Comparaison de quatre méthodes d'imputation des revenus mobiliers

Modou DIA, CEPS-INSTEAD,
Esch-sur-Alzette (LUXEMBOURG)

JMS-2012, Colloque du 24-26 janvier 2012 à Paris.

I) Problématique & Définition

1) Problématique sur le but et la nature de l'imputation

2) Définition du type de variable à imputer

1) Problématique

- Sur la base des limites constatées sur une méthode antérieure d'imputation des revenus mobiliers (imputation multiple de Michigan) : il s'agit de trouver une autre alternative;

- Expérimenter quatre autres méthodes d'imputation afin d'en choisir la moins mauvaise.

2) Définition des revenus mobiliers

C’est un revenu brut agrégé collecté sur tout le ménage-logement avec les composantes suivantes:

- Bénéficié d’intérêts d’épargne;
- Dividendes;
- Bénéfices tirés d’investissements en capital.

3) Statistiques sur les données manquantes

Tableau 1 : Statistiques sur les données manquantes ou observées des revenus mobiliers

FLAGS / Indicateurs	Revenus Mobiliers observés	Revenus Mobiliers manquants	Total
Fréquences non pondérées	1214	740	1954*
Fréquences pondérées	70449	40917	111365**
Pourcentages non pondéres	62,13%	37,87%	100%
Pourcentages pondéres	63,26%	36,74%	100%

Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

II) Considérations générales sur les données

1) Faut-il pondérer ou non ?

2) Choix et découpage des variables explicatives

3) Traitement des valeurs extrêmes

1) Faut-il pondérer?

- La réponse est «OUI» car:
  - Chaque année, un échantillon supplémentaire, prélevé dans la population arrivée depuis la dernière vague d'enquête, est ajouté à l'échantillon à cause de l'attrition;
  - La distribution pondérée, de certaines variables pertinentes par rapport à la perception de revenus mobiliers, est plus stable que leur distribution non pondérée (voir tableau 2 ci-dessous)

Tableau 2 : Distribution de variables relatives à la possession de revenus mobiliers ou de variables influentes sur les revenus mobiliers

Variable	Modalité	Vague d'enquête	Pourcentage non pondéré	Pourcentage pondéré
CLASSE D'AGES	16_34 ans	6	23,3	15,5
		7	18,5	15,4
		8	15,4	15,0
	35_49 ans	6	36,0	34,7
		7	35,8	34,9
		8	35,6	35,1
	50_64 ans	6	24,7	26,5
		7	27,2	27,5
		8	29,7	28,0
	65 ans ou +	6	16,0	23,3
		7	18,5	22,2
		8	19,3	21,9
REVENUS MOBILIERS	NON	6	48,3	41,0
		7	43,8	40,9
		8	41,4	40,5
	OUI	6	51,7	59,0
		7	56,2	59,1
		8	58,6	59,5
STATUT D'OCCUPATION	Locataire avec loyer < prix du marché	6	4,0	4,8
		7	4,0	5,2
		8	3,6	3,3
	Locataire avec loyer au prix du marché	6	34,5	21,4
		7	27,3	24,8
		8	23,1	30,4
	Occupant (e) à titre gratuit	6	3,4	3,3
		7	3,0	3,2
		8	2,4	1,9
	Propriétaire	6	58,2	70,5
		7	65,8	66,8
		8	70,9	64,3

Source Enquête EU-SILC 2006-2008, CEPS/INSTEAD-STATEC

2) Choix et découpage des variables explicatives

- Les outils utilisés pour ce choix sont la matrice des corrélations et les méthodes classification;

- Le recoupement de leurs solutions donne les variables suivantes:
  - le revenu total brut du ménage (quintiles);
  - la classe d'âge du chef de ménage(*16_34 ans, 35_49 ans, 50_64 ans, 65 ans ou +)*;
  - les coûts du logement ("Une charge importante" , "Une charge moyennement importante" , "Une charge pas du tout importante").

