- Si les unités regroupées ne sont pas dans l'échantillon et/ou ont des poids différents, le total n'est pas conservé
- La méthode du partage des poids pourrait permettre de mieux les traiter en faisant en sorte que le total soit conservé