



Précisions et limites des enquêtes par sondage



le cas des Enquêtes Ménages Déplacements

SOMMAIRE

SOMMAIRE	3
1 > Les erreurs présentes dans une EMD	4
1.1. Les erreurs de mesure.....	4
1.1.1. Les erreurs sur l'échantillon :.....	4
1.1.1.1. Erreurs lors du tirage de l'échantillon.....	4
1.1.1.2. Erreurs liées au non respect des ordres de tirage	5
1.1.1.3. Erreurs liées à la représentativité de l'échantillon final	5
1.1.2. Erreurs des personnes enquêtées.....	6
1.1.2.1. Erreurs de mémoire.....	6
1.1.2.2. Erreurs d'appréciation	6
1.1.3. Erreurs de traitement.....	7
1.1.3.1. Erreurs de l'enquêteur.....	7
1.1.3.2. Erreurs de saisie	7
1.1.3.3. Erreurs d'analyse.....	7
1.2. Les erreurs statistiques.....	8
1.2.1. Les notions de seuil de confiance et de marge d'erreur	8
1.2.2. Estimation de la marge d'erreur	9
1.2.2.1. Estimation de l'erreur sur une proportion.....	9
1.2.2.2. Estimation de l'erreur sur une moyenne	12
1.2.3. Sondage stratifié	13
1.2.4. Effet de grappe	14
2 > Approche des limites d'exploitation d'une enquête ménages déplacements	17
2.1. Les impératifs.....	17
2.1.1. Respecter l'échelon du tirage aléatoire.....	17
2.1.2. Pas d'analyse dans une matrice pour moins de 10 déplacements enquêtés.....	17
2.1.3. Analyse en regroupement des questions pour lesquelles l'erreur de mémoire du ménage enquêté est forte.....	18
2.1.4. Une méthode d'analyse fiable.....	18
2.2. Les préconisations	19

1 > Les erreurs présentes dans une EMD

Une enquête ménage est un sondage. On interroge une partie de la population puis on extrapole le résultat obtenu à l'ensemble de la population. Ce type d'enquête par sondage présente donc un certain nombre d'erreurs dans les résultats qu'elle produit. Ces erreurs sont de deux types : les erreurs "statistiques" (liées au principe du sondage par échantillon et qui sont quantifiables grâce à des lois mathématiques) et les erreurs de "mesure" (erreurs, souvent liées à l'humain et qui interviennent aux différents stades de l'enquête).

1.1. Les erreurs de mesure

Comme dans toutes les enquêtes, les enquêtes ménages déplacements sont sujettes à un certain nombre d'erreurs de "mesure". Ces erreurs sont principalement des erreurs liées directement à l'activité humaine dans l'enquête. Elles peuvent intervenir à différents stades depuis le tirage de l'échantillon jusqu'à l'exploitation finale des résultats. Pour la plupart d'entre elles, elles ne sont pas quantifiables, c'est la rigueur tout au long de l'enquête qui permet d'en limiter les répercussions.

Nous allons en énoncer un certain nombre ci-dessous en essayant pour chacune de faire le point sur les méthodes utilisées pour en limiter l'impact et en estimant aussi l'erreur finale qu'elles peuvent induire sur les résultats.

1.1.1. Les erreurs sur l'échantillon :

1.1.1.1. Erreurs lors du tirage de l'échantillon

Pour qu'un sondage soit représentatif, l'échantillon interrogé doit être choisi de manière totalement aléatoire dans l'ensemble de la population que l'on cherche à caractériser. Or dans la réalité cette étape n'est pas toujours facile à réaliser.

Dans le cas d'une enquête ménages, l'échantillon est un ensemble de ménages. Pour le tirer au sort il faut donc pouvoir disposer d'un fichier dans lequel on retrouve l'ensemble des ménages résidents sur le secteur à enquêter. Or il n'existe plus de source de ce type disponible en France. Il faut donc trouver un autre fichier. Deux possibilités sont offertes :

- soit on tire les ménages dans la liste des abonnés au téléphone (comme pour l'enquête de Lyon en 2006), mais comme cette liste ne représente en aucun cas l'ensemble des ménages (à ce jour moins de 60% des ménages apparaissent dans ce fichier dans les grandes agglomérations) on choisit au final d'enquêter le voisin de gauche de chaque ménage tiré au sort;
- soit on utilise le fichier des propriétés bâties de la direction générale des impôts qui lui recense tous les logements, quelque soit leur vocation actuelle (vacant, résidence principale ou résidence secondaire). C'est ce fichier qui a été choisi pour les enquêtes Grenoble et Saint-Etienne.

Il existe donc une erreur dès le tirage de l'échantillon. En effet, comme le fichier choisi recense des résidences principales, des résidences secondaires mais aussi des logements vacants, il ne représente pas les ménages résidants dans le secteur. Le tirage aléatoire choisira forcément des résidences secondaires ou des logements vacants que l'on ne veut pas enquêter.

Par exemple, si dans un secteur on veut enquêter 100 ménages, pour respecter au mieux les règles de la statistique et limiter l'erreur finale aux simples marges que ces règles énoncent, il faudrait tirer au hasard 100 ménages et enquêter ces ménages. Or le tirage dans le fichier des impôts fait que sur les 100 logements tirés 10 seront peut être inoccupés et 5 seront des résidences secondaires. Pour enquêter réellement 100 ménages, il faudra tirer au sort plus de logements (au moins le double). Au final, l'échantillon n'est donc plus vraiment aléatoire puisque l'on tire en fait 100 ménages dans une liste aléatoire de 200 logements.