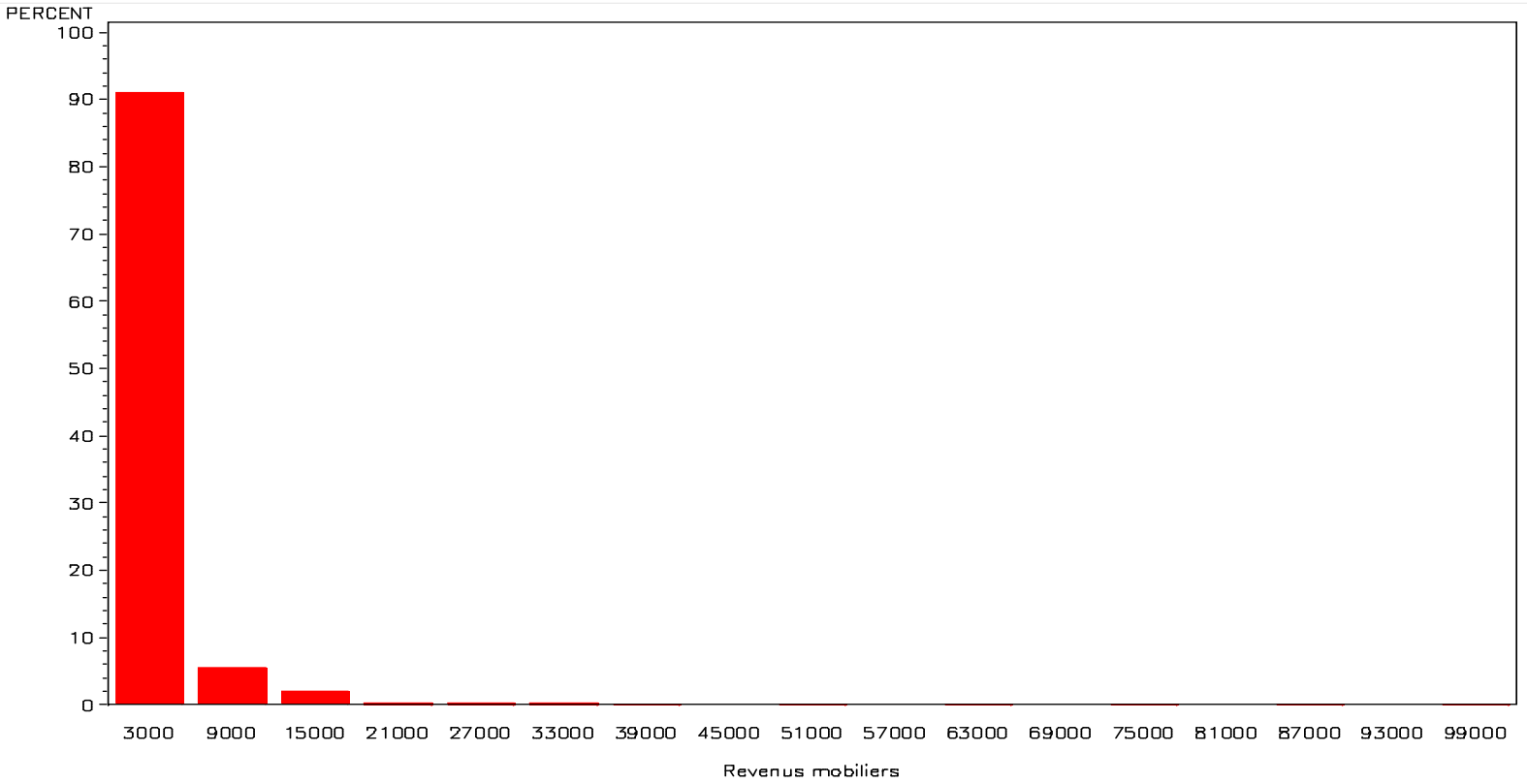
  Soit 60 classes d'imputation.

3) Traitement des valeurs extrêmes

- Pas de censure de valeurs extrêmes vu la distribution très asymétrique des revenus mobiliers dans le graphique 1 ci-dessous;

- L'absence de censure permet de ne pas désavantager dans la comparaison les méthodes éventuellement insensibles aux valeurs extrêmes.

