

MÉTHODE NON-PARAMÉTRIQUE DE DÉTECTION ET DE TRAITEMENT DES INDIVIDUS ATYPIQUES

—

CAS DE L'ENQUÊTE AUPRÈS DES VISITEURS RÉSIDANT HABITUELLEMENT HORS DE FRANCE (EVE)

Claire JACOD (), Julien VALENTINO (**)*

() Banque de France, Direction de la Balance des Paiements
(**) MEEDDAT, Commissariat Général au Développement Durable,
prédécesseur de Claire Jacod*

La ligne « voyages » de la Balance des paiements recense, en débit, les dépenses touristiques des résidents français à l'étranger et, en crédit, les dépenses touristiques des non-résidents lors de leur séjour en France. Avec l'arrivée de l'euro en 2001, la méthodologie de la ligne « voyages » de la Balance des Paiements a nécessité une complète révision. Après une phase transitoire, les chiffreages fondés sur une nouvelle méthodologie ont été définitivement adoptés et publiés à l'occasion du rapport annuel de la Balance des Paiements¹ de 2006.

Concernant le poste des recettes de la ligne « voyages »², une enquête complexe a été mise en place dès 2004, qui constitue la seule source de données disponible : l'Enquête auprès des Voyageurs Etrangers (EVE). Trimestrielle, elle est réalisée en partenariat avec l'Insee³ et sous-traitée à TNS-Sofres. Les objectifs de cette enquête sont multiples : estimer les dépenses touristiques des non-résidents en séjour en France, mais aussi estimer par nationalité le nombre de touristes, le volume de nuitées, ... L'enjeu est de taille, la France étant le premier pays au monde en termes d'arrivées de touristes, avec plus de 80 millions d'arrivées de touristes en 2007⁴, soit 9 % des arrivées de touristes internationaux dans le monde.

Ce document expose un des principaux traitements appliqué aux données : la détection des individus atypiques en termes de dépenses. Les caractéristiques de l'enquête et de ses résultats nous ont poussés à utiliser une méthodologie originale que nous souhaitons présenter ci-après. La suite des traitements sur les données d'EVE visent à l'élaboration du poste recettes de la ligne « Voyages » : traitement des atypismes restant, prise en compte des séjours non terminés ou à cheval sur plusieurs trimestres, ... Par souci de concision, ces traitements supplémentaires ne seront pas exposés ici.

Mots clés :

EVE, enquête, atypique, dépense, densité, seuil, estimation, réplcation, arrondi

¹ Ce rapport est disponible sur le site de la Banque de France à l'adresse suivante : http://www.banque-france.fr/fr/stat_conjoncture/balance/bdppof/rabp2006.htm.

² Les termes spécifiques à ce domaine particulier sont regroupés et définis dans un rapide glossaire en fin de document.

³ A l'origine, l'enquête était réalisée en collaboration avec la Direction du Tourisme, dont la mission sur les enquêtes touristiques a rejoint l'Insee début 2009.

⁴ Données publiées par la Direction du Tourisme, dans la collection *Chiffres clés du Tourisme*, disponible à cette adresse : http://www.tourisme.gouv.fr/fr/z2/stat/chiffres/att00009212/chiffres_cle-FR.pdf.

1. Présentation d'EVE et des problèmes d'individus atypiques

1.1. Présentation de l'enquête

L'enquête est basée sur un questionnaire qui s'articule, en ce qui concerne les dépenses touristiques, en deux parties : les dépenses réalisées avant le départ et les dépenses réalisées sur place. Pour chacune de ces deux parties, l'interrogation porte non seulement sur le montant total dépensé, mais également sur le moyen de paiement utilisé : par exemple carte bancaire professionnelle ou personnelle. On s'intéresse également aux dépenses de transport international, car elles ne font pas partie du champ et sont donc à exclure du montant total pour l'élaboration des recettes de voyages. Compte-tenu du public auquel ils s'adressent, les questionnaires doivent être disponibles dans une dizaine de langues⁵. L'interrogation est réalisée par des enquêteurs⁶ à différents points de sortie du territoire qui doivent prendre en compte à la fois les différents modes de transport (air, rail, route, mer) et leur répartition géographique. Par exemple, pour le mode aérien, les interrogations ont lieu dans les aéroports de Charles de Gaulle, Orly, Beauvais, Lyon, Nice, et dans les autres aéroports de province⁷. Pour le mode route, pour la frontière espagnole, selon qu'on se situe à l'est ou à l'ouest, les personnes qui traversent la frontière n'ont pas les mêmes caractéristiques : on trouve essentiellement des Français et des Espagnols à l'ouest, tandis qu'à l'est, les pays de résidence sont plus variés, les résidents du nord de l'Europe (Allemands Néerlandais, ...) étant nombreux. On fait donc la distinction entre la frontière espagnole est et la frontière espagnole ouest parmi les interrogations.

Mais les réponses aux questionnaires sont insuffisantes à l'élaboration des recettes de la ligne « voyages ». En effet, il faut les extrapoler à la population d'intérêt, qui correspond aux non-résidents quittant la France quelle que soit la durée de leur séjour. **Or, et c'est ce qui fait la principale spécificité d'EVE, cette population n'est pas connue.** Elle est évaluée à partir des statistiques de trafic de sortie du territoire, dont le recueil est réalisé par le Centre d'Études Techniques de l'Équipement (CETE) d'Aix-en-Provence ou qui sont directement fournies par des organismes ou des sociétés de transport⁸, et d'une estimation de la part des non-résidents dans ce trafic. Ces dernières statistiques sont obtenues à partir de comptages effectués par TNS-Sofres, et correspondent à une part non négligeable du travail sur l'enquête. On distingue également les excursionnistes des touristes puisque leurs comportements de dépenses sont différents. La part des excursionnistes dans le nombre de non-résidents quittant la France est estimée, soit à partir des données disponibles dans le questionnaire pour les modes de transport international non routier, soit grâce à une clé de répartition⁹ datant des Enquêtes aux Frontières (EAF) pour le mode de transport routier. Pour ce dernier mode, la ventilation touristes/excursionnistes tient compte de critères comme la saison ou l'importance du point de sortie du territoire.

Le fait de ne pas connaître la population d'intérêt rend difficile l'estimation de la qualité des données. Il existe une incertitude due aux étapes préliminaires d'estimation de la population de référence que l'on ne peut pas évaluer. Cette incertitude est indépendante des traitements réalisés ultérieurement mais impacte les données publiées dans la Balance des paiements.

La mission confiée à TNS-Sofres est d'estimer la population d'intérêt, de traiter les données de l'enquête EVE et de les extrapoler à cette population. A cette étape, la contribution de la Banque de France aux traitements consiste à détecter les individus atypiques en termes de dépenses, suivant une méthode que nous avons choisi de vous présenter. Un ensemble de traitements est réalisé par la suite par la Banque de France, qui vise à l'élaboration du poste recettes de la ligne « voyages ». Ces traitements ne sont pas présentés ici.

⁵ Les questionnaires sont disponibles en allemand, anglais, chinois, espagnol, français, grec, italien, japonais, néerlandais et portugais.

⁶ On compte environ 600 vacations d'enquête par trimestre.

⁷ Tous les aéroports de province ne sont pas enquêtés mais, en moyenne, le taux de couverture du trafic vers l'international est de 92%.

⁸ Par exemple la DGAC pour le transport aérien, Thalys, Eurostar, SnCF pour le train, ...

⁹ Pour simplifier la présentation un singulier est ici employé. Dans la réalité, cette clé n'est pas unique. Elle est fonction de la nationalité des non résidents, de leur mode de transports, ...

Pour des raisons essentiellement budgétaires, le nombre de questionnaires a diminué en 2007, suite au passage à une deuxième vague de l'enquête portant jusqu'à fin octobre 2010. On est passé de 120 000 questionnaires par an à 80 000. La réduction du nombre de questionnaires impacte la qualité de l'enquête. La méthode de détection des individus atypiques utilisée par la Banque de France a été renforcée, car la baisse du nombre de questionnaires avait posé des problèmes lors de son utilisation. C'est cette dernière méthode qu'on détaille ici.

1.2. Impact des certains questionnaires en termes de dépenses

Depuis le lancement de l'enquête, on constate pour chaque période la rentrée de questionnaires dont les montants déclarés sont particulièrement importants et donc fortement contributifs aux dépenses totales extrapolées à la population des visiteurs étrangers. En effet, sur la dernière année disponible, allant du 4^{ème} trimestre 2007 au 3^{ème} trimestre 2008, 10 questionnaires affichent des dépenses déclarées par des non-résidents supérieures à un demi-million d'euros. Ces 10 réponses pèsent pour 55 % des dépenses de voyages extrapolées de la période (cf. tableau ci-dessous).

Seuil	Nombre de questionnaires dont les dépenses totales sont supérieures au seuil	Dépenses totales des questionnaires dont les dépenses sont supérieures au seuil (en milliards d'euros)	Part dans les dépenses totales des questionnaires dont les dépenses totales sont supérieures au seuil
10 000 €	663	241,5	68,8 %
30 000 €	164	227,9	64,9 %
50 000 €	94	220,8	62,9 %
100 000 €	45	214,0	60,9 %
200 000 €	22	203,7	58,0 %
500 000 €	10	193,4	55,1 %

Or, en regardant de plus près les questionnaires concernés¹⁰, il ressort que les dépenses mentionnées dans ces questionnaires ne sont pas à prendre en compte pour l'élaboration du poste dépenses de la ligne « voyages » de la balance des paiements. En effet, les montants sont erronés (erreur de lecture ou de conversion des devises), fantaisistes (les réponses aux questionnaires sont incohérentes) ou ne correspondent tout simplement pas à des « dépenses de tourisme » (achat de chalet à la montagne, de maison, d'appartement¹¹, ...). Les montants des premiers questionnaires peuvent être corrigés des erreurs de lecture ou de conversion. Mais il convient de déceler l'ensemble de ces questionnaires et de corriger leurs dépenses, soit de façon directe, quand l'erreur est visible et corrigable, soit de façon indirecte, en remplaçant le montant par une estimation le cas échéant. Notre propos est ici de déterminer une méthode permettant de détecter ces questionnaires atypiques.