Pour rendre quasiment imperceptible cette erreur sur le résultat final, il existe une méthode. En fait le tirage au sort est effectué sur 3 à 4 fois l'échantillon souhaité, mais à chaque logement tiré au sort, on affecte un numéro d'ordre croissant. Le premier logement est donc numéroté 1 puis ainsi de suite. Lors de l'enquête on interroge ensuite les vrais ménages (résidences principales) issus de cette liste de logements en les prenant dans l'ordre de leur numéro. Cette pratique permet de se rapprocher fortement d'un échantillon aléatoire. Les erreurs finales liées au tirage de l'échantillon peuvent donc être considérées comme étant non significatives.

1.1.1.2. Erreurs liées au non respect des ordres de tirage

Un autre problème qui se pose lors des enquêtes ménages est que l'on ne peut pas obliger les ménages à y répondre. Or, pour respecter le critère aléatoire de l'échantillon, il faudrait absolument interroger les X premiers ménages (où X est la taille de l'échantillon souhaitée). Dans la réalité, il y a toujours un certain nombre de ménages qui ne souhaitent ou ne peuvent pas participer. Il faut alors se décaler dans l'échantillon et enquêter des ménages qui au départ ne faisaient pas parti des X premiers ménages de la liste tirée au sort.

Au final, l'échantillon enquêté n'est donc plus le vrai échantillon aléatoire.

Pour limiter le biais lié à ce phénomène, la seule solution consiste à faire tous les efforts possibles pour ne pas trop se décaler dans l'échantillon. Il faut donc insister lourdement auprès des ménages. L'erreur résultante est difficile à quantifier, mais si tous les efforts sont faits pour ne pas trop décaler l'échantillon, on estime qu'elles sont négligeables face aux erreurs statistiques induites par le principe du sondage.

1.1.1.3. Erreurs liées à la représentativité de l'échantillon final

Un autre problème lié aux EMD tient au mode de recueil. L'enquête de ménage est réalisée par une interview au domicile du ménage en présence de tous les membres

du ménage tiré au sort. De ce principe naît un autre biais : il est beaucoup plus difficile d'enquêter des gros ménages que des petits. En effet il est logiquement plus difficile de réunir 5 personnes à la même heure dans la même pièce qu'une seule. La répercussion de ce phénomène est simple : dans l'échantillon effectivement enquêté, les petits ménages sont surreprésentés et les gros ménages sont sous-représentés. Pour corriger ce problème, le redressement (étape qui consiste à définir un coefficient multiplicateur pour calculer les caractéristiques de l'ensemble des ménages sur la base des réponses des ménages enquêtés) est calculé en tenant compte dans chaque secteur de la répartition des ménages par taille. Ainsi les réponses d'un ménage d'une personne compteront pour 100 ménages (coefficient de redressement de 100), alors que les réponses d'un ménage de 5 personnes compteront pour 130. Grâce à ce redressement on estompe les erreurs liées à ce phénomène.

1.1.2. Erreurs des personnes enquêtées

D'autres types d'erreur des enquêtes ménages sont directement liés aux personnes enquêtées. C'est le cas de tous les sondages, mais dans le cas d'une EMD ou le questionnaire est particulièrement volumineux et complexe, elles peuvent être plus nombreuses. On peut les classer en 2 catégories :

1.1.2.1. Erreurs de mémoire

Une personne enquêtée peut ne pas se souvenir exactement de ses déplacements effectués la veille. Il est fréquent que certains soient oubliés. Pour palier à ces erreurs il faut compter sur le travail de l'enquêteur, mais aussi sur le travail des gestionnaires qui relisent les questionnaires. Grâce à ces contrôles on peut corriger la majorité de ces erreurs. Il suffit pour les déplacements de suivre leur enchaînement. Si une personne est partie d'un point A pour se rendre au point B le matin. Puis qu'elle vous dit ensuite qu'elle s'est rendue d'un point C au point A le soir, elle a visiblement omis le déplacement de B vers C. Ce type d'erreur est facilement corrigé au cours de l'enquête. Certaines erreurs de ce type peuvent néanmoins subsister, mais si le travail des enquêteurs et de leur gestionnaire est bien fait, elles doivent être non significatives dans le résultat final.

1.1.2.2. Erreurs d'appréciation

D'autres erreurs peuvent aussi être produites par une difficulté d'appréciation des personnes enquêtées. Ces erreurs ne concernent que certaines questions. Ces questions sont en générales connues. On peut, par exemple, citer les questions liées au temps et aux heures. En effet si l'on est capable de dire avec certitude que l'on est allé hier matin d'un point A vers un point B et que l'on a utilisé sa voiture, il devient plus difficile de dire précisément son heure de départ et son heure d'arrivée et donc d'estimer le temps du déplacement. En général les personnes donneront des tendances ("je suis partie vers 8h30"). Les erreurs finales sur ce type de question sont donc importantes. On peut réaliser des contrôles afin de supprimer les incohérences mais l'approximation restera grande. La seule solution pour limiter l'impact de ces

erreurs est donc de ne réaliser des traitements qu'en utilisant des plages de temps ou d'horaires suffisamment importantes.

1.1.3. Erreurs de traitement

Les erreurs de traitement sont directement liées au déroulé de l'enquête. Elles peuvent apparaître à différents moments.

