RENOVATION DE L'ENQUETE CAMME : de l'échantillonnage au calcul de précision, en passant par l'allocation optimale entre les enquêteurs.

Marc CHRISTINE (*), Laurent DAVEZIES (**)[1], Matthieu MORANDO (***) et Sylvie ROUSSEAU (*)

(*) INSEE, DSDS, Unité Méthodes Statistiques
(**) Direction de l'Evaluation et de la Prospective (DEP)
(***) INSEE, Direction Régionale d'Alsace

Introduction

Depuis la fin des années cinquante, l'Insee mène régulièrement une enquête de conjoncture auprès des ménages. Cette enquête a pour objectif de recueillir l'opinion que se font les ménages de la situation économique générale de la France et de leur situation financière personnelle, ainsi que leurs intentions en matière d'épargne et d'achats de biens d'équipement. Ces informations sont destinées à fournir une aide au diagnostic conjoncturel de l'économie française et européenne. Dans ses éditions récentes (i.e., depuis 1989), cette enquête, baptisée CAMME, est réalisée mensuellement.

Une refonte de cette enquête a été conduite en 2003, sous le triple objectif d'adapter le questionnaire aux nouvelles demandes européennes, d'améliorer les traitements statistiques amont et aval et de moderniser la chaîne informatique devenue obsolète.

Une partie introductive s'attachera à la présentation du scénario d'enquête, dans son contexte européen, et de l'organisation de la collecte. Nous décrirons ensuite successivement les phases d'échantillonnage, d'allocation des fiches-adresses aux enquêteurs et de traitements statistiques aval. L'avant-dernière partie sur les calculs de précision effectués, tenant compte des particularités du plan de sondage et des traitements aval, nous amènera, en conclusion, à évoquer rapidement les perspectives d'amélioration de cette enquête, centrale sur le plan du diagnostic conjoncturel, et les travaux d'étude envisagés à court et moyen terme.

[1] Etait à l'UMS au moment de la réalisation de ce travail.

1. Présentation générale de l'enquête

Afin de suivre, voire d'anticiper le comportement de consommation des ménages, l'Union européenne demande aux Etats-Membres de réaliser chaque mois une enquête de conjoncture auprès des ménages.

1.1. Historique des enquêtes de conjoncture à l'Insee

L'enquête actuelle de ***Conjoncture Auprès des Ménages MEnsuelle (CAMME***), ou enquête de conjoncture auprès des consommateurs, selon la désignation européenne, s'inscrit dans un dispositif mis en place par l'Insee en 1958. Les enquêtes de conjoncture auprès des ménages ont poursuivi dès l'origine des objectifs complémentaires : d'une part, recueillir l'opinion des ménages sur certains indicateurs conjoncturels ; étudier, d'autre part, le parc et les intentions d'achats de biens d'équipement des ménages, et, enfin, les vacances des Français. Semestrielles à l'origine, puis quadrimestrielles à partir de 1965, ces enquêtes sont réalisées de façon harmonisée dans le cadre communautaire européen depuis 1972.

A la demande de la Commission Européenne a été lancée en 1986 une enquête mensuelle de conjoncture - baptisée alors CAMME - dont le questionnaire est une reprise de la partie conjoncturelle de l'enquête quadrimestrielle, adaptée à la demande européenne. Expérimentale jusqu'en mai 1989, cette enquête a acquis à cette date le statut d'opération permanente de l'Insee (cf. [5]). Les moyens de collecte et d'exploitation, définis et mis en place pour une opération expérimentale, ont fait l'objet d'une réflexion approfondie conduite jusqu'à la fin de 1990. L'enquête rénovée est réalisée à partir de janvier 1991. C'est cette enquête qui produit aujourd'hui, mensuellement, les statistiques mensuelles d'opinion des ménages sur la conjoncture économique.

1.2. Utilisation de l'enquête mensuelle de conjoncture

La Commission Européenne (CE) intègre depuis 1972 l'enquête CAMME parmi les 11 enquêtes de conjoncture réalisées par l'Insee au sein du Système Européen Harmonisé des enquêtes de conjoncture. La signature d'un nouveau contrat avec la CE, prenant effet en 2003, a conduit l'INSEE à approfondir l'harmonisation des questionnaires et à introduire de nouvelles questions. Ainsi, les questions de conjoncture, au cœur du questionnaire, sont une traduction française du questionnaire européen mis au point par la Commission. L'Insee fournit à la CE, dans le cadre d'un calendrier précis, les données anonymisées de l'enquête, sous forme de tableaux et de fichiers détaillés.

Parallèlement, la division des Comptes Trimestriels de l'Insee réalise certains traitements sur les données issues de CAMME : solde d'opinion, désaisonnalisation. La publication mensuelle dans « Informations Rapides », fréquemment commentée dans les media sous la désignation de « moral des Français », est sous sa responsabilité. La division des Comptes Trimestriels utilise ces informations pour le diagnostic conjoncturel et les prévisions à court terme de consommation des ménages.

Enfin, l'enquête CAMME est le support de « micro-plates-formes » composées de 10 questions, permettant d'aborder des thèmes d'actualité dans des délais rapides. Ces plates-formes sont utilisées, en général, deux à trois fois par an. Les dernières plates-formes ont porté sur des thèmes aussi divers que « les Français et l'euro », la détention d'armes à feu, l'utilisation des crédits d'impôts, les intentions de départs en vacances des Français. La prochaine plate-forme, associée à l'enquête d'avril 2005, portera sur l'opinion des salariés sur leurs salaires, dans le cadre de la préparation d'une enquête plus vaste sur ce thème.

1.3. La refonte de l'enquête CAMME

Comme il a été dit, l'un des objectifs de la refonte de l'enquête était d'adapter le questionnaire aux nouvelles demandes européennes.

Le nouveau questionnaire est composé de quatre parties :

- une partie introductive, le « module de contact avec le répondant », qui tient compte de la particularité du mode de collecte téléphonique de cette enquête (point précisé en infra, § 1.5).

- le module « conjoncture », au cœur du dispositif, s'articule autour d'une partie « opinion sur la conjoncture générale en France », sur les 12 mois passés et sur les 12 mois à venir (situation économique générale, chômage, inflation…) et d'une partie « opinion sur la situation économique du ménage » également passée et à venir (niveau de vie, opportunité d'épargner, intentions d'achat d'un logement, d'une voiture…).

- le module « socio-démographique » décrit succinctement le ménage et les caractéristiques du répondant.

- éventuellement, une plate-forme de dix questions, utilisée ponctuellement pour éclairer des questions d'actualité ou liées à la conjoncture.

C'est dans le module « conjoncture » qu'ont été introduites les principales modifications : alignement des périodes de référence à 12 mois, plus grande proximité des libellés à la version anglaise, et surtout ajout de deux questions quantitatives sur les prix : opinion chiffrée en % sur l'évolution des prix au cours des 12 mois passés, et au cours des 12 mois à venir.

1.4. Mise en œuvre et éléments de méthode

En France, le scénario de l'enquête CAMME est invariant depuis plus d'une dizaine d'années. ***Il s'agit de la seule enquête Insee intégralement réalisée par téléphone.***

L'échantillon des personnes contactées est constitué à partir de numéros de téléphone tirés dans la base des abonnés à France-Télécom. L'échantillon interrogé mensuellement est constitué de trois sous-parties. Le tiers entrant (d'effectif 1.100) correspond aux ménages contactés pour la première fois, le tiers médian à ceux qui répondent pour la deuxième fois, enfin le tiers sortant est composé des ménages enquêtés pour la troisième et dernière fois.

Les ménages sont ainsi interrogés trois mois de suite par des enquêteurs assermentés de l'Insee. Les résultats de l'enquête, obligatoire et reconnue d'intérêt général, sont protégés par le secret statistique et couverts par la loi relative à l'informatique, aux fichiers et aux libertés (loi de 1978 modifiée).

Le mode d'élaboration de chacun des 15 indicateurs de conjoncture retenus repose sur le calcul d'un solde entre le pourcentage d'opinions positives et le pourcentage d'opinions négatives. Par exemple, pour apprécier la perception de l'évolution passée du niveau de vie, on additionne les réponses faisant état d'une amélioration et l'on soustrait celles qui font état d'une détérioration. Les réponses neutres, du type niveau de vie stationnaire, ne sont pas directement prises en compte dans le calcul. Les résultats transmis à la Commission Européenne sont déclinés par sexe, tranches d'âges, catégorie sociale ou encore tranche de revenu.

Enfin, un indicateur composite, calculé par la Division des Comptes Trimestriels comme moyenne arithmétique de l'opinion sur le niveau de vie passé et futur, de la situation financière passée et future et de l'opportunité de faire des achats importants, est élaboré de façon à disposer d'une vision synthétique du « moral des Français ».

1.5. Organisation de la collecte

Les enquêtes sont réalisées chaque mois sauf en août. La collecte se déroule sur 2 à 3 semaines en début du mois et est gérée seulement par 8 Directions Régionales de l'Insee. Le calendrier de l'enquête est fixé et transmis en décembre de l'année précédente. Les dates de fin de collecte fixées sont à respecter impérativement. Cette enquête fait en effet l'objet d'un contrat avec la Commission Européenne, à qui l'Insee s'engage à fournir les données relatives à chaque mois d'enquête avant la fin de celui-ci.

La panélisation d'un ménage se fait à partir du numéro de téléphone à la première vague d'interrogation. Le ménage panélisé sera le ménage auquel appartient le titulaire de la ligne téléphonique correspondant au numéro échantillonné. La personne interrogée est alors obligatoirement le titulaire de cette ligne, ou son éventuel conjoint. De plus, le numéro de téléphone de contact devra être celui de la résidence principale du ménage. Dans le cas d'un transfert de la ligne téléphonique de la résidence principale vers la ligne de la résidence secondaire ou vers un téléphone mobile, le ménage sera cependant interrogé. C'est l'objet du « module de contact avec le répondant » évoqué plus haut (§ 1.3) de bien déterminer le répondant.

De manière générale, la question que se pose l'enquêteur pour savoir si la personne doit être interrogée est la suivante : « est-ce la seule ligne fixe accessible et couramment utilisée ? » En particulier, si un ménage dispose de deux lignes (privée et professionnelle, ou principale et secondaire…), il ne faut l'interroger que si c'est son numéro privé en résidence principale qui a été échantillonné. Le problème n'est pas tant le risque que le ménage soit échantillonné deux fois (c'est-à-dire que les deux numéros soient tirés), que la génération d'un biais statistique, de tels ménages ayant une probabilité deux fois plus grande de faire partie de l'échantillon.

1.6. Caractéristiques particulières de l'enquête

La méthodologie de l'enquête mensuelle de conjoncture auprès des ménages comporte certaines particularités, liées à son mode de collecte et à son caractère conjoncturel, qui méritent d'être soulignées, d'autant que quelques unes d'entre elles impliquent une définition particulière de certaines règles de collecte.

1.6.1. Particularités liées à la collecte par téléphone.

Les ménages à enquêter sont tirés dans l'annuaire électronique du téléphone. De ce fait :

- certaines catégories de ménages ne seront jamais enquêtées, à savoir :

  . les ménages qui ne possèdent pas le téléphone
  . les ménages sur liste d'opposition (rouge ou orange)
  . les ménages disposant uniquement d'un téléphone GSM.

- le ménage est identifié par son numéro de téléphone et non plus par son logement comme c'est le cas des enquêtes réalisées en face à face dont la base de sondage est constituée par les listes de logements issues des recensements de population (et complétées par les logements « neufs »).

- l'identification du chef de ménage[2] doit se faire au sein de la « cellule familiale » titulaire de la ligne téléphonique et non de l'ensemble du ménage comme c'est le cas en ce qui concerne l'identification de la personne de référence. Le chef de ménage sera donc obligatoirement soit le titulaire de la ligne soit son conjoint, même dans le cas où le titulaire de la ligne ne serait pas la personne de référence du ménage au sens du recensement. Les différences entre les concepts devraient toutefois être minimes.

[2] Notion préférée à celle de personne de référence en raison du type de collecte.

1.6.2. Particularités liées au caractère conjoncturel de l'enquête.

- l'utilisation des résultats de l'enquête à des fins d'analyse conjoncturelle implique une certaine stabilité des échantillons de ménages enquêtés. Chaque ménage enquêté est ainsi panélisé sur trois mois, chaque mois d'interrogation constituant une vague d'enquête. Les échantillons sont renouvelés par tiers à chaque enquête mensuelle, un tiers de ménages nouvellement tirés entrant dans l'échantillon en remplacement d'une proportion équivalente de ménages ayant déjà fait l'objet de trois interrogations successives.

- la nature conjoncturelle de l'enquête et son statut d'enquête d'opinion impliquent, plus que dans tout autre enquête ménages, un strict respect du libellé des questions lors de l'interrogation du ménage, sans reformulation. La réponse « ne sait pas », éventuellement générée par une mauvaise compréhension de la question, est ici une modalité de réponse à part entière. Elle n'est cependant pas proposée par l'enquêteur comme modalité possible, mais sélectionnée en dernier recours.

- en moyenne, cinq jours seulement séparent la fin de la collecte de la transmission des résultats aux utilisateurs ; les délais de traitement de l'enquête sont donc serrés, et le calendrier de collecte doit être strictement respecté par les enquêteurs et les gestionnaires.

2. Spécifications statistiques de l'échantillonnage.

2.1. Taille de l'échantillon.

Pour une année de collecte, 11 échantillons mensuels[3] de 1.100 numéros de téléphone constituent la fraction « entrante ». Au total, l'échantillon mensuel théorique est donc de 3.300 unités. Notons que, le module socio-démographique n'étant posé que lors de la première interrogation, les non-répondants (absents de longue durée, inaptes, refus…) de la première vague ne sont pas réintroduits lors des deux vagues suivantes. Il faut donc compter sur environ 2.600 numéros interrogés, et entre 2.000 et 2.200 questionnaires renseignés par mois.

Les tirages de ces échantillons sont effectués chaque mois à partir de la base de données Wanadoo Data (filiale de France-Télécom en charge de la commercialisation des bases d'abonnés), par leurs soins et selon les spécifications de l'Insee.

La base de sondage Wanadoo Data, riche d'environ 18 millions de foyers de particuliers joignables, est constituée à partir des données annuaires de France-Télécom (pages blanches), expurgée notamment des listes d'opposition (5 millions d'abonnés en liste rouge et 830.000 en liste orange).

Cette base de données, sur le champ restreint de la France métropolitaine, est alimentée par les mouvements quotidiens enregistrés par « France-Télécom annuaires » à partir des créations, radiations et modifications de lignes transmises par l'ensemble des agences commerciales. Elle est expurgée au préalable des listes d'opposition, des doublons (même numéro de téléphone avec deux titulaires distincts, en général les deux membres d'un même couple…). Il s'agit bien évidemment d'une base « ménages » ou « foyers », ne comportant aucun n° d'entreprise[4] ni aucun n° de fax, et aucun n° de téléphone mobile.