L'observation de la distribution¹² des dépenses totales déclarées nous amène à plusieurs conclusions :

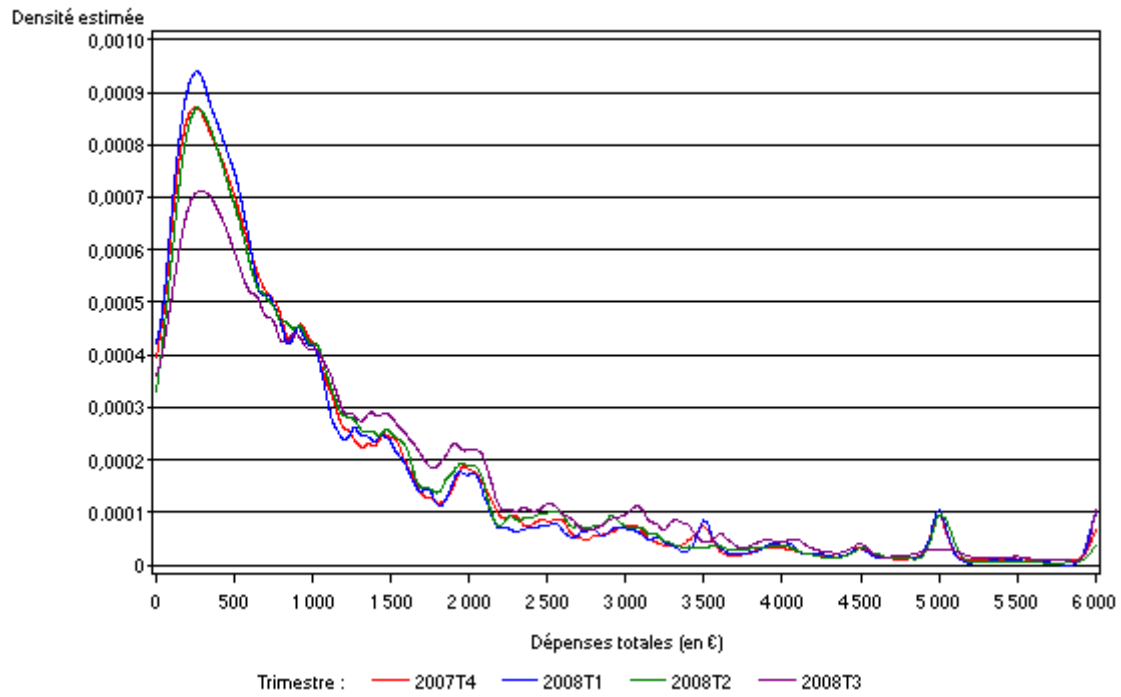
- les distributions ont sensiblement la même forme pour chaque trimestre ;
- il existe des « pics locaux » situés sur les valeurs « rondes » (tous les 500 ou 1 000 euros) ;

¹⁰ Les questionnaires sont scannés, on peut donc les visualiser directement.

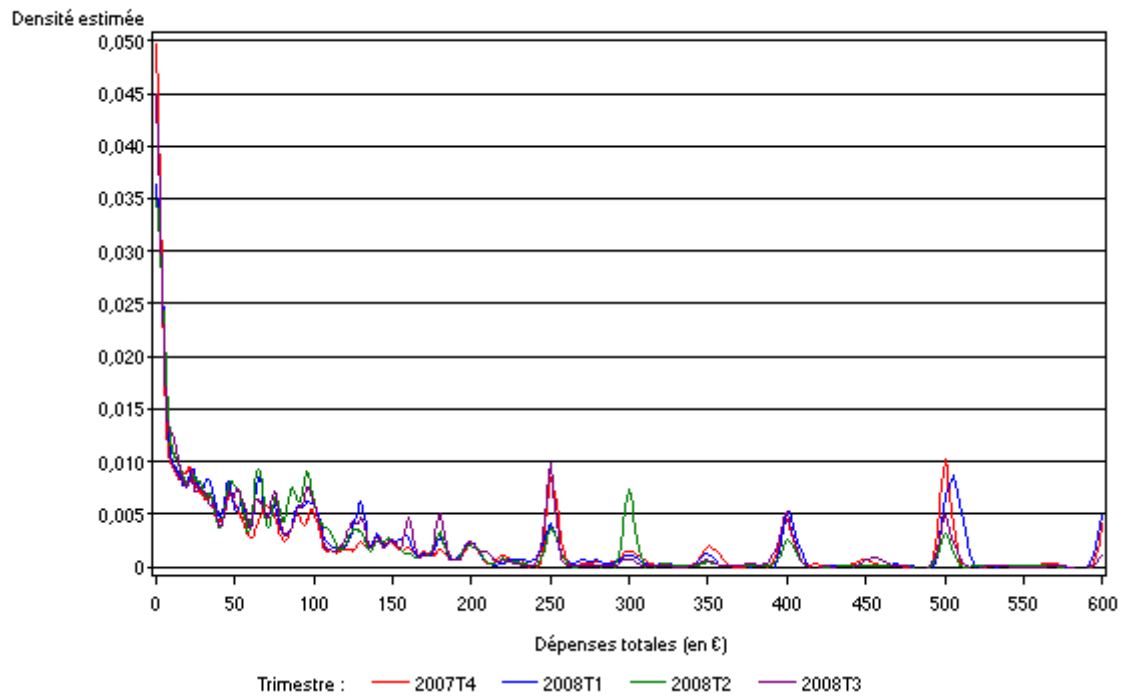
¹¹ En principe, ces flux sont déjà comptabilisés via d'autres sources dans d'autres lignes de la Balance des paiements, comme les investissements directs à l'étranger.

¹² On utilise un estimateur à noyau de Tukey ou biweight, qui effectue un lissage (cf. annexe 1).

Densité des dépenses totales des touristes

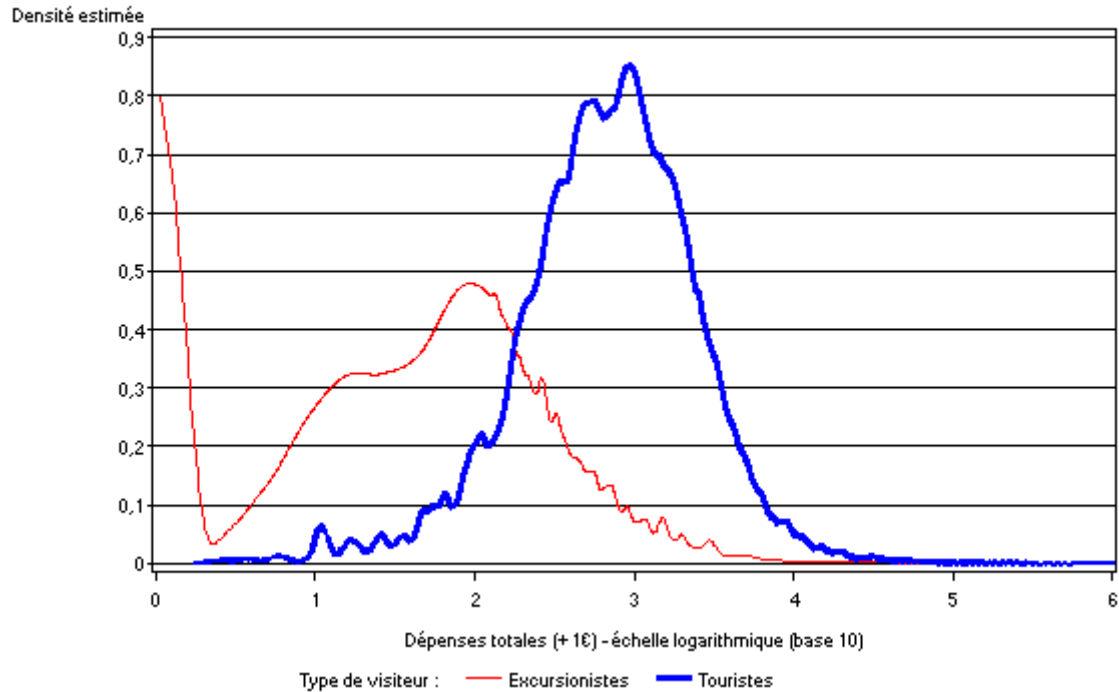


Densité des dépenses totales des excursionnistes



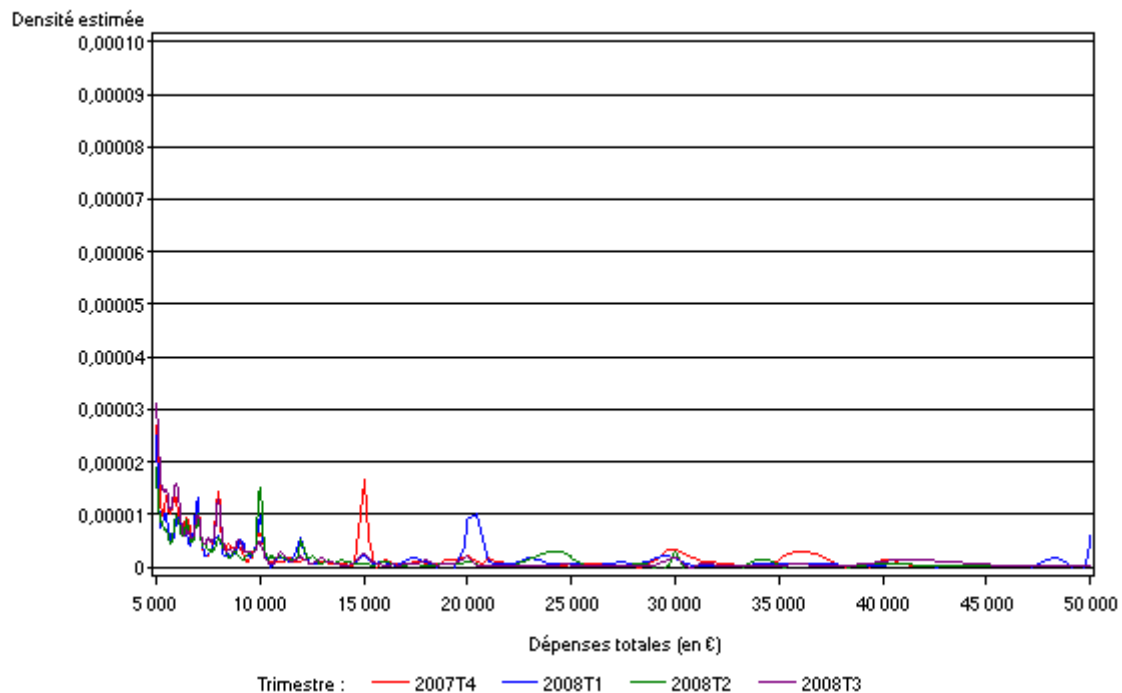
- les excursionnistes, qui font l'aller-retour en France dans la journée, et les touristes, qui passent au moins une nuit sur place, ont des profils de dépenses très différents : il conviendra donc de faire une analyse séparée pour les touristes et pour les excursionnistes ;