1.1.3.1. Erreurs de l'enquêteur

Il s'agit d'erreurs de deux types :

- des erreurs liées à une mauvaise application de la méthodologie qui aboutiraient par exemple à ne pas comptabiliser correctement certains déplacements (les déplacements à pied souvent). Grâce au suivi des gestionnaires et au contrôle du Cete lors de l'enquête, ces erreurs doivent très vite disparaître et ne doivent pas se retrouver dans le fichier final.
- des erreurs de codage lors du remplissage du questionnaire. En effet lorsque l'enquêteur remplit les questionnaires il peut faire des erreurs en renseignant les codes. Par exemple il peut vouloir coder un motif à l'origine en "domicile" (normalement codé "01") et écrire "11" sur la fiche, ce qui correspond à un motif travail. Là aussi ce type d'erreur doit être repéré par le gestionnaire qui relit l'intégralité des questionnaires. Elles peuvent être corrigées grâce au chaînage des déplacements. On identifie aussi ces erreurs plus tard en réalisant des tests dit d'apurement sur les fichiers informatiques. Ces tests visent à identifier les incohérences entre les différentes questions afin de supprimer toutes les erreurs.

1.1.3.2. Erreurs de saisie

Il s'agit ici d'erreurs qui peuvent intervenir lors de la mise sous informatique des fiches d'enquête. Elles peuvent être liées à une mauvaise lecture de la machine automatique, à une mauvaise action de l'opérateur, ou encore être liées à une écriture trop peu lisible. Là aussi ces erreurs sont, pour la grande majorité d'entre elles, corrigées grâce au test d'apurement.

1.1.3.3. Erreurs d'analyse

Cette fois ci, il s'agit d'erreurs liées à l'opérateur final, celui qui traite l'enquête. Vu la complexité des fichiers, il est, par exemple, tout à fait possible que lors d'un tri croisé une partie des déplacements soit oubliée. Dans ce cas tous les pourcentages calculés seraient alors faux.

Ce type d'erreur est le plus facile à corriger. Il suffit que l'opérateur soit rigoureux et qu'il mette en place des procédures de vérification et de validation de tous ses résultats. Malheureusement ce n'est pas toujours le cas (par manque de temps) et il faut savoir que ce type d'erreurs est souvent bien plus impactant que toutes les erreurs listées ci-dessus mais aussi que celles qui vont suivre.

1.2. Les erreurs statistiques

Le principe du sondage, qui consiste à interroger une partie d'une population afin de connaître les caractéristiques recherchées de l'ensemble de la population, est possible grâce à la loi statistique dite des "grands nombres". Cette loi nous explique que : si l'échantillon sondé est choisi de manière aléatoire et est au moins de 30 individus, alors, la variable analysée, sera représentative de la valeur réelle sur l'ensemble de la population, avec une certaine marge d'erreur que l'on peut calculer. Plus la valeur de la variable calculée est forte et l'échantillon important, plus cette marge d'erreur est faible (elle peut même devenir non significative). A l'inverse, une valeur faible sur un petit échantillon implique une forte marge d'erreur qui rend alors la valeur calculée inexploitable.

Les résultats sur l'ensemble de la population sont donc fournis avec une certaine erreur qui est exprimée par 2 termes : le seuil de confiance et la marge d'erreur.

1.2.1. Les notions de seuil de confiance et de marge d'erreur

Ces deux notions sont essentielles pour qualifier les résultats obtenus. Elles n'ont de sens que si elles sont données ensemble. Par abus de langage, on ne parle souvent que de marge d'erreur, or une marge d'erreur n'est en fait exprimée que de paire avec un seuil de confiance.

Le seuil de confiance détermine en pourcentage la probabilité que le résultat fournit soit donné avec une certaine marge d'erreur. Dans la pratique on utilise 3 seuils différents : 80%, 95% ou 99%.

La marge d'erreur, qui peut être aussi appelée précision, erreur relative, ou intervalle de confiance, est exprimée soit en pourcentage soit en nombre et définit donc l'erreur à un certain seuil de confiance. Cette marge d'erreur est calculée grâce à la loi des grands nombres. Elle dépend de plusieurs critères : le seuil de confiance choisi, le type de la variable sur laquelle on travaille, la valeur de cette variable (calculée après redressement), le nombre d'individus effectivement enquêtés. Plus le seuil de confiance augmente, plus la marge d'erreur augmente. Elle augmente aussi quand la valeur calculée et/ou le nombre d'individus enquêtés diminue. A l'inverse, la marge d'erreur diminue quand la valeur calculée et/ou le nombre d'individus enquêtés augmente.

A titre d'exemple, dans une enquête ménages où la part de marché de la voiture sur l'ensemble de l'enquête est de 60%, on devrait dire : au seuil de confiance de 95%, la part de marché de la voiture sur l'ensemble de l'enquête est de 60% à X% près (où X est alors une erreur relative ou une précision). Cela signifie que la part de marché de la voiture a 95% de chance d'être comprise entre $60 - (60 * X\%)$ et $60 + (60 * X\%)$. Pour simplifier l'explication, nous partons du principe que $X = 3\%$, ceci n'est pas une valeur universelle puisque X varie pour chaque enquête et pour chaque variable de l'enquête.

On dira alors que la part de marché de la voiture sur l'ensemble de l'enquête a 95% de chance d'être comprise entre 58,2% et 61,8% (quand on exprime ces deux chiffres, on parle alors de l'intervalle de confiance).

Si, pour la même variable, on choisit d'exprimer le résultat au seuil de confiance de 80%, la valeur de X serait alors plus petite. On dirait alors par exemple que la part de marché de la voiture à 80% de chance d'être comprise entre 59% et 61%.

A l'inverse, si on souhaite exprimer les résultats avec un seuil de confiance de 99%, la valeur de X sera plus grande. On dirait alors que la part de marché de la voiture a 99% de chance d'être comprise entre 57,7% et 62,3%.

De manière générale et pour tout ce qui va suivre nous utiliserons un seuil de confiance de 95%.

1.2.2. Estimation de la marge d'erreur

Pour pouvoir estimer l'erreur il faut d'abord répartir les variables en 2 classes :

- **les variables de proportion** : elles sont bâties sur le principe d'un rapport entre un nombre d'individus (au sens statistique) sur l'ensemble des individus. Dans le cas d'une enquête ménages ce sera par exemple le pourcentage de ménages ne possédant pas de voiture, le pourcentage de ménages comprenant 2 personnes, le pourcentage de personnes s'étant déplacé la veille de l'enquête, le pourcentage de déplacements par mode ou par motif, le pourcentage de déplacements émis par une zone, etc. On considérera aussi comme une proportion un rapport construit sur une partie seulement de l'ensemble des individus statistiques comme par exemple le pourcentage de personnes ne disposant pas de voiture et ayant utilisé le transport collectif la veille.

- **Les variables de moyenne** : de manière simple elles englobent tout ce que l'on ne peut pas classer dans les variables de proportion : taux de motorisation (nombre de VP/nombre de ménages), la mobilité par personne (nombre de déplacements/nombre de personnes), la taille moyenne des ménages (nombre de personnes/nombre de ménages), etc.

1.2.2.1. Estimation de l'erreur sur une proportion

Les erreurs sur les proportions sont les plus faciles à calculer. Grâce à la loi des grands nombres on peut définir les formules suivantes pour lesquelles "p" est la proportion calculée (0,1 pour 10% par exemple), "n" est le nombre d'individus statistiques enquêtés et "ε" est l'erreur relative (0,1 pour 10% par exemple). On peut ainsi calculer :

L'intervalle de confiance au seuil de confiance de 95% :

$$\text{intervalle de confiance au seuil de confiance de 95\%} = p \pm 1,96 \sqrt{\frac{p(1-p)}{n}}$$

L'erreur relative au seuil de confiance de 95% :

$$\varepsilon = 1,96 \sqrt{\frac{1-p}{np}}$$

On peut directement connaître le nombre de personnes à enquêter à partir d'une erreur relative définie et d'une proportion estimée :

$$n \geq \frac{1,96^2(1-p)}{\varepsilon^2 p}$$

L'ensemble de ces formules nous permet de réaliser le tableau ci-dessous. Grâce à ce tableau on peut obtenir la relation entre le pourcentage calculé par l'enquête et son erreur relative en fonction d'un nombre d'individus enquêtés. On peut aussi retourner le tableau pour savoir le nombre d'individus que l'on doit enquêter si l'on souhaite obtenir une erreur définie pour une proportion estimée.

Tableau 1 : relation entre proportion, erreur relative et nombre d'individus à enquêter au seuil de confiance de 95%

Proportion mesurée \ Erreur relative	Erreur relative							
	1%	5%	10%	15%	20%	25%	30%	50%
1%	3 803 184	152 127	38 032	16 903	9 508	6 085	4 226	1 521
5%	729 904	29 196	7 299	3 244	1 825	1 168	811	292
10%	345 744	13 830	3 457	1 537	864	553	384	138
15%	217 691	8 708	2 177	968	544	348	242	87
20%	153 664	6 147	1 537	683	384	246	171	61
25%	115 248	4 610	1 152	512	288	184	128	46
30%	89 637	3 585	896	398	224	143	100	36
35%	71 344	2 854	713	317	178	114	79	29
40%	57 624	2 305	576	256	144	92	64	23
45%	46 953	1 878	470	209	117	75	52	
50%	38 416	1 537	384	171	96	61	43	
55%	31 431	1 257	314	140	79	50	35	
60%	25 611	1 024	256	114	64	41	28	
65%	20 686	827	207	92	52	33		
70%	16 464	659	165	73	41			
75%	12 805	512	128	57				
80%	9 604	384	96	43				
85%	6 779	271	68					
90%	4 268	171	43					
100%								

A titre d'exemple, dans une enquête ménages déplacements, on pourrait lire le tableau de cette manière : si sur un ensemble de secteurs de tirage on obtient que 70% des ménages a une voiture et si sur l'ensemble de ces secteurs, 659 ménages ont été enquêtés, alors au seuil de confiance de 95%, l'erreur relative sur mon résultat est

de 5%. Autrement dit, dans ces conditions, il y a 95% de chance que le pourcentage de ménages ayant une voiture soit compris entre 67% et 74%.

Ce tableau nous montre bien que plus la proportion mesurée va être faible, plus l'erreur va augmenter ou plus il va falloir enquêter d'individus. Si l'on reste dans l'exemple ci-dessus, mais que l'on veut exprimer cette fois que 30% des ménages n'ont pas de voiture. On a toujours enquêté 659 ménages et d'après le tableau on voit donc que cette fois l'erreur est comprise entre 10 et 15%. On saura donc qu'il y a 95% de chance que le pourcentage de ménage n'ayant pas de voiture soit compris entre 27% et 34%.