Graphique 3 : Distribution des revenus mobiliers collectés



Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

III) Les Quatre Méthodes d'Imputation Retenues

- **Méthode du hot deck aléatoire intra-classe ;**
- **La méthode de la médiane intra-classe;**
- **Méthode du mode intra-classe;**
- **Méthode de la moyenne intra-classe.**

1) Méthode du hot deck aléatoire intra-classe

- Soit X_{ijkl}, une observation i de la classe d'imputation C_{jkl} de n_i observations avec :

  - j = 1… 5 quintiles du revenu total
  - k= 1….4 classes d'âge du chef de ménage ;
  - L= 1….3 modalités des coûts du logement.

- Elle consiste à prendre aléatoirement un répondant X_{hjkl} avec remise ou sans remise sur l'ensemble des répondants s_h dans la classe C_{jkl} contenant n_h répondants : $X_{hjkl} = X_{ijkl}$, h $\square$ s_h , tel que $P(X_{ijkl} = X_{hjkl}) = 1/ n_h$

2) La méthode de la médiane intra-classe

- Soit X_{hjkl} une observation de la classe C_{jkl} de répondants d'effectif n_h;

- Si n_h est impair, la médiane retenue comme donneur dans cette classe d'imputation est égale à $X_{jkl(nh+1)/2)}$;

- Si n_h est pair, la médiane retenue comme donneur dans cette classe d'imputation est égale à $(X_{jkl(nh/2)} + X_{jkl[(nh/2)+1])})/2$

3) La méthode du mode intra-classe

- Elle consiste à choisir la valeur renseignée la plus fréquente dans la classe d'imputation pour imputer une donnée manquante y appartenant:
  - En cas de solution multiple, c'est la solution par défaut du logiciel SAS qui est adoptée;
  - La médiane de la classe correspondante est utilisée pour pallier l'absence de mode.

4) La méthode de la moyenne mode intra-classe

- Soit la classe d'imputation C_{jkl} de n_i observations X_{ijkl} , alors la moyenne dans cette classe est égale à $a_{0i}=(\Sigma_{ni} X_{ijl}/ n_i)$;

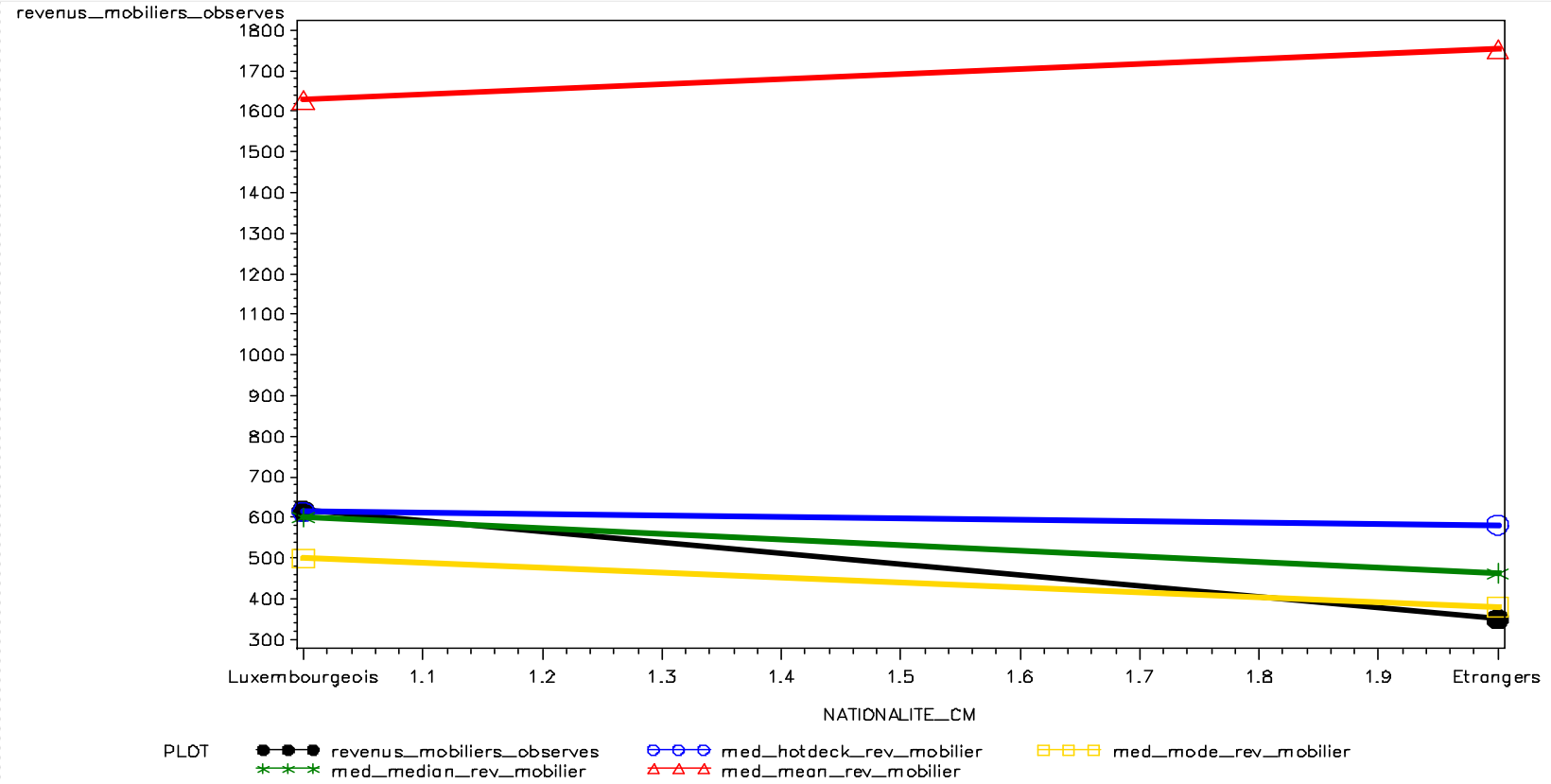
- C'est la moyenne pondérée qui est effectivement utilisée. Les poids W_i ne sont pas intégrés dans la formule par souci d'éviter une lourdeur.

IV)Tests de cohérence interne des revenus mobiliers

Examen des profils des médianes des différents revenus mobiliers estimés en fonction:

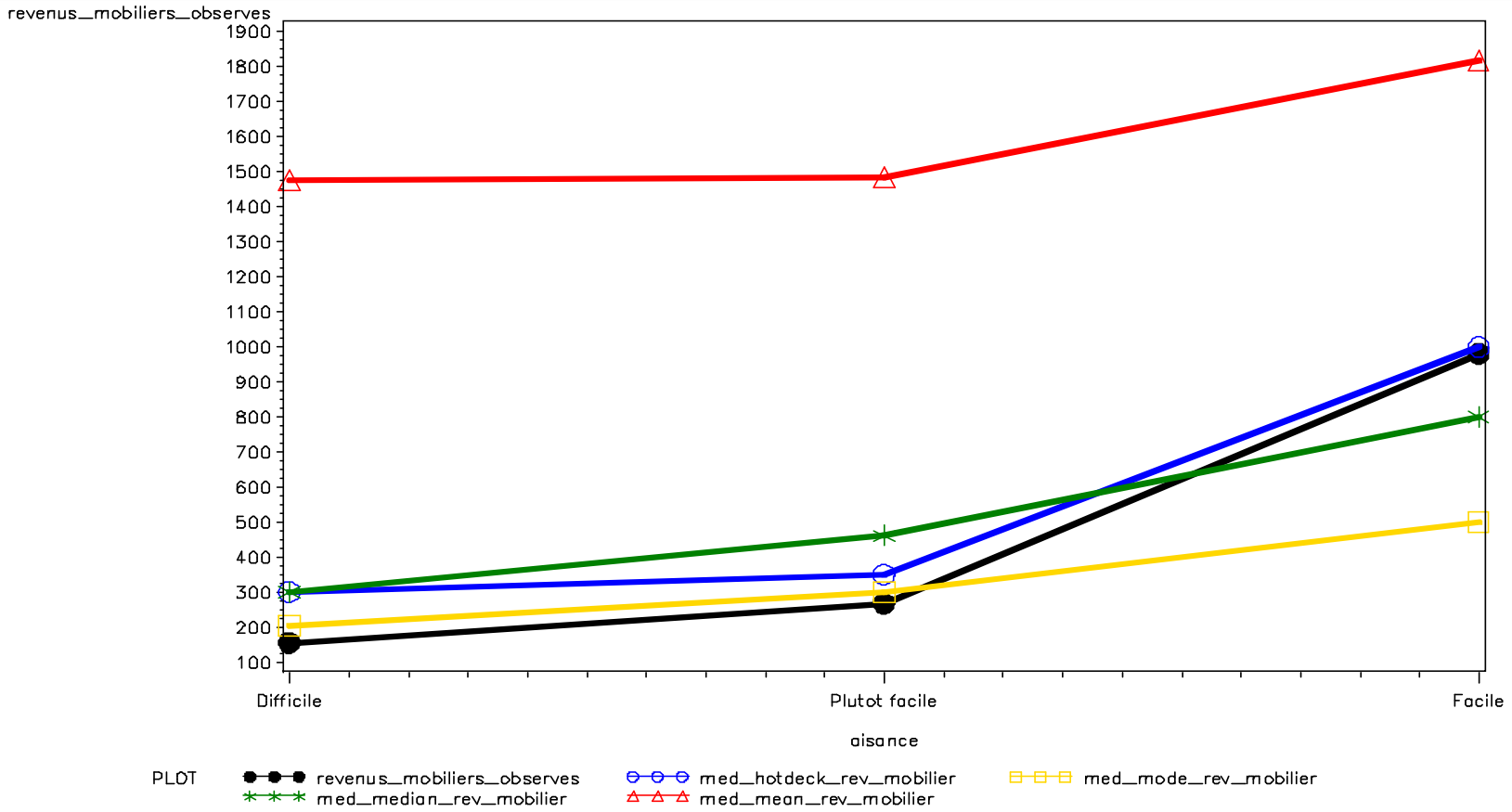
- de la nationalité du chef de ménage;
- de l'aisance de vie du ménage;
- du statut d'occupation du ménage;
- du loyer générique c'est à dire réel ou fictif.