[3] On rappelle que l'enquête n'a pas lieu en août.
[4] Les professions libérales, ayant un même numéro à caractère personnel et professionnel, sont bien entendu conservées dans le champ de l'enquête.

2.2. Discussion sur la base de sondage et le mode de collecte

Il faut insister sur le fait, évoqué au § 1.6.1, que nos échantillons sont tirés dans une base « incomplète », au sens où ***les abonnés sur liste orange ou rouge (listes d'opposition), les non-abonnés ou les titulaires d'une ligne mobile uniquement ne peuvent être sélectionnés.***

L'hypothèse est faite que les biais qui résultent de la non-exhaustivité de la base peuvent être corrigés par des corrections de structure, c'est-à-dire qu'à catégorie donnée (définie par les variables socio-démographiques habituelles), les comportements des « ménages » interrogés sont identiques, qu'ils figurent ou pas sur la base des abonnés de Wanadoo Data. Les ménages absents de la base de données sont donc supposés pouvoir être représentés par les unités tirées. Cette hypothèse est évidemment forte mais incontournable, dans le cadre d'une enquête légère, rapide et pas trop coûteuse.

Par ailleurs, il est possible pour un même ménage de disposer de plusieurs lignes téléphoniques, soit à son domicile principal, soit dans celui-ci et dans une résidence secondaire : en théorie, ceci entraîne une probabilité d'inclusion plus élevée pour de tels ménages. Pour résoudre ce problème, un filtre dans le questionnaire permet de vérifier que le n° de téléphone composé correspond bien à une ligne installée dans une résidence principale et non dans une résidence secondaire. *Nous avons convenu en effet de ne pas utiliser la variable « résidence secondaire », récupérée dans la base Wanadoo Data mais jugée peu fiable.*

Plus généralement, c'est la notion même d'*unité statistique* qui pose des problèmes de définition ou, en tout cas, d'identification, dans le cas d'un tirage sur une base de numéros de téléphone. Ce phénomène est bien connu et avait été pointé dès l'expérimentation initiale de CAMME : l'unité statistique est l'ensemble des personnes pouvant être jointes à un numéro de téléphone ; elle ne se réduit pas nécessairement à un ménage au sens statistique traditionnel, du fait notamment de l'individualisation du mode de possession du téléphone, accentuée et amplifiée ces dernières années par la montée en charge de la téléphonie mobile.

Une solution à ce problème pourrait être d'ajouter au questionnaire un module visant à préciser le contour de l'unité statistique (en interrogeant sur le nombre de lignes utilisées par les membres du ménage), et de redresser les réponses en conséquence.

L'ensemble de ces difficultés ne sont pas méconnues. Cependant, elles sont difficilement contournables dans le cas d'une enquête téléphonique qui doit rester légère. Les alternatives seraient en effet coûteuses.

La solution la plus satisfaisante, répondant tant à l'incomplétude de la base de sondage qu'aux difficultés de définition de l'unité statistique, consisterait à renoncer au téléphone au profit d'une enquête en face-à-face classique (au moins pour la première interrogation). Une solution intermédiaire pourrait consister à tirer l'échantillon dans le Recensement (ou l'échantillon-maître), puis rechercher les numéros de téléphone dans l'annuaire, interroger par téléphone les ménages que l'on retrouve ainsi et par visite les autres (sachant que, parmi ces derniers, il y en a une partie qui auront déménagé après le dernier Recensement…). Un intérêt annexe de cette méthode « mixte » serait de pouvoir vérifier les biais dus aux listes d'opposition en comparant les réponses des deux sous-populations.

Toutefois, le tirage d'un échantillon dans le Recensement (ou l'échantillon-maître), solution naturelle pour une enquête en face-à-face, présenterait en contrepartie aussi des contraintes par rapport à un échantillon téléphonique : notamment, le tirage d'un degré intermédiaire d'unités primaires permet une certaine concentration des points d'enquête favorisant une réduction des frais de déplacement (mais pas utile dans le cadre d'une enquête téléphonique…), mais il génère de la variance supplémentaire et ne paraît donc pas opportun dans ce cadre. De surcroît, la possibilité de disposer d'une base téléphonique mise à jour en permanence dispense de rechercher des solutions complémentaires de type « logements neufs », qui sont nécessaires quand le Recensement devient ancien par rapport à la date de collecte.

Nous avons dans un premier temps renoncé à étudier ces alternatives, compte-tenu des coûts qui en résulteraient et des délais de la refonte (imposés par la Commission européenne), mais aussi parce que les résultats obtenus jusqu'à aujourd'hui à partir de CAMME sont en fait satisfaisants, notamment dans l'optique de l'analyse conjoncturelle et pour l'emploi de ses résultats dans des modèles économétriques : ceci suggère ainsi que les biais éventuels introduits tant par la base de sondage que par l'unité statistique n'ont que peu d'impact sur les variables principales d'intérêt, c'est-à-dire les évolutions des soldes d'opinion. Celles-ci sont en effet très fortement corrélées à la conjoncture économique, et constituent un bon prédicteur des variations de la consommation. Néanmoins, il pourra être souhaitable d'examiner, dans l'avenir, des solutions alternatives à ces problèmes.

2.3. Critères de stratification de l'échantillon.

2.3.1. Une stratification explicite autour de deux critères :

2.3.1.1. Typologie de commune.

On utilise le ***code TU99*** habituellement utilisé[5] dans les enquêtes Insee construites à partir de l'échantillon-maître. Pour cela, Wanadoo Data a implémenté dans ses bases une table de correspondance commune ✶ TU99 livrée par l'Insee. La stratification utilise in fine les modalités d'un code agrégeant le code TU99 en 6 modalités.

2.3.1.2. Verticalité

Un code « Verticalité de l'habitat » est obtenu par Wanadoo Data en cumulant le nombre d'abonnés (intégrant bien entendu ceux qui sont en liste d'opposition) par point de distribution. Il modélise le nombre d'abonnés habitant à une même adresse, et permet donc de distinguer par proxy habitat individuel (1 seul n° à l'adresse) et habitat collectif (plusieurs n°s).

2.3.2. Une stratification implicite.

Comme le tirage des échantillons est réalisé par Wanadoo Data, qui conserve bien évidemment la maîtrise et la propriété de la base de données, on a cherché à simplifier les procédures, en évitant des choix peut-être meilleurs sur le plan statistique mais plus complexes, comme par exemple imposer des conditions d'équilibrage qui requerraient l'utilisation de la macro CUBE...

Les tirages se font donc strate par strate grâce à une procédure de tirage systématique dans un fichier trié au préalable selon certaines variables « secondaires » qui vont servir de stratification « implicite ». Deux critères, parmi ceux disponibles dans la base Wanadoo Data, sont pris en compte, par importance décroissante :

2.3.2.1. en majeur, un critère géographique :

Le fichier est trié par code région puis, à l'intérieur de chaque région, les communes sont classées par valeurs croissantes du taux de pénétration de la liste rouge[6].

Un tel tri permet d'assurer une bonne dispersion géographique ainsi qu'une bonne représentation de l'ensemble des zones, en fonction de l'intensité de pénétration de la liste rouge. La dispersion

[5] Une autre *variable « Typologie de commune » figure dans le fichier Wanadoo Data, qui reprend le zonage de l'Insee en « aires urbaines et zones d'emploi pour l'espace rural » (7 classes, distinguant pôles urbains, pôles ruraux, périphéries des pôles ruraux…). Mais elle n'est pas utilisée pour le tirage de l'échantillon.*

[6] Par extension de langage, on a désigné dans ce papier par « liste rouge » l'ensemble des listes d'opposition.

géographique permet aussi de faciliter la prise en compte d'un critère de proximité lors de l'affectation des n°s échantillonnés aux enquêteurs (module ALLOCAMME, décrit en troisième partie).

2.3.2.2. en mineur, un code reflétant le niveau de revenu :

Wanadoo Data emploie un code typologique constitué d'une cinquantaine de classes. Il s'agit d'une typologie au numéro dans la voie, qui intègre des données socio-démographiques, urbanistiques et géographiques. Cette variable est codée sur trois positions dont les deux extrêmes reflètent des variables déjà prises en compte dans la stratification (catégorie de commune et type d'habitat). A chaque modalité élémentaire de la variable est affectée une estimation de revenu moyen. C'est cette estimation qui sert de critère de tri.

Comme pour le premier critère, l'utilisation de cette variable agrégée comme critère de tri secondaire permet d'opérer une bonne dispersion entre les différentes modalités, de façon à améliorer la représentation du niveau de revenu.

2.4. Allocation de l'échantillon entrant entre les strates, détermination des probabilités d'inclusion des unités et calcul des poids à affecter aux unités tirées.

Ces trois questions sont liées et ne se confondent pas. On est en effet dans une situation où il faut bien différencier :

- la probabilité de sélection d'une unité dans la base de sondage
- et le poids que l'on attribuera à chaque unité tirée (il s'agit ici du poids initial, en amont de toute procédure de redressement).

Dans l'application CAMME antérieure à la refonte de 2003, le tirage était fait en deux phases : une première phase réalisée par Wanadoo Data assurait la fourniture d'un échantillon de taille importante destiné à alimenter tous les échantillons mensuels entrants d'une année, ces derniers étant tirés lors d'une seconde phase mise en œuvre chaque mois par l'Insee.

Comme ce tirage en deux phases ne se justifiait plus, ni sur le plan théorique, ni sur le plan pratique, le plan de sondage a été restreint à une phase, chaque tirage mensuel étant assuré directement par Wanadoo Data. Ceci permet en outre de disposer à chaque tirage de la base de données la plus à jour possible, et, partant, d'éviter une augmentation régulière au cours de l'année du nombre de n°s non attribués.

Les efforts mis en œuvre lors de cette refonte pour améliorer le plan de sondage par rapport aux éditions antérieures de CAMME ont porté principalement sur la correction des défauts de structure de l'échantillon dus à la non-couverture des listes d'opposition.

➔ Pour bien comprendre comment les choses se passent, considérons l'exemple d'une strate où le taux de pénétration de la liste rouge serait constant, égal à τ .

• Supposons tout d'abord que l'on réalise un sondage aléatoire simple sans remise de n unités dans une base $\mathcal{P}_1$ d'effectif N_1 (représentant la base des abonnés joignables, c'est-à-dire qui ne figurent pas en liste d'opposition). On suppose que la population totale (ensemble des abonnés) dans la strate est N, et on fait l'hypothèse que les N_1 unités de la base de sondage représentent correctement la population totale $\mathcal{P}$. On a donc : $\frac{N_1}{N} = 1-\tau$. La probabilité de tirage d'une unité sera donc $\frac{n}{N_1} = \Pi$, mais le poids à affecter à cette unité sera de $\frac{N}{n}$, soit $\omega = \frac{N}{N_1} \times \frac{N_1}{n} = \frac{1}{\Pi(1-\tau)}$.

• Plus généralement, si l'on tire dans la population $\mathcal{P}_1$ n unités, chacune avec une probabilité de tirage Π_i, le poids affecté à ces unités devra être $\frac{1}{(1-\tau)\,\Pi_i}$.

Considérons alors l'estimateur :

$$\hat{T}_1 = \sum_{i\in s} \frac{1}{\Pi_i} \cdot Y_i = \sum_{i\in P_1} \frac{Y_i}{\Pi_i} \mathbf{1}_{i\in s}$$

Son espérance (vis-à-vis de la loi de tirage) sera alors :

$$E\hat{T}_1 = \sum_{i\in P_1} Y_i = T_1(Y)$$

où T_1 (Y) est le total de Y sur la population $\mathcal{P}_1$.

L'hypothèse d'homogénéité des comportements entre $\mathcal{P}$ et $\mathcal{P}_1$ ou de bonne « représentation » de $\mathcal{P}$ par $\mathcal{P}_1$ peut se traduire par l'identité des moyennes $\frac{1}{N}\sum_{i\in P} Y_i$ et $\frac{1}{N_1}\sum_{i\in P_1} Y_i$.

Un estimateur sans biais $\hat{T}$ du total de Y sur $\mathcal{P}$ doit donc être relié à l'estimateur de ce même total sur $\mathcal{P}_1$ par la relation $\hat{T} = \frac{N}{N_1}\hat{T}_1$

Si $\hat{T}_1$ est un estimateur sans biais de T_1 (Y), on aura : $E\hat{T} = \frac{N}{N_1} T_1(Y) = T(Y)$.

Un estimateur sans biais de T_1 (Y) sera donc :

$\hat{T} = \frac{1}{1-\tau}\sum_{i\in s} \frac{Y_i}{\Pi_i}$, puisque : $\frac{1}{1-\tau} = \frac{N}{N_1}$.

➔ On comprend aisément comment ceci peut se généraliser au cas où les taux de pénétration de la liste rouge dépendent de l'unité i ou, plus précisément, de l'adresse, de la commune ou de l'IRIS d'appartenance de l'unité i, et comment on va pouvoir en déduire des règles d'allocation de l'échantillon ou de calcul des poids.

Notons τ_i le taux de pénétration de la liste rouge pour l'unité i.

Si chaque unité i est tirée avec une probabilité Π_i, le poids qui lui sera affecté sera $\boxed{\omega_i = \frac{1}{\Pi_i(1-\tau_i)}}$

Cela signifie que, à probabilité d'inclusion donnée, on augmentera le poids des unités en fonction du taux de pénétration de la liste rouge qui leur est affecté : plus ce dernier est important, plus grand sera le poids.

Intuitivement, cela revient à dilater « localement » la variable d'intérêt pour extrapoler de la population des lignes joignables à celle de l'ensemble des lignes.

Un estimateur du total T(Y) relatif à l'ensemble de la population sera donc :

$$\hat{T} = \sum_{i\in s} \omega_i Y_i = \sum_{i\in s} \frac{1}{\Pi_i(1-\tau_i)} Y_i = \sum_{i\in P_1} \frac{1}{\Pi_i(1-\tau_i)} \mathbf{1}_{i\in s} Y_i$$

Son espérance sera :

$$E\hat{T} = \sum_{i \in P_1} \frac{1}{\Pi_i (1-\tau_i)} Y_i \underbrace{E(\mathbf{1}_{i \in s})}_{=\Pi_i} = \sum_{i \in P_1} \frac{Y_i}{(1-\tau_i)}$$

L'hypothèse d'homogénéité des comportements dans $\mathcal{P}$ et $\mathcal{P}_1$ se traduira par le fait que $\sum_{i \in P_1} \frac{Y_i}{(1-\tau_i)} = \sum_{i \in P} Y_i$, ce qui assure que : $E\hat{T}$ = T (Y).

➔ Finalement, la solution préconisée sera donc :

- tirer des échantillons s_h dans chacune des strates $\mathcal{P}_{1,h}$ de la population $\mathcal{P}_1$, avec des probabilités d'inclusion Π_i
- affecter aux unités les poids $\omega_i = \frac{1}{\Pi_i (1-\tau_i)}$
- et considérer l'estimateur : $\hat{T} = \sum_{i \in s} \omega_i Y_i$.