Le cas des matrices

Les matrices origine/destination sont des sorties fréquemment utilisées dans les EMD. Dans ces dernières, on n'exprime pas des proportions, mais des nombres de déplacements. Pourtant, techniquement parlant, les matrices sont bien à considérer comme des variables de proportion (on ramène alors le nombre de déplacements de chaque OD sur le nombre total de déplacements de la matrice) et c'est donc bien les règles ci-dessus qu'il faut appliquer. Néanmoins, pour les tous petits nombres les proportions sont tellement faibles que la loi des grands nombres ne peut plus s'appliquer. On utilise alors les règles suivantes :

- Si le nombre d'individus statistiques est inférieur à 10, alors l'erreur est tellement importante que toute exploitation est impossible
- Si le nombre total d'individus statistiques émis par une zone (vers toutes les autres zones) est inférieur ou égale à 1% de la somme des individus statistiques de la matrice, on utilise alors le tableau suivant:

Tableau 2 : précision et intervalle de confiance au seuil de confiance de 95% pour les matrices OD si le total des déplacements émis par une zone est inférieur à 1% du total des déplacements de la matrice.

Nombre de déplacements enquêtés depuis la zone l (en ligne) vers la zone c (en colonne)	10	12	15	20	30	40	50	80	100	200	300	400	500
intervalle de confiance	4 à 16	5 à 19	7 à 23	11 à 29	19 à 41	27 à 53	36 à 64	62 à 98	80 à 120	172 à 228	264 à 336	360 à 440	455 à 545
marge d'erreur	63%	58%	52%	45%	37%	32%	28%	22%	20%	14%	12%	10%	9%

- Si le nombre total d'individus statistiques émis par une zone (vers toutes les autres zones) est supérieur à 1% de la somme des individus statistiques de la matrice, on utilise alors les mêmes tableaux et règles que pour les proportions.

1.2.2.2. Estimation de l'erreur sur une moyenne

L'estimation de l'erreur sur une moyenne est beaucoup plus difficile à calculer. Pour pouvoir réaliser ces calculs, il faudrait connaître le rapport $\frac{\sigma}{M}$ où σ est l'écart type de la variable sur l'ensemble de la population et M la moyenne de cette variable calculée aussi sur l'ensemble de la population. Il est donc évident que, dans le cas d'une enquête ménages, ces deux valeurs sont impossibles à calculer. Cependant les études du CERTU ont démontré que d'une EMD à l'autre on observe une relative stabilité de la valeur $\frac{\sigma}{M}$ de chaque variable. Grâce aux valeurs de $\frac{\sigma}{M}$ fournies par le CERTU et à la formule :

$$n \geq \frac{1.96^2}{\epsilon^2} \left(\frac{\sigma}{m} \right)^2$$

où "n" est le nombre d'individus enquêtés, "ε" est l'erreur relative et $\frac{\sigma}{m}$ l'estimation de $\frac{\sigma}{M}$ donnée par le CERTU, on peut bâtir le tableau suivant :

Tableau 3 : relation entre l'erreur relative et le nombre de ménages enquêtés pour les variables de moyenne au seuil de confiance de 95%

Erreur relative \ Variable	1%	5%	10%	20%	30%	50%	Unité
Taille moyenne des ménages	13 830	553	138	35	15	6	Ménage
Nombre d'actifs par ménage	27 756	1 110	278	69	31	11	Ménage
Nombre de VP par ménage	38 416	1 537	384	96	43	15	Ménage
Mobilité par ménage tous modes tous motis	24 586	983	246	61	27	10	Ménage

Les estimations d'erreur sur les moyennes pour d'autres variables basées sur les personnes ou les déplacements sont possibles. Pour ne pas compliquer l'explication nous ne les fournissons que dans le paragraphe 1.2.4 à l'issue de la prise en compte des nouveaux paramètres développés dans ce dernier.

1.2.3. Sondage stratifié

Les formules mentionnées ci-dessus valent pour un sondage basé sur un échantillon aléatoire unique. Dans certains cas, on peut stratifier l'échantillon afin d'augmenter la précision du résultat. Cette stratification consiste à découper le fichier en plusieurs groupes d'individus pour ensuite réaliser autant de tirages aléatoires qu'il y a de groupes. Cette démarche permet par la suite d'avoir des résultats exploitables sur ces sous groupes. On pourrait par exemple prendre le cas d'un sondage sur les habitudes de consommation. Au final on souhaitera faire des exploitations des résultats en fonction des catégories sociales des personnes enquêtées. Si on réalise un tirage aléatoire sur l'ensemble de la population alors les résultats obtenus par catégorie sociale auront une erreur accentuée. A l'inverse on peut choisir de classer d'abord son fichier de population, puis d'effectuer ensuite un tirage aléatoire dans chaque groupe (cadres, ouvriers, etc.). On obtiendra ainsi des résultats pour chaque catégorie sur lesquels on peut appliquer les estimations d'erreurs définis aux paragraphes 1.2.2.1 et 1.2.2.2.

Dans le cas d'une enquête ménage, on souhaite obtenir des informations au niveau de l'ensemble du périmètre enquêté, mais on souhaite aussi pouvoir réaliser des exploitations sur des plus petits périmètres (une communauté de communes, un secteur de l'agglomération). **On choisit donc de stratifier l'échantillon par secteurs géographiques que l'on appelle secteurs de tirage.** C'est à l'intérieur de chacun de ces secteurs que l'on effectue donc le tirage aléatoire des ménages. On obtient ainsi autant de sondages qu'il y a de secteurs de tirage et des résultats exploitables et avec une marge d'erreur calculable pour chaque secteur de tirage. La question qui se pose alors est de savoir si, en agrégeant tous ces sondages, on obtient des réponses valables statistiquement pour l'ensemble de l'enquête. Les travaux menés par le CERTU à ce jour permettent d'avoir une opinion assez fiable sur ce point. Après plusieurs tests réalisés, il n'existe, à priori, pas d'incidence significative de cette stratification. Les chiffres produits par la somme des résultats de plusieurs secteurs de tirage sont totalement similaires à des chiffres qui auraient été calculés sur un échantillon unique et surtout les marges d'erreurs restent identiques (à 3 chiffres après la virgule près).

Le découpage en secteurs de tirage n'est donc pas considéré comme une vraie stratification. Son incidence sur les erreurs finales est nulle. On peut donc réaliser des exploitations sur des ensembles "variables" de secteurs de tirage. Attention toutefois, pour que les lois de la statistique continuent à s'appliquer, on ne devrait en principe utiliser que des sommes de secteurs de tirages dans leur ensemble.

1.2.4. Effet de grappe

L'effet de grappe est un autre paramètre à prendre en compte, surtout dans les enquêtes ménages.

Comme nous l'avons vu depuis le début de ce chapitre, l'ensemble des lois qui nous permettent d'estimer les erreurs est lié à un contexte de base : **l'échantillon doit être aléatoire**. Dans les enquêtes ménages déplacements, on effectue donc un tirage de ménages. Hormis les erreurs de mesures citées au paragraphe 1.1.1, cet échantillon doit donc bien être considéré comme aléatoire et, pour tous les résultats de proportion et de moyenne sur les ménages, on peut donc appliquer l'ensemble des formules citées aux paragraphes 1.2.2.1 et 1.2.2.2 .

Le problème se pose pour les résultats sur les personnes et sur les déplacements. En effet, **les deux échantillons personnes et déplacements ne sont pas aléatoires**. Les personnes interrogées l'ont été parce que leur ménage a été tiré au sort et les déplacements enquêtés le sont parce qu'ils sont réalisés par des personnes dont le ménage a été tiré au sort. C'est ce que l'on appelle l'effet de grappe. **Les personnes sont dans la première grappe, les déplacements dans une seconde**. Les règles de calcul de l'erreur ne peuvent donc pas s'appliquer directement.

Néanmoins, grâce à des tests, on peut calculer ce que l'on appelle des facteurs de grappe. Ce travail a été mené par le CERTU pour les enquêtes ménages. Les conclusions en sont les suivantes :

- **Pour les personnes, l'effet de grappe est peu sensible**. Le facteur de grappe est en moyenne de 1,1. Il faut donc multiplier toutes les erreurs sur les estimations vues aux paragraphes 1.2.2.1 et 1.2.2.2 par ce facteur.
- **Pour les déplacements, l'effet de grappe est beaucoup plus important** (ce qui est logique puisque nous sommes dans une seconde grappe). On obtient ainsi un facteur de grappe **de l'ordre de 2**. Toutes les estimations d'erreurs des paragraphes 1.2.2.1 et 1.2.2.2 doivent donc être multipliées par 2. Par répercussions, l'échantillon minimum nécessaire pour obtenir une précision voulue doit lui être multiplié par 4 (2^2). Pour les déplacements, le tableau du paragraphe 1.2.2.1 (sur les variables de proportions) devient donc le tableau ci-contre :

Tableau 4 : relation entre proportion, erreur relative et nombre de déplacements à enquêter au seuil de confiance de 95% dans les cas spécifique d'une EMD avec prise en compte de l'effet de grappe

Proportion mesurée \ Erreur relative	Erreur relative							
	1%	5%	10%	15%	20%	25%	30%	50%
1%	15 212 736	608 509	152 127	67 612	38 032	24 340	16 903	6 085
5%	2 919 616	116 785	29 196	12 976	7 299	4 671	3 244	1 168
10%	1 382 976	55 319	13 830	6 147	3 457	2 213	1 537	553
15%	870 763	34 831	8 708	3 870	2 177	1 393	968	348
20%	614 656	24 586	6 147	2 732	1 537	983	683	246
25%	460 992	18 440	4 610	2 049	1 152	738	512	184
30%	358 549	14 342	3 585	1 594	896	574	398	143
35%	285 376	11 415	2 854	1 268	713	457	317	114
40%	230 496	9 220	2 305	1 024	576	369	256	92
45%	187 812	7 512	1 878	835	470	300	209	
50%	153 664	6 147	1 537	683	384	246	171	
55%	125 725	5 029	1 257	559	314	201	140	
60%	102 443	4 098	1 024	455	256	164	114	
65%	82 742	3 310	827	368	207	132		
70%	65 856	2 634	659	293	165			
75%	51 221	2 049	512	228				
80%	38 416	1 537	384	171				
85%	27 117	1 085	271					
90%	17 074	683	171					
100%								

