Pilotage de l'Enquête Nationale sur les Transports et les Déplacements 2007/08 : traitements post-collecte, mise à disposition d'une base provisoire. Phase 4

S. Roux, Jean-Paul Hubert, Jimmy Armoogum

Institut national de recherche sur
les transports et leur sécurité

Rapport de phase 4 : *Traitements post-collecte, mise à disposition d'une base provisoire*

Projet : Pilotage de l'Enquête Nationale sur les Transports et les Déplacements 2007/08

Subvention Ministère – DAEI : N°07 MT S018°:

N° INRETS : F05 – 57

Sophie Roux
Jean-Paul Hubert
Jimmy Armoogum

INRETS –DEST

SOMMAIRE

I. Contrôle de la cohérence des données de l'Enquête Nationale sur les Transports et les Déplacements 2007-08

Avoir un questionnaire rempli entièrement ou en parti n'implique pas obligatoirement que les informations fournies sont correctes. Une partie des réponses peut se révéler aberrante d'une question à l'autre. Il est donc indispensable de vérifier la qualité et la fiabilité des données saisies en effectuant des contrôles de cohérences. Ces contrôles peuvent se diviser en deux étapes. La première consiste à vérifier que tous les individus ayant répondu aux questionnaires sont présents dans la base de donnés et dans les différents modules. La seconde consiste à contrôler la cohérence des réponses fournies par l'enquêté. Cette étape ne peut s'appliquer que sur deux types d'erreurs : les erreurs de codifications ou de saisies des variables et les erreurs de mesure.

La première partie du rapport expliquera pourquoi il est indispensable de contrôler la cohérence des données avant de les corriger. La seconde partie proposera des approches théoriques pour vérifier la cohérence des données et la dernière partie sera réservée à une application sur l'Enquête Nationale sur les Transports et les Déplacements 2007-08. Dans un premier temps, nous expliquerons comment les incohérences ou non-réponse ont été mises en évidences et ensuite nous proposerons des méthodes pour corriger les données incohérentes ou manquantes.

1. Le contrôle de la cohérence des données : un passage obligatoire avant toutes corrections de la non-réponse

Le contrôle du nombre de répondant est essentiel dans une enquête par sondage. Le statisticien doit obligatoirement s'assurer :

- Que l'on puisse identifier tous les individus ou ménages ayant répondu à l'enquête ;
- Qu'une unité ne peut avoir qu'un seul identifiant ;
- Qu'il n'y a pas de double compte. L'identifiant d'une personne, ménage ou tout autre chose, ne doit être présent qu'une seule fois dans l'enquête et identique d'une base à l'autre ;
- Et que l'enquêté a répondu à tous les modules du questionnaire s'appliquant à lui ou à sa situation.

Le contrôle de la cohérence des données à l'intérieur d'un même questionnaire consiste à vérifier l'exactitude des informations recueillies auprès d'un même ménage ou individu tout au long du questionnaire. Évidemment, il est impossible pour le statisticien de savoir si les réponses fournies par l'enquêté sont exactes ou non. Ce dernier se basera donc sur le postulat suivant : « les observations disponibles dans le fichier d'exploitation correspondent pour chaque unité statistique à la « vraie » valeur » (CARON, 1993). Pour contrôler les données, il cherchera uniquement à mettre en évidence les illogismes qui peuvent apparaître dans un même questionnaire. Il essaiera également d'identifier les erreurs relevant d'une mauvaise codification ou saisie des variables de celles des erreurs de mesure volontaire. La vérification n'est pas un outil de correction des données. Son objectif est de s'assurer de l'exhaustivité des données, de réduire les incohérences et les valeurs impossibles, de fournir de l'information sur la qualité des données et d'indiquer s'il est nécessaire d'apporter de nombreuses vérifications ou non.

Ces étapes sont essentielles car la mise en place des méthodes de corrections de la non-réponse est décidée à partir des informations contenues dans la base. Quel que soit le type de données, le contrôle de cohérence est donc une étape préliminaire pour toutes les enquêtes. Il faut d'une part distinguer les non-répondants des unités hors champs, puis les non-réponses partielles des non-réponses totales, et d'autre part, contrôler les informations fournies par les enquêtés. Cette dernière étape ne doit en aucun cas être négligée et doit être faite avant toutes corrections de la non-réponse car si une réponse recueillie dans un questionnaire semble erronée ou suspecte, celle-ci peut être

invalidée et devient par conséquent manquante. Ceci crée « artificiellement » de la non-réponse partielle qui se traite par les mêmes méthodes de correction que celles obtenues spontanément.

Les techniques utilisées pour corriger la non-réponse partielle sont en général celles de l'imputation. L'imputation consiste à remplacer une donnée manquante par une (ou plusieurs) donnée(s) déduite(s) ou calculée(s) en fonction des renseignements obtenus pour l'unité défaillante et/ou pour les unités qui lui sont proches (ARMOOGUM, MADRE, 1997). Elle peut résulter :

- de calculs directs à partir d'autres informations sur la même unité ;
- de relations plus formalisées estimées en général par régression sur les observations complètes (par exemple, vitesse fonction de la distance permettant d'imputer la durée) ;
- d'un (ou plusieurs) « donneur(s) », c'est-à-dire d'une (ou plusieurs) observation(s) dont les caractéristiques sont voisines, que l'on rapproche de l'observation incomplète.

L'objectif des procédures d'imputation est l'obtention d'une matrice de données complète (on parle dans ce cas de « clean data matrix »). Ceci s'avère très utile lors des analyses multivariées qui ne peuvent se réaliser sur des données entachées de valeurs manquantes.

Lorsque l'on procède à une imputation, il est nécessaire non seulement de décrire toutes les procédures d'imputation utilisées, mais aussi de créer des variables indicatrices, dites "flag", qui permettraient de marquer dans le fichier les données imputées. Ceci laisserait alors la possibilité au statisticien de juger de l'influence ou non des données imputées et de changer éventuellement de méthodologie d'imputation, mais aussi d'en tenir compte lors du calcul des intervalles de confiance.

2. La vérification des données : approche théorique

La vérification des données est basée sur un ensemble de règles. Ces règles sont formulées en fonction du domaine d'études, du questionnaire, et parfois d'autres enquêtes antérieures ou annexes (pour permettre la comparaison) et portent sur l'ensemble des questions du questionnaire. Chaque unité statistique est soumise à toutes les règles de vérification qui s'appliquent à elle. Pour une unité donnée, une règle peut être satisfaite, dans ce cas aucun traitement n'est requis, ou non satisfaite, dans ce cas une vérification est nécessaire.

Deux types de contrôle sont possibles pour appliquer ces règles (ARMOOGUM, MADRE, 1997) :

- Les micro-contrôles : ils visent à vérifier la cohérence interne du questionnaire. Ils s'appliquent sur une unité, c'est-à-dire qu'ils doivent garantir la validité et la cohérence des enregistrements individuels et des relations entre les enregistrements relatifs au ménage.
- Les macro-contrôles : ils consistent à repérer les erreurs ayant un impact non négligeable sur un résultat global. Ils s'appliquent uniquement sur un ensemble d'unités (population totale ou sous-groupe).

Les opérations de vérification des données peuvent se faire manuellement par le statisticien ou l'opérateur de saisie, à l'aide d'un programme informatique, ou en combinant les deux méthodes.

a) Des exemples de vérifications

Certains types de contrôles sont communs à toutes les enquêtes, d'autres dépendent du sujet d'études, de la longueur du questionnaire, de la taille de l'échantillon ou encore de la méthodologie de collecte mise en place. La réflexion qui suit proposera uniquement des exemples non exhaustifs et non ordonnés de vérifications possibles.

1. Le contrôle du nombre de répondant

Le contrôle du nombre de répondant est essentiel. Il est impératif que l'enregistrement d'une personne, d'un ménage, ou toute autre chose, ainsi que celui de ses données, ne soit fait qu'une seule et unique fois pour éviter les doubles comptes.

2. Le contrôle de la structure du questionnaire

Une autre vérification est celle du contrôle de la structure du questionnaire. Dans un questionnaire, il n'est pas rare qu'un module ne soit pas posé à une partie de l'échantillon parce qu'il ne s'applique pas à lui ou à sa situation. Il est donc important de respecter la spécification de ces filtres et de les intégrer dans les règles de contrôle.
Une des règles possibles est d'admettre que si une unité a échappé à un filtre alors toutes les réponses ayant trait à ce filtre sont effacées. A l'inverse, si une unité a été filtrée par erreur, ces réponses devront être imputées avec des méthodes identiques aux non réponses partielles.