L'estimateur sous la forme écrite ci-dessus n'est pas un estimateur de HORVITZ-THOMSON en les variables Y_i mais il en devient un si l'on considère que la variable d'intérêt est en fait : $\frac{Y_i}{1-\tau_i}$.

<u>Quel type d'échantillon constituer au sein de chaque strate ?</u>

Plusieurs possibilités :

a) sondage aléatoire simple de n_h unités parmi N_{1h}

$\Pi_i = \frac{n_h}{N_{1h}}$ pour $i \in \mathcal{P}_{1,h}$

Avantage : faire réaliser un tirage à probabilités égales dans chaque strate. C'est également une solution plus simple à mettre en œuvre par Wanadoo Data.

Inconvénient : les poids des unités sont différents : $\omega_i = \frac{N_{1h}}{n_h} \frac{1}{(1-\tau_i)}$.

Dans ce cadre, l'allocation à réaliser entre les strates peut être :

- soit de type *allocation proportionnelle* (par rapport à l'effectif de la base de sondage, c'est-à-dire des n°s hors liste rouge, ou de la population d'inférence, c'est-à-dire de l'ensemble des n°s).

- soit de type *NEYMAN*, laquelle devrait être calculée en prenant en compte, dans le calcul de la variance des estimateurs du total d'une variable directrice donnée, les poids (non aléatoires) ω_i et non les inverses des probabilités d'inclusion.

Ceci pose toutefois deux problèmes :

- quelles variables choisir pour le calcul de variance ?
- il faudrait pouvoir disposer des (vraies) variances des dites variables en intra dans chaque strate. Comme on n'a jamais utilisé de telle stratification, c'est le principal obstacle : en fait, il faudrait collecter des données tests dans le cadre de cette stratification pour estimer les variances et adapter ensuite l'allocation.

b) sondage à probabilités inégales à l'intérieur de chaque strate.

Problème : comment choisir alors les probabilités de sélection ? Un choix possible consiste à chercher à *assurer l'identité des poids des unités tirées.*

Soit ω_h le poids (constant) dans la strate $\mathcal{P}_{1,h}$.

On cherche alors à réaliser : $\omega_h = \frac{1}{\Pi_i (1-\tau_i)}$ pour i ∈ $\mathcal{P}_{1,h}$.

D'où $\Pi_i = \frac{1}{\omega_h (1-\tau_i)}$ pour i ∈ $\mathcal{P}_{1,h}$.

Avantages et inconvénients « duaux » de ceux de la solution précédente.
Inconvénient additionnel : les probabilités inégales rendent plus compliqué le calcul de la variance des estimateurs.

Avantage additionnel : avec cette solution, il est possible de réaliser une *allocation de l'échantillon qui rendent égaux tous les poids finaux.*

En effet, à l'intérieur de la strate $\mathcal{P}_{1,h}$, on a la relation classique : $n_h = \sum_{i \in P_{1,h}} \Pi_i$, d'où, si l'on note ω la valeur commune du poids final : $n_h = \frac{1}{\omega} \sum_{i \in P_{1,h}} \frac{1}{(1-\tau_i)}$.

Par sommation sur toutes les strates, on obtiendra : $n = \frac{1}{\omega} \sum_h \sum_{i \in P_{1,h}} \frac{1}{(1-\tau_i)}$.

Cette équation donne alors la valeur de ω en fonction de la taille fixée d'échantillon, n, soit :

$$\omega = \frac{1}{n} \sum_k \sum_{i \in P_{1,k}} \frac{1}{(1-\tau_i)}.$$

Par ailleurs, on en déduit l'allocation :

$$n_h = n \frac{\sum_{i \in P_{1,h}} \frac{1}{(1-\tau_i)}}{\sum_k \sum_{j \in P_{1,k}} \frac{1}{(1-\tau_j)}}$$

Les probabilités d'inclusion Π_i sont ensuite données par :

$$\Pi_i = n \frac{1}{1-\tau_i} \frac{1}{\sum_k \sum_{j \in P_{1,k}} \frac{1}{1-\tau_j}}$$

Ces formules générales peuvent se simplifier en utilisant le fait que les taux de pénétration de la liste rouge ne sont pas, en réalité, définis séparément sur chaque unité de la base, mais sont constants pour toutes les unités appartenant à une « zone » élémentaire (adresse, îlot, IRIS, commune ...).

En effet, supposons que : $\tau_i = \bar{\tau}_l$ pour $i \in Z_l$, où Z_l est une zone élémentaire et $\bar{\tau}_l$ est le rapport du nombre d'abonnés sur liste rouge dans la zone au nombre total de numéros dans la zone, soit :

$$\bar{\tau}_l = \frac{\underline{R}_l}{\underline{N}_l},$$

en notant $\underline{R}_l$ le nombre d'abonnés sur liste rouge dans la zone l (le soulignement rappelle qu'il s'agit d'effectifs calculés au niveau d'une zone, par opposition aux effectifs de strates).

Remarque : dans les formules ci-dessus, il ne faut évidemment pas que certaines unités de la base des n°s joignables puissent se voir affectées d'une valeur de $\bar{\tau}_l$ égale à 1 : mais cela voudrait dire que tous les n°s de la zone à laquelle appartiendrait une unité de ce type seraient en liste d'opposition, donc y compris celle que l'on analyse, alors que, par hypothèse, cette unité est dans une strate $\mathcal{P}_{1,h}$ ***, donc est un n° joignable.***

Ceci étant, il ne faut pas raisonner à un niveau trop fin pour la constitution de ces zones : si l'on raisonne à l'adresse, on risque en effet, pour les petites adresses (en nombre de n°s recensés), d'avoir des valeurs discrètes des $\bar{\tau}_l$, assez différentes des valeurs moyennes (par exemple : 0.5, 0.33, 0,25...), donc générant des distorsions des probabilités d'inclusion ou des poids. Concrètement, l'adresse doit être remplacée par une zone plus agrégée (cf. § 2.5). De surcroît, l'hypothèse d'homogénéité des comportements a plus de « chances » d'être vérifiée si l'on raisonne sur des zones plus agrégées à l'intérieur desquelles le taux de pénétration de la liste rouge est supposé constant.

Alors, en faisant l'hypothèse que, pour chaque modalité h de la stratification (correspondant à un croisement donné des variables typologiques retenues), la population totale $\mathcal{P}_h$ (i.e. l'ensemble des numéros) se partitionne en zones élémentaires, on aura, pour toute strate $\mathcal{P}_{1,h}$ (de la base de sondage) :

$$\sum_{i \in P_{1,h}} \frac{1}{(1-\tau_i)} = \sum_{l / Z_l \subset P_h} \sum_{i \in Z_l \cap P_{1h}} \frac{1}{1-\frac{\underline{R}_l}{\underline{N}_l}}$$

$$= \sum_{l / Z_l \subset P_h} \sum_{i \in Z_l \cap P_{1h}} \frac{\underline{N}_l}{\underline{N}_l - \underline{R}_l}$$

$$= \sum_{l / Z_l \subset P_h} (\underline{N}_l - \underline{R}_l) \frac{\underline{N}_l}{\underline{N}_l - \underline{R}_l}$$

[car les unités de la strate $\mathcal{P}_{1,h}$ qui sont dans la zone Z_l sont les numéros hors liste rouge de cette zone, d'effectif $\underline{N}_l - \underline{R}_l$]

$$= \sum_{l / Z_l \subset P_h} N_l$$

$= N_h$.

On en déduit que : $\sum_h \sum_{i \in P_{1,h}} \frac{1}{(1-\tau_i)} = \sum_h N_h$ = N.

Finalement, on obtient :

$n_h = n \frac{N_h}{N}$: effectif d'échantillon proportionnel à la taille de la population totale (*et non à celle de la base)* ***en termes de lignes téléphoniques.***

$\omega = \frac{N}{n}$

$\Pi_i = \frac{n}{N} \frac{1}{1-\tau_i}$

Pour chaque échantillon entrant, n = 1.100, et N_h et N représentent, respectivement, le nombre de lignes téléphoniques ménages dans la strate h et le nombre total de lignes. Ces données doivent être fournies par Wanadoo Data et révisées annuellement à date fixe.

Nota : une telle allocation n'est pas possible avec l'option du tirage aléatoire simple à l'intérieur de chacune de strates, car les poids finaux sont forcément inégaux. Néanmoins, on peut obtenir une allocation du même type en imposant que *les moyennes des poids sur l'ensemble de la base de sondage soient égales à l'intérieur de chacune des strates*.

2.5. Mise en œuvre pratique.

- Dans chaque strate, les n°s seront tirés par Wanadoo Data avec des **probabilités inégales** données par la formule :

$$\Pi_i = \frac{n}{N} \frac{1}{1-\tau_i}.$$

N et n ont été définis ci-dessus.

τ_i est le taux de pénétration de la liste rouge pour l'unité i.

Idéalement, il est défini comme suit : chaque n° i appartient à une commune, les plus grandes communes ayant été découpées en IRIS. Le taux de pénétration de la liste rouge est alors un taux moyen calculé au niveau de l'IRIS d'appartenance du n° i ou, à défaut, au niveau de la commune si celle-ci n'est pas découpée en IRIS. Ceci permet de prendre en compte les différentiels de taux de pénétration qui peuvent intervenir à des niveaux assez fins au sein d'une ville.

Cependant, dans la pratique, pour certaines communes découpées en IRIS, certains n°s ne comportent pas dans la base Wanadoo Data la désignation de l'IRIS d'appartenance, notamment dans les cas d'habitat individuel. ***On a donc retenu au final de ne calculer les taux de pénétration de la liste rouge qu'au niveau de la commune d'appartenance.***

Ces données sont actualisées une fois par an par Wanadoo Data, à partir de l'intégration dans leur base de l'information sur les listes d'opposition, et conserveront donc des valeurs constantes jusqu'à la prochaine révision.

- Le poids à affecter aux n°s échantillonnés à l'issue de la phase d'échantillonnage est in fine le même pour tous, soit : $\boxed{\omega = \frac{N}{n}}$, avec N et n tels que définis au § 2.4.

  Ce poids est affecté ex-post par l'Insee.

- Pour des raisons de disponibilité des données, le démarrage de la nouvelle application CAMME s'est faite en recourant à un proxy dans les formules précédentes, conduisant à *remplacer les nombres totaux de lignes qui figurent dans les formules précédentes (N_h et N) par des estimations des nombres totaux de logements*, tels qu'issus de l'enquête Emploi.

  Il s'agit d'une approximation dont l'incidence devrait être faible et qui a pour intérêt de permettre une extrapolation à l'ensemble des ménages (et pas seulement à l'ensemble des ménages raccordés à une ligne téléphonique fixe). Ceci a un impact sur le calcul de l'allocation de l'échantillon entre les strates, sur celui des probabilités de sélection des n°s et sur le poids attribué aux n°s échantillonnés.

  Au final, l'allocation de l'échantillon entre les différentes strates s'est faite proportionnellement au nombre de logements principaux estimé à partir des données de l'enquête Emploi en continu, relatives, pour l'initialisation du processus, à l'année 2002 (moyenne simple des 4 trimestres pondérés). Cette allocation devrait être révisée chaque année.

Tableau 0 : répartition de l'échantillon mensuel entrant.

TUAG → / Type d'habitat ↓	0	1	3	5	7	8	TOTAL
Individuel	236	100	80	72	98	39	625
Collectif	22	25	49	74	157	148	475
TOTAL	**258**	**125**	**129**	**146**	**255**	**187**	**1100**

Observations.

- Le code agrégé TUAG est défini comme suit à partir du code TU99 (tranche de taille d'unité urbaine au recensement de 1999) :

TU99	TUAG
0 : communes rurales	0
1, 2 : unités urbaines de moins de 10.000 hab.	1
3, 4 : unités urbaines de 10.000 à moins de 50.000 hab.	3
5, 6 : unités urbaines de 50.000 à moins de 200.000 hab.	5
7 : unités urbaines de 200.000 hab. ou plus (sauf Paris)	7
8 : unité urbaine de Paris	8

- La variable « type d'habitat » est construite comme suit à partir des modalités de la variable TH de l'enquête Emploi :

TH	Type d'habitat
1 : ferme	Individuel
2 : hôtel, pension de famille, garni	Collectif
3 : construction provisoire, habitation de fortune	Individuel
4 : maison individuelle	Individuel
5 : maison ou immeuble comportant plusieurs logements entièrement ou principalement à usage d'habitation	Collectif
6 : immeuble principalement à usage industriel, commercial ou administratif comportant au moins un logement à usage d'habitation (usine, atelier, immeuble de bureau magasin, école, collège, hôpital, mairie, gare, bureau de poste, stade...)	Collectif

Dans la base de sondage des n°s de téléphone éligibles, on approchera cette variable à l'aide du code « verticalité » : *1 n° à l'adresse = individuel, sinon collectif.*

3. Allocation de l'échantillon entre les enquêteurs : un algorithme réalisant la répartition d'une population en classes minimisant un critère de distance.

Comme on l'a dit précédemment, la gestion de l'enquête CAMME est confiée à 8 Directions régionales (DR) de l'Insee : contrairement au principe habituel des enquêtes ménages selon lequel chaque DR gère la collecte relevant de son champ territorial, il convient dans le cas présent de répartir l'échantillon des n°s à interroger entre les 8 DR gestionnaires, puis entre les différents enquêteurs au sein de chacune d'entre elles.

Bien que l'enquête CAMME ne procède pas par déplacement d'enquêteur, puisqu'il s'agit d'une interrogation téléphonique, il a été jugé utile et intéressant sur le plan méthodologique de construire un algorithme assurant la répartition de l'échantillon des n°s de téléphone à interroger entre les différents enquêteurs.

Cet algorithme est fondé sur deux principes :

- assurer une certaine « proximité » entre enquêteur et enquêté. Celle-ci est obtenue en cherchant à minimiser la somme des distances entre enquêteurs et enquêtés ;

- tout en respectant une contrainte de charge d'enquêtes par enquêteur (appelée *charge utile*).

Cet algorithme, baptisé « ALLOCAMME » permet donc une répartition « équitable » et rationnelle des n°s enquêtés entre les 8 Directions Régionales de l'Insee gestionnaires. Son principe pourrait s'appliquer plus généralement à n'importe quelle enquête où il apparaîtrait nécessaire de minimiser sous contrainte de charge par enquêteur les déplacements enquêteur / enquêté. Il pourrait par ailleurs trouver des applications dans un domaine autre que la statistique, par exemple dans l'optimisation d'un réseau de distribution…

3.1. Rappel du contexte.

Il s'agit de répartir l'échantillon entrant de chaque vague mensuelle de l'échantillon CAMME (soit 1.100 n°s de téléphone) entre les 8 DR de gestion (puis ensuite entre les enquêteurs gérés au sein de chaque DR). Les charges d'enquêtes supportées par chacun des enquêteurs constituent une contrainte intangible (c'est un paramètre défini, géré et mis à jour par les DR elles-mêmes) ; le « principe de territorialité » vise in fine à attribuer un n° échantillonné à une DR « proche » géographiquement, cette contrainte devant être gérée de manière souple[7].