Le tableau ci-dessus nous montre donc le fort impact du tirage en grappe sur les déplacements (comparé au *Tableau 1 : relation entre proportion, erreur relative et nombre d'individus à enquêter au seuil de confiance de 95%* *Tableau 1*). Cet impact n'est donc pas à négliger lors des exploitations, cependant on peut en relativiser l'impact général sur une EMD. En effet, si on a 5 000 ménages enquêtés dans une EMD, soit l'échantillon aléatoire, on a enquêté environ 11 000 personnes (dans la grappe 1) et on peut estimer à au moins 30 000 la taille de l'échantillon de déplacements (dans la grappe 2). On a donc en principe un échantillon de déplacements qui est au moins 6 fois plus gros que l'échantillon de ménages. Comme l'erreur statistique nous oblige à multiplier l'échantillon de déplacements par 4 pour conserver la même erreur, on en déduit donc qu'a priori on peut réaliser des exploitations aussi poussées sur les déplacements que sur les ménages, dans la même enquête, sans pour autant faire grimper l'erreur statistique. Toute la question est de se fixer des limites pour ne pas pousser trop loin l'exploitation sur les déplacements.

Pour les estimations d'erreur sur les variables de moyenne il faut alors utiliser la même méthode qu'énoncée au paragraphe 1.2.2.2 . On peut donc bâtir le tableau suivant qui tient compte de l'effet de grappe.

Tableau 5 : relation entre l'erreur relative et le nombre de déplacements enquêtés au seuil de confiance de 95%, pour les variables de moyenne, avec prise en compte de l'effet de grappe

Variable \ Erreur relative	Erreur relative						Unité
	1%	5%	10%	20%	30%	50%	
Taux d'occupation des VP pour les déplacements liés au travail	19 915	797	199	50	22	8	Déplacement
Taux d'occupation des VP pour les déplacements tous modes	48 189	1 928	482	120	54	19	Déplacement
Mobilité par personne tous modes tous motifs	29 749	1 190	297	74	33	12	Personne
Mobilité par personne tous motifs en VP	134 337	5 373	1 343	336	149	54	Personne
Mobilité par personne tous motifs TC	390 925	15 637	3 909	977	434	156	Personne
Mobilité par personne tous modes motif travail	167 805	6 712	1 678	420	186	67	Personne

D'après ce tableau, sur un échantillon d'environ 300 personnes (à peu près deux secteurs de tirage) on pourra donc obtenir la mobilité tous modes tous motifs par personne avec une marge d'erreur de 10% au seuil de confiance de 95%. Donc par exemple cette mobilité (calculée à 3,35depts/j/per) aura 95% de chance d'être comprise entre 3 et 3,7 déplacements par jour par personne. Sur le même secteur, la mobilité en TC aura une marge d'erreur comprise entre 30% et 50% au seuil de confiance de 95%. On peut donc estimer qu'une mobilité calculée de 0,25 dept/j/per aura en fait 95% de chance d'être comprise entre 0,15 et 0,35 déplacements par jour et par personne.

2 > Approche des limites d'exploitation d'une enquête ménages déplacements

D'après tous ce que nous venons de voir ci-dessus, c'est tout au long d'une enquête ménages déplacements que la fiabilité des résultats est mise à l'épreuve. Pour toutes les erreurs de mesure, à l'exception des erreurs de mémoire et des erreurs d'analyse, nous devons partir du principe que c'est au cours de la phase d'enquête que toutes les précautions ont été prises pour en limiter les impacts. L'erreur finale liée à la somme de ces erreurs doit donc être considérée comme insignifiante (à partir du moment où l'enquête est labélisée par le CERTU).

Pour les erreurs statistiques le problème est différent. Il faut donc se fixer des règles lors des analyses afin de conserver des résultats significatifs. Ceci n'est pas chose facile et doit souvent être réalisé au cas pas cas en fonction des variables et de la zone étudiée.

Il existe cependant certaines règles qui restent toujours valables. Pour les autres nous essayerons de donner des indications qui permettent de rester dans des marges d'erreurs acceptables. Cette notion de marge d'erreur acceptable est importante puisque dans la quasi-totalité des cas les chiffres issus de l'enquête ménages seront communiqués sans les notions de seuil et d'intervalle de confiance. Ils seront donc pris pour argent comptant par les lecteurs.

2.1. Les impératifs

2.1.1. Respecter l'échelon du tirage aléatoire

Comme nous l'avons vu plus haut, la règle de base pour que les lois de la statistique s'appliquent, est que l'échantillon doit être aléatoire. Dans une EMD l'échelon sur lequel cette règle est garantie est le secteur de tirage. **Il ne faut donc jamais descendre en dessous du secteur de tirage** pour réaliser des exploitations.

De la même manière, **2 demi-secteurs de tirage ne font pas un nouveau secteur de tirage**. Même si l'on obtient un nombre de ménages identique que sur un secteur, l'échantillon n'est en effet plus aléatoire.

2.1.2. Pas d'analyse dans une matrice pour moins de 10 déplacements enquêtés

Comme nous l'avons vu au paragraphe 1.2.2.1, dans une matrice origine/destination, **il ne faut pas utiliser sans regroupement les résultats pour lesquels le nombre de déplacements bruts est inférieur à 10**.

2.1.3. Analyse en regroupement des questions pour lesquelles l'erreur de mémoire du ménage enquêté est potentiellement forte

Dans le cas des questions sur les heures, on sait que les erreurs du ménage peuvent être importantes. Pour limiter l'impact de ces erreurs il convient **donc d'utiliser lors de l'exploitation des regroupements importants : par heures, voire demi-heure et en extrême limite au quart d'heure** uniquement pour les déplacements de courte durée qui sont, en général, mieux appréciés. Il conviendra aussi de prendre des précautions lors de l'analyse.

2.1.4. Une méthode d'analyse fiable

Comme nous l'avons vu, une simple erreur lors de la construction de nouveaux tableaux à partir des résultats de l'EMD peut avoir des répercussions en terme d'erreur beaucoup plus importantes que toutes les autres erreurs. Il faut donc faire preuve d'une grande rigueur lors de l'exploitation de ces tableaux. Pour les exploitations depuis les tableaux de l'exploitation standard, la plus grosse partie du risque est déjà écartée puisque les tris croisés sont déjà réalisés. Il convient quand même de mettre en place un petit processus de vérification qui peut être simplement de vérifier qu'à l'issue de son traitement les sommes en ligne et colonne correspondent bien aux totaux du tableau original.

Pour les exploitations réalisées directement à partir de la base de données, le risque d'erreur est beaucoup plus important. Il faut donc mettre en place des processus de vérification. Le plus simple et qui est relativement efficace est le suivant : lorsque l'on travail sur la base on réalise souvent des tris croisés ou des réductions de la base (pour ne travailler que sur un secteur par exemple). Le principe est de continuer à travailler sur l'ensemble de base en ne réalisant pas une réduction, mais uniquement une identification de la partie qui nous intéresse. Grâce à cela, on peut, à chaque étape, vérifier que le nombre d'enregistrements observés est toujours le bon et que la valeur redressée totale est toujours correcte.

2.2. Les préconisations

Pour faire des préconisations nous allons essayer de définir ce que l'on peut considérer comme des erreurs maximums acceptables. A partir de là, nous allons déterminer un nombre minimum d'individus statistiques à prendre en compte pour obtenir cette erreur maximum. Bien entendu partout où il est possible d'augmenter la taille de l'échantillon cela devra être fait.

L'ensemble de ces préconisations appliqué aux variables de proportion est rassemblé dans le tableau ci-dessous.

Tableau 6 : Erreur maximum acceptable au seuil de confiance de 95% et répercussions sur le nombre minimum de ménages, personnes et déplacements de l'échantillon

Proportion mesurée	Erreur maximum acceptable	Intervalle de confiance		Nombre de ménages minimum de l'échantillon	Nombre de personnes minimum de l'échantillon	Nombre de déplacements minimum de l'échantillon
		Min	Max			
5%	50%	2,50%	7,50%	292	353	1 168
10%	50%	5,00%	15,00%	138	167	553
15%	40%	9,00%	21,00%	136	165	544
20%	30%	14,00%	26,00%	171	207	683
25%	25%	18,75%	31,25%	184	223	738
30%	25%	22,50%	37,50%	143	174	574
40%	20%	32,00%	48,00%	144	174	576
50%	15%	42,50%	57,50%	171	207	683
60%	10%	54,00%	66,00%	256	310	1 024
70%	10%	63,00%	77,00%	165	199	659
80%	5%	76,00%	84,00%	384	465	1 537

D'après le tableau ci-dessus on voit donc qu'il faut compter entre 140 et 300 ménages pour obtenir les erreurs acceptables. Hors, dans les EMD on a environ 75 ménages dans chaque secteur de tirage. Il faut donc compter environ 2 à 3 secteurs de tirage pour obtenir l'échantillon minimum de ménage. Dans cette hypothèse on peut envisager d'avoir entre 300 et 500 personnes dans 2 ou 3 secteurs de tirage. On est donc bien ici dans la fourchette de l'échantillon minimum affiché dans le tableau. De la même manière on peut estimer entre 1 100 et 1 600 déplacements pour 2 ou 3 secteurs de tirage ce qui nous place aussi dans la fourchette du tableau.

En fonction de cela on peut en conclure que dans une enquête ménages déplacements :

• pour traiter des variables de proportion concernant les ménages, il faut travailler au minimum sur un ensemble de deux secteurs de tirages et si cela est possible plutôt sur des ensembles de 3 secteurs de tirages,

• pour traiter des variables de proportion concernant les personnes ou les déplacements, il faut travailler au minimum sur un ensemble de deux secteurs de tirage.

• Au vu des tableaux des paragraphes 1.2.2.2 et 1.2.4 on peut aussi en conclure qu'il faut bien 2 et si possible 3 secteurs de tirages pour traiter des variables de moyenne. On est alors sur des erreurs de l'ordre de 10% au seuil de confiance de 95%.

• On doit garder à l'esprit que pour réaliser des exploitations satisfaisantes, il faut avant redressement au moins : 150 ménages, 300 personnes et environ 1 000 déplacements.

Enquête

Ménages déplacements



CONSEIL GÉNÉRAL
Haute-Loire

Conseil général
LOIRE
EN RHÔNE-ALPES
Région
Rhône-Alpes

LOIRE
Agglomération
FOREZ

SAINT-ÉTIENNE
métropole



Enquête Ménages Déplacements du bassin de vie stéphanois

epures

L'Agence d'urbanisme
de la région stéphanoise

46 rue de la télématique

BP 40801 – 42952 Saint-Etienne CEDEX 9

Tél : 04 77 92 84 00 fax 04 77 92 84 09

Mail : epures@epures.com – Web : www.epures.com

Sources : CERTU : guide méthodologique des
enquêtes ménages déplacements

CERTU : éléments pour évaluer la
précision obtenue dans l'estimation des indicateurs
donnés par les EMD – Décembre 2006

Crédit photographiques : epures

Imprimé par nos soins