3. Le contrôle du format des réponses

Le format, les valeurs et les ordres de grandeurs des réponses attendues doivent également être vérifiés. Cet examen permet de mettre en avant les erreurs de saisies ou de codifications. A la question « quel est votre âge ? », la règle à établir est que la réponse soit au format numérique et compris entre 0 et 120 ans maximum. Pour une question fermée à réponse unique avec, par exemple, 4 modalités codées de A à D, la règle à mettre en place est que la réponse à la question doit obligatoirement être A ou B ou C ou D. Toute autre réponse sera mise à blanc.

4. Le contrôle de cohérence des réponses lors de multiples bases

Dans une enquête complexe ou tout simplement longue, il est fréquent que les réponses aux questions soient réparties sur plusieurs bases de données, chacune regroupant les réponses à un module spécifique. Les réponses globales sont généralement séparées de celles qui les détaillent. Prenons le cas où le nombre de personnes appartenant au ménage est égal à 3 dans une première table et le nombre d'enfants et d'adultes appartenant au ménage est respectivement égal à 2 et 3 dans une seconde table. Prises séparément, ces réponses satisfont les règles de format, valeur et d'ordre de grandeur. Aucun traitement n'est donc requis. En agrégeant les deux tables, la réponse à la question du nombre de personnes appartenant au ménage ne satisfait pas la règle de l'addition du nombre d'enfants appartenant au ménage avec le nombre d'adultes appartenant au ménage. Il est donc utile de corriger chaque base mais il est dangereux de les corriger séparément sans tenir compte de l'ensemble des données.

5. Le contrôle de cohérence des réponses lors d'une enquête par panel

Dans une enquête sous forme de panel ou tout du moins régulière, il est recommandé de comparer les réponses des individus d'une enquête à l'autre. Cette confrontation permet de détecter toute modification importante par rapport à l'enquête précédente et de signaler ainsi les valeurs aberrantes.

6. Le contrôle de l'ensemble des données

Une fois les vérifications faîtes et les données corrigées, un dernier contrôle consiste à comparer et analyser le jeu complet des données. Cette opération permet entre autre de supprimer les valeurs extrêmes ou aberrantes qui nuiraient à la qualité des résultats de l'enquête.

3. La vérification des données : application à l'ENTD 2007-08

a) Un premier contrôle sur le terrain

L'échantillon des ménages de l'Enquête Nationale sur les Transports et les Déplacements 2007-08 est celui de la France métropolitaine. La mise en œuvre de l'enquête, ainsi que son déroulement sur l'année, nécessite une mise à disposition d'une partie des personnels de toutes les directions régionales de l'Insee. Leur rôle est essentiel car ils ont la capacité de faire baisser le nombre de non-réponses ou de réponses erronées et sont les premiers à contrôler la cohérence des données.
Au contact direct des enquêteurs, dont ils gèrent la paie, ils ont aussi la responsabilité de les former à l'enquête et répondre aux attentes et questions de ces derniers en cours de collecte afin d'éviter les réponses manquantes ou fausses. Les responsables font, entre autre, des mises à jour régulières destinées aux enquêteurs regroupant les réponses aux questions les plus soulevées. Ils ont également pour mission de vérifier le travail des enquêteurs et sont parfois chargés de saisir les informations du « carnet véhicule ». Une fois le questionnaire validé par l'enquêteur, ils contrôlent une partie de la cohérence des données (non exhaustivement et en fonction du nombre de remarques écrites par les enquêteurs). Un laps de temps réduit entre la passation du questionnaire et la mise à disposition des données est un avantage majeur lors de la vérification puisqu'en cas de doute sur certaines réponses, le correcteur peut se permettre de rappeler l'enquêteur.

Après vérification et validation des informations par les responsables d'enquêtes, toutes les directions régionales doivent envoyer dans un laps de temps imparti les données aux pôles en charge de la reprise des variables de profession et de la concaténation de l'ensemble des données. Les variables de profession (actuelle, antérieure, et du conjoint décédé) sont les premières à être contrôlées exhaustivement. L'examen se restreint uniquement à savoir si les enquêtés ont déclaré ou non une réponse. Les questions relatives aux professions sont des questions ouvertes, de fait l'enquêteur écrit textuellement ou librement la réponse de l'enquêté. Un logiciel embarqué (Sicore) dans l'ordinateur portable de l'enquêté permet de coder automatiquement en Professions et Catégories Socioprofessionnelles les informations saisies, soit environ 88% des déclarations de l'ENTD 2007-08. Le reste des réponses est repris manuellement par un personnel spécialisé. La totalité de ces travaux ont été effectuées vague par vague, soit 6 fois.

Une fois ce travail terminé, tout le questionnaire a été soumis à une batterie de tests de cohérence. Cette étape particulièrement importante détermine en grande partie la qualité des données recueillies. La correction des données de l'enquête a commencé dès la réception de la première vague dans le but de diminuer le temps d'attente entre la fin de l'enquête et les bases définitives. Certains contrôles de cohérence ont donc été effectués vague par vague, d'autres sur l'ensemble des 6 vagues. Les tests concernent aussi bien le nombre de répondants à chaque module que la cohérence interne de chaque volet du questionnaire et son ensemble, ainsi que les problèmes de chronologie, les erreurs de saisie ou de codage.

La vérification du nombre d'unité présent dans chaque module

La difficulté de traitement de la cohérence des données de l'ENTD 2007-08 est multipliée par comparaison aux enquêtes classiques, du fait de l'organisation très hiérarchisée du questionnaire et de la présence de nombreux tirages et sous-échantillonnages. Ces tirages avaient pour but d'optimiser la réponse à certains modules (voyages à longue distance et carnet véhicule nécessitant une stratification des individus ou des véhicules) ou de limiter le temps d'interview (tirage aléatoire des enfants scolarisés, des voitures ou des vélos). Selon les spécifications du concepteur de l'ENTD, le Service national informatique de Lille retourne à la Division Conditions de Vie 44 bases de données « aval » à chaque vague d'enquête. Ces dernières ne sont pas toutes construites sur le même niveau. Le questionnaire ne s'adressant pas uniquement à un seul individu du ménage, la structure des tables différencie le niveau du répondant. Il existe quatre niveaux : le niveau ménage, le niveau individu

appartenant aux ménages, le niveau individu « kish » et le niveau véhicule « kish ». Outre les réponses aux questions des enquêtés, ces 44 tables SAS comprennent des variables de contrôle générées automatiquement par le logiciel CAPI sur les dates et durées des questionnaires, le tirage de l'individu « kish » et du véhicule et sur quelques variables issues du recensement.
Ces tables sont en outre dimensionnées à la taille maximale des tableaux de variables. Ainsi, chaque variable individuelle existe en 20 occurrences, nombre maximum prévu d'individu dans un ménage, chaque variable voiture en 15, etc.
Afin de limiter la répétitivité du questionnaire sur les déplacements et les voyages en visite 2 , le data-model permettait aussi de sauter certaines questions en demandant simplement si telle ou telle condition avait changé ou non par rapport à un autre enregistrement, ou si un enregistrement était, en bloc, identique à autre. Il importait donc d'avoir des tables « aval » encore proches de la structure du data-model afin de ne pas multiplier les corrections à faire par les liens existant entre les enregistrements.
Une chaîne de correction a donc été développée, qui part des tables « aval » brutes livrées par Lille et se déroule en 6 étapes.

1ère étape : « élagage des tables » pour y enlever des variables inutiles, scories du data-model.

2ème étape : corrections individuelles, réalisées à partir des remarques des enquêteurs et des multiples vérifications de cohérence. Ces corrections s'accumulent en permanence, et sont prises en compte à chaque exécution de la chaîne.

3ème étape : Réorganisation des tables pour sortir de la structure du data-model.

4ème étape : corrections automatiques sur la base de règles de cohérences.

5ème étape : enrichissement des tables avec des données extérieures, notamment coordonnées géographiques et distances.

6ème étape : recollement des tables des 6 vagues et mise en forme définitive, pour réalisation des imputations.