Le problème a été formalisé sur le plan théorique afin de construire un véritable algorithme de répartition spatiale sous une contrainte de charge et satisfaisant un critère d'optimalité.

Cet algorithme est mis en œuvre sous forme d'un programme informatique (intégré à la chaîne de préparation des échantillons de l'enquête), qui réalise automatiquement l'affectation de chaque échantillon entrant aux différents enquêteurs. Par ailleurs, les échantillons en réinterrogation sont affectés aux mêmes enquêteurs que ceux auxquels ils ont été affectés en vague entrante. Ils ne sont donc pas concernés par cet algorithme (hors n°s impossibles à joindre, inaptes et refus, qui sont sortis de l'échantillon des vagues suivantes).

[7] Bien que la mode soit à la délocalisation des services, il a été jugé peu pertinent, par exemple, qu'un enquêteur de Marseille interroge un n° de Strasbourg et vice-versa.

3.2. Formalisation théorique du problème.

Notations :

s = échantillon (ou, plus généralement, population de référence à répartir) de taille n.

E = ensemble des enquêteurs, de taille K.

Pour chaque enquêteur $e_k \in E$, on définit une charge utile C_k.

On suppose : $\sum_{k=1}^{K} C_k = n.$

On définit une distance entre chaque élément i de la population et chaque enquêteur e_k, soit :

$d(i, e_k)$.

Le problème est d'affecter chaque individu i de la population à un et un seul enquêteur e_k, de façon à minimiser le critère égal à la somme des distances de chacun des individus à l'enquêteur auquel il est affecté, tout en respectant la contrainte de charge utile pour chacun des enquêteurs : un enquêteur ne peut avoir plus d'individus affectés que sa charge utile ne le lui permet.

→ Mathématiquement, le problème peut se formaliser ainsi :

Chercher une partition d'un ensemble s en K classes notées A_k (chaque classe A_k représentant le sous-ensemble des individus affectés à l'enquêteur k), telles que :

$\forall\ k \in \{1, 2, ...K\}$: card $(A_k) = C_k$

et minimisant la fonction : $\sum_{k=1}^{K} \sum_{i \in A_k} d(i, e_k)$.

3.3. Algorithme proposé.

3.3.1. Etape préliminaire.

On construit les classes A_k^*, pour tout $k \in \{1, 2, ...K\}$, définies comme suit :

$A_k^* = \{i \in s ; d(i, e_k) = \underset{e_j \in E}{Min}\ d(i, e_j)\}$.

Ceci signifie qu'on commence à affecter chaque individu i à l'enquêteur e_k le plus proche.

On peut supposer, pour simplifier, que, pour i fixé dans s, les valeurs d (i, e_k) [où $e_k \in E$], sont toutes distinctes[8], ce qui entraîne que, pour chaque i dans s, il y a un seul e_k qui réalise le minimum de d (i, e_j). Ainsi, il n'y a pas d'ambiguïté dans la définition des classes A_k^*, et celles-ci formeront bien une partition de s.

Si cette condition n'était pas réalisée, il faudrait modifier légèrement la définition des A_k^* en introduisant par exemple une contrainte d'ordre lexicographique :

$A_k^* = \{i \in s ; d(i, e_k) = \underset{e_j \in E}{Min}\ d(i, e_j)$ *et :* $\forall\ j < k : d(i, e_j) > d(i, e_k)\}$.

k est donc le plus petit indice dans l'ordre alphabétique pour lequel d (i, e_k) est minimale.

On fabrique donc une partition de s en K classes, qui fournit un optimum absolu au critère de distance, mais celle-ci ne satisfait évidemment pas les contraintes de charge.

Notons :

$E^+ = \{ k \in \{1, 2, ...K\}, \text{card}(A_k) \leq C_k \}$: c'est l'ensemble des enquêteurs en *sous-charge*, c'est-à-dire avec une contrainte de charge non saturée, donc une capacité d'absorption positive.

$E^- = \{ k \in \{1, 2, ...K\}, \text{card}(A_k) > C_k \}$: c'est l'ensemble des enquêteurs en *surcharge*, c'est-à-dire avec un déficit de capacité.

3.3.2. Initialisation du processus de transfert.

A l'initialisation, les ensembles E^+ et E^- sont définis à partir des parties A_k^*.

On va réaliser des transferts d'unités de l'échantillon affectées aux enquêteurs en surcharge vers d'autres enquêteurs, de façon à « rééquilibrer » les charges, en détériorant le moins possible le critère de distance.

Le transfert d'une unité i de la classe A_k^*, où $k \in E^-$, vers une classe A_j^* va faire augmenter le critère de distance de la quantité d (i, e_j) - d (i, e_k) (positive ou nulle par construction, éventuellement nulle si l'individu i est à égale distance des deux enquêteurs e_j et e_k). *On va imposer que cette augmentation soit la plus petite possible.*

On va donc chercher un indice $k_0 \in E^-$, une unité $i_0 \in A_{k_0}^*$ et un indice j_0 (différent de k_0), réalisant :

[8] En revanche, on peut avoir d (i_1, e_k) = d (i_2, e_k), avec i_1 et i_2 distincts, pour un k donné.

$$\underset{\substack{k \in E^- \\ i \in A_k^* \\ 1 \le j \le K,\ j \ne k}}{Min} \quad [d(i, e_j) - d(i, e_k)].$$

Concrètement, on détermine donc dans laquelle des classes déficitaire, soit $A^*_{k_0}$ et, à l'intérieur de celle-ci, lequel des éléments i_0 va être transféré à une nouvelle classe $A^*_{j_0}$, tout en faisant en sorte que l'augmentation résultante du critère global soit minimale.

Ayant déterminé l'unité i_0 et les indices k_0 et j_0, on modifie alors la composition des ensembles A^*_k, en posant :

$A_{k_0} = A^*_{k_0} - \{ i_0 \}$

$A_{j_0} = A^*_{j_0} \cup \{ i_0 \}$

$A_k = A^*_k$, pour tout $k \neq k_0$ et $k \neq j_0$.

On redéfinit ensuite les ensembles E^+ et E^- à partir de ces nouveaux ensembles A_k et on itère le processus.

3.3.3. Etape courante.

On recommence l'opération de transfert à partir des nouveaux ensembles A_k définis à l'étape précédente.

Comme on détruit l'optimalité globale à partir de la construction initiale des ensembles A^*_k, il faut éviter de faire deux transferts successifs qui ramèneraient à la même composition de ces ensembles et ferait boucler l'algorithme indéfiniment.

On est sûr d'éviter ce risque si l'on impose que le critère de distance ne puisse qu'*augmenter* à chaque étape de transfert. Cela signifie qu'on ne cherchera une classe d'accueil pour un transfert que parmi celles qui assurent une telle augmentation.

A chaque étape, on cherchera donc des indices $k_0 \in E^-$, $i_0 \in A_{k_0}$ et $j_0 \in \{1, 2, ...K\}$, réalisant :

$$\underset{\substack{k \in E^- \\ i \in A_k \\ 1 \le j \le K \,/\, d(i,e_j) > d(i,e_k)}}{Min} \quad [d(i, e_j) - d(i, e_k)].$$

Remarque : pourquoi ne faut-il pas astreindre les classes d'accueil lors d'un transfert à n'être choisies que parmi les classes en sous-charge (c'est-à-dire à rechercher j_0 seulement dans E^+ au lieu de {1, 2, ..., K} tout entier) ?

La réponse est qu'on peut améliorer le critère en faisant deux transferts successifs, d'abord d'une unité vers une classe déjà saturée, ce qui augmente le déficit de ladite classe, puis d'un autre élément pris dans cette dernière vers une classe en sous-charge, plutôt qu'en transférant directement l'unité considérée vers une classe en sous-charge.

L'exemple suivant illustre cette situation.

	ENQUÊTEURS		
UNITES ECHANTILLONS	**1**	**2**	**3**
1	10	20	90
2	10	40	100
3	70	10	60
4	30	30	10
Charge utile	**1**	**1**	**2**

Chaque case représente la distance entre une unité i et un enquêteur e_k .

On voit clairement la composition des ensembles A_k^* initiaux :

$A_1^* = \{1, 2\}$, $A_2^* = \{3\}$, $A_3^* = \{4\}$, avec un optimum global de 40.

D'où il résulte que l'enquêteur 1 (ou classe 1) est en surcharge et l'enquêteur 3 en sous-charge. On voit donc que si l'on ne transférait des unités que d'une classe en surcharge vers une classe en sous-charge, on transférerait l'unité 1 de la classe 1 à la classe 3 ; c'est ce transfert qui est le moins pénalisant pour le critère à minimiser, dont la valeur devient alors : 120.

Mais il est préférable de faire un premier transfert de l'unité 1 de la classe 1 vers la classe 2 (augmentation du critère de 10), ce qui met provisoirement la classe 2 en surcharge, puis de l'unité 3 de la classe 2 vers la classe 3 (augmentation du critère de 50), ce qui rééquilibre la classe 2 et conduit à la répartition finale :

$A_1 = \{2\}$, $A_2 = \{1\}$, $A_3 = \{3,4\}$, avec un critère de distance de valeur : 100.

➔ L'algorithme s'arrête quand $E^+ = s$ et $E^- = \varnothing$.

3.4. Mise en oeuvre opérationnelle.

La mise en oeuvre de cet algorithme s'est faite par l'intermédiaire d'une macro SAS dénommée « ALLOCAMME ».

Chaque n° échantillonné est repéré géographiquement par le chef-lieu du département de l'adresse à laquelle il correspond et chaque enquêteur par le chef-lieu de la DR de gestion dont il dépend. La distance entre un n° et un enquêteur a été calculée dans un premier temps à partir des distances kilométriques routières les plus courtes entre les chefs-lieux de département et de région ainsi définis. Dans un second temps, il est apparu plus satisfaisant d'utiliser une matrice de distance construite à partir des « centroïdes » des départements, d'abord parce que la distance « à vol d'oiseau » est plus intuitive s'agissant de n°s de téléphone, ensuite et surtout parce que l'on peut reconstruire facilement la matrice en cas de rajout d'une DR de gestion.

Naturellement, on pourrait procéder de manière analogue avec une information géométrique définie au niveau de la commune ou de la région.

On a ainsi pris comme convention d'affecter une distance nulle entre les enquêteurs d'une DR de gestion et un n° enquêté dont l'adresse appartient au département dont le chef-lieu est le siège de cette DR (exemple : un n° au HAVRE est à une distance nulle des enquêteurs rattachés à ROUEN). En revanche, un n° affecté à une DR de gestion est à une distance non nulle des enquêteurs de cette DR s'il n'est pas situé dans le même département (exemple : un n° à SEILLANS, dans le Var, est à une distance non nulle des enquêteurs de MARSEILLE).

Les charges utiles sont des paramètres fournis par les Directions Régionales concernées et susceptibles d'une mise à jour périodique. ***Il faut noter que la macro, dans la version où elle a été conçue, ne peut fonctionner que si la somme des charges utiles est rigoureusement égale à la taille de l'échantillon (i.e., 1.100 dans le cas présent).***

Ainsi, les données figurant en entrée de la macro ALLOCAMME sont d'une part, le fichier de l'échantillon entrant des 1.100 n°s de téléphone, avec l'indication du département de rattachement, et d'autre part, les deux tables de paramètres que sont la matrice des distances et la table des charges par enquêteur.

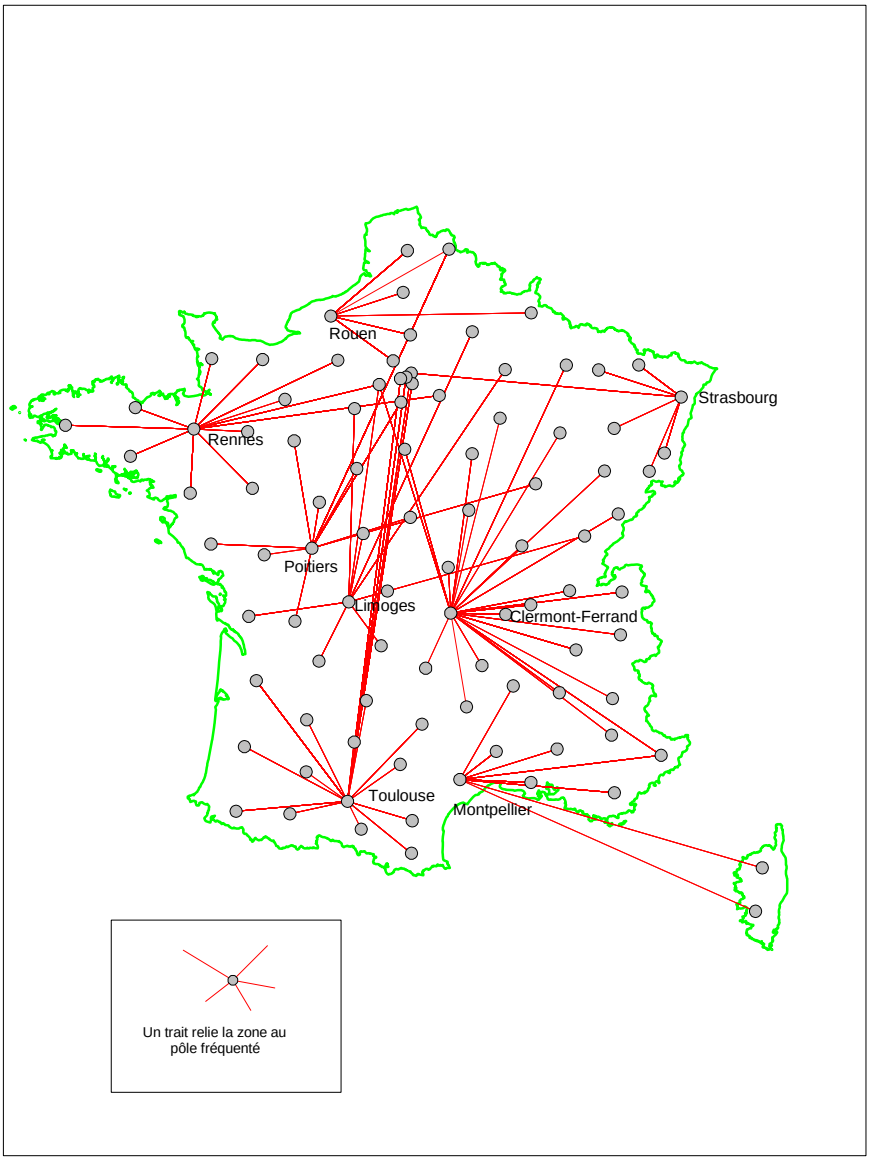
La macro fonctionne alors en réalisant dans un premier temps l'allocation de l'échantillon entre les différentes DR de gestion, la table des charges par enquêteur étant convertie en une table de charge par DR ; puis, dans un second temps, l'échantillon alloué à une DR donnée est réparti entre les différents enquêteurs pris aléatoirement, tout en respectant leur charge (concrètement, les enquêteurs de la DR sont triés par charge croissante et l'échantillon, trié dans un ordre aléatoire, est affecté, enregistrement par enregistrement, à chaque enquêteur jusqu'à saturation de sa charge).