Densité des dépenses totales des visiteurs - période 2007T4-2008T3

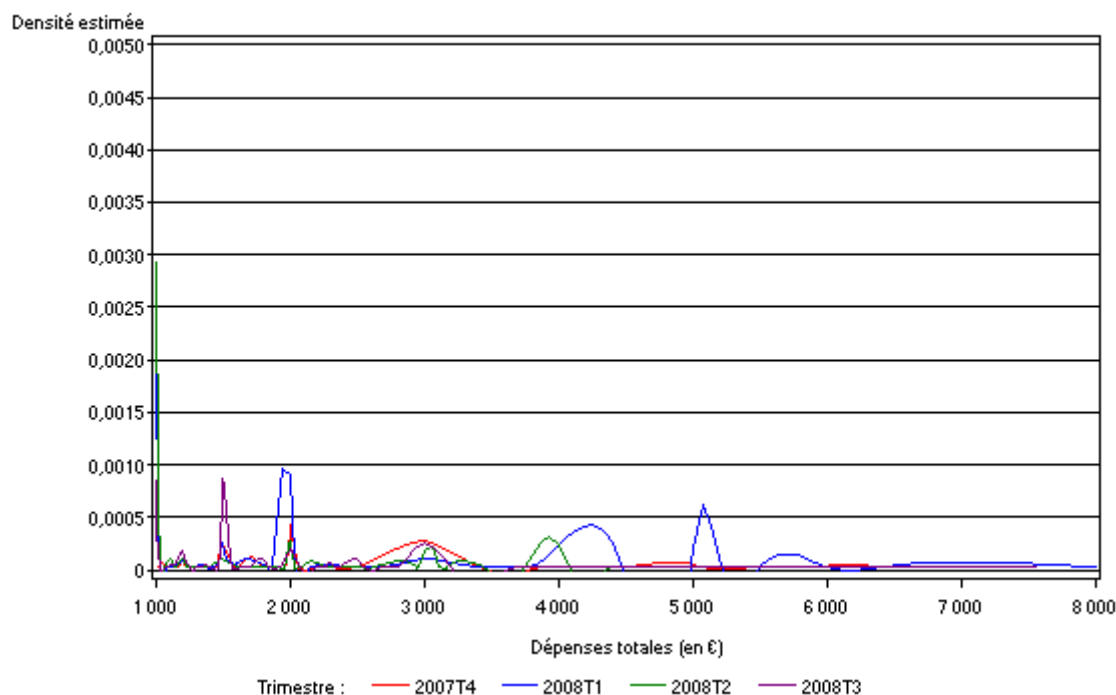


- pour les excursionnistes, deux modes apparaissent : un premier autour de 0 euro, un second autour de 100 euros ; pour les touristes, on observe un mode autour de 1 000 euros ;
- les queues de distributions, dans les valeurs élevées, ont une forme particulière : la densité s'annule puis « sursaute » plusieurs fois, jusqu'à des valeurs très élevées.

Densité des dépenses totales des touristes - queues de distribution



Densité des dépenses totales des excursionnistes - queues de distribution



Etant données les caractéristiques des distributions observées (multimodales, comportant des pics dans des valeurs très élevées), il n'a pas été possible de les normaliser. De plus, la population d'intérêt est mal connue et très hétérogène et le questionnaire n'apporte pas d'informations sociodémographiques sur les visiteurs, ce qui laisse peu de choix sur les méthodes auxquelles il est possible de recourir pour traiter les questionnaires aberrants en termes de dépenses. Par exemple, la nationalité des visiteurs est insuffisante pour caractériser leurs dépenses et déterminer leur atypie. Ainsi, l'application de méthodes itératives en retirant les questionnaires aux plus fortes dépenses un par un, ou l'utilisation de transformation de type Box-Cox¹³ se sont révélées infructueuses. Les tests se basant sur ces hypothèses de normalité (test de Grubbs¹⁴, test de Rosner¹⁵) n'ont donc pas pu être utilisés.

Le test de Walsh¹⁶ a la particularité de ne pas exiger de conditions de normalité, et de pouvoir détecter la présence d'un nombre quelconque de points atypiques. Cependant, réalisé de manière itérative, ce test n'a pas cessé de détecter un nombre croissant de questionnaires atypiques en termes de dépenses, ce test a été stoppé alors qu'il détectait plus de 50% des questionnaires comme atypiques.

Ainsi, n'ayant pas trouvé de solution satisfaisante dans les méthodes « classiques », inadaptées aux données ou trop contraignantes, nous avons exploré une méthode alternative.

¹³ Appliquer une transformation Box-Cox de paramètre λ à une donnée x , c'est la transformer en $f_\lambda(x)$ où :

$$f_\lambda : t \rightarrow \begin{cases} \frac{t^\lambda - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \ln(t) & \text{si } \lambda = 0 \end{cases}$$

¹⁴ Le test de Grubbs est utilisé par la Direction du Tourisme et TNS-Sofres pour la détection des durées de séjour atypiques dans l'enquête SDT. Cf. l'annexe 2 pour de plus amples détails.

¹⁵ Le test de Rosner est présenté en annexe 2.

¹⁶ Le test de Walsh est présenté en annexe 2.

2. La détection des dépenses atypiques

2.1. Objectifs et contraintes de la méthode

2.1.1. Objectifs

L'objectif principal de la méthode recherchée est de déterminer si les dépenses d'un questionnaire donné sont aberrantes ou non.

Compte-tenu des profils différents entre les visiteurs, l'analyse doit être menée séparément sur les touristes et les excursionnistes. Etant donné le régime trimestriel de production et la saisonnalité de l'objet étudié, l'analyse doit être faite par trimestre.

Un questionnaire peut être déclaré atypique en raison non seulement du montant de ses dépenses totales, mais aussi de certaines caractéristiques du séjour. Ainsi, un questionnaire dont les dépenses totales s'élèvent à 500 000 euros est très atypique. Mais, un autre dont les dépenses totales s'élèvent à 20 000 euros pour un touriste n'ayant passé qu'une nuit en France est tout aussi atypique.

Schématiquement on peut considérer a priori que les dépenses des visiteurs non résidents sont de deux types :

- une partie « fixe », dans laquelle l'achat d'une maison ou d'un appartement viendrait « polluer » nos estimations ;
- une partie « variable », fonction du nombre de nuitées pour les touristes ou du nombre de personnes pour les excursionnistes, pour laquelle les dépenses déclarées doivent être un tant soit peu en cohérence avec des caractéristiques « physiques » du questionnaire (nombre de personnes, de nuitées ...).

De ce fait, l'atypie doit être détectée au niveau des dépenses totales des questionnaires, mais aussi au niveau des dépenses par nuitée (pour les touristes) et par personne (pour les excursionnistes).

2.1.2. Contraintes

On se fixe les cinq principales contraintes suivantes :

Si un questionnaire est considéré comme atypique pour ses dépenses totales, un autre questionnaire dont les dépenses totales lui seraient très proches (quelques euros d'écart par exemple) doit être lui aussi considéré comme atypique.

Les questionnaires qualifiés d'atypiques (respectivement non atypiques) doivent le demeurer après les opérations suivantes :

- Multiplication des dépenses totales de tous les questionnaires par une même constante positive : cela signifie, par exemple, que si l'on convertit toutes les données dans une autre devise, le seuil au-delà duquel les questionnaires seront qualifiés d'atypiques s'en trouvera inchangé (à la conversion dans la « nouvelle » monnaie près) ;
- Addition d'une même constante aux dépenses totales de tous les questionnaires : l'instauration d'une taxe fixe (droit d'entrée, ...) ne modifiera pas le seuil au-delà duquel les questionnaires seront qualifiés d'atypiques (au décalage près de la valeur de la taxe).

On souhaite que l'atypie des questionnaires dépende du trimestre étudié :

- pour respecter la saisonnalité des dépenses ;
- pour prendre en compte l'effet de la conjoncture sur la distribution des dépenses.

La méthode retenue doit être simple à implémenter et à utiliser.

Elle doit être compatible avec le travail d'imputation réalisé par TNS-Sofres.

2.2. Méthode non-paramétrique choisie

Les méthodes classiques n'ont pas pu être appliquées au cas de l'enquête EVE, notamment en raison de la forme des queues de distributions. Nous avons donc souhaité utiliser cette particularité pour caractériser les individus atypiques en termes de dépenses et les détecter.

Pour chaque type de visiteur (touriste ou excursionniste) et pour chaque trimestre, on s'intéresse à la densité des dépenses totales et des dépenses par nuitée (pour les touristes) ou par personne (pour les excursionnistes). **On considère que tous les « soubresauts » suivant le premier point d'annulation de chaque densité correspondent à des dépenses atypiques.**

Ainsi, pour chaque type de visiteur et pour chaque trimestre, on construit deux seuils :

- un premier seuil en termes de dépenses totales, calculé comme étant la première valeur d'annulation de la densité des dépenses totales ;
- un second en termes de dépenses par nuitée ou par personne, calculé comme étant la première valeur d'annulation de la densité des dépenses par nuitée ou par personne.

Ensuite, sont considérés comme atypiques tous les questionnaires dont les dépenses sont supérieures au premier ou au second seuil.

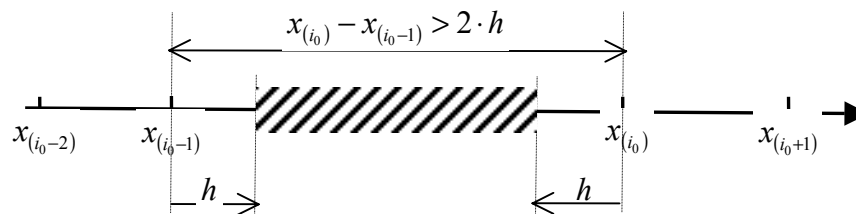
2.2.1. Présentation théorique

Compte-tenu des formules de calcul de l'estimation à noyau de Tukey de la densité¹⁷, le premier point où la densité s'annule se situe entre les deux premières observations (en les classant par ordre croissant) significativement espacées. Ce qui se traduit mathématiquement par une différence supérieure à deux fois une certaine valeur h : la fenêtre. Pour le choix du calcul de cette fenêtre, nous avons utilisé celle préconisée par Lejeune (cf. [1]), Elle a l'avantage d'être simple, rapide à calculer, et peu sensible aux valeurs extrêmes par rapport à celle de P. Deheuvels¹⁸ (cf. [4]) ou celle de type validation croisée qui ne donne pas une expression directe de la fenêtre.

On trie les n observations : $x_{(1)} < x_{(2)} < \dots < x_{(n)}$. La valeur de la fenêtre h est donnée par

l'expression : $h = 0,75 \cdot \left(\frac{n}{100}\right)^{-1/5} \cdot \min_{i=1, \dots, \lfloor \frac{n}{2} \rfloor} \left(x_{\left(i + \lfloor \frac{n}{2} \rfloor\right)} - x_{(i)} \right)$. Si $x_{(i_0)} - x_{(i_0-1)} > 2 \cdot h$, l'estimation de la

densité est nulle sur l'intervalle $[x_{(i_0-1)} + h, x_{(i_0)} - h]$ qui apparaît en zébré sur l'axe ci-dessous, et les valeurs $x_{(i_0)}$, $x_{(i_0+1)}$, ... $x_{(n)}$ seront considérées comme atypiques.



D'un point de vue pratique, le seuil pourra être choisi n'importe où dans l'intervalle $]x_{(i_0-1)}, x_{(i_0)}[$ de façon à ce qu'il soit « rond » (c'est-à-dire sans décimale, arrondi à l'unité ou à la dizaine près).

La méthode proposée revient à détecter la première plage de valeurs de dépenses suffisamment large pour laquelle il n'y a pas de questionnaire correspondant. C'est donc la première rupture en termes de continuité des dépenses : où que l'on se situe avant le seuil, on peut toujours trouver des questionnaires dont les dépenses y sont inférieures, tandis qu'au niveau du seuil il y a un « trou ».

Le seuil se situe au niveau du premier « plat » de la fonction de répartition.

¹⁷ Ces formules sont détaillées en annexe 1.

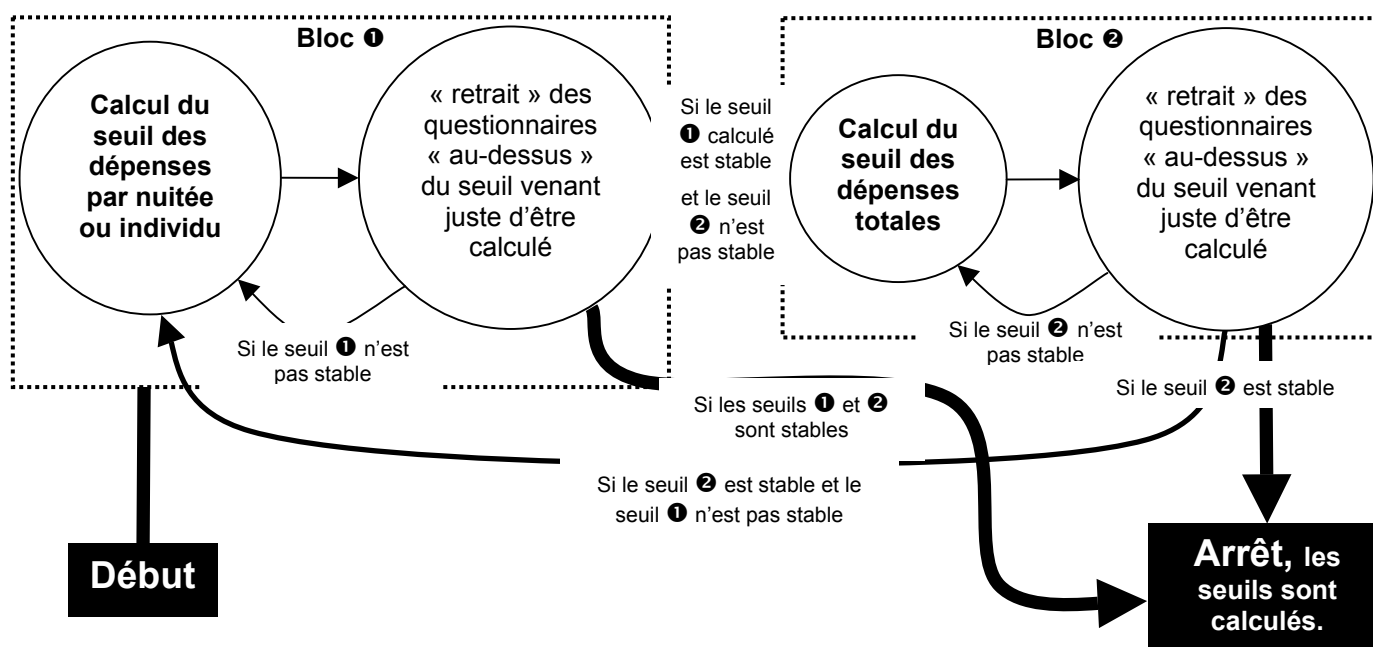
¹⁸ La méthode de P. Deheuvels nécessite une estimation de la variance de la densité, ainsi que le calcul de l'intégrale du carré de la dérivée seconde de la densité supposée Gaussienne.

2.2.2. Présentation technique

Comme annoncé précédemment, **pour chaque trimestre et chaque type de visiteur, deux seuils sont calculés** : l'un en termes de dépenses totales, l'autre en termes de dépenses par nuitée ou personne. Pour l'élaboration de ces seuils, on enchaîne et réitère les calculs jusqu'au moment où les seuils obtenus sont stables.

En effet, les enchaînements des calculs des deux types de seuils permettent des « interactions » entre les opérations : ainsi lors du calcul du seuil des dépenses totales par exemple, les questionnaires atypiques en termes de dépenses par nuitée ou individu auront déjà été supprimés. Le seuil calculé ne prendra donc pas en compte les questionnaires déjà qualifiés d'atypiques. Les répétitions des calculs, c'est-à-dire le calcul de seuil des dépenses totales jusqu'à ce qu'il soit stable, donnent un seuil plus robuste que s'il n'avait été calculé qu'une seule fois et que seuls les questionnaires les plus atypiques avaient été détectés.

Pour chaque trimestre, et chaque type de visiteur, le schéma suivant décrit l'ensemble des opérations réalisées :



Ainsi, comme le décrit le schéma précédent, l'ensemble des deux blocs 1 et 2 sont réalisés successivement jusqu'à ce que les deux seuils calculés soient tous les deux « stables ». La stabilité des deux seuils signifie que le seuil des dépenses par individu ou par nuitée est identique à la fin des deux dernières itérations du bloc 1, et le seuil des dépenses totales est identique à la fin des deux dernières itérations du bloc 2.

Le bloc 1 calcule le seuil des dépenses par individu (pour les excursionnistes) ou par nuitée (pour les touristes). Le calcul est réalisé suivant les formules du paragraphe précédent, à la nuance près que le seuil « choisi » dans l'intervalle $]x_{(i_0-1)} + h, x_{(i_0)} - h[$ est $x_{(i_0-1)} + h$. Ensuite le seuil est recalculé après élimination des questionnaires dont les dépenses par individu ou nuitée sont supérieures au dernier seuil de dépenses par individu ou nuitée calculé. Le bloc 1 se termine dès que le seuil calculé est identique lors des deux dernières itérations.

De la même façon, le bloc 2 calcule le seuil des dépenses totales. Le seuil est calculé une première fois, puis recalculé après élimination des questionnaires dont les dépenses totales sont supérieures au dernier seuil des dépenses totales calculé. Le bloc 2 se termine dès que le seuil calculé est identique lors des deux dernières itérations.

Le tableau suivant illustre l'enchaînement des calculs des seuils sur la population des excursionnistes du 3^{ème} trimestre 2008¹⁹.

n° d'itération	calcul du seuil en termes de :	calcul du seuil en termes de dépenses par personne		calcul du seuil en termes de dépenses totales		nombre de questionnaires au-delà d'au moins un des deux seuils
		valeur de la fenêtre "h"	valeur du seuil (non arrondi)	valeur de la fenêtre "h"	valeur du seuil (non arrondi)	
1	dép. par personne	9,86	545,85			78
2		9,16	545,85			78
3				20,19	943,14	96
4	dép. totales			19,83	866,15	96
5				19,68	866,15	96
6	dép. par personne	9,05	385,43			124
7		9,07	385,43			124
8	dép. totales			19,14	469,14	173
9				18,64	469,14	173
10	dép. par personne	8,97	278,97			196
11		8,92	278,97			196
12	dép. totales			18,56	469,14	196

Unité : euros.

Remarques :

- Commencer la procédure par le bloc ② au lieu du bloc ① ne modifie pas les résultats obtenus.
- En pratique, après implémentation sous le logiciel SAS, la stabilité des seuils est vite obtenue, et les résultats s'obtiennent en quelques minutes.
- Il est nécessaire de prendre en compte la multimodalité des distributions. En effet, les questionnaires aux dépenses nulles sont très nombreux (cf. tableau ci-dessous) et risquent d'influencer le calcul de la fenêtre. En effet, la densité au voisinage de zéro étant très variable, la fenêtre pourrait être sous-estimée. Afin de s'affranchir de ce problème-là, une routine de troncature est lancée avant chaque calcul de seuil. Celle-ci détermine une approximation du premier minimum local de la densité²⁰. Les dépenses inférieures à ce minimum local sont exclues avant de procéder au calcul des seuils.

Type de visiteur	Trimestre	Nombre de questionnaires	Part de questionnaires dont :	
			les dépenses par nuitée ou par personne sont inférieures à 1 €	les dépenses totales sont inférieures à 1 €
Touristes	2007T4	11 262	0,6%	0,6%
	2008T1	11 599	0,7%	0,7%
	2008T2	11 790	0,5%	0,5%
	2008T3	11 612	0,6%	0,6%
Excursionnistes	2007T4	4 000	30,5%	30,5%
	2008T1	3 662	28,0%	28,0%
	2008T2	3 471	26,2%	26,2%
	2008T3	3 649	26,9%	26,9%

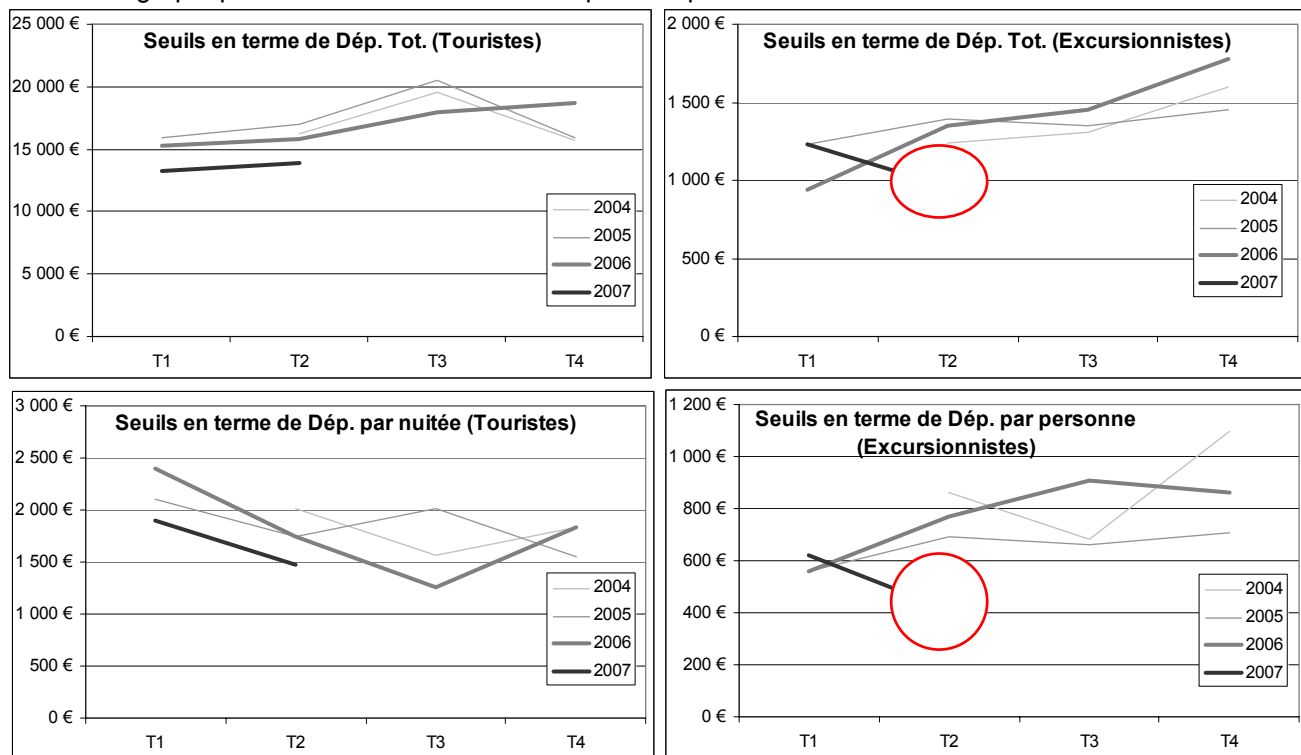
¹⁹ Les calculs portent sur les dépenses brutes sans ajout de bruit blanc et donc sans itération de l'ensemble du processus (cf. infra). Les résultats présentés ici sont très inférieurs aux seuils finaux exposés plus loin.