On arrive alors à 20 tables :

13 tables pour la visite 1 :

- 2 tables de niveau ménage (TCM+TCM logement et Q_Menage)
- 4 tables véhicules, recensement du parc, une par type de véhicule
- 4 tables véhicules, une par type, description fine avec échantillonnage des voitures et des vélos
- 2 tables individus (TCM et Q_Individu)
- 1 table lieux de travail, étude ou garderie, dépendante de Q_individu

5 tables pour la visite 2 :

- 2 tables de niveau Kish : K_mobilité et une table méthodologique concernant notamment la remise et remplissage des documents de collecte et du GPS
- 1 table déplacements locaux et quotidiens
- 1 table voyages à plus de 100 km du domicile
- 1 table déplacements des voyages à longue distance des 4 dernières semaines

2 tables pour le véhicule-carnet : une sur le carnet, une sur les déplacements du carnet

1. L'attribution d'un numéro identifiant le ménage

Afin de distinguer tous les ménages, individus, véhicules, voyages et déplacements, il est indispensable de leur attribuer un numéro unique qui les suivra tout au long du questionnaire. L'identifiant dit « ménage » n'étant pas généré automatiquement par CAPI, les enquêteurs doivent rentrer manuellement au début du questionnaire un numéro défini au préalable par l'Insee. CAPI se

charge par la suite de donner automatiquement à chacun des individus, véhicules, voyages et déplacements un identifiant. Par conséquent, le statisticien n'est pas à l'abri d'une erreur de la part de l'enquêteur et plus rarement du CAPI.
Ainsi, il n'est pas surprenant d'observer à chaque vague qu'un ou deux identifiants renseignés par l'enquêteur aient été inversés avec un autre ménage nécessitant une reprise de tous les identifiants des ménages concernés. Ce qui l'est moins est d'avoir des réponses en visite 2 et ne pas savoir qui y a répondu. En effet, le protocole de l'enquête impose qu'un individu soit normalement tiré au sort à la fin de la visite 1 pour être le répondant de la visite 2. Pour limiter les échecs, il est initialement demandé aux ménages de citer les personnes qui ne pourront pas être jointes pour la visite 2. Il est possible que toutes les personnes soient citées, c'est-à-dire que personne n'est joignable. Or, il arrive que parmi ces ménages, certains comportent une visite 2. Nous sommes donc en face d'une contradiction, personne n'est censé être là et pourtant, il y a un questionnaire, rempli. Pour les ménages concernés, il est impossible d'identifier le répondant. Entre 7 et 16 ménages ont été concernés à chaque vague. Les enquêteurs ont été contactés pour récupérer le nom du répondant à la visite 2. Tous ont pu être joints et tous ont fourni le nom du répondant à une exception près. Son identifiant a dû être récupéré en utilisant la méthode de l'imputation déductive : des déplacements en voiture ont été déclarés en visite 2 et une seule personne du ménage possédait le permis de conduire.

2. Les doubles comptes

Nous n'avons pas observé de double compte dans les bases de l'ENTD 2007-08 sauf à une exception : il a été attribué à une unité deux numéros d'identification. En effet, un ménage a arrêté prématurément sa première visite mais a souhaité aller jusqu'au bout de l'enquête en prenant un autre rendez-vous. Lors de ce second rendez-vous, l'enquêteur a reposé toutes les questions de la première visite alors qu'il aurait dû reprendre le questionnaire à l'endroit où il s'était arrêté en première visite. Le logiciel CAPI lui interdisant de rentrer un identifiant identique à un autre déjà enregistré dans son ordinateur portable, l'enquêteur n'a pas eu d'autre choix que celui de créer un second identifiant relatif au même ménage. Le premier identifiant a donc été effacé des tables puisqu'il contenait moins d'informations sur l'enquêté.

3. La présence des enquêtés dans les modules

Il est extrêmement important de savoir si tous les individus ont répondu à l'ensemble des modules les concernant pour pouvoir corriger la non réponse totale. Seuls les enquêtés ayant répondu au moins à une question du module apparaissent dans les bases originales de l'enquête. Il est donc nécessaire de faire la différence entre les individus non concernés par un ou plusieurs modules de ceux n'ayant pas répondu à l'ensemble du module. Les modules du questionnaire individuel portant sur la mobilité régulière et la « motilité » (pratiques de mobilité, possession d'abonnements de transport en commun, permis de conduire, accidents) étaient filtrés en fonction de l'âge ou de l'activité, ce qui ne constitue pas une non-réponse, mais ils pouvaient également être posés à une autre personne du ménage (un proxy) qui pouvait se déclarer incapable de répondre à la place de cet individu, d'où une non-réponse sur ce module. Certaines visites ont également pu être interrompues brutalement, laissant le dernier module incomplet. Des enquêtes validées où trop de modules de la première visite étaient en non réponse ont pu ainsi être déclassées en défectueuses.

L'absence de véhicule « kish » dans certains modules

Prenons l'exemple du module 1.9 « usage des véhicules actuels ». Le questionnement consacré aux véhicules à disposition du ménage est en deux parties. Il commence par un recensement de tous les véhicules utilisés depuis 12 mois, sauf les vélos d'enfant et se poursuit par la description des véhicules. La description de l'ensemble des véhicules de types deux-roues à moteur, voiturettes, quads, tricycles à moteur, tracteurs agricoles est obligatoire. En revanche, la description est limitée à une voiture, véhicule utilitaire léger (VUL) ou camping car (CC), ou parfois deux si le véhicule « kish »

choisi pour recevoir le carnet est de ce type. Idem pour la description des vélos. Le travail de cohérence consiste à vérifier que tous les véhicules devant être décrits le sont réellement. Pour ce faire, il est indispensable de jongler entre toutes les tables concernées, soit au total 14. Finalement, il s'est avéré que seuls treize véhicules n'ont pas été décrits : 5 véhicules de types deux-roues à moteur, 3 vélos et 5 voitures. Il peut s'agir d'une erreur de saisie du nombre de véhicules, générant des fiches impossibles à compléter. Ces fiches seront soit corrigées par imputation au vu du peu d'effectif soit effacées s'il s'agit d'une erreur de saisie du nombre de véhicules.

La présence d'individus non concernés par un ou plusieurs modules

Inversement, il peut arriver que des unités apparaissent dans certaines tables alors qu'elles ne devraient pas y être soit parce que les filtres n'ont pas joué leur rôle soit parce que les enquêteurs ont rentré des informations sur le ménage alors que ce dernier n'était pas présent. En effet, lors du repérage du logement, il est demandé à l'enquêteur de se renseigner sur la résidence lorsque l'enquêté n'est pas joignable. Ces informations, utiles au statisticien pour différencier les résidences hors champs des refus, nécessitent de la part des enquêteurs de s'informer auprès des voisins, de la mairie…etc. Par mégarde, il est donc possible que le compilateur de l'ensemble des données intègre ces informations aux données d'enquêtes. En effet, dans la pratique, les informations collectées par les enquêteurs lorsque les enquêtés ne sont pas dans leur résidence ont été ajoutées aux données d'enquêtes. Ces ménages doivent en réalité être considérés comme de la non-réponse totale à l'enquête ou des unités hors champs (la résidence principale au RP99 est devenue une résidence secondaire, etc…) et doivent donc être enlevés.

De même, il s'est avéré que certaines tables réservées exclusivement aux réponses du « kish » en visite 2 comportaient des identifiants d'individus n'ayant pas fait cette visite. Ils doivent en réalité être observés comme de la non-réponse totale en seconde visite et être retirés des tables. Inversement, des « kish » ayant fait une visite 2 n'étaient pas présents dans les tables de déplacements locaux car le compilateur a pensé que les individus déclarant ne pas s'être déplacés ne devaient pas y apparaître. Ce sont en réalité des répondants même s'ils ne se déplacent pas et doivent être ajoutés aux tables des données.

La vérification des réponses des enquêtés

Contrairement aux enquêtes nationales transports précédentes, la collecte n'a pas été réalisée sur un support papier mais à l'aide d'un CAPI. Ce procédé permet une automatisation des tests de cohérence en présence de l'enquêté et donc une réduction du temps de contrôle pour le statisticien à trois conditions : que les contrôles de cohérence aient pu être testés par les enquêteurs afin de s'assurer qu'ils ne risquent pas de bloquer l'entretien ni d'interrompre sans cesse le questionnement, ce qui s'avèrerait contre-productif, que les consignes du questionnaire numérique soient exactement les mêmes que celles du questionnaire papier et que des contrôles de cohérences aient été programmés. Au vu de la complexité des spécifications et de la nécessité de réorganiser profondément le questionnaire après le test CAPI pour réduire la longueur du questionnaire, il serait utopique de penser que le passage du format papier au format numérique a été irréprochable. Le manque de temps et la complexité de l'enquête ont été des freins à l'ajout de tous les contrôles de cohérences souhaités.