Graphique 2 : Évolution de la médiane des revenus mobiliers collectés en fonction de la nationalité du chef de ménage



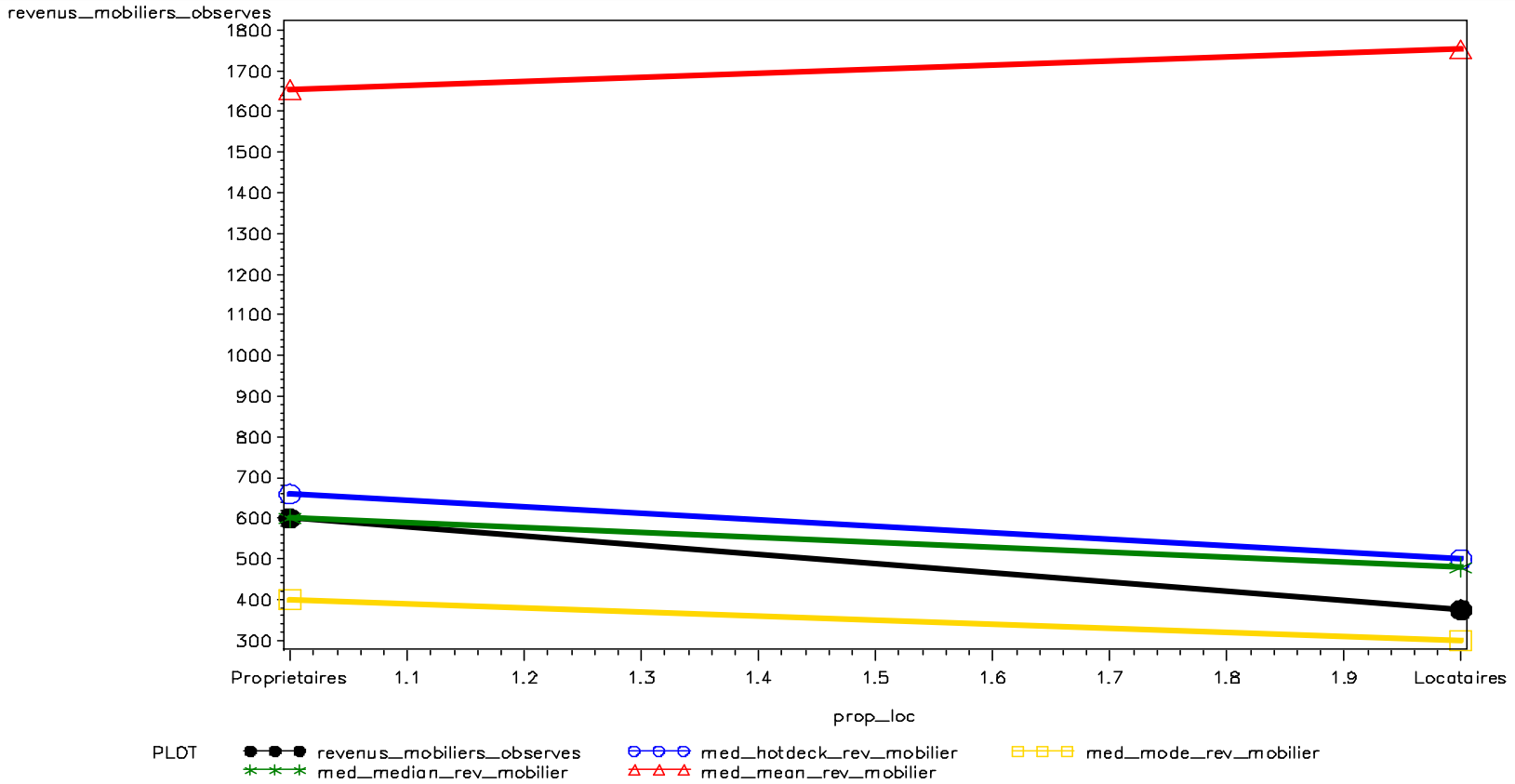
Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Graphique 3 : Évolution de la médiane des revenus mobiliers collectés en fonction de l'aisance de vie du ménage par rapport à son revenu