²⁰ Ce minimum local est approximé par la première valeur en laquelle la dérivée de la densité devient positive.

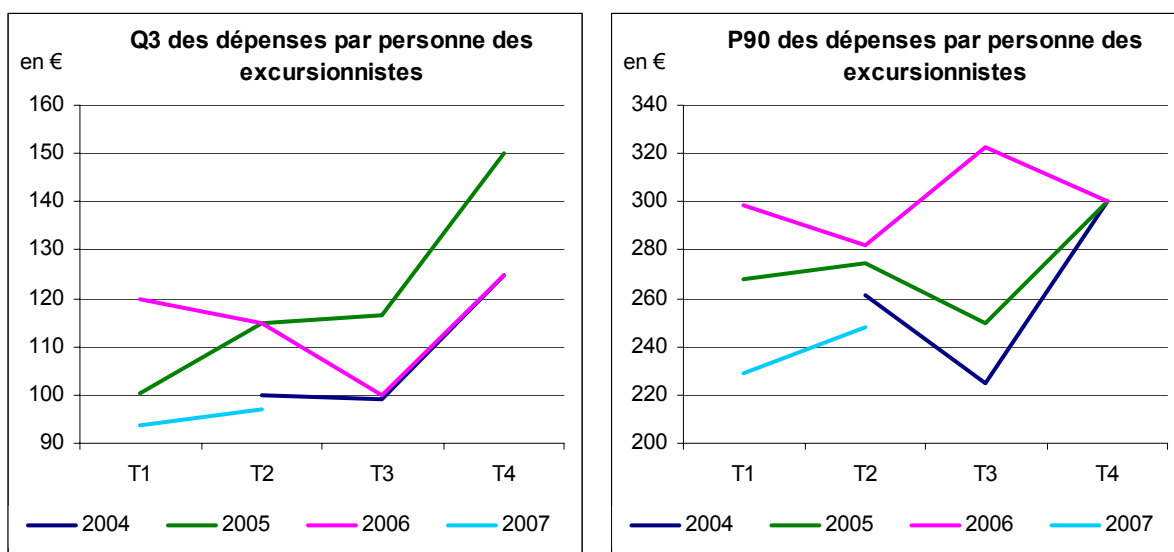
2.3. Amélioration de la robustesse de la méthode

2.3.1. Problème du 2^e trimestre 2007

A partir du deuxième trimestre 2007, le nombre des questionnaires a été réduit, pour des raisons essentiellement budgétaires. **La réduction des données a entraîné une très forte baisse du seuil en termes de dépenses par personne pour les excursionnistes.** Par contre, pour les touristes, qui contribuent à hauteur de 90 % aux dépenses totales, aucun problème particulier ne s'est posé les seuils. Les graphiques suivants retracent l'historique des quatre seuils :



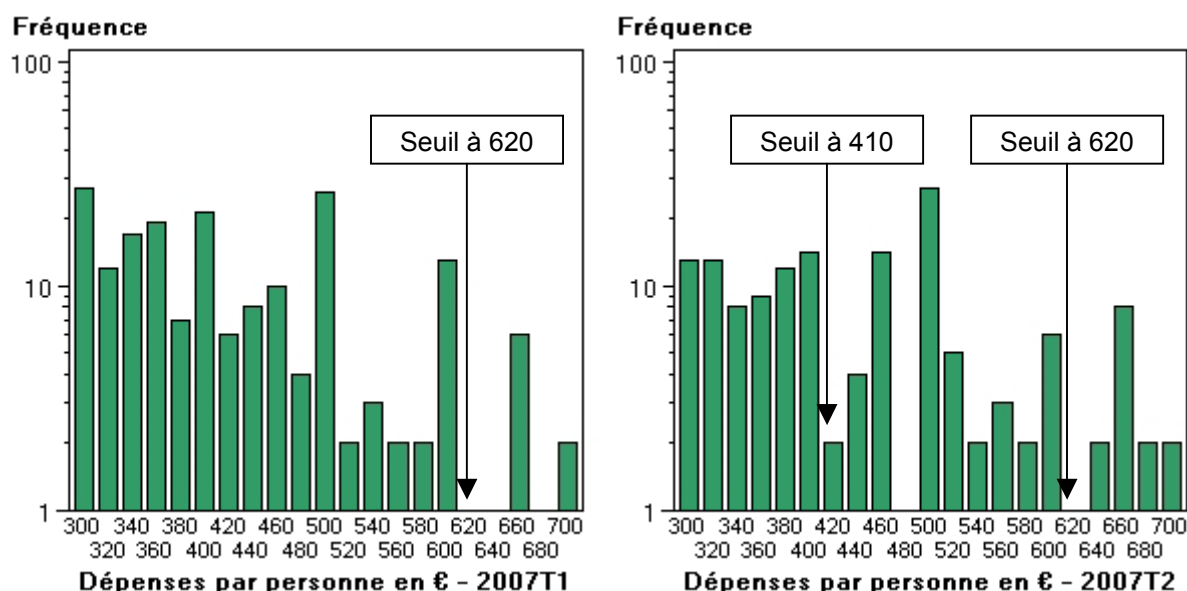
Comme le montrent les graphiques suivants, cette très forte baisse du seuil en termes de dépenses par personne pour les excursionnistes n'est pas justifiable par des caractéristiques usuelles que sont le troisième quartile et le dernier décile, qui augmentent entre le premier et le deuxième trimestre 2007.



Champ : excursionnistes - questionnaires aux dépenses non nulles.

Des histogrammes avec des largeurs d'intervalles bien choisies permettent d'expliquer le bas niveau de ce seuil du deuxième trimestre 2007.

Voici les histogrammes des dépenses par personne non nulles des excursionnistes interrogés au premier trimestre et au deuxième trimestre 2007 :



Compte-tenu de la baisse du nombre de questionnaires exploitables (soit -35 % entre les premiers semestres 2006 et 2007), le calcul des seuils s'effectue désormais sur moins de « points ». De plus, le calcul des différentes valeurs d'annulation des densités (principe retenu pour la détermination des seuils) se base sur les dépenses arrondies au centime d'euro près. Cela fait donc moins de valeurs de dépenses qu'il n'y a de questionnaires exploitables, d'autant plus que certaines valeurs rondes (comme 10 €, 20 €, 50 €, 100 €, ...) apparaissent souvent. Ainsi pour le 2007T2, les seuils des excursionnistes sont calculés avec 44 % d'individus en moins²¹ qu'au 2006T2.

Pour faire face à cette baisse de quantité de « matière première », on se propose de tenir compte de l'incertitude dans les réponses des interviewés, qui ont tendance à arrondir leurs dépenses, dans le calcul des seuils tel que présenté dans le paragraphe 2.2.

2.3.2. Amélioration de la méthode grâce à l'introduction d'un aléa

Afin d'avoir davantage de « matière première », la dépense totale de chaque questionnaire est multipliée par $1 + \varepsilon$ où ε est un bruit blanc²². Ensuite le processus habituel de calcul des quatre seuils est réalisé.

Afin que le seuil calculé ne soit pas entièrement soumis aux aléas d'un seul tirage, l'opération décrite précédemment est renouvelée un certain nombre de fois²³ sur l'ensemble des touristes ou des excursionnistes du trimestre considéré. Ensuite on calcule la médiane pour chacun des quatre seuils.

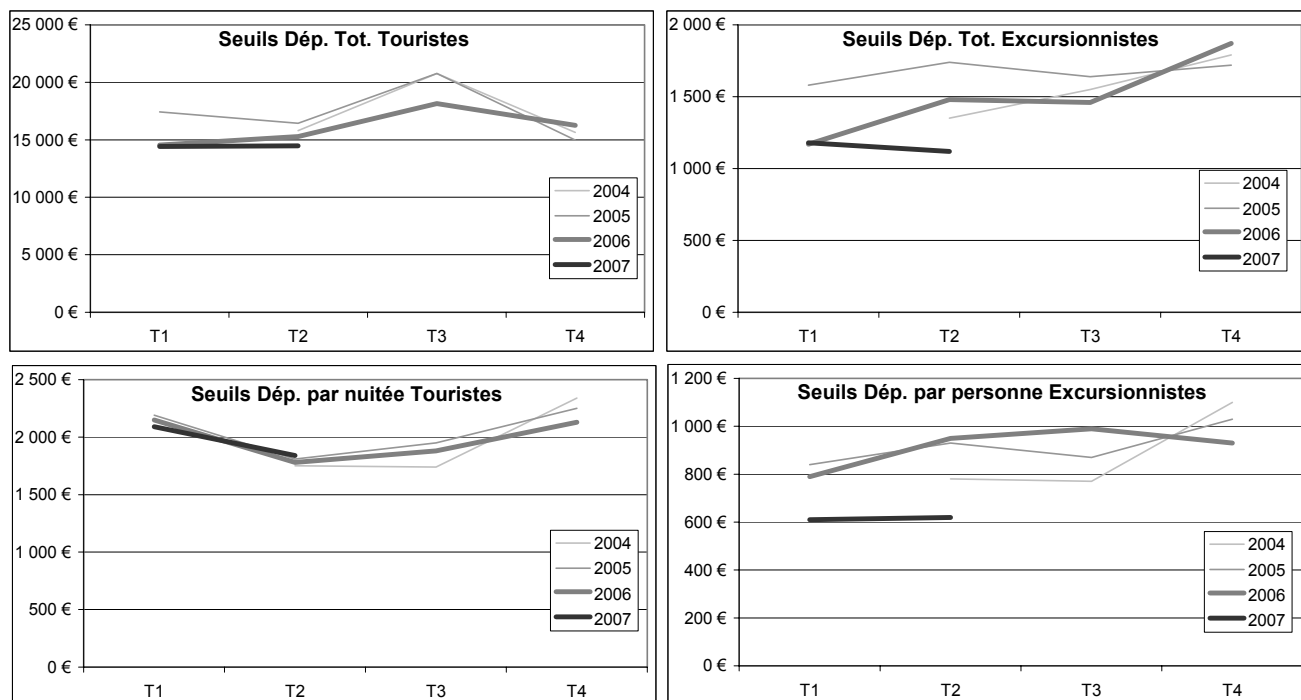
²¹ Soit sur la base de 889 dépenses différentes (au centime d'euro près), alors qu'en 2005 et 2006 ce nombre n'avait jamais été inférieur à 1391.