Avant la mise en place du CAPI, la saisie des réponses du format papier vers un format numérique pouvait entraîner des erreurs de codification. Depuis son utilisation, le contrôle de la codification des réponses relatives aux questions fermées est accéléré. En effet, l'ordinateur affiche désormais les modalités de réponses correspondant à la question posée. L'enquêté ne peut donc pas coder de modalités autres que celles proposées par l'ordinateur. Les erreurs de formats, d'unités et plus généralement de codification sont rendues impossibles pour les questions de ce type. Cela n'empêche pas pour autant l'enquêté de pouvoir donner une réponse fausse ou l'enquêteur de rentrer

une réponse autre que celle donnée par l'enquêté. L'avantage ici de l'utilisation du logiciel CAPI est de pouvoir supprimer l'étape du contrôle de la codification. En revanche, CAPI ne proposant pas, par définition, de modalités de réponses aux questions ouvertes, le contrôle de la codification des réponses à ces questions est toujours nécessaire.

Dans l'ENTD 2007-08, la très grande majorité des questions ouvertes concernent les lieux de destinations ou des données chiffrées (date de départ et d'arrivée du déplacement, date d'immatriculation de la voiture, date de la première immatriculation de la voiture, nombre de voyages à longues distances effectués, nombre de véhicules dans le parc du ménage, etc.). Ces questions ouvertes sont essentielles dans une enquête sur la mobilité et nécessitent donc un contrôle poussé.

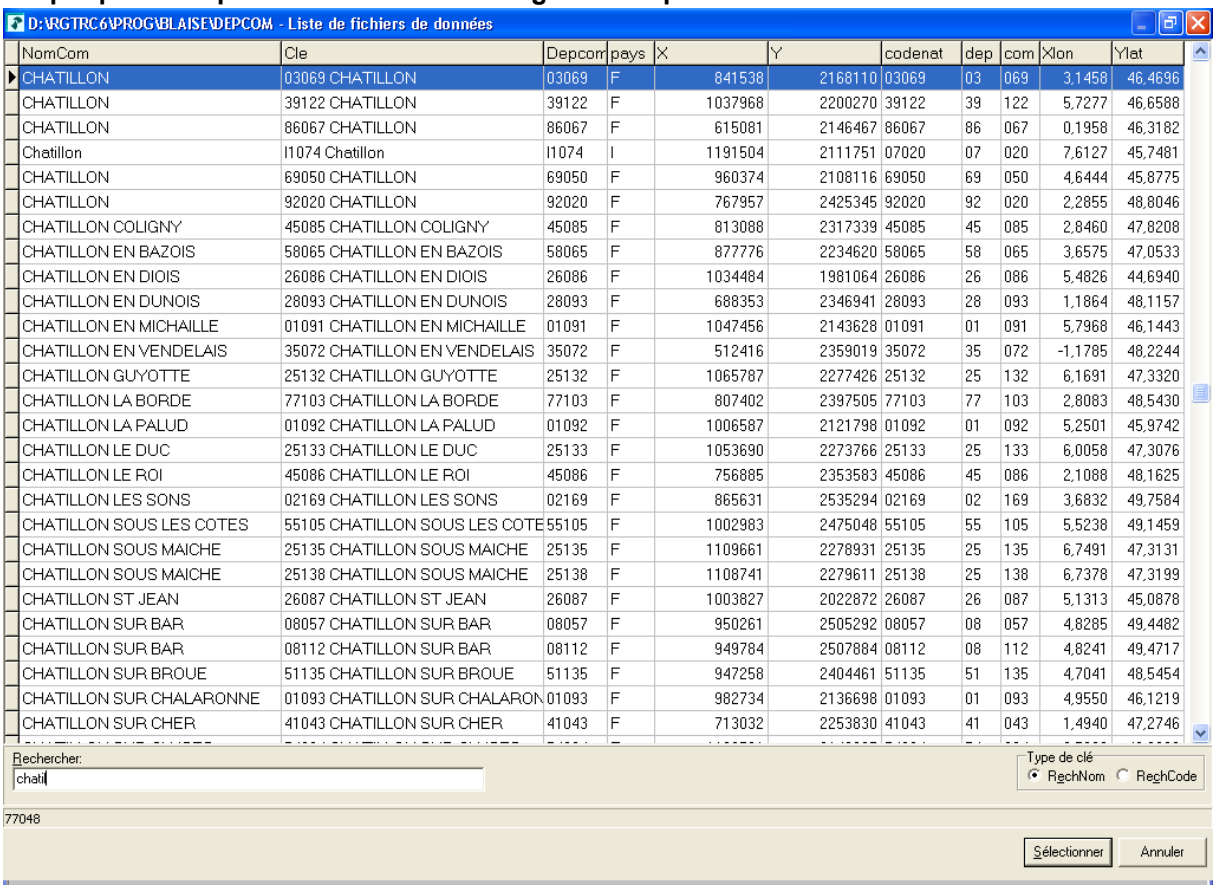
1. Le contrôle des variables de lieux

Une méthode de codage des variables de lieux a été ajoutée au CAPI pour limiter les erreurs de codification et la durée de passation du questionnaire. L'enquêteur a deux possibilités pour coder les lieux de destinations du déplacement de l'enquêté. La première, recommandée par les concepteurs, consiste à chercher dans une table géocodée le nom de la commune de destination de l'enquêté. La recherche est simplifiée puisqu'il suffit de rentrer les premières lettres du nom de la commune pour qu'une liste avec des propositions de noms de communes apparaisse. En utilisant cette méthode, les variables telles que le code Insee de la commune, le département et le pays sont imputées directement par le logiciel CAPI et ne nécessitent donc pas de contrôle de la part du statisticien. La seconde méthode est réservée au cas ou l'enquêteur ne trouve pas la commune de destination dans la table géocodée. Il doit alors remplir en clair le nom de la commune, ainsi que les variables pays, département (si la commune est en France), ou ville la plus proche (lorsque la destination est à l'étranger). Cette seconde méthode a l'inconvénient d'augmenter le temps de passation du questionnaire et d'obliger le statisticien à coder lui-même les noms de communes.

Les réponses aux variables de lieux ont fait l'objet de micro et macro-contrôles. Au cours de ces contrôles, il a été établi que les erreurs rencontrées résultaient aussi bien d'erreurs volontaires qu'involontaires de la part des enquêteurs. Les erreurs de mesure involontaires correspondent aux lieux mal orthographiés par l'enquêteur qui ne peuvent par conséquent être codés par ce dernier puisque n'apparaissant pas dans la liste géocodées. Cette vérification, de type micro-contrôle, consistait pour l'essentiel à regarder les noms de communes écrites en clair et les remarques annexes formulées par les enquêteurs sur les variables de lieux. Lors de la vérification de ces remarques, des erreurs de mesure volontaires de la part de l'enquêteur ont également été mises en évidence. En effet, il n'est pas rare de trouver deux destinations différentes pour une même question : une codée à l'aide du CAPI et une seconde écrite sur les remarques annexes. Ainsi, lorsque l'enquêteur ne trouvait pas le nom de la commune de destination dans la table géocodée, ce dernier n'écrivait pas en clair le nom de la commune mais préférait entrer une réponse fausse pour diminuer le temps de l'entretien soit en choisissant la première commune de la liste géocodée soit une commune dont les 3 ou 4 premières lettres étaient identiques à celle qu'il aurait souhaité coder. Une autre erreur, dont il est difficilement vérifiable de connaître son origine volontaire ou involontaire, est celle du choix du code Insee de la commune. En effet, la table géocodée ne propose pas uniquement des noms de communes. Elle présente également à l'enquêteur les codes Insee correspondant. Ces deux éléments sont nécessaires à la table car il existe en France plusieurs communes portant le même nom mais dans des départements différents. Ainsi lorsque l'enquêteur entre un nom de commune, il doit également demander dans certains cas à l'enquêté le département auquel appartient la commune. Or, il s'avère que parfois l'enquêteur entre en réponse la première commune qui apparaît dans la table sans porter attention au code Insee correspondant. La mise en évidence de ces erreurs est rendue possible à l'aide d'un macro-contrôle. En vérifiant la variable de lieux en fonction de la distance, de la durée, du mode, de la vitesse, de la distance à vol d'oiseau du déplacement, etc. des trajets vers Lille en Belgique se sont avérés être des déplacements vers Lille en France, de même pour des villes

comme La Rochelle (deux départements possibles), ou encore Châtillon (40 possibilités réparties entre la France, l'Italie et les noms composés (Graphique 3)).