3.5. Exemple.

Le tableau et la carte suivants présentent l'allocation réelle obtenue pour un échantillon de 1.100 n°s entrants vers les huit Directions Régionales gestionnaires de l'enquête. Nous pouvons souligner par exemple le déséquilibre généré par le fait que la DR d'Ile de France n'est pas gestionnaire : la méthode retenue conduit à affecter ces n°s à plusieurs DR délégataires, éventuellement éloignées.

DR gestion → / Région du n°	Alsace	Auvergne	Bretagne	Haute-Normandie	Languedoc-Roussillon	Limousin	Midi-Pyrénées	Poitou-Charentes	Total
Ile de France	12	11	30	19		4	108	27	**211**
Champagne-Ardenne		5		4		10			**19**
Picardie				21		11			**32**
Haute-Normandie			12	21					**33**
Centre		11				13		20	**44**
Basse-Normandie			27						**27**
Bourgogne		21						9	**30**
Nord-Pas-de-Calais				25				37	**62**
Lorraine	31	7							**38**
Alsace	33								**33**
Franche-Comté	4	11				5			**20**
Pays de la Loire			40					21	**61**
Bretagne			56						**56**
Poitou-Charentes						10		21	**31**
Aquitaine						6	51		**57**
Midi-Pyrénées							48		**48**
Limousin						16			**16**
Rhône-Alpes		107			4				**111**
Auvergne		25							**25**
Languedoc-Roussillon		1			34		13		**48**
Provence-Alpes-CA		16			77				**93**
Corse					5				**5**
Total	**80**	**215**	**165**	**90**	**120**	**75**	**220**	**135**	**1100**

Répartition d'un échantillon Camme juillet 2003



© IGN - Insee
Source : IGN, Wanadoo Data juillet 2003

4. Traitements statistiques aval.

La refonte de l'enquête CAMME a conduit à revoir et améliorer la phase des traitements statistiques aval : correction de la non-réponse, redressements et calages sur marge. La logique retenue, basée sur la simplicité des traitements et la robustesse des modèles d'imputation utilisés, permet de répondre aux contraintes fortes imposées en termes de délai d'élaboration des résultats.

4.1. Traitement de la non-réponse partielle

Seule la partie socio-démographique est concernée par un traitement de la non-réponse partielle. La nature même du questionnaire de conjoncture impose en effet l'obtention d'une réponse pour chaque question posée, éventuellement en utilisant la modalité « Ne sait pas », non proposée en première lecture.

Pour bon nombre de variables, le « data model » sous CAPI impose d'obtenir une réponse : il s'agit de questions peu sensibles, portant sur la taille du ménage, le nombre d'enfants et le nombre d'actifs, l'exercice d'une profession, la nature de l'emploi et la classification professionnelle du titulaire et, éventuellement, de son conjoint. Le sexe, ainsi que l'âge en cas de non-réponse, sont renseignés par l'enquêteur en fin d'entretien. Cette simplification sur l'âge est permise par le très faible taux de non-réponse dans le cadre de cette enquête téléphonique et par le fait que cette variable sert uniquement à une ventilation en quatre tranches d'âge des résultats transmis à la Commission Européenne.

Un traitement simplifié concerne la codification de la CS, lorsqu'elle est manquante (soit en raison d'un refus de répondre de la part de l'enquêté à l'une des questions permettant sa codification automatique par SICORE, soit en raison d'un échec de cette codification non résolu lors de la phase de reprise manuelle). S'il est précisé un statut d'indépendant, le titulaire est réputé être artisan ; pour tous les autres statuts, la CS agrégée « employé » est retenue pour les femmes, celle d' « ouvrier » pour les hommes. Il s'agit donc d'affecter la modalité la plus fréquente (modale) en fonction de l'information disponible. Là encore, c'est le faible taux de non-réponse qui permet ce traitement léger et rend acceptable la déformation de la distribution induite. Un traitement par Hot-Deck aurait cependant été préférable. In fine, la CS est agrégée en cinq modalités, toujours pour les besoins de la ventilation des résultats transmis à la Commission.

Concernant la question sur le revenu global du ménage, l'enquêté a la possibilité de répondre soit en donnant un montant précis, soit en le situant dans l'une des onze tranches de revenu proposées. Un premier traitement consiste à choisir aléatoirement un montant précis de revenu pour les réponses données en tranches. L'observation des réponses donnant un revenu en euros montre qu'en général les montants sont arrondis, à 50 € près jusqu'à 1 000 €, à 100 € près au-delà. L'existence de points d'accumulation sur les montants arrondis pose des problèmes lors de l'imputation des valeurs manquantes, basée sur une modélisation du revenu. Un deuxième traitement consiste alors à rajouter à chaque réponse en euros un aléa, tiré dans une loi uniforme (intervalle de 50 € pour les revenus inférieurs à 1 000 €, de 100 € au-delà).

L'imputation des revenus manquants est alors basée sur une modélisation (procédure GLM de SAS) du revenu en fonction des variables suivantes :
CS agrégée ; tranche d'âge ; CS ∗ nombre d'actifs (en quatre tranches).

Une deuxième modélisation plus simple, faisant intervenir uniquement la CS agrégée, permet de traiter d'éventuels cas résiduels (non-réponse d'un ménage croisant des modalités rares des variables explicatives.)

Une solution plus satisfaisante concernant l'imputation des revenus manquants aurait consisté à se baser sur une modélisation des revenus de la population totale, plutôt que de partir de l'échantillon. Néanmoins, cette solution alternative n'était pas immédiate à mettre en place (notamment parce qu'il n'existe pas de distribution standard des revenus toute faite), et finalement peu justifiée : actuellement, cette variable sert exclusivement à ventiler par quartiles de revenu les réponses aux questions de conjoncture pour les besoins de la Commission européenne.

4.2. Calage sur marges

Pour des raisons de continuité des séries historiques, la procédure de calage sur marges n'est mise en œuvre que pour les sorties européennes. Pour une période transitoire, la division des Comptes Trimestriels continue en effet de produire et diffuser des séries en utilisant des données non calées, c'est-à-dire munies de leurs pondérations initiales (identiques, par construction de l'échantillon).

Les programmes relatifs à la phase de calage utilisent des marges issues de l'Enquête Emploi en Continu (EEC). Le protocole de mise à jour des marges de calage prévoit de procéder simultanément à l'actualisation de ces marges et des données utilisées lors de la stratification, dans la mesure où l'on utilise en grande partie les mêmes variables pour ces deux étapes. En pratique, les résultats sont calés sur une année civile (les résultats de l'EEC pour les quatre trimestres de l'année n-2, puis de l'année n-1 dès que les quatre trimestres sont disponibles).

Les variables utilisées pour le calage (CALMAR sous SAS) sont les suivantes :

- taille du ménage, en 4 modalités (1, 2, 3, et plus de quatre personnes)
- âge du titulaire, en 4 tranches (16-29 ans ; 30-49 ans, 50-65 ans et 65 ans et plus)
- « CS agrégée UE », en 7 modalités (artisans, agriculteurs, employés, ouvriers qualifiés, ouvriers non qualifiés, chômeurs et inactifs)
- code « verticalité », en deux modalités (habitat individuel vs collectif, par proxy : cf. § 2.3.1.2.)
- code « typologie de commune », en 6 modalités (§ 2.3.1.1.)

Les deux dernières variables sont celles utilisées en amont lors de la stratification de l'échantillon. Elles permettent de corriger la déformation éventuellement générée par le calage sur les premières variables.

En pratique, le calage est effectué en deux étapes. Un premier calage est lancé, utilisant la méthode du « raking ratio », qui garantit des poids toujours positifs, mais non bornés supérieurement. Cette étape sert uniquement à récupérer une première structure de poids. Dans un second temps, une procédure de calage utilisant le modèle LOGIT permet de borner la distribution des rapports de poids en éliminant les 10 % de poids les plus faibles, et les 5 % de poids les plus forts. L'intérêt de la première étape consiste à éviter de construire les bornes inférieures et supérieures des rapports de poids par approximations successives, et donc d'automatiser la procédure.

Notons enfin que la variable « sexe » n'est pas utilisée dans la procédure de calage, du fait de la structure des répondants selon cette variable (deux tiers des répondants sont de sexe féminin, l'explication principale étant qu'il s'agit d'une enquête téléphonique.)

4.3. Tabulation et éditions des tableaux mensuels

Les résultats sont tabulés et édités en fin de mois selon deux procédures bien distinctes, pour répondre aux attentes respectives de la Commission Européenne d'une part, et de la Division des Comptes Trimestriels d'autre part. Dans les deux cas, la mise en œuvre utilise une édition de tableaux HTLM via un ODS, à partir des tabulations effectuées sous SAS.

Pour la Commission, un tableau de résultats statistique donne, pour chaque question de conjoncture, la distribution en pourcentage des réponses selon les différentes modalités possibles (en général, « augmentation importante » ; « augmentation modérée » ; « stabilité » ; « faible diminution » ; et « diminution forte », complétée de la modalité « ne sait pas »). Ces résultats sont ventilés selon les critères suivants :

- quartile de revenu ;
- CS agrégée UE (voir le paragraphe précédent)
- quotité de travail pour les CDI (travail à temps partiel ou temps complet)
- niveau d'éducation (primaire, secondaire, supérieur)
- tranches d'âge
- sexe

Ces résultats sont complétés d'un tableau de fréquence des répondant selon les variables d'analyse, et d'un fichier détail anonymisé reprenant l'ensemble des réponses et précisant leurs pondérations respectives.

Parallèlement, un fichier est édité pour la Division des Comptes Trimestriels. Comme dit plus haut, le fichier est construit à partir de données non repondérées, et ne propose aucune ventilation des résultats selon des critères socio-démographiques. Pour chaque question de conjoncture, le « solde d'opinion » est calculé, par différence entre les réponses positives et négatives. Après désaisonnalisation, ces soldes servent notamment au calcul de l'indice agrégé, commenté sous la dénomination de « moral des Français ». Ils alimentent aussi les modèles de prévision de la consommation de biens manufacturés et, de manière générale, le diagnostic conjoncturel.

Il est utile, à ce stade, de rappeler que le niveau du solde n'a pas vraiment de signification, et que seule son évolution est à commenter. Cette précaution étant rappelée, nous pouvons souligner, comme cela a été le cas récemment dans la presse, que l'appellation « moral des Français » déborde largement le champ d'investigation de cette enquête, concentrée sur l'environnement économique des ménages interrogés. Pourtant, ex-post, l'analyse justifie cette appellation, le solde agrégé étant sensible à des évènements a priori éloignés des pures réalités économiques : il réagit ainsi fortement à des évènements aussi divers que les résultats d'élection, les évènements climatiques majeurs, mais aussi les résultats sportifs d'envergure… bref, au « moral des Français ! »

5. Calculs de précision

Nous traitons d'abord dans ce paragraphe de la précision de l'estimation d'un solde pour une opinion et un mois donnés. Nous répondons par exemple à une question du type « avec quelle précision connaît-on le solde de l'opinion des Français en octobre 2004 sur l'évolution future du nombre de chômeurs ? » ou, plus généralement, « quelle est la précision, un mois donné, de l'indicateur résumé du moral des Français ? ». Nous mesurons ensuite la précision des évolutions entre deux dates, en utilisant les résultats obtenus pour les niveaux des soldes. Nous répondons ici, par exemple, à « la baisse du moral des Français entre mars et juillet 2004 était-elle significative ? ».

5.1. Préambule

5.1.1. Quelques éléments de cadrage sur l'échantillon

Tableau 1 : répartition par strate des effectifs échantillonnés et répondants d'octobre 2004

Strate h	*Nombre de ménages*[9] M_h	*Nombre théorique d'unités enquêtées chaque mois* n'_h	*Nombre de répondants en oct. 04* r_h	*Taux de réponse brut en oct. 04*	*Nombre de répondants par vague*		
					1ère	*2ème*	*3ème*
Communes rurales, individuel	5 244 899	708	430	61%	*155*	*136*	*139*
Communes rurales, collectif	484 102	66	43	65%	*17*	*14*	*12*
UU de 5 000 à 9 999 hab., individuel	2 226 357	300	188	63%	*70*	*54*	*64*
UU de 5 000 à 9 999 hab., collectif	553 240	75	47	63%	*17*	*14*	*16*
UU de 10 000 à 49 999 hab., individuel	1 781 036	240	146	61%	*54*	*51*	*41*
UU de 10 000 à 49 999 hab., collectif	1 089 935	147	85	58%	*34*	*26*	*25*
UU de 50 000 à 1 999 999 hab., individuel	1 603 488	216	147	68%	*55*	*45*	*47*
UU de 50 000 à 1 999 999 hab., collectif	1 633 225	222	132	59%	*53*	*40*	*39*
UU de plus de 2 millions d'hab., individuel	2 172 294	294	194	66%	*71*	*68*	*55*
UU de plus de 2 millions d'hab., collectif	3 492 737	471	291	62%	*112*	*92*	*87*
UU de Paris, individuel	867 286	117	78	67%	*31*	*26*	*21*
UU de Paris, collectif	3 299 377	444	239	54%	*97*	*79*	*63*
TOTAL	**24 447 976**	**3 300**	**2 020**	**61%**	***766***	***645***	***609***

En théorie, le nombre d'unités interrogées chaque mois et dans chaque strate s'obtient simplement en triplant les allocations d'un échantillon entrant. En réalité, moins d'unités sont enquêtées car celles qui n'ont pas répondu en 1ère vague aux questions socio-démographiques ne sont pas réinterrogées les mois suivants. De fait, le taux de réponse brut, calculé sur la base des allocations théoriques, minore le taux de réponse apparent constaté sur l'échantillon réellement interrogé. En moyenne, le taux de réponse brut vaut environ 60% et le taux de réponse apparent avoisine 80%.

Dans les deux cas, les taux de réponse les plus faibles s'observent généralement dans l'habitat collectif ainsi que dans les zones les plus urbanisées, spécialement dans l'unité urbaine (UU) de Paris. En conséquence, la structure par strate de l'échantillon des répondants n'est plus proportionnelle à celle de la population des ménages, contrairement à ce qui avait prévalu pour calculer les allocations des échantillons entrants.