²² En faisant l'hypothèse que les interviewés donnent leur dépenses à plus ou moins 10% près, on peut imposer à ε de suivre une loi normale d'espérance nulle et d'écart type égal à 1/20. Ainsi dans 95% des cas,

$\varepsilon \in [-1,96 \quad 1,96]$ d'où $\varepsilon/20 \in [-10\% \quad 10\%]$.

²³ Dans le cas présent, le processus a été renouvelé 31 fois.

Les graphiques suivants montrent une simulation de cette nouvelle méthode jusqu'au deuxième trimestre 2007.

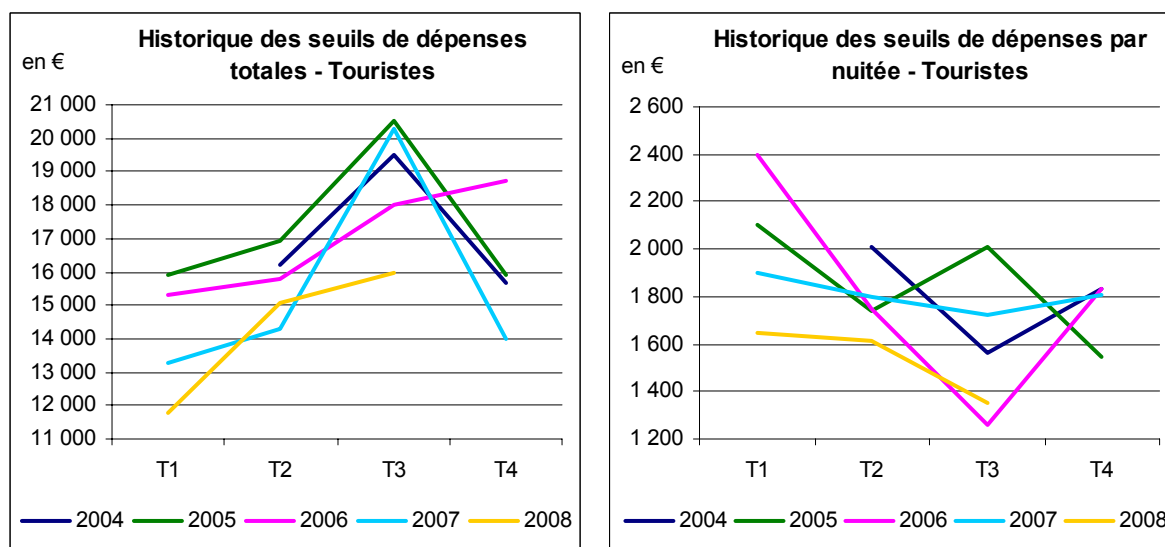


Ainsi cette nouvelle méthode a l'avantage d'être plus robuste. Les seuils sont moins volatils, et leur saisonnalité paraît ainsi plus forte. De plus, le seuil des dépenses par personne des excursionnistes du deuxième trimestre 2007 est plus élevé (620 € au lieu de 410 €). Certes les deux premiers seuils de 2007 sont plus bas que ceux des années précédentes, mais ils sont conformes au niveau historiquement bas de la distribution.

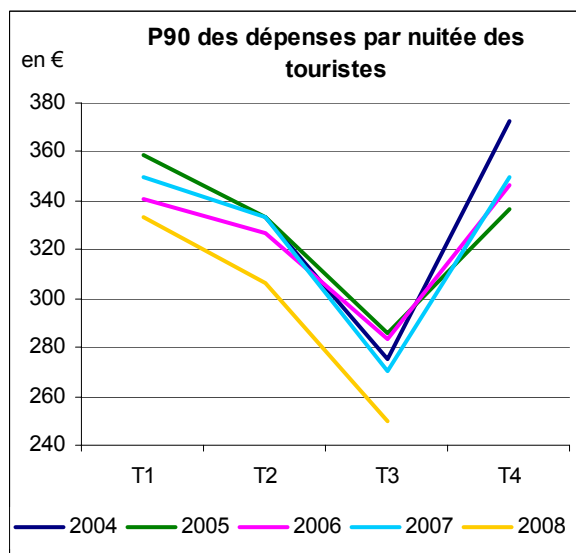
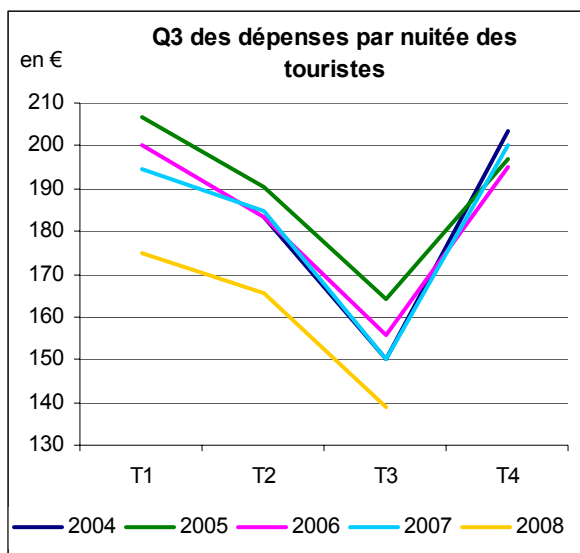
2.4. Résultats sur les dernières périodes

2.4.1. Seuils

Pour les touristes, la tendance de l'ensemble des seuils est à la baisse dès le dernier trimestre 2007.

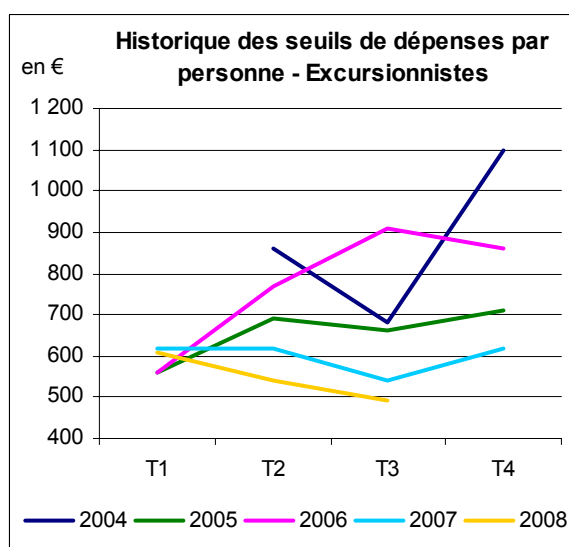
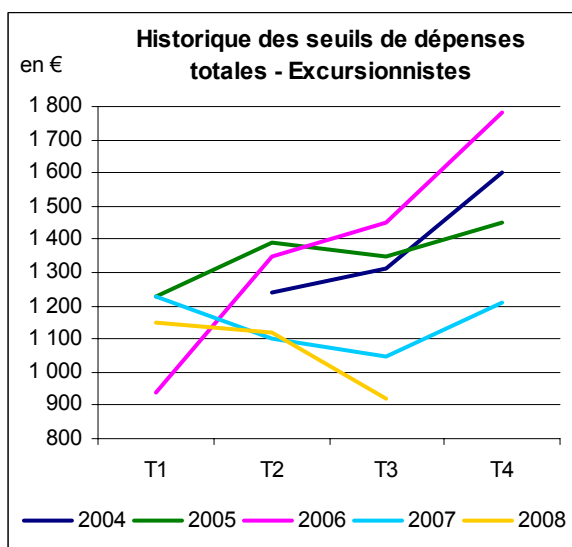


La baisse du seuil en termes de dépenses par nuitée des touristes suit la baisse du dernier quartile et dernier décile de la distribution de ces mêmes dépenses dès le premier trimestre 2008. Les dépenses par nuitée prenant moins fréquemment des valeurs extrêmes, il en ressort que le seuil calculé est plus faible, sans que le nombre de questionnaires considérés comme atypiques soit plus important (cf. infra).

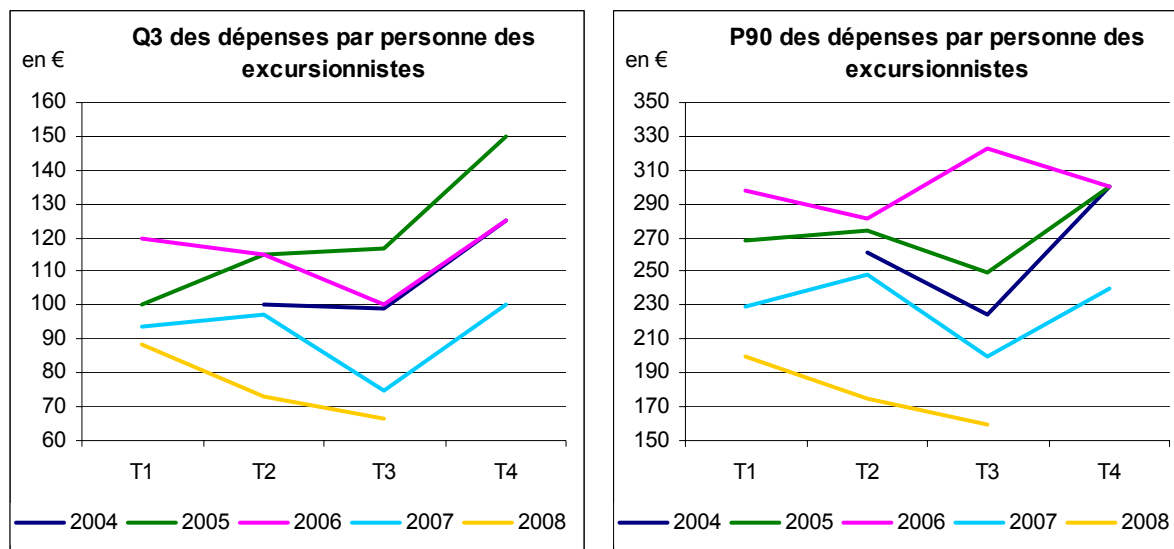


Champ : touristes - questionnaires aux dépenses non nulles.

Pour les excursionnistes, cette diminution des seuils a été entamée dès le deuxième trimestre 2007 en termes de dépenses totales, et dès le premier trimestre 2008 en termes de dépenses par personne.



Comme précédemment pour les seuils concernant les touristes, et comme l'illustrent les graphiques suivants, la baisse du seuil en termes de dépenses par personne pour les excursionnistes suit la baisse du troisième quartile et du dernier décile.



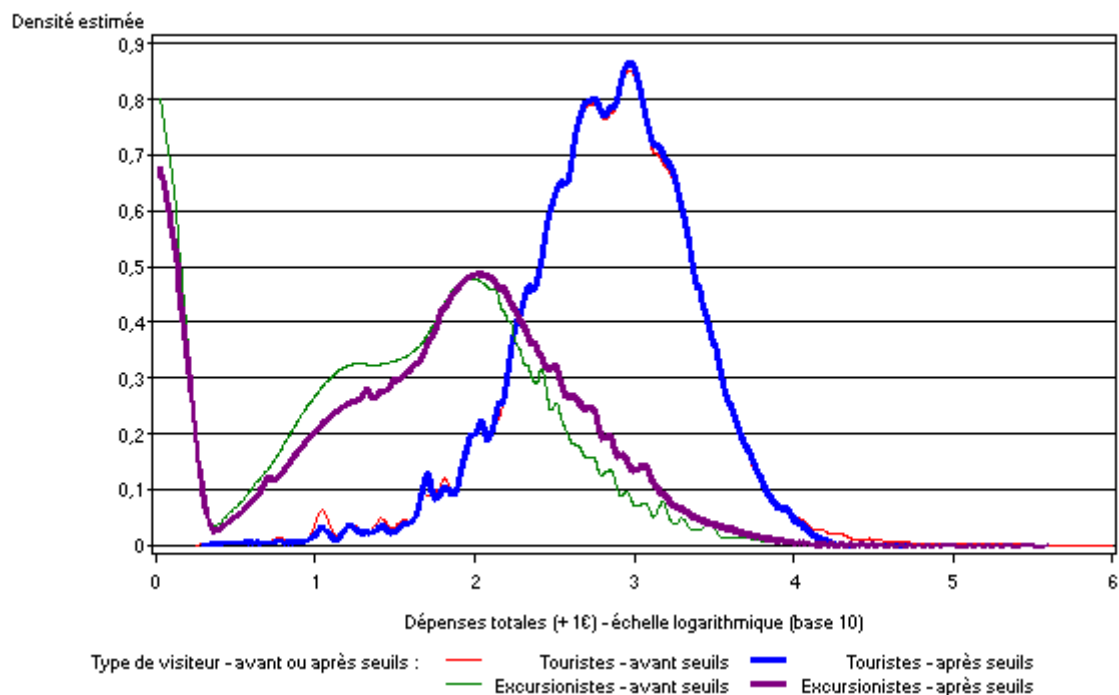
Champ : excursionnistes - questionnaires aux dépenses non nulles.

2.4.2. Impact de la ré-imputation des individus atypiques

Les questionnaires aux dépenses atypiques ne sont pas supprimés de la base de données. Ils sont détectés et imputés au même titre que l'ensemble des questionnaires présentant de la non-réponse, avec comme contrainte de ne pas dépasser les seuils fixés.

2.4.2.1. Sur les distributions

Densité des dépenses totales des visiteurs - avant et après seuils - période 2007T4-2008T3



2.4.2.2. Sur les dépenses totales

Type de visiteur	Trimestre	Dépenses totales extrapolées (en milliards d'euros)		écart (%)
		Avant seuils	Après seuils ²⁴	
Excursionnistes	2007T4	14,0	3,2	77,3
	2008T1	7,9	2,5	68,0
	2008T2	11,5	2,3	79,7
	2008T3	5,2	3,6	30,0
Touristes	2007T4	33,9	16,0	52,8
	2008T1	26,3	15,8	39,7
	2008T2	201,4	26,6	86,8
	2008T3	51,0	40,3	21,0
	Total	351,1	110,3	68,6

Note de lecture : pour le quatrième trimestre de 2007, pour les excursionnistes, l'extrapolation des dépenses totales avant calcul des seuils est de 14,0 milliards d'euros. Après ré-imputation des dépenses par TNS-Sofres en tenant compte des seuils, elle s'élève à 3,2 milliards d'euros. La baisse due à la ré-estimation des questionnaires dépassant l'un des seuils est donc de 77,3 %.

L'impact est particulièrement fort sur les dépenses des touristes du deuxième trimestre 2008. En effet, sur les quatre questionnaires les plus extrêmes pour ce trimestre-là, les dépenses déclarées sont particulièrement élevées : un questionnaire à 20 millions d'euros, correspondant à un achat immobilier, et un questionnaire à 100 millions d'euros correspondant à des réponses complètement fantaisistes. L'imputation de ces questionnaires conduit à fortement réduire les dépenses de ces questionnaires. Donc les dépenses extrapolées après calcul des seuils en ressortent nettement diminuées.

2.4.3. Qualité du calcul des seuils et respect des contraintes

2.4.3.1. Qualité des seuils

On s'intéresse dans un premier temps à la variabilité induite par le calcul des seuils en tant que tels. On observe donc la variabilité induite par le facteur aléatoire introduit dans le calcul des seuils. Pour cela, on considère 50 échantillons bootstrap²⁵, sur les trois premiers trimestres 2008, pour lesquels on calcule les différents seuils selon la méthode expliquée précédemment.

Les coefficients de variation varient peu selon les trimestres, en dehors du deuxième trimestre pour les excursionnistes, qui sont nettement plus élevés. La méthode de calcul des seuils est donc plutôt stable.

Seuils sur les dépenses totales

Type de visiteur	Trimestre	Nombre d'échantillons bootstrap	Moyenne	Ecart-type	Minimum	Maximum	CV
excursionnistes	2008T1	50	1 124	31	1 050	1 150	2,8 %
	2008T2	50	1 088	65	890	1 180	6,0 %
	2008T3	50	903	35	850	980	3,9 %
touristes	2008T1	50	12 301	448	11 200	13 300	3,6 %
	2008T2	50	14 209	478	13 500	15 400	3,4 %
	2008T3	50	15 085	868	13 300	16 500	5,8 %

Unité : euros.

²⁴ Il ne s'agit pas ici des résultats définitifs des recettes de la ligne voyages, puisque d'autres traitements statistiques, comme ceux mentionnés en début de document, sont ensuite appliqués aux données.

²⁵ Après tirage des 50 échantillons, sur chacun d'entre eux, les non-réponses sont imputées de manière indépendante.

Seuils sur les dépenses par nuitée ou par individu

Type de visiteur	Trimestre	Nombre d'échantillons bootstrap	Moyenne	Ecart-type	Minimum	Maximum	CV
excursionnistes	2008T1	50	602	23,3	570	680	3,9 %
	2008T2	50	542	46,6	460	580	8,6 %
	2008T3	50	509	23,0	449	550	4,5 %
touristes	2008T1	50	1 695	44,6	1 630	1 800	2,6 %
	2008T2	50	1 623	45,8	1 485	1 700	2,8 %
	2008T3	50	1 395	30,8	1 340	1 480	2,2 %

Unité : euros.

On s'intéresse également à l'impact des seuils sur les dépenses ré-extrapolées par TNS-Sofres. Pour les 50 échantillons bootstrap précédents, sur le premier trimestre 2008, les dépenses ont été à nouveau imputées par TNS-Sofres.

Si la variabilité des dépenses des touristes est assez stable sur le trimestre, ce n'est pas le cas des dépenses des excursionnistes. Cependant, les questionnaires des excursionnistes sont trois fois moins nombreux que ceux des touristes.

Dépenses totales du 1^{er} trimestre 2008, après ré-imputation par TNS-Sofres

	Nombre d'échantillons bootstrap	Moyenne	Ecart-type	Minimum	Maximum	CV
excursionnistes	50	2 761	325	2 217	4 352	11,8%
touristes	50	16 168	463	15 301	17 208	2,9%

Unité : millions d'euros.

Dépenses totales par individu du 1^{er} trimestre 2008, après ré-imputation par TNS-Sofres

	Nombre d'échantillons bootstrap	Moyenne	Ecart-type	Minimum	Maximum	CV
excursionnistes	50	1 476	182	1 161	2 026	12,3%
touristes	50	7 449	149	7 024	7 746	2,0%

Unité : millions d'euros.

2.4.3.2. Respect des contraintes et conclusions

Conformément aux objectifs énoncés, la méthode proposée permet la détection des questionnaires dont les dépenses sont exceptionnellement élevées. L'étude est bien réalisée par trimestre et type de visiteur. De plus, l'atypie est étudiée en termes absolu (seuil des dépenses totales) et relatif (seuil des dépenses par personne pour les excursionnistes et par nuitée pour les touristes).

L'écart entre le « dernier » questionnaire admis et le premier « refusé » en termes de dépenses totales n'est pas de l'ordre de quelques centimes d'euros, mais de plusieurs euros pour les excursionnistes et de plusieurs centaines d'euros pour les touristes, comme le montre le tableau suivant :

Type de visiteur	Trimestre	écart en euros
Excursionnistes	2007T4	12,31
	2008T1	61,07
	2008T2	40,90
	2008T3	22,95
Touristes	2007T4	218,00
	2008T1	311,99
	2008T2	121,24
	2008T3	303,35

Ainsi, on ne peut pas avoir le cas d'un questionnaire rejeté s'il est « proche » d'un questionnaire conservé.

Enfin, de par la formule détaillée en 2.2.1., les questionnaires qualifiés d'atypiques le restent après multiplication des dépenses totales par même nombre réel positif, ou par addition d'un même nombre réel.

Type de visiteur	Trimestre	Nombre de questionnaires	Nombre de questionnaires au-dessus du seuil en termes de dépenses totales	Nombre de questionnaires au-dessus du seuil en termes de dépenses par nuitée ou par personne	Nombre de questionnaires au-dessus d'un des seuils	Pourcentage de questionnaires au-dessus d'un des seuils
Excursionnistes	2007T4	4 077	59	80	88	2,2 %
	2008T1	3 662	69	58	80	2,2 %
	2008T2	3 541	60	79	85	2,4 %
	2008T3	4 021	95	90	108	2,7 %
Touristes	2007T4	11 631	104	80	153	1,3 %
	2008T1	11 599	108	106	169	1,5 %
	2008T2	11 918	71	82	126	1,1 %
	2008T3	12 645	82	69	124	1,0 %

Les avantages de la méthode couplée, c'est-à-dire de l'utilisation conjointe d'un seuil de dépenses totales et d'un seuil de dépenses par personne ou par nuitée, sont multiples. Comme annoncé dans les objectifs et contraintes de la méthode, les questionnaires aux fortes dépenses de façon « absolues » ou « relatives » sont détectés. De plus, la « lisibilité » des seuils est meilleure puisqu'ils sont plus stables d'un trimestre à l'autre (à saisonnalité comparable). Enfin, les deux approches sont complémentaires, l'une privilégiant plus les atypies des dépenses des touristes et l'autre des dépenses des excursionnistes.

Cependant, dans les estimations finales des dépenses, d'autres atypismes persistent : dépenses pondérées élevées, non homogénéité des strates, ... De ce fait, les différents estimateurs utilisés se doivent de corriger ces effets. A l'heure actuelle, s'ils s'avèrent satisfaisants en termes de dépenses, ils le sont moins en termes de nuitées. Pour cette raison ainsi que par souci de concision, ils ne sont pas présentés ici, mais le seront sûrement lors de prochaines JMS !

Annexe 1 : présentation sommaire de l'estimateur à noyau de Tukey (ou biweight)

L'estimateur à noyau de Tukey²⁶ est un estimateur non paramétrique de la densité d'une variable aléatoire continue. Méthode dérivée de l'histogramme « classique », la densité au point x est proportionnelle au nombre de points pondérés au voisinage de x . Le voisinage, appelé fenêtre, est l'intervalle $[x-h, x+h]$ où h représente la *largeur de fenêtre* (même si la largeur « véritable » de l'intervalle vaut $2 \cdot h$). La pondération est donnée par une fonction, appelée noyau, en fonction de la distance au point x .

Une estimation au point x de la densité sur la base des observations x_1, x_2, \dots, x_n notée $\hat{f}_n(x)$ (la « vraie » valeur de la densité au point x est $f(x)$) vérifie :

$$\hat{f}_n(x) = \frac{1}{n \cdot h} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)$$

Dans le cas d'un noyau de Tukey, le noyau est défini par la fonction :

$$K : u \rightarrow \begin{cases} \frac{15}{16}(1-u^2)^2 & \text{si } u \in [-1, 1] \\ u \rightarrow 0 & \text{si } u \notin [-1, 1] \end{cases}$$

M. Lejeune propose dans [1] la formule suivante pour le calcul de la fenêtre :

$$h = 0,75 \cdot \left(\frac{n}{100}\right)^{-1/5} \cdot \min_{i=1, \dots, \lfloor \frac{n}{2} \rfloor} \left(x_{\left(i + \lfloor \frac{n}{2} \rfloor\right)} - x_{(i)} \right)$$

où $x_{(i)}$ désigne le i -ème élément des n observations x_1, x_2, \dots, x_n triées dans l'ordre croissant

et $\lfloor y \rfloor$ la partie entière du nombre réel y .

☆ La présence dans la formule du terme en $n^{-1/5}$ peut s'expliquer par les faits suivants :

- ✓ La minimisation de la variance de la densité estimée à un point x donné, conduit à une

fenêtre optimale $h_{opt 1}$ suivante :
$$h_{opt 1} = \left[\frac{f(x) \cdot \int_{\mathbb{R}} K(u)^2 du}{f''(x)^2 \cdot \left[\int_{\mathbb{R}} u^2 K(u) du \right]^2} \right]^{1/5} \cdot n^{-1/5}$$
. Ainsi, il

y a deux termes : une composante dépendant uniquement de grandeurs issues de la densité f et du point x , mais qui n'est pas calculable spontanément, car f est inconnue. La seconde composante est $n^{-1/5}$.

- ✓ De même, la minimisation de l'erreur intégrée quadratique moyenne (et donc sans se placer à un point particulier comme au point précédent) donne une fenêtre optimale

²⁶ Cette annexe présente les points essentiels de l'estimation présentée dans l'ouvrage [1] référencé en bibliographie.

$h_{opt\ 2}$ à deux composantes. L'une « fonction de f » et l'autre de $n^{-1/5}$. $h_{opt\ 2}$ vérifie :

$$h_{opt\ 2} = \left[\frac{\int_{\mathbb{R}} K(u)^2 du}{\left(\int_{\mathbb{R}} f''(x)^2 du \right) \cdot \left[\int_{\mathbb{R}} u^2 K(u) du \right]^2} \right]^{1/5} \cdot n^{-1/5}.$$

☆L'autre partie de l'expression de la fenêtre est expliquée par M. Lejeune de la façon suivante « pour une diversité de lois et pour $n = 100$, la valeur optimale reste proche de 0,75 fois la largeur de l'intervalle de probabilité 0,5 autour du mode. Cette méthode est simple à mettre en œuvre et effectue généralement un lissage adéquat ». (cf. [1] page 190).

Annexe 2 : présentation rapide de quelques tests de détection d'individus atypiques.

Le test de Grubbs²⁷

Le test de Grubbs est utilisé par la Direction du Tourisme et TNS-Sofres pour la détection des durées de séjour atypiques dans l'enquête Suivi de la Demande Touristique (SDT). Il permet de détecter la présence d'une valeur aberrante en comparant chaque valeur extrême à la moyenne de la population. Réalisé itérativement jusqu'à ce qu'aucun individu atypique ne soit détecté, cela permet de détecter tous les individus atypiques.

Dans le cas précis de l'enquête SDT, la variable codant le nombre de nuitée est $nb\text{nuit}$. La transformation logarithmique $1 + \log(1 + nb\text{nuit})$ permet de « normaliser » la distribution de cette variable. Puis, un test est réalisé sur la statistique $Z = \frac{|\text{max} - \text{moyenne}|}{\text{écart} - \text{type}}$, afin de détecter si la valeur maximale peut être qualifiée d'atypique ou non.

Le test de Rosner²⁸

Le test de Rosner permet de détecter au plus 10 individus atypiques dans une population. De plus, après suppression de ces individus, la variable doit être distribuée selon une loi normale. Ce test revient à comparer chacun des individus atypiques suspectés à la moyenne de l'ensemble de la population sans les individus plus extrêmes que l'individu étudié.

Le test de Walsh²⁹

Le test de Walsh permet de détecter la présence d'un nombre quelconque d'individus atypiques. Il ne nécessite aucune condition sur les données (normalité). Il est basé sur la détection d'un grand écart entre deux valeurs consécutives.

Pour n observations triées dans l'ordre croissant : $x_{(1)} < x_{(2)} < \dots < x_{(n)}$, et $1 \leq r < n$ le test de Walsh teste l'hypothèse H_0 : les r observations $x_{(n-r+1)}, x_{(n-r+2)}, \dots, x_{(n)}$ sont issues de la même loi que les autres données, contre l'hypothèse alternative H_1 : les r observations $x_{(n-r+1)}, x_{(n-r+2)}, \dots, x_{(n)}$ ne sont pas issues de la même loi que les autres données.

Pour $60 < n \leq 220$, le risque de rejeter à tort H_0 (le risque de première espèce) est $\alpha = 10\%$.

Pour $220 < n$, le risque de rejeter à tort H_0 (le risque de première espèce) est $\alpha = 5\%$.

On calcule alors les données suivantes (avec $[y]$ la partie entière de y):

$$c = 1 + \lfloor \sqrt{2n} \rfloor, \quad b^2 = 1/\alpha, \quad a = \frac{1 + b \sqrt{\frac{c-b^2}{c-1}}}{c-b^2-1}$$

Les r observations $x_{(n-r+1)}, x_{(n-r+2)}, \dots, x_{(n)}$ sont alors déclarées atypiques (au risque α de les déclarer atypiques à tort) si l'inégalité suivante est vérifiée :

$$x_{(n-r+1)} - x_{(n-r)} > a \cdot (x_{(n-r)} - x_{(n-r+1-c)})$$

²⁷ Ce test est décrit dans une note d'Abdel Khiati référencée [2] dans la bibliographie.

²⁸ Ce test est décrit pages 17, 21 et 22 dans le document référencé [3] dans la bibliographie.

²⁹ Ce test est décrit pages 17 et 20 dans le document référencé [3] dans la bibliographie.

Glossaire

Liste et description des principaux termes « spécifiques » au domaine étudié.

Clé de répartition touristes/excursionnistes :

Désigne l'ensemble des ratios permettant de déterminer les proportions de touristes et d'excursionnistes à partir du flux de visiteurs. Cette ventilation est essentiellement fonction de la nationalité et du mode de transport utilisé pour quitter la France. Pour le mode « route », elle est difficilement observable car les excursionnistes empruntent moins facilement les grands axes que les touristes, et ne s'arrêtent pas non plus obligatoirement sur des aires d'autoroutes. Pour les autres modes, elle est directement obtenue à partir des questionnaires.

Enquête aux Frontières (EAF) :

Ancienne enquête mise en place par la Direction du Tourisme (DT) au fil des années. Cette enquête était destinée à observer le tourisme des visiteurs non-résidents en France. Cette opération de grande envergure était menée aux points de sortie du territoire, ce qui permettait d'interroger les non-résidents au moment où ils quittaient le territoire métropolitain : ports, aéroports, gares ferroviaires ou routières, shuttle, frontières routières et autoroutières. Elle a été réalisée en 1963, 1976, 1982, 1989, 1991, 1993-1994, 1996, enfin en 2000-2001. Entre deux enquêtes aux frontières, la DT estimait le nombre de touristes non-résidents à l'aide de régressions économétriques reposant notamment sur les recettes de la ligne « voyages » et les résultats des enquêtes de fréquentation hôtelière.

Enquête auprès des Voyageurs Étrangers (EVE) :

Enquête trimestrielle qui a remplacé les EAF en 2004. Elle est réalisée en partenariat par la Direction du Tourisme et la Banque de France. Les non-résidents sont enquêtés par questionnaires auto-administrés lors de leur sortie du territoire (dans des salles d'embarquement, aires d'autoroutes, trains, ...).

Excursionniste :

Non-résident en visite en France, arrivant sur le territoire français et repartant le même jour.

Ligne « Voyages » de la Balance des paiements :

Poste de la Balance des paiements qui recense durant une période donnée :

- en débit, les dépenses touristiques des résidents français à l'étranger ;
- en crédit, les dépenses touristiques des non-résidents lors de leur séjour en France.

Le champ précis est défini dans le 5^{ème} manuel du FMI relatif à la Balance des Paiements.

Nuitée :

Désigne le nombre de nuits « individuelles » passées en France. Ainsi, un couple de non-résidents avec deux enfants visitant la France et y séjournant trois nuits, aura un nombre de nuitées égal à 12 (4 personnes * 3 nuits).

Touriste :

Non-résident en visite en France et y demeurant au minimum une nuit.

Visiteur :

Non-résident en visite en France, qu'il soit touriste ou excursionniste.

Bibliographie

[1] Lejeune M., *Statistique - La théorie et ses applications*, Springer, chap. 8, p. 184-190, 2004.

[2] Abdel Khiati (Direction du Tourisme) et TNS-Sofres, « Données atypiques sur la durée de séjour – deuxième bilan de nos travaux », document de travail, septembre 2004.

[3] Swdiv et Efa West, *Handbook for statistical analysis of environmental background data*, Naval Facilities Engineering Command, https://portal.navfac.navy.mil/portal/page/portal/navfac/navfac_ww_pp/navfac_nfesc_pp/environmental/erb/lab/hndbk-sw.pdf, Box 2.6-2.11, p. 17- 23, July 1999.

[4] Deheuvels P (1977) Estimation Non Paramétrique de la Densité par histogrammes généralisés. *Revue de Statistique Appliquée* vol. XV : 5-42
http://archive.numdam.org/ARCHIVE/RSA/RSA_1977_25_3/RSA_1977_25_3_5_0/RSA_1977_25_3_5_0.pdf