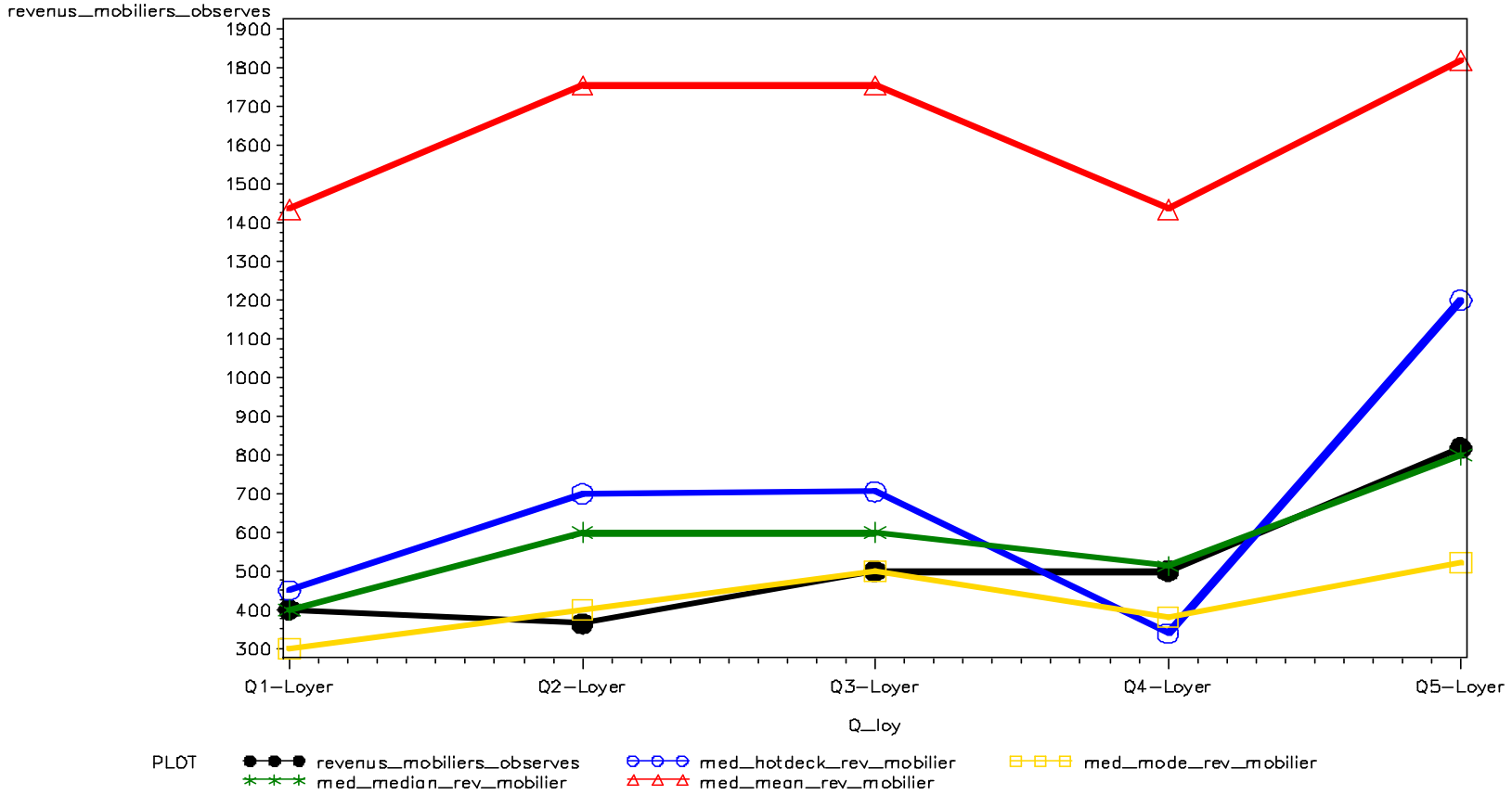
Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Graphique 4 : Évolution de la médiane des revenus mobiliers collectés en fonction du statut d'occupation du logement



Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

Graphique 5 : Évolution de la médiane des revenus mobiliers collectés en fonction du niveau de loyer générique



Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

V) Test de validation sur les résidus

- Grâce à la méthode de validation par l'échantillon-test étendu à l'ensemble des observations renseignées:
  - Calcul de la moyenne pondérée des résidus;
  - Calcul des écarts absolus moyens pondérés.

Tableau 3 : Moyenne pondérée des résidus et
Écarts absolus moyens pondérées des résidus des estimations

Indicateurs / Méthodes	Moyenne pondérée des écarts d'erreurs	Ecarts absolus Moyens pondérés
Hot deck	-1135,36	3657,71
Moyenne	-325,30	1912,02
Mode	1045,99	1946,88
Médiane	1049,58	1918,48

Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

VI)Test de validation par l'analyse des paramètres des distributions des données observées et des données imputées

Comparaison entre la distribution des revenus mobiliers observés et celle des revenus mobiliers imputés avec les 4 méthodes sur la base des paramètres suivants (voir tableau 4 ci-dessous):

- L'étendue;
- La moyenne;
- La médiane;
- L'écart-type.

Tableau 4 : Statistiques sur la distribution des données originales et sur les données estimées avec les quatre méthodes d'estimation.

Indicateurs / Méthodes	Minimum	Maximum	Moyenne	Médiane	Ecart-type
Valeurs observées	1	100000	1904	500	32894
Hot deck	5	72000	2989	606	42684
Moyenne	33	53845	2811	1662	26979
Mode	20	53845	1979	400	27953
Médiane	33	53845	1873	566	27612

Source Enquête EU-SILC 2008, CEPS/INSTEAD-STATEC

VII) Conclusions

- Le premier test sur les résidus consacre la supériorité des performances de la méthode du hot deck;

- Dans le second test sur la plausibilité selon l'hypothèse MAR des distributions des revenus mobiliers, la méthode du hot deck et la méthode du mode donnent des résultats plus cohérents;

- Enfin dans le dernier test, l'avantage revient à la méthode du hot deck, les performances de la méthode de la moyenne sont jugées largement inférieures à celles des autres méthodes.

- La méthode du hot deck est choisie comme la moins mauvaise parmi les 4 méthodes testées;

- L'impact de la méthode du hot deck serait une relative surestimation des revenus mobiliers en se fiant au critère de la médiane le plus adapté pour juger de la qualité de l'estimation d'une variable dont la distribution est fortement affectée par les valeurs extrêmes;

- On note une surestimation importante pour l'estimateur Horvitz-Thompson du total des revenus mobiliers bruts sur la base de la différence significative entre la moyenne des revenus mobiliers observés et celle des revenus mobiliers estimés avec la méthode du hot deck.