Graphique 1 : Capture d'écran de la table géocodée pour la commune de Chatillon


D:\RGTRC6\PROG\BLAISE\DEPCOM - Liste de fichiers de données

NomCom	Cle	Depcom	pays	X	Y	codenat	dep	com	Xlon	Ylat
CHATILLON	03069 CHATILLON	03069	F	841538	2168110	03069	03	069	3,1458	46,4696
CHATILLON	39122 CHATILLON	39122	F	1037968	2200270	39122	39	122	5,7277	46,6588
CHATILLON	86067 CHATILLON	86067	F	615081	2146467	86067	86	067	0,1958	46,3182
Chatillon	I1074 Chatillon	I1074	I	1191504	2111751	07020	07	020	7,6127	45,7481
CHATILLON	69050 CHATILLON	69050	F	960374	2108116	69050	69	050	4,6444	45,8775
CHATILLON	92020 CHATILLON	92020	F	767957	2425345	92020	92	020	2,2855	48,8046
CHATILLON COLIGNY	45085 CHATILLON COLIGNY	45085	F	813088	2317339	45085	45	085	2,8460	47,8208
CHATILLON EN BAZOIS	58065 CHATILLON EN BAZOIS	58065	F	877776	2234620	58065	58	065	3,6575	47,0533
CHATILLON EN DIOIS	26086 CHATILLON EN DIOIS	26086	F	1034484	1981064	26086	26	086	5,4826	44,6940
CHATILLON EN DUNOIS	28093 CHATILLON EN DUNOIS	28093	F	688353	2346941	28093	28	093	1,1864	48,1157
CHATILLON EN MICHAILLE	01091 CHATILLON EN MICHAILLE	01091	F	1047456	2143628	01091	01	091	5,7968	46,1443
CHATILLON EN VENDELAIS	35072 CHATILLON EN VENDELAIS	35072	F	512416	2359019	35072	35	072	-1,1785	48,2244
CHATILLON GUYOTTE	25132 CHATILLON GUYOTTE	25132	F	1065787	2277426	25132	25	132	6,1691	47,3320
CHATILLON LA BORDE	77103 CHATILLON LA BORDE	77103	F	807402	2397505	77103	77	103	2,8083	48,5430
CHATILLON LA PALUD	01092 CHATILLON LA PALUD	01092	F	1006587	2121798	01092	01	092	5,2501	45,9742
CHATILLON LE DUC	25133 CHATILLON LE DUC	25133	F	1053690	2273766	25133	25	133	6,0058	47,3076
CHATILLON LE ROI	45086 CHATILLON LE ROI	45086	F	756885	2353583	45086	45	086	2,1088	48,1625
CHATILLON LES SONS	02169 CHATILLON LES SONS	02169	F	865631	2535294	02169	02	169	3,6832	49,7584
CHATILLON SOUS LES COTES	55105 CHATILLON SOUS LES COTE	55105	F	1002983	2475048	55105	55	105	5,5238	49,1459
CHATILLON SOUS MAICHE	25135 CHATILLON SOUS MAICHE	25135	F	1109661	2278931	25135	25	135	6,7491	47,3131
CHATILLON SOUS MAICHE	25138 CHATILLON SOUS MAICHE	25138	F	1108741	2279611	25138	25	138	6,7378	47,3199
CHATILLON ST JEAN	26087 CHATILLON ST JEAN	26087	F	1003827	2022872	26087	26	087	5,1313	45,0878
CHATILLON SUR BAR	08057 CHATILLON SUR BAR	08057	F	950261	2505292	08057	08	057	4,8285	49,4482
CHATILLON SUR BAR	08112 CHATILLON SUR BAR	08112	F	949784	2507884	08112	08	112	4,8241	49,4717
CHATILLON SUR BROUE	51135 CHATILLON SUR BROUE	51135	F	947258	2404461	51135	51	135	4,7041	48,5454
CHATILLON SUR CHALARONNE	01093 CHATILLON SUR CHALARON	01093	F	982734	2136698	01093	01	093	4,9550	46,1219
CHATILLON SUR CHER	41043 CHATILLON SUR CHER	41043	F	713032	2253830	41043	41	043	1,4940	47,2746

2. Le contrôle des variables de dates

Des contrôles de cohérence ont également été établis pour les questions relatives aux dates lors de la programmation du questionnaire. Ils ont porté sur le format (JJ/MM/AAAA) et sur la chronologie des événements lorsqu'il est demandé à l'enquêté de décrire ses déplacements tout au long de la journée ou de la semaine. En revanche, aucun contrôle n'a été fait sur la valeur prise par ces dates. De fait, un grand nombre d'erreurs de codifications a été observé. Elles résultent généralement d'erreurs involontaires de la part de l'enquêteur puisque les réponses aberrantes sont souvent des erreurs de saisie. Ainsi, il n'est pas rare de trouver des dates avec des années égales à 1008, 3008 ou encore 2088. Par conséquent, lorsque ce déplacement est suivi d'un autre et que l'année du déplacement précédent est supérieure à 2008, le déplacement suivant l'est aussi à cause du contrôle dû à l'obligation de chronologie. Les erreurs volontaires de codifications sont ainsi multipliées. Si les erreurs sur les années des déplacements nécessitent uniquement une vérification de type micro-contrôle, les erreurs de saisies sur les mois ou les jours de déplacements relèvent eux d'un macro-contrôle. Lors de cette vérification, les dates de déplacements ont été contrôlées en fonction de la vague d'enquête, de la durée, du motif, de l'ordre du déplacement, des remarques des enquêteurs, etc.

Une autre explication possible aux problèmes de date peut être due au fait que les consignes concernant les dates pour le module des déplacements « mobilité locale » et celui des déplacements « carnet » ne sont pas identiques. En effet, dans le premier module, les enquêtés doivent décrire leurs déplacements de 4h00 du matin à 3h59 le lendemain. De fait, même passé minuit, la date du déplacement reste la même. A l'inverse, pour le carnet véhicule, les déplacements se comptabilisent

de 0h00 du matin à 23h59 le même jour. De fait, passé minuit, la date du déplacement change. Les enquêtés ainsi que les enquêteurs ont pu être troublés et confondre les deux consignes.

La vérification de la non-réponse des enquêtés

Avoir de la non-réponse à une question n'implique pas obligatoirement que celle-ci soit volontaire. Elle peut également être involontaire. Pouvoir faire la part entre les deux est important du point de vue de la méthodologie et de la conception du questionnaire même si statistiquement les non-réponses seront traitées de façon identique, c'est-à-dire par une méthode d'imputation. Les causes de la non-réponse volontaire sont multiples (refus de répondre aux questions sensibles ou à l'ensemble du questionnaire, etc.). En revanche, une seule raison est généralement évoquée lorsque l'on parle de non-réponse involontaire : celle d'une erreur de programmation de filtre.

1. La non-réponse involontaire : une vérification défaillante des filtres programmés

Dans le questionnaire de l'ENTD 2007-08, il peut arriver que certains individus ne soient pas concernés par quelques questions, voire une partie du questionnaire. Des filtres ont donc été ajoutés lors de la programmation du CAPI. Ils sont essentiels car ils permettent de diminuer le temps de passation du questionnaire (rappelons que l'enquête dure en moyenne 115 minutes). Ils ont également pour avantage de faire gagner du temps au statisticien puisque ce dernier n'a pas besoin de filtrer manuellement les réponses des enquêtés. Ces filtres doivent être programmés et vérifiés avec le plus grand soin car ils peuvent soit créer des réponses supplémentaires inutiles, soit être la cause de réponses manquantes ce qui serait particulièrement pénalisant. L'absence de filtre est moins dommageable pour le statisticien qu'une mauvaise programmation. Il préférera mettre à blanc des réponses plutôt que de devoir les imputer.
Vingt-cinq versions du data-model ont été envoyées par l'équipe Blaise-Capi de Lille, auxquelles s'ajoutent trois révisions en cours de collecte. Elles ont toutes été testées par l'équipe de conception (CPOS, concepteur, contrôleur de la division condition de vie affecté à l'ENTD) ainsi que par le pôle enquête ménage de Nancy. Malgré cela, certains bugs ont été décelés trop tard pendant la collecte, voire après.

Le contrôle des données de l'ENTD 2007-08 a permis de repérer une absence de filtres sur plusieurs questions, de même que de mauvaises programmations favorisant l'augmentation des réponses manquantes. Pour comprendre les conséquences de ces erreurs, prenons deux exemples distincts : l'un sur une erreur de programmation du filtre, le second sur l'ajout d'un filtre.

Une erreur de programmation sur la variable de durée BTRAVTEMPSR

Dans ce premier exemple, l'erreur porte sur la durée du trajet retour entre le domicile et le travail de l'ensemble des individus du ménage.

Combien de temps (PRENOM) met-il/elle habituellement pour faire

BTRAVTEMPSA **Le trajet aller ?**	__| __| h __| __| mn

Si BTRAVTEMPSA >= 5 minutes, poser **:**

BTRAVTEMPSR **Le trajet retour ?**	__| __| h __| __| mn

Sur la version papier du questionnaire, lorsque la variable BTRAVTEMPSA est supérieure ou égale à 5 minutes, il faut demander à l'enquêté quel est le temps de son trajet retour. Après vérifications des données, on observe que toutes les réponses sont manquantes lorsque la case « mn » de la variable BTRAVTEMPSA est comprise entre 00 et 04 et ce quel que soit la valeur prise par « h ». La

programmation du filtre est donc erronée car son fonctionnement dépend uniquement de la variable « mn ». Il aurait fallut que le filtre porte à la fois sur les variables « h » et « mn ». Une imputation des réponses manquantes sera donc nécessaire.

Tableau 1 : Pourcentage d'imputation à effectuer sur la variable BTRAVTEMPSR en raison de la programmation incorrecte du filtre selon les vagues

Vague 1	Vague 2	Vague 3	Vague 4	Vague 5	Vague 6
9,5%	8,8%	9,5%	8,3%	10,0%	7,5%

Source : INSEE Enquête Nationale sur les Transports et les Déplacements 2007-08
Lecture : parmi les individus de la vague 1 ayant un temps de trajet aller compris entre 1h00 et 1h04, 2h00 et 2h04, etc., 9,5% des réponses devront être imputées sur la variable BTRAVTEMPSR en raison de la programmation incorrecte du filtre.

Cette erreur de programmation est moins dommageable qu'il n'y paraît bien qu'il s'agisse d'une variable relative à la mobilité. En effet, pour corriger cette erreur, le statisticien peut envisager deux solutions de type imputation déductive. La première méthode consiste à regarder les données déjà existantes de l'individu « kish » pour le ménage concerné. En effet, ce dernier a décrit auparavant ses déplacements un jour de semaine. Il suffit alors d'imputer la donnée manquante par son temps de trajet retour en espérant qu'il soit parti vers son lieu de travail habituel ce jour là en utilisant les mêmes modes de transports. La seconde méthode consiste à déduire le temps de trajet retour en fonction de la destination, du motif et du mode de transport.

L'ajout d'un filtre sur la variable QDECRIT

Dans ce second exemple, l'erreur porte sur le module des accidents de la circulation.

QDECRIT
Question enquêteur : l'accident a-t-il déjà été décrit par un autre membre du ménage?

1. Oui *poser la question QSOIN, puis passez à l'accident suivant ou à la personne suivante*	1
2. Non	2

QSOIN
L'accident a-t-il nécessité des soins pour (PRENOM) ?

1. Soins sans intervention d'un médecin (ni agent hospitalier)	1
2. Soins avec intervention d'un médecin, sans hospitalisation	2
3. Soins avec hospitalisation de moins de 24h	3
4. Soins avec hospitalisation d'un jour ou plus	4
5. Aucun soin	5

Q22P
Est-ce que les forces de l'ordre ont été alertées et sont venues ?

1. Oui	1
2. Non	2

Sur la version numérique du questionnaire, le filtre de la question QDECRIT se déclenche lorsque la réponse à la question QDECRIT est positive. En revanche, lorsque la réponse à la question QDECRIT est négative, la question QSOIN n'est pas posée et l'enquêté passe directement à la question. QSOIN n'est donc demandé que si l'accident a déjà été décrit. De fait, ce filtre supplémentaire provoque l'imputation de 100% des réponses des enquêtés ayant eu un accident non décrit par un autre membre du ménage. Cette erreur a été découverte pendant la quatrième vague de l'enquête. Le CPOS et le concepteur avaient alors le choix entre deux mauvaises solutions : modifier le data-model

pour la vague 6 ou informer les enquêteurs et leur donner la consigne de poser la question hors de CAPI pendant les vagues 5 et 6, ce qui fut choisi. Les enquêteurs avaient alors pour consigne en vague 5 et 6 de poser la question QSOIN lorsque la réponse à la question QDECRIT est négative et d'entrer la réponse dans la case « remarque ».

Tableau 2 : Pourcentage d'imputation à effectuer sur la variable QSOIN en raison d'un filtre supplémentaire selon les vagues

Vague 1	Vague 2	Vague 3	Vague 4	Vague 5	Vague 6
100,0%	100,0%	100,0%	100,0%	86,9%	91,1%

Source : INSEE Enquête Nationale sur les Transports et les Déplacements 2007-08

Lecture : parmi les individus ayant eu un accident non décrit par un autre membre du ménage en vague 5, 86,9% des réponses devront être imputées sur la variable QSOIN en raison d'un filtre supplémentaire.

Le contrôle de la cohérence des données est essentiel puisque laisser des données incohérentes biaiseraient les estimations. Il est donc nécessaire de les transformer en non-réponses partielles. Ces données, de même que les données manquantes, se corrigent par imputation. Dans l'Enquête Nationale sur les Transports et les Déplacements 2007-08, les principales méthodes utilisées pour l'imputation sont de deux types :

- déductives, c'est-à-dire lorsque l'information manquante est remplacée selon une règle déterministe, utilisant les variables disponibles sur cette même unité ;
- par hot-deck, c'est-à-dire lorsque la donnée manquante est remplacée par la réponse observée d'une unité répondante qui lui est proche.

II. Correction des erreurs d'échantillonnage et de la non-réponse par calage sur marges.

Les techniques de calage sur marges ont été développées afin d'améliorer l'estimation de la variable d'intérêt, car nous réduisons la variance de notre estimateur lorsque les variables auxiliaires de calage sont corrélées avec la variable d'intérêt. Cette technique consiste à faire coïncider les marges de quelques variables de l'échantillon à celles de la population en modifiant la pondération. Lorsque les variables auxiliaires sont qualitatives, cette approche ne nécessite pas la connaissance, dans la population, du croisement de ces variables auxiliaires. La méthode de calage sur marges la plus connue en matière de sondage est la 'méthode itérative du quotient' ou Raking Ratio introduite par Deming et Stephan (1940) et Stephan (1942). Deville et Särndal (1992) ont généralisé l'approche des estimateurs par régression en introduisant les estimations par 'calage sur marges'. Nous pouvons ainsi corriger les différents types d'erreurs à condition d'utiliser les variables expliquant le mécanisme de la non-réponse lors du calage sur marges.
Cette étape est essentielle pour assurer une bonne représentativité de l'échantillon et la comparabilité avec d'autres sources statistiques (enquêtes de l'INSEE). Nous profiterons de cette étape pour introduire des contraintes temporelles lorsque cela sera nécessaire.

1. Calage sur les marges du recensement de la population

L'analyse du mécanisme de réponse montre une corrélation entre le fait de répondre et le niveau de motorisation du ménage. Cette variable étant connue dans le recensement de la population qui sera disponible qu'à partir du mois de mars 2009. C'est pour cela que nous proposons un redressement provisoire à l'aide d'un calage sur les marges de l'Enquête Emploi (et nous referons cette exercice lorsque nous disposerons des marges du recensement de la population).

2. Variables de calage pour le niveau ménages-individus

Afin de construire une pondération de la première visite, nous avons utilisé les variables de calage suivantes pour le niveau ménages – individus :

Niveau national

PCS de la personne de référence

- Agriculteurs (actifs occupés et anciens actifs)
- Indépendants ou professions libérales (actifs occupés et anciens actifs)
- Professeurs, instituteurs et assimilés, professions intermédiaires de la santé, autres cadres ou professions intermédiaires (actifs occupés)
- Professeurs, instituteurs et assimilés, professions intermédiaires de la santé, autres cadres ou professions intermédiaires (anciens actifs)
- Employés administratifs, employés du commerce et des services aux personnes, ouvriers (actifs occupés)
- Employés administratifs, employés du commerce et des services aux personnes, ouvriers (anciens actifs)
- Autres (inactifs, chômeurs n'ayant jamais travaillé,…)

Sexe x Âge de la personne de référence

- **Nombre d'hommes (personne de référence) :**
  - de moins de 30 ans
  - de 30 à 39 ans
  - de 40 à 49 ans
  - de 49 à 50 ans
  - de 60 à 69 ans
  - plus de 70 ans

- **Nombre de femmes (personne de référence) :**
  - de moins de 30 ans
  - de 30 à 39 ans
  - de 40 à 49 ans
  - de 49 à 50 ans
  - de 60 à 69 ans
  - plus de 70 ans

Type du ménage

- Ménage d'une personne
- Couple sans enfant
- Famille monoparentale
- Couple avec 1 enfant
- Couple avec 2 enfants ou plus
- Autres

Type d'immeuble

- Immeuble individuel
- Autres

Nationalité (personne de référence)

- Française
- Autre

Zone de résidence des ménages

- Commune rurale (TUU=0)
  - Commune monopolarisée ou commune multipolarisée selon le ZAU 99
  - Espace à dominante rurale selon le ZAU 99
- Commune urbaine (TUU>0) -
  - Ville-Centre
  - Banlieue
- Unité urbaine de moins de 19.999 habitants
  - Ville-Centre
  - Banlieue
- Unité urbaine de 20.000 à 49.999 habitants
  - Ville-Centre
  - Banlieue
- Unité urbaine de 50.000 à 99.999 habitants
  - Ville-Centre
  - Banlieue
- Unité urbaine de 100.000 à 199.999 habitants
  - Ville-Centre
  - Banlieue
- Unité urbaine de 200.000 à 1.999.999 habitants
  - Ville-Centre
  - Banlieue
- Unité urbaine de Paris
  - Ville-Centre
  - Banlieue

Nombre d'individus

- **Nombre d'hommes**
  - de moins de 25 ans
  - de 25 à 34 ans
  - de 35 à 49 ans
  - de 50 à 64 ans
  - de plus de 65 ans
- **Nombre de femmes**
  - de moins de 25 ans
  - de 25 à 34 ans
  - de 35 à 49 ans
  - de 50 à 64 ans
  - de plus de 65 ans

Vague de l'enquête (6 vagues)

Région Ile-de-France

- **Âge de la personne de référence :**
  - de moins de 30 ans
  - de 30 à 39 ans
  - de 40 à 49 ans
  - de 49 à 50 ans
  - de 60 à 69 ans
  - plus de 70 ans
- **Taille du ménage**
  - 1 personne
  - 2 personnes
  - 3 personnes
  - 4 personnes
  - 5 personnes et +
- **Nationalité (personne de référence)**
  - Française
- **Diplôme le plus élevé obtenu**
  - Aucun diplôme ou CEP ou diplôme équivalent ou diplôme non déclaré
  - CAP, BEP ou autre diplôme de ce niveau ou BEPC seul
  - Bac ou brevet professionnel ou diplôme de ce niveau ou Bac+2 ans,
- **Type d'immeuble**
  - Immeuble individuel

Région Bretagne

- **Âge de la personne de référence :**
  - de moins de 30 ans
  - de 30 à 39 ans
  - de 40 à 49 ans
  - de 49 à 50 ans
  - de 60 à 69 ans
  - plus de 70 ans

- **Taille du ménage**
  - 1 personne
  - 2 personnes
  - 3 personnes
  - 4 personnes
- **Diplôme le plus élevé obtenu**
  - Aucun diplôme ou CEP ou diplôme équivalent ou diplôme non déclaré
  - CAP, BEP ou autre diplôme de ce niveau ou BEPC seul
  - Bac ou brevet professionnel ou diplôme de ce niveau ou Bac+2 ans,
- **Type d'immeuble**
  - Immeuble individuel
- **Zone de résidence des ménages**
  - Commune rurale (TUU=0)
    - Commune monopolarisée ou commune multipolarisée selon le ZAU 99
    - Espace à dominante rurale selon le ZAU 99
  - Commune urbaine (TUU>0) -
    - Ville-Centre
  - Unité urbaine de 200.000 à 1.999.999 habitants
    - Ville-Centre

Région Pays de la Loire

- **Âge de la personne de référence :**
  - de moins de 30 ans
  - de 30 à 39 ans
  - de 40 à 49 ans
  - de 49 à 50 ans
  - de 60 à 69 ans
  - plus de 70 ans
- **Taille du ménage**
  - 1 personne
  - 2 personnes
  - 3 personnes
  - 4 personnes
- **Diplôme le plus élevé obtenu**
  - Aucun diplôme ou CEP ou diplôme équivalent ou diplôme non déclaré
  - CAP, BEP ou autre diplôme de ce niveau ou BEPC seul
  - Bac ou brevet professionnel ou diplôme de ce niveau ou Bac+2 ans,
- **Type d'immeuble**
  - Immeuble individuel
- **Zone de résidence des ménages**
  - Commune rurale (TUU=0)
    - Commune monopolarisée ou commune multipolarisée selon le ZAU 99
    - Espace à dominante rurale selon le ZAU 99
  - Commune urbaine (TUU>0) -
    - Ville-Centre
  - Unité urbaine de 200.000 à 1.999.999 habitants
    - Ville-Centre
    - Banlieue

Région Midi-Pyrénées

- **Âge de la personne de référence :**
  - de moins de 30 ans
  - de 30 à 39 ans
  - de 40 à 49 ans
  - de 49 à 50 ans
  - de 60 à 69 ans
  - plus de 70 ans
- **Taille du ménage**
  - 1 personne
  - 2 personnes
  - 3 personnes
  - 4 personnes
- **Diplôme le plus élevé obtenu**
  - Aucun diplôme ou CEP ou diplôme équivalent ou diplôme non déclaré
  - CAP, BEP ou autre diplôme de ce niveau ou BEPC seul
  - Bac ou brevet professionnel ou diplôme de ce niveau ou Bac+2 ans,
- **Type d'immeuble**
  - Immeuble individuel
- **Zone de résidence des ménages**
  - Commune rurale (TUU=0)
    - Commune monopolarisée ou commune multipolarisée selon le ZAU 99
    - Espace à dominante rurale selon le ZAU 99
  - Commune urbaine (TUU>0) -
    - Ville-Centre
  - Unité urbaine de 200.000 à 1.999.999 habitants
    - Ville-Centre
    - Banlieue

Région Languedoc-Roussillon

- **Âge de la personne de référence :**
  - de moins de 30 ans
  - de 30 à 39 ans
  - de 40 à 49 ans
  - de 49 à 50 ans
  - de 60 à 69 ans
  - plus de 70 ans
- **Taille du ménage**
  - 1 personne
  - 2 personnes
  - 3 personnes
  - 4 personnes
- **Diplôme le plus élevé obtenu**
  - Aucun diplôme ou CEP ou diplôme équivalent ou diplôme non déclaré
  - CAP, BEP ou autre diplôme de ce niveau ou BEPC seul
  - Bac ou brevet professionnel ou diplôme de ce niveau ou Bac+2 ans,
- **Type d'immeuble**
  - Immeuble individuel

- **Zone de résidence des ménages**
  - Commune rurale (TUU=0)
    - Espace à dominante rurale selon le ZAU 99
  - Commune urbaine (TUU>0) -
    - Ville-Centre
  - Unité urbaine de 100.000 à 199.999 habitants
    - Ville-Centre
  - Unité urbaine de 200.000 à 1.999.999 habitants
    - Ville-Centre

3. Calage des individus Kish

Niveau national

PCS de la personne « Kish » (population des individus de 6 ans et plus)

- Agriculteurs (actifs occupés et anciens actifs)
- Indépendants ou professions libérales (actifs occupés et anciens actifs)
- Professeurs, instituteurs et assimilés, professions intermédiaires de la santé, autres cadres ou professions intermédiaires (actifs occupés)
- Professeurs, instituteurs et assimilés, professions intermédiaires de la santé, autres cadres ou professions intermédiaires (anciens actifs)
- Employés administratifs, employés du commerce et des services aux personnes, ouvriers (actifs occupés)
- Employés administratifs, employés du commerce et des services aux personnes, ouvriers (anciens actifs)
- Autres (inactifs, chômeurs n'ayant jamais travaillé,…)
- Individus de 6 à 15 ans

Sexe x Âge de la personne (population des individus de 6 ans et plus)

- Hommes
  - de 6 à 25 ans
  - de 25 à 34 ans
  - de 35 à 49 ans
  - de 50 à 64 ans
  - plus de 65 ans
- Femmes
  - de 6 à 25 ans
  - de 25 à 34 ans
  - de 35 à 49 ans
  - de 50 à 64 ans
  - plus de 65 ans

Taille du ménage (population des individus de 6 ans et plus)

- 1 personne
- 2 personnes
- 3 personnes
- 4 personnes
- 5 personnes et +

Zone de résidence (population des individus de 6 ans et plus)

- Commune rurale (TUU=0)
  - Commune monopolarisée ou commune multipolarisée selon le ZAU 99
  - Espace à dominante rurale selon le ZAU 99
- Commune urbaine (TUU>0) -
  - Ville-Centre
  - Banlieue
- Unité urbaine de moins de 19.999 habitants (1,2,3)
  - Ville-Centre
  - Banlieue
- Unité urbaine de 20.000 à 49.999 habitants (4)
  - Ville-Centre
  - Banlieue
- Unité urbaine de 50.000 à 99.999 habitants (5)
  - Ville-Centre
  - Banlieue
- Unité urbaine de 100.000 à 199.999 habitants (6)
  - Ville-Centre
  - Banlieue
- Unité urbaine de 200.000 à 1.999.999 habitants (7)
  - Ville-Centre
  - Banlieue
- Unité urbaine de Paris (8)
  - Ville-Centre
  - Banlieue

Jour de la semaine

Vague de l'enquête

Niveau régional (pour les 5 régions)
PCS de la personne « Kish » (population des individus de 6 ans et plus)
Sexe x Âge de la personne (population des individus de 6 ans et plus)

CONCLUSION

Dans toute enquête, pour éviter la non-réponse lors du recueil des données, il est crucial que les concepts utilisés soient clairs et compris de manière identique, tant par les personnes interrogées que par les enquêteurs. L'accès au logement des enquêtés étant de plus en plus difficile, de nouvelles méthodes de recueil des données doivent être développées. Nous devons aussi profiter des développements technologiques. L'utilisation du récepteur GPS dans l'Enquête Nationale sur les Transports et les Déplacements 2007-08 laissent présager la réalisation d'enquête presque totalement automatique ne nécessitant que peu de contact avec l'enquêteur. Les Enquêtes Nationales sur les Transports sont les seules à donner une description détaillée des déplacements des ménages résidant en France et de leur usage des moyens de transport tant collectifs qu'individuels. Les nouveautés méthodologiques et technologiques apportées pour à l'édition 2007-08 sont des atouts supplémentaires qui permettent d'augmenter la quantité et la qualité des informations recueillies.

Dans la théorie de la correction par repondération, la non-réponse totale ajoute une phase supplémentaire à l'échantillonnage : le mécanisme de réponse. Nous pouvons l'identifier à l'aide d'une modélisation qui fait apparaître les probabilités de réponse. Deville et Särndal (1992) ont généralisé les méthodes d'estimation utilisant de l'information auxiliaire à l'aide d'estimateurs par calage sur marges. Le calage sur marges permet de corriger les erreurs liées à la non-réponse et aussi d'augmenter la précision des estimateurs.

L'apurement de l'enquête se poursuivra ensuite par la correction de la non-réponse partielle et des erreurs de mémoire des enquêtés. Une erreur de mémoire est une omission ou une réponse erronée produite involontairement par personne interrogée. En effet, un volet rétrospectif est présent dans l'Enquête Nationale sur les Transports et les Déplacements 2007-08. Une personne est interrogée sur ses voyages à longue distance (plus de 80 kilomètres à vol d'oiseau) effectués entre la date de visite 1 et les trois derniers mois, en faisant appel à sa mémoire. Nous suspectons donc des erreurs de mémorisation du nombre de voyages faits et des erreurs de datation qu'il faudra corriger.

Les transports vont encore considérablement évoluer au XXIe siècle. L'étude de l'histoire des habitudes passées et des changements de comportement est indispensable pour éclairer les changements à venir. Les Enquêtes Nationales sur les Transports ont été menées au niveau national à cinq reprises donnant cinq photographies de la mobilité des Français à un moment donné. Elles sont riches en informations mais ne permettent pas, d'une part, de mesurer les changements individuels de comportement puisque les données sont recueillies environ tous les 10 ans et d'autre part, d'appréhender le récit d'une histoire de la mobilité. L'édition 2007-08 innove par rapport aux précédentes par sa volonté d'esquisser une perspective historique en proposant un questionnaire biographique, outil de recueil ordinairement réservé à la démographie.
L'analyse des biographies combinées à celles des ENT va nous permettre approfondir ces questions de changements de comportements.

BIBLIOGRAPHIE

OUVRAGES

Ardilly, P., 2006, *Les techniques de sondage*, Technip, 675p.

Armoogum, J., 2002, *Correction de la non-réponse et de certaines erreurs de mesures dans une enquête par sondage : Application à l'enquête Transports et Communications 1993-94*, Rapport INRETS, n 239, 173p.

CERTU, 2008, L'enquête ménages déplacements "Standard Certu", 204p.

Little, R.J.A., Rubin, D.B., 1987, Statistical Analysis with Missing Data, John Wiley, New York, 304p.

Platek R., Pierre-Pierre, F.K. et Stevens, P., 1985, Élaboration et conception des questionnaires d'enquête, Statistique Canada, Division des méthodes de recensement d'enquêtes-ménages, 83p.

Richardson, A.J., Ampt, E.S. et Meyburg, A., 1995, Survey Methods for Transport Planning, Eucalyptus Press, Melbourne, ??p.

ARTICLES

Armoogum, J., Madre, J.-L., 1997, « Du redressement des non-réponses totales aux contrôles sur la cohérence des réponses », Recherche Transport et Sécurité, N° 57, pp.67-77.

Bonnel P., Armoogum J., 2005, « National transport surveys –What can we learn from international comparisons », European Transport Conference 2005, Strasbourg.

Berthier, C., Dupont F., 1999, « L'incidence du caractère obligatoire des enquêtes », Insee Méthodes N°69-70-71, pp. 131-146.

Bilocq, F., 1996, « Conception et évaluation de questionnaires », Insee Méthodes N°69-70-71, pp. 77-92.

Brion, P., Caron, N., Pietri-Bessy, P., 2005, « Redresser la non-réponse totale dans les enquêtes auprès des entreprises : les pièges à éviter. Illustration avec l'enquête innovation », Document INSEE, Actes des Journées de Méthodologie Statistique, 9p.

Brilhault, G., Caron, N., 2004, « Correction de la non-réponse totale : par imputation ou par repondération ? », Documents INSEE - DSE, N° E2004/01, 76p.

Caron, N., 1993, « Réflexion sur les erreurs de mesure : l'exemple de l'enquête conjoncture auprès des ménages », Documents INSEE - DSDS, N° F9308.

Caron, N., 2005, « La correction de la non-réponse par repondération et par imputation », Série Documents de Travail INSEE Méthodologie Statistique, N° M0502, 48p.

Deming, W.E., Stephan, F.F., 1940, « On a least squares adjustment of a sampled frequency table when the exal totals are known », Annals of Mathematical Statistics, Vol. 11, pp. 427-444.

Deville, J.–C., 1998, « La correction de la non-réponse par calage ou par échantillonnage équilibré », Actes du colloque de la Société Statistique du Canada, Sherbrooke, Canada.

Deville, J.-C., Särndal, C.E., 1992, « Calibration estimators and generalized raking techniques in survey sampling », Journal of the Americain Statistical Association, Vol. 87, pp. 376-382.

Deville, J.-C. , Särndal, C.E., Sautory, O., 1993, « Generalised raking procedures in survey sampling », Journal of the Americain Statistical Association, Vol. 88, pp. 1013-1020.

Deville, J.-C. et Särndal, C.E. (1994), « Variance estimation for the regression imputed Horvitz-Thompson estimator », Journal of Official Statistics, Vol. 10, N° 4 pp. 381-394.

Deville, JC., Dupont, F., 1996, « Non-réponse : principes et méthodes », Insee Méthodes N°56-57-58, pp. 53-69.

Desrosieres A., 2003, « Historiciser l'action publique. L'Etat, le marché et les statistiques », in Laborier, P., Trom, D. (dir.), Historicités de l'action publique, Paris, PUF, pp. 207-221.

Dupont, F., 1994, « Imputation procedures for quantitative and qualitative variables », document INSEE - DSDS, N° F9406.

Glaude, M., 2000, « Les enquêtes auprès des ménages à l'Insee : petit bilan et perspective », Courrier des statistiques N°95-96, pp. 39-52.

Le Guennec, J. et Sautory, O., 2003, « La macro Calmar2, manuel d'utilisation », document interne INSEE.

Oh, H. L., Scheuren, F. J., 1983, « Weighting Adjustment for Unit non-response », in W. G. Madow, I. Olkin, and D. B. Rubin (eds), Incomplete data in Sample Surveys, Vol. 2: Theory and Bibliographies, New York: Academic Press, pp. 143-184.

Platek, R., Gray, G.B., 1983, « Imputation methodology : Total survey error », in W. G. Madow, I. Olkin, and D. B. Rubin (eds), Incomplete Data in Sample Surveys, Vol II : Theory and Bibliographies, New York: Academic Press, pp. 249-333.

Sautory, O., 1993, « La macro CALMAR : redressement d'un échantillon par calage sur marges », INSEE Document de travail, N° F9310, 56p.

Stephan, F. F., 1942, « An iterative method of adjusting sample frequency tables when expected marginal totals are known », Annals of Mathematical Statistics, Vol. 13, pp. 166-178.