[9] *Source :* enquête Emploi en continu 2002, France métropolitaine, moyenne simple des 4 trimestres pondérés.

5.1.2. L'estimation d'un solde d'opinion

L'objectif est d'évaluer le solde de l'opinion des Français sur un item *y* :

$$\bar{Y} = \frac{1}{M}\sum_{k\in U} y_k$$

où *U* désigne la population métropolitaine composée de *M* ménages et où *y* représente la variable d'intérêt correspondant à la réponse à une question d'opinion, par exemple « *Pensez-vous que, dans les douze prochains mois, le nombre de chômeurs va … fortement augmenter / un peu augmenter / rester stationnaire / un peu diminuer / fortement diminuer ? ».*

Selon l'opinion émise (positive, neutre ou négative), la réponse d'une unité *k* est recodée : $y_k = \begin{cases} +1 \\ 0 \\ -1 \end{cases}$.

Pour chaque item, la division des Comptes Trimestriels de l'INSEE publie dans « Informations Rapides » un solde brut d'opinion calculé par différence entre les pourcentages de réponses positives et négatives (un solde corrigé des variations saisonnières est également diffusé). Le solde brut s'exprime donc comme la moyenne empirique des réponses obtenues sur l'échantillon S_2 des *r* unités répondantes :

$$\boxed{\hat{\bar{Y}} = \frac{1}{r}\sum_{k\in S_2} y_k}.$$

Cet estimateur, auquel nous nous référerons toujours dans la suite, n'est donc pas l'estimateur d'Horvitz-Thompson puisque les poids d'extrapolation à la population d'inférence ne valent pas l'inverse des probabilités d'inclusion.

5.2. Calcul de précision pour un solde d'opinion

5.2.1. Méthodologie adoptée

5.2.1.1. Préambule

Pour évaluer la précision de l'estimateur $\hat{\bar{Y}}$, nous avons envisagé successivement plusieurs approches, par niveau de complexité croissante, en sériant les difficultés liées :

- au tirage à probabilités inégales réalisé dans une base de sondage incomplète,
- à la présence de non-réponse,
- au caractère rotatif de l'échantillon, et spécialement à la règle de non-réinterrogation d'une unité non répondante en 1ère vague.

Cette démarche progressive nous a permis de distinguer les effets sur la précision du tirage à probabilités inégales et de la non-réponse. Finalement, nous concluons sur un scénario qui retrace la réalité de l'échantillonnage de manière à la fois satisfaisante et simple (scénario 6 infra).

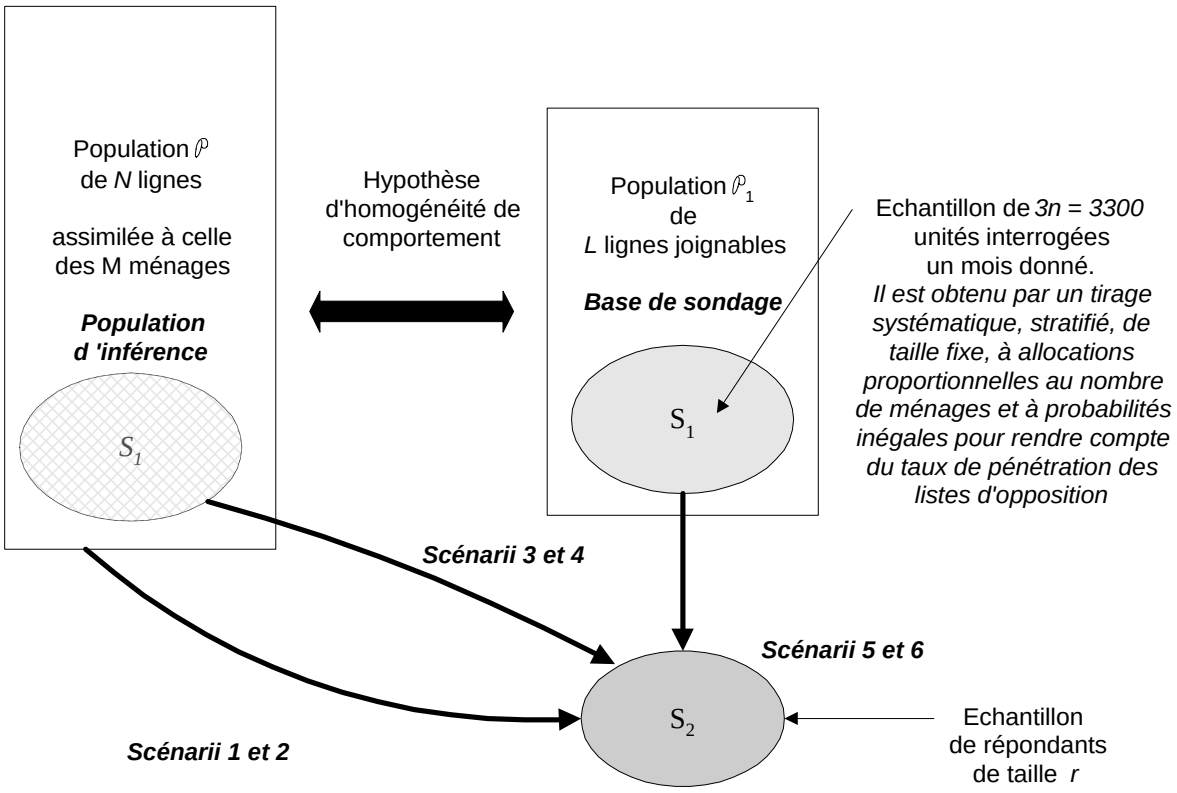
5.2.1.2. Les différentes approches envisagées

Dans un souci de simplification, nous avons considéré que les unités non répondantes à la 1ère vague n'auraient pas répondu non plus aux vagues suivantes (numéros toujours impossibles à joindre, inaptes, refus délibéré, etc.). De fait, nous avons raisonné sur les allocations théoriques et avons considéré le plan de taille fixe, égale à 3 300 abonnés. En toute rigueur, il faudrait distinguer les vagues d'interrogation et écrire la probabilité de réponse à la 2ème vague (resp. à la 3ème) comme produit de la probabilité de répondre à la 1ère interrogation multipliée par celle de répondre à la 2ème (resp. la 3ème) sachant que l'unité a répondu à la 1ère vague. Cependant, nous aurions alors dû estimer ces probabilités à partir d'effectifs parfois très faibles (par exemple, seules 22 unités sont enquêtées à la 1ère vague dans la strate « habitat collectif des communes rurales »).

Nous avons aussi négligé l'effet de lassitude des enquêtés et les biais de sélection qui en découlent. De même, nous avons ignoré les biais de conditionnement dus aux deux ré-interrogations (les individus qui acceptent de répondre à trois reprises ne sont pas nécessairement « représentatifs » de la population ou ont pu modifier leur comportement entre les vagues).

Nous avons envisagé successivement les six scénarii illustrés par la figure 1.

Figure 1 : Les différents scénarii envisagés



Dans un premier temps, nous avons ignoré l'aléa dû à la non-réponse et supposé que la population des non-répondants a le même comportement que celle des répondants pour les variables d'intérêt de l'enquête. Nous avons également assimilé la population des ménages à celle des lignes, comme cela avait été fait au moment de l'échantillonnage. Nous avons ainsi considéré deux scénarii conduisant directement **de la population $\mathcal{P}$ des *M* ménages à l'échantillon S_2 des *r* répondants** :

1. Plan simple sans remise de taille fixe de *r* répondants parmi *M* unités (***scénario 1***).

2. Plan stratifié avec sondage aléatoire simple sans remise dans chaque strate *h*, de taille fixe r_h répondants parmi M_h unités (***scénario 2***).

Puis, nous avons modélisé la non-réponse en nous plaçant dans le cadre d'un sondage en deux phases. Le principe est le suivant :

- La 1ère phase sélectionne l'échantillon S_1 de taille fixe $n' = 3n = 3\ 300$ unités interrogées un mois donné.
- La 2ème phase permet de prendre en compte la présence de non-réponse totale. La méthode consiste à considérer que l'échantillon S_2 des répondants résulte d'un tirage dans S_1 avec, comme probabilités d'inclusion, les probabilités de réponse. Nous avons modélisé le mécanisme de non-réponse par un tirage poissonnien et estimé la probabilité de réponse à partir de l'échantillon des répondants.

Plus précisément, nous avons considéré les 4 scénarii suivants :

3. **de l'échantillon S_1, pris dans la population $\mathcal{P}$, à l'échantillon S_2** avec un plan de 1ère phase stratifié avec sondage aléatoire simple sans remise dans chaque strate de taille fixe n'_h parmi M_h unités, suivi par :

   a. Un tirage de 2ème phase poissonnien avec, comme probabilité de réponse estimée, le ratio *r/n'* égal au nombre de répondants divisé par le nombre d'interrogés (***scénario 3***).

   b. Un tirage de 2ème phase poissonnien stratifié avec, comme probabilité de réponse estimée dans une strate *h* quelconque, le ratio r_h/n'_h égal au nombre de répondants divisé par le nombre d'interrogés de la strate (***scénario 4***).

4. **de l'échantillon S_1, pris dans la base de sondage incomplète $\mathcal{P}_1$, à l'échantillon S_2** avec un plan de 1ère phase stratifié avec un tirage systématique à probabilités inégales dans chaque strate, suivi par :

   a. Un tirage poissonnien en 2ème phase avec, comme probabilité de réponse estimée, le ratio *r/n'* égal au nombre de répondants divisé par le nombre d'interrogés (***scénario 5***).

   b. Un tirage poissonnien en 2ème phase stratifié avec, comme probabilité de réponse estimée dans une strate *h* quelconque, le ratio r_h /n'_h égal au nombre de répondants divisé par le nombre d'interrogés de la strate (***scénario 6***).

5.2.2. Mise en œuvre

Pour chacun de ces scénarii, nous avons estimé la variance de l'estimation du solde. Les expressions des différents estimateurs sont exposées en annexe 2.

D'autre part, comme les calculs de précision devaient pouvoir s'intégrer facilement dans la chaîne de production pré-existante, nous avons entrepris de développer quelques programmes SAS rapides et aisés à mettre en œuvre.

5.2.3. Résultats numériques pour quelques items

Tableau 2 : précision obtenue pour quelques soldes publiés en mars, juillet ou octobre 2004
(cf. annexe 3 pour les résultats détaillés par scénario)

Indicateur	*Mois d'enquête 2004*	*Solde estimé*	*Intervalle de confiance à 95%*		*Effet de sondage par rapport au scénario 1 (SAS dans P)*		
			Scénario1	*Scénario 6*	*Scénario2*	*Scénario 4*	*Scénario 6*
Evolution passée du niveau de vie en France	Mars	**-0,5600**	± 0,02585	**± 0,03533**	1	1,29	1,87
	Juillet	**-0,5073**	± 0,02679	**± 0,03543**	1	1,30	1,75
	Octobre	**-0,4985**	± 0,02768	**± 0,03559**	1	1,25	1,65
Perspectives d'évolution du niveau de vie en France	Mars	**-0,2798**	± 0,02753	**± 0,03013**	1	1,07	1,20
	Juillet	**-0,2506**	± 0,02812	**± 0,03056**	1	1,07	1,18
	Octobre	**-0,2906**	± 0,0277	**± 0,03092**	1	1,09	1,25
Evolution passée de la situation financière des ménages	Mars	**-0,1634**	± 0,02575	**± 0,02675**	1	1,03	1,08
	Juillet	**-0,1468**	± 0,02605	**± 0,0272**	1	1,04	1,09
	Octobre	**-0,1584**	± 0,02674	**± 0,02782**	1	1,03	1,08
Perspectives d'évolution de la situation financière des ménages	Mars	**0,0178**	± 0,02395	**± 0,02424**	1	1,01	1,02
	Juillet	**-0,0359**	± 0,02559	**± 0,02594**	1	1,01	1,03
	Octobre	**-0,0059**	± 0,02532	**± 0,02573**	0,99	1,01	1,03
Opportunité de faire des achats importants	Mars	**-0,1333**	± 0,02839	**± 0,02907**	1	1,02	1,05
	Juillet	**-0,0506**	± 0,02934	**± 0,02948**	1	1,01	1,01
	Octobre	**-0,1327**	± 0,02902	**± 0,02983**	1	1,02	1,06
Perspectives d'évolution de la situation économique générale	Mars	**-0,1616**	± 0,03065	**± 0,03157**	1	1,02	1,06
	Juillet	**-0,1403**	± 0,03282	**± 0,03405**	0,99	1,03	1,08
	Octobre	**-0,2678**	± 0,03206	**± 0,03464**	1	1,06	1,17
Perspectives d'évolution du nombre de chômeurs	Mars	**-0,4738**	± 0,03072	**± 0,0371**	1	1,15	1,46
	Juillet	**-0,5473**	± 0,02959	**± 0,0387**	1	1,28	1,71
	Octobre	**-0,5762**	± 0,02855	**± 0,03853**	1	1,32	1,82

Les résultats (cf. tableau 2) mettent d'abord en évidence la relative incertitude avec laquelle le niveau des soldes est connu un mois donné. Par exemple, en octobre 2004, à la question portant sur l'évolution future du nombre de chômeurs, la différence entre le pourcentage de Français « pessimistes » et celui d'« optimistes » valait - 57,6%. Ce solde est connu à ± 3,85 points près (selon le scénario n° 6 le plus complet) : il serait donc compris entre - 61,5% et - 53,8% avec un niveau de confiance d'environ 95%.

Il s'avère, sur ces questions, que la stratification ne semble pas générer de gain substantiel de précision. D'une part, les allocations proportionnelles, calculées sur le nombre total de ménages dans les strates, sont altérées du fait de la non-réponse. D'autre part, les critères de stratification semblent peu influer sur l'opinion. De fait, les strates paraissent relativement homogènes entre elles sur les variables d'opinion.

Pour certaines questions, les différentes approches que nous avons envisagées conduisent à des niveaux de précision voisins. C'est le cas par exemple des opinions sur les perspectives d'évolution

de la situation financière individuelle ou sur l'opportunité de faire des achats importants. Par contre, pour d'autres items (par exemple sur les perspectives d'évolution du nombre de chômeurs ou sur l'évolution passée du niveau de vie en France), l'effet de sondage est plus élevé. La prise en compte de la non-réponse et surtout du tirage à probabilités inégales affecte la précision de ces items.

Nous pouvons aussi observer que, pour une question donnée, la date d'enquête semble peu influer sur la précision du niveau d'un solde comme sur l'effet de sondage.

5.3. Précision de l'indicateur résumé « le moral des Français »

Cet indicateur mensuel est la moyenne arithmétique des 5 soldes suivants :

- évolution passée du niveau de vie en France (item y_1)
- perspectives d'évolution du niveau de vie en France (y_2)
- évolution passée de la situation financière personnelle (y_3)
- perspectives d'évolution de la situation financière personnelle (y_4)
- opportunité de faire des achats importants (y_5)

En reprenant les notations précédentes, cet indicateur peut donc s'écrire :

$$\boxed{\hat{\bar{Y}} = \frac{1}{5}\sum_{j=1}^{5}\hat{\bar{Y}}_j = \frac{1}{5r}\sum_{k\in S_2}\left(y_{1,k}+y_{2,k}+y_{3,k}+y_{4,k}+y_{5,k}\right) = \frac{1}{r}\sum_{k\in S_2} t_k}$$

où $t_k = \frac{1}{5}\left(y_{1,k}+y_{2,k}+y_{3,k}+y_{4,k}+y_{5,k}\right)$.

On calcule $\boxed{\hat{Var}(\hat{\bar{Y}}) = \hat{Var}(\hat{T})}$ selon les mêmes principes que précédemment en remplaçant les quantités y_k par les t_k.

On peut a priori escompter une plus grande précision pour l'indicateur résumé que pour chacun des cinq soldes d'opinion qui le construisent car on a toujours $V(\hat{\bar{Y}}) \le \frac{1}{5}\sum_{j=1}^{5} V(\hat{\bar{Y}}_j)$ (plus les corrélations entre les cinq soldes élémentaires seront faibles, plus la variance du « moral des Français » sera faible).

Le tableau 3 fournit la précision de l'indicateur résumé en mars, juillet et octobre 2004. Ainsi, en octobre 2004, le pourcentage de Français « pessimistes » dépassait la proportion des plus « optimistes » de 21,7 points. Cet écart serait environ compris entre -23,7% et -19,8%.

Tableau 3 : précision obtenue pour le « moral des Français » de mars, juillet ou octobre 2004
(cf. annexe 4 pour les résultats détaillés par scénario)

Indicateur	*Mois d'enquête 2004*	*Solde estimé*	*Intervalle de confiance à 95%*		*Effet de sondage par rapport au scénario 1 (SAS dans P)*		
			Scénario1	*Scénario 6*	*Scénario2*	*Scénario 4*	*Scénario 6*
Le moral des Français	Mars	**-0,2237**	± 0,01619	**± 0,0188**	1	1,12	1,35
	Juillet	**-0,1983**	± 0,01675	**± 0,0192**	1	1,13	1,31
	Octobre	**-0,2172**	± 0,01672	**± 0,0194**	1	1,13	1,35

5.4. Précision d'une évolution par différence

Les soldes étant commentés en évolution, nous sommes naturellement appelés à évaluer la précision de l'évolution d'un solde entre deux dates. Deux cas de figure sont à distinguer :

- Comparer deux soldes calculés sur des échantillons disjoints, aux mois *m* et *m'* (où *m'-m ≥3)*
- Comparer deux soldes mensuels consécutifs (du mois *m* au mois *m+1*) ou espacés d'un mois (du mois *m* au mois *m+2*). Ces soldes sont calculés sur un échantillon partiellement renouvelé.

5.4.1 Comparaison d'une évolution entre deux soldes établis à plus de deux mois d'intervalles sur des échantillons disjoints

La précision de l'évolution d'un solde entre les mois *m* et *m'* (avec *m'-m ≥3)*, notée $\hat{\bar{Y}}_{m'} - \hat{\bar{Y}}_{m}$, peut s'obtenir simplement à partir de la précision de chaque solde mensuel, pourvu de supposer que les échantillons des mois *m* et *m'* sont indépendants. Ceci n'est pas rigoureusement exact puisqu'une même ligne ne peut être sélectionnée à plusieurs reprises. Cependant, cette hypothèse se conçoit assez naturellement vu les tailles d'échantillon en jeu (la probabilité pour qu'un abonné joignable soit sélectionné vaut environ 0,02%).

Nous obtenons alors :

$$\boxed{\hat{Var}\left(\hat{\bar{Y}}_{m'} - \hat{\bar{Y}}_{m}\right) = \hat{Var}\left(\hat{\bar{Y}}_{m'}\right) + \hat{Var}\left(\hat{\bar{Y}}_{m}\right)}.$$

Comme remarqué précédemment, la dispersion d'une variable s'avère assez stable dans le temps sur les exemples que nous avons étudiés. Il s'avère donc que :

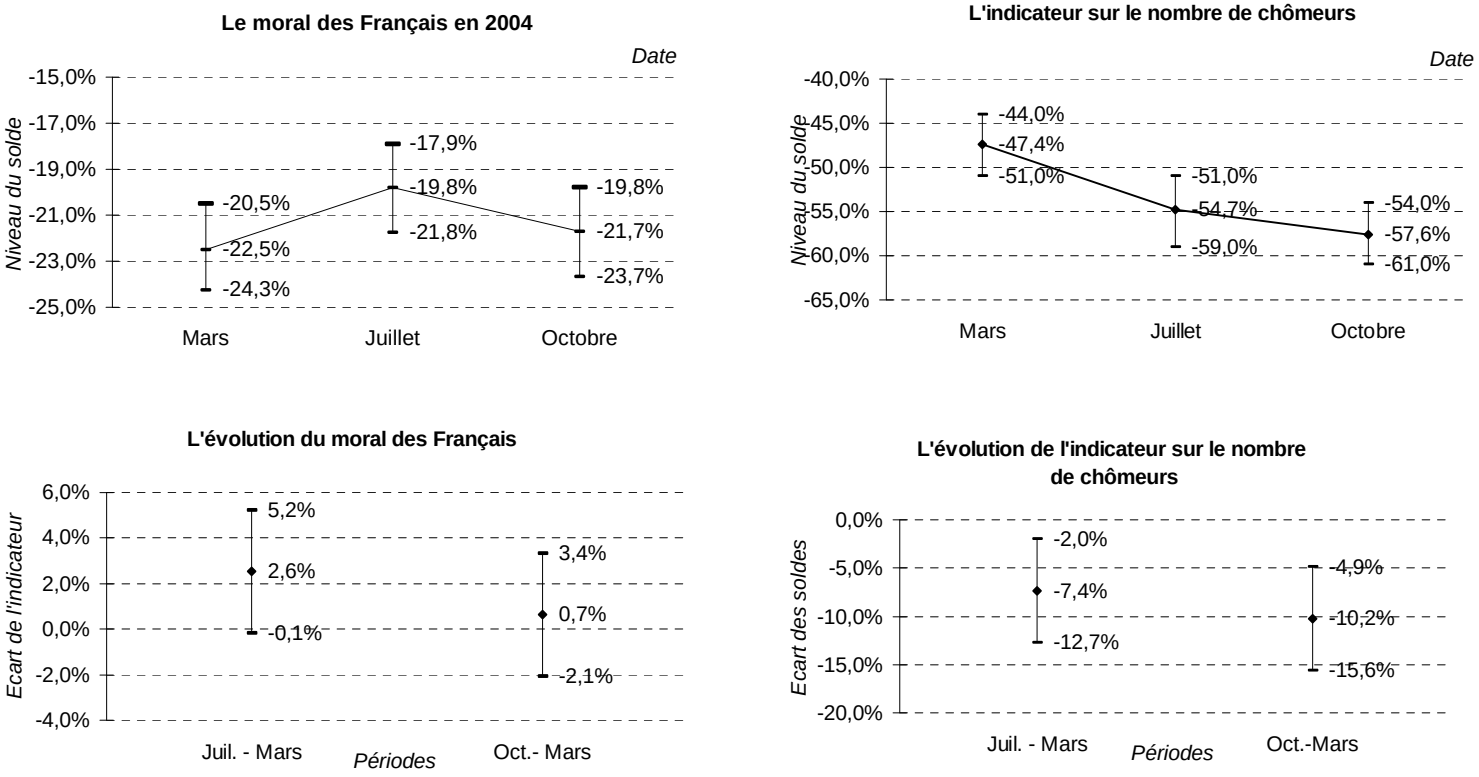
$$\hat{Var}\left(\hat{\bar{Y}}_{m'} - \hat{\bar{Y}}_{m}\right) \cong 2 \times \hat{Var}\left(\hat{\bar{Y}}\right).$$

Les résultats exposés dans le tableau 4 et la figure 2 se rapportent aux deux périodes allant de mars à juillet 2004 et de mars à octobre 2004. Ces dates rendent compte des différences d'opinion intervenues à un trimestre ou un semestre d'intervalle, avant l'été et après la rentrée.

Tableau 4 : précision d'évolutions entre mars et juillet 2004 et entre mars et octobre 2004
(cf. annexe 5 pour les résultats détaillés par scénario)

Indicateur	*Périodes de 2004*	*Différence entre les soldes sur la période (en points)*	*Intervalle de confiance à 95% (en point) (Scénario 6)*	*Effet de sondage par rapport au scénario 1 (SAS dans P)*
Evolution passée du niveau de vie en France	de mars à juil.	5,3	± 5,0	1,81
	de mars à oct.	6,2	± 5,0	1,75
Perspectives d'évolution du niveau de vie en France	de mars à juil.	2,9	± 4,3	1,19
	de mars à oct.	-1,1	± 4,3	1,22
Evolution passée de la situation financière des ménages	de mars à juil.	1,7	± 3,8	1,08
	de mars à oct.	0,5	± 3,9	1,08
Perspectives d'évolution de la situation financière des ménages	de mars à juil.	-5,4	± 3,6	1,03
	de mars à oct.	-2,4	± 3,5	1,03
Opportunité de faire des achats importants	de mars à juil.	8,3	± 4,1	1,03
	de mars à oct.	0,1	± 4,2	1,05
Perspectives d'évolution de la situation économique générale	de mars à juil.	2,1	± 4,6	1,07
	de mars à oct.	-10 ,6	± 4,7	1,11
Perspectives d'évolution du nombre de chômeurs	de mars à juil.	-7,4	± 5,4	1,58
	de mars à oct.	-10,2	± 5,3	1,62
Le moral des Français	de mars à juil.	2,6	± 2,6	1,33
	de mars à oct.	0,7	± 2,7	1,35

Figure 2 : Précision de l'évolution de l'indicateur résumé et de l'opinion sur le nombre de chômeurs de mars à juillet 2004 et de mars à octobre 2004



Entre juillet et mars 2004, le moral des Français affichait une augmentation voisine de 3 points, en passant de –22,5 points à –19,8 points environ. Ce regain de confiance semble probant, quoique à la limite de la significativité puisque l'augmentation de l'indicateur résumé au cours de ce trimestre serait comprise entre 0 et 5 points environ. La situation semble s'être inversée un peu plus tard puisque nos résultats ne permettent pas de distinguer d'évolution significative, dans un sens ou dans un autre, de mars à octobre 2004, l'indicateur variant entre –2,1 points et +3,4 points. Sur ces deux périodes, l'évolution de l'indicateur résumé est connue à environ ± 2,6 points près.

Les différences sont plus marquées pour l'indicateur relatif à l'évolution du nombre de chômeurs. Deux phénomènes co-existent : d'une part, les évolutions, qu'elles soient calculées sur un trimestre ou un semestre, sont toujours négatives et témoignent d'une morosité toujours accrue. D'autre part, ces évolutions sont connues avec une incertitude relativement importante, proche de ± 5 points. Par exemple, de mars à octobre 2004, les Français étaient de plus en nombreux à ne pas envisager d'amélioration sur la situation du chômage. La baisse de l'indicateur sur cette période se situerait entre 5 et 15 points environ.

5.4.2. Comparaison de l'évolution entre deux soldes consécutifs ou espacés d'un mois

Dans ce cas de figure, l'échantillon des répondants est partiellement renouvelé entre les mois *m* et *m'* (*m'-m=1,2).* La précision de la différence entre les deux soldes tient donc compte de l'évolution des opinions individuelles au cours du laps de temps considéré. Plus précisément, on a :

$$Var\left(\hat{\bar{Y}}_{m'} - \hat{\bar{Y}}_{m}\right) = Var\left(\hat{\bar{Y}}_{m}\right) + Var\left(\hat{\bar{Y}}_{m'}\right) - 2Cov\left(\hat{\bar{Y}}_{m}, \hat{\bar{Y}}_{m'}\right)$$

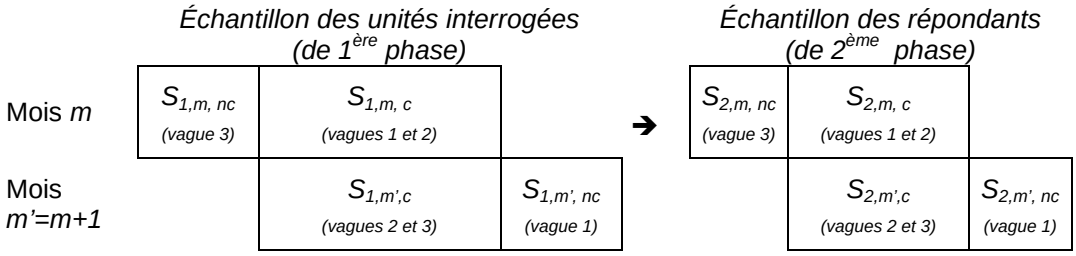
L'estimation de cette évolution est a priori plus précise que celle ayant trait à une période plus longue (puisqu'on bénéficie, en général, d'une corrélation positive entre les grandeurs observées pour un même individu à deux dates consécutives). La corrélation temporelle de l'opinion se confirme ici : par exemple, entre octobre et juillet 2004, le coefficient de corrélation linéaire varie entre 0,3 et 0,5 selon les items considérés (cf. tableau 5 infra).

Une des difficultés ici tient au fait que la non-réponse peut être différenciée selon les vagues d'interrogations : certains individus enquêtés à deux ou trois reprises n'ont répondu qu'une seule fois ou deux fois.

Pour lever cette difficulté, nous avons d'abord considéré le cadre d'un tirage aléatoire simple de taille fixe (scénario 1) pour lequel nous avons plus facilement obtenu une première estimation de la précision. Nous avons ensuite appliqué l'effet de sondage obtenu au paragraphe 5.4.1 (relativement constant dans le temps pour un item donné) afin d'obtenir une estimation de la précision de meilleure qualité sous le scénario 6, plus fidèle au plan de sondage.

Les parties communes et non communes de l'échantillon des unités interrogées et de l'échantillon des répondants peuvent être représentées comme sur la figure 3. Elles sont respectivement indicées « c » et « nc » dans toutes les notations.

Figure 3 : composition des échantillons dans le dispositif rotatif dans le cas où *m'=m+1*



Définissons les taux de chevauchement de l'échantillon des répondants aux mois *m* et *m',* notés θ_m et $\theta_{m'}$, de la manière suivante :

$$\theta_m = \frac{r_c}{r_m} \text{ et } \theta_{m'} = \frac{r_c}{r_{m'}}$$

où r_m (resp. $r_{m'}$) compte le nombre de répondants à la date *m* (resp. *m'*) et r_c le nombre de répondants communs aux deux mois *m* et *m'.*

L'évolution du solde peut s'écrire :

$$\hat{\bar{Y}}_{m'} - \hat{\bar{Y}}_{m} = \frac{1}{r_{m'}} \sum_{k \in S_{2,m'}} y_k - \frac{1}{r_m} \sum_{k \in S_{2,m}} y_k$$

$$= (1-\theta_{m'}) \cdot \hat{\bar{Y}}_{m',nc} + \theta_{m'} \cdot \hat{\bar{Y}}_{m',c} - (1-\theta_m) \cdot \hat{\bar{Y}}_{m,nc} - \theta_m \cdot \hat{\bar{Y}}_{m,c}$$

où $\hat{\bar{Y}}_{m',c} = \frac{1}{r_c} \sum_{k \in S_{2,m',c}} y_k$; $\hat{\bar{Y}}_{m',nc} = \frac{1}{r_{m'} - r_c} \sum_{k \in S_{2,m',nc}} y_k$; $\hat{\bar{Y}}_{m,c} = \frac{1}{r_c} \sum_{k \in S_{2,m,c}} y_k$ et $\hat{\bar{Y}}_{m,nc} = \frac{1}{r_m - r_c} \sum_{k \in S_{2,m,nc}} y_k$.

En faisant l'hypothèse que les échantillons $S_{2,m,nc}$, $S_{2,m',nc}$ et $S_{2,m,c}\left(=S_{2,m',c}=S_{2,c}\right)$ sont indépendants, ce qui est légitime vu les tailles d'échantillons considérées et le mécanisme de rotation, nous obtenons donc :

$$Cov\left(\hat{\bar{Y}}_{m'},\hat{\bar{Y}}_{m}\right)=Cov\left(\theta_{m'}\cdot\hat{\bar{Y}}_{m',c},\theta_{m}\cdot\hat{\bar{Y}}_{m,c}\right).$$

Si de plus les échantillons de répondants étaient obtenus par sondage aléatoire simple de taille fixe pris dans la population des ménages (scénario 1), on aurait :

$$Var\left(\hat{\bar{Y}}_{m'}-\hat{\bar{Y}}_{m}\right)=Var\left(\hat{\bar{Y}}_{m'}\right)+Var\left(\hat{\bar{Y}}_{m}\right)-2\cdot\theta_{m}\cdot\theta_{m'}\cdot Cov\left(\hat{\bar{Y}}_{m',c},\hat{\bar{Y}}_{m,c}\right).$$

Ce qui nous permet d'obtenir une première estimation de la variance des évolutions d'un solde pour des échantillons « chevauchant » :

$$\hat{Var}_{scenario\_1}\left(\hat{\bar{Y}}_{m'}-\hat{\bar{Y}}_{m}\right)=\hat{Var}_{scenario\_1}\left(\hat{\bar{Y}}_{m'}\right)+\hat{Var}_{scenario\_1}\left(\hat{\bar{Y}}_{m}\right)-2\cdot\theta_{m}\cdot\theta_{m'}\cdot\hat{Cov}_{scenario\_1}\left(\hat{\bar{Y}}_{m',c},\hat{\bar{Y}}_{m,c}\right).$$

Les taux de sondage étant très faibles et les dispersions relativement stables dans le temps, nous obtenons comme approximation de la variance :

$$\hat{Var}_{scenario\_1}\left(\hat{\bar{Y}}_{m'}-\hat{\bar{Y}}_{m}\right)\cong\frac{\hat{s}_{m}^{2}}{r_{m}}+\frac{\hat{s}_{m'}^{2}}{r_{m'}}-2\theta_{m}\theta_{m'}\frac{\hat{s}_{m,m'}}{r_{c}}\cong\hat{s}^{2}\left(\frac{1}{r_{m}}+\frac{1}{r_{m'}}-2\theta_{m}\theta_{m'}\frac{\hat{\rho}}{r_{c}}\right)$$

où $\hat{\rho}$ estime le coefficient de corrélation de la variable étudiée entre les deux périodes considérées.

Comme le design-effect entre le scénario et le scénario 6 est relativement constant dans le temps pour un item donné, nous déduisons de cette première estimation une estimation de meilleure qualité définie par :

$$\boxed{\hat{Var}\left(\hat{\bar{Y}}_{m'}-\hat{\bar{Y}}_{m}\right)_{scenario\_6}=\hat{D}_{6/1}\times\hat{Var}\left(\hat{\bar{Y}}_{m'}-\hat{\bar{Y}}_{m}\right)_{scenario\_1}}$$

où $\hat{D}_{1/6}$ désigne l'effet de sondage obtenu pour l'estimation de la variance du solde considéré sous le scénario 6 (sondage systématique à probabilité inégale suivi d'une deuxième phase poissonnienne.

Le tableau 5 livre les résultats obtenus pour la période allant de juillet à octobre pour laquelle une vague d'interrogation est commune (les individus entrants de juillet).

Tableau 5 : Précision de l'évolution des soldes entre juillet et octobre 2004 et coefficients de corrélation entre les opinions à ces deux dates

Indicateur	***Différence de solde (en points)***	***Amplitude de l'intervalle de confiance à 95% (en points)***	***Coefficient de corrélation linéaire entre les opinions***
Evolution passée du niveau de vie en France	0,9	4,71	0,40
Perspectives d'évolution du niveau de vie en France	-4,0	4,12	0,34
Evolution passée de la situation financière des ménages	-1,2	3,64	0,41
Perspectives d'évolution de la situation financière des ménages	3,0	3,43	0,39
Opportunité de faire des achats importants	-8,2	3,99	0,32
Perspectives d'évolution de la situation économique générale	-12,8	4,57	0,38
Perspectives d'évolution du nombre de chômeurs	-2,9	5,18	0,34
Le moral des Français	-1,9	2,51	0,50

Ces chiffres permettent donc de conclure par exemple que la tendance à la baisse du moral des Français ne semble pas significative entre juillet et octobre 2004.

5.5. Conclusion

Différents prolongements de nos travaux sont envisagés :

- Utiliser une approximation plus juste des probabilités d'inclusion d'ordre deux. Il nous faudrait pour cela disposer des probabilités d'ordre un cumulées pour chaque unité échantillonnée. Cette grandeur est calculable par Wanadoo Data.
- Estimer la variance en supposant que le tirage de l'échantillon de 1ère phase s'effectue avec remise dans la base des abonnés joignables. Le très faible taux de sondage intercède en faveur de cette simplification.
- Utiliser la méthode de partage des poids et adapter le questionnaire en demandant aux unités interrogées le nombre de lignes qu'elles possèdent.

D'autre part, les soldes publiés par la division des Comptes Trimestriels pourraient à l'avenir être calés, sur les marges de l'enquête Emploi, au même titre que les données transmises à l'Union Européenne. Pour calculer la précision de l'estimateur obtenu après calage, nous pourrions, de manière assez classique, adapter nos calculs en raisonnant cette fois non pas sur la variable d'opinion elle-même, mais sur le résidu de la régression de cette variable sur les variables de calage.

Cependant, il convient de tenir compte ici des difficultés particulières que présente le calage de l'enquête CAMME:

- le concept d'unité statistique diffère entre la base de sondage et la source qui donne les marges de référence: l'échantillon provient d'une base d'abonnés joignables et la population d'inférence est celle des ménages.
- la définition de la variable « habitat collectif ou individuel » de Wanadoo Data n'est pas identique à celle du concept INSEE.
- la personne de référence peut différer du titulaire de la ligne.

Soulignons pour finir qu'une étude en cours recherche les déterminants de l'opinion des ménages et qu'un bénéfice serait de pouvoir en déduire une stratification plus efficace pour l'enquête.

6. Conclusions et perspectives d'évolution.

6.1. Le bilan de la refonte de CAMME.

Ce bilan met en évidence un premier paradoxe : alors qu'il s'agit (sans doute ...) de l'une des enquêtes les plus simples de l'Insee (en termes de questionnement, de traitement, de rapidité d'exploitation ...), la refonte de CAMME a mobilisé des moyens importants, tant en méthodologie statistique, en informatique, en conduite de projet qu'en management des différentes équipes concernées. Et sans doute l'une des difficultés de la refonte a-t-elle été la multiplicité des acteurs, y compris, une fois n'est pas coutume, en amont dès l'échantillonnage, contrairement à toutes les autres enquêtes de l'Insee.

C'est dire que cette refonte constitue un investissement important dont la valeur doit être jugée à l'aune des bénéfices qu'elle procure. En même temps, toutes les innovations introduites dans l'enquête constituent d'un certain point de vue un modèle réduit de tous les types de traitements et de toutes les étapes de la mise en œuvre et de l'exploitation d'une enquête standard. Et malgré son apparente simplicité, l'enquête CAMME a cristallisé des difficultés qu'une enquête plus complexe ne renierait pas, notamment en matière de calcul de précision : c'est aussi là que réside l'intérêt du présent papier, au-delà de la seule enquête CAMME.

Au final, on peut dire que ces investissements, même sur une enquête réputée modeste, sont parfaitement justifiés compte tenu de l'intérêt que présentent ses résultats et de leur très bonne couverture médiatique.

Ceci étant, la refonte a permis de nombreuses améliorations sur de nombreux domaines et permettent de mesurer le chemin parcouru depuis le premier lancement de l'enquête expérimentale CAMME en 1986 :

- Satisfaction complète des exigences européennes en matière de questionnement.

- Maîtrise des différents maillons de la chaîne de production, notamment en matière d'échantillonnage, ce qui a permis à la fois de clarifier les relations avec le prestataire et de mieux contrôler les spécifications de la conception et du tirage proprement dit.

- Plus grande rigueur dans le respect des calendriers, à toutes les étapes de la production.

- Amélioration et raffinement des procédures de traitement aval, même si celles-ci ne procurent pas complètement tous les avantages attendus (notamment l'impact de la procédure de calage).

- Amélioration du taux de réponse de l'ordre de 10 points par rapport à l'ancienne version ; essentiellement, il s'agit de la baisse des IAJ résultant d'une base de sondage mise à jour en permanence (alors que dans l'ancienne version, la base vieillissait au cours de l'année), mais aussi de la baisse des refus. De plus, la gestion des rappels est facilitée pour les enquêteurs grâce à l'environnement CAPI.

- Clarification du rôle des (nombreux) acteurs du processus.

6.2. La poursuite des travaux mis en œuvre dans le présent papier.

Il s'agit essentiellement de parachever les calculs de précision sur les indicateurs *calés* qui, à terme, devraient être la référence unique en matière de publication des résultats.

A terme, un module de calculs des intervalles de confiance devrait être mis à disposition de la Division des Comptes trimestriels, afin que celle-ci puisse effectuer à sa guise les calculs de précision. On peut gager au demeurant que ce calcul ne sera pas nécessaire chaque mois, les résultats devant présenter une certaine robustesse qui a d'ailleurs été vérifiée sur les quelques mois qui ont servi de banc d'essai.

6.3. Les perspectives d'amélioration à court et moyen terme.

Afin de finaliser complètement la refonte, quelques travaux doivent encore être menés dans les prochains mois.

- En termes d'échantillonnage :

Un problème subsiste concernant le calcul, par Wanadoo Data, des taux de pénétration au niveau de l'IRIS des numéros en liste rouge, pour certaines adresses en logement individuel. Pour une première livraison, les taux de pénétration ont donc été calculés au niveau communal seulement.

Pour les livraisons suivantes, nos préconisations sont les suivantes :

- pour les communes partitionnées en IRIS :
  . on partitionne les n°s d'une commune en deux strates (ce qui est d'ailleurs déjà prévu) : collectif / individuel
  . on calcule des taux de pénétration liste rouge par IRIS pour tous les n°s en collectif
  . on calcule des taux de pénétration liste rouge par commune pour tous les n°s en individuel.

- pour les petites communes non partitionnées en IRIS, les taux de pénétration sont calculés au niveau de la commune.

- En termes de traitement statistique aval :

Un certain nombre d'améliorations pourraient être apportées à la modélisation des revenus, qui n'auront cependant qu'un effet très limité sur les résultats produits, du fait que l'on ne diffuse les résultats que par quartile de revenus... Les traitements d'imputation et de traitements de la non-réponse ont en effet été récemment présentés à David Haziza (Statistique Canada), lors d'une séance de formation au traitement de la Non-Réponse. Celui-ci a validé les principes retenus, mais suggéré quelques pistes méthodologiques d'amélioration (régression avec résidus aléatoires…), sous réserve qu'elles soient compatibles avec les délais tendus de publication des résultats et que leur mise en œuvre ne soit pas trop coûteuse en temps.

- En termes de production informatique :

La mise en production sur le Host du Centre Informatique d'Orléans (CNIO) des traitements statistiques, prévue à l'origine en juin 2004, a été reportée au premier semestre 2005. Elle ne devrait pas poser de difficultés, les chaînes informatiques ayant déjà été testées sur gros système. On y gagnera en automatisation, donc en délais, mais on perdra cependant une certaine souplesse, qui a pu être utile récemment : ainsi, la Commission européenne avait-elle demandé de séparer les résultats avant et après le 11 mars 2004 (attentats de Madrid), ce qui a pu être fait en filtrant à l'entrée sur la date... Ce type de traitements exceptionnels sera plus difficile à mettre en œuvre sur le CNIO.

Si cette mise en production ne doit pas être remise en cause, il serait intéressant que le pôle des Enquêtes nationales Ménages de Nancy garde cependant l'outil actuel de traitement statistique en maintenance, au moins pendant un certain temps, pour pouvoir répondre à des demandes spécifiques.

6.4. Travaux d'étude à venir.

Les indicateurs issus de l'enquête CAMME servent, comme il a été dit précédemment, à l'analyse conjoncturelle et sont introduits dans des équations économétriques de prévision de la consommation. En revanche, peu d'études ont été faites, jusqu'à présent, sur la compréhension de l'opinion exprimée.

Plusieurs axes d'étude pourraient donc être entrepris dans un futur proche.

- Une première gamme d'analyse tourne autour des *facteurs explicatifs de l'opinion*. Les premiers calculs de précision menés sur les indicateurs recalés avaient en effet tendance à mettre en évidence un impact relativement faible de la procédure de calage ; ceci peut s'expliquer par le fait que les facteurs socio-démographiques traditionnels jouent peu dans la détermination de l'opinion. Certes les variables dont on dispose dans cette enquête sont en relativement faible nombre mais il faudrait élargir la gamme, en faisant de manière systématique des régressions logistiques sur l'ensemble des réponses aux questions d'opinion.

- Au-delà des facteurs explicatifs de l'opinion, on a vu que c'était la variation des indicateurs résumés qui donnait son sens aux soldes d'opinion : d'où la question de savoir qui « fait » l'opinion, au sens de quels sont les facteurs explicatifs de la variation d'opinion : est-elle le fait plutôt de telle catégorie que de telle autre ? de quel ordre de grandeur est la variation de l'opinion en fonction des catégories ? A partir de quand une variation d'opinion est-elle significative ? y a-t-il des variations importantes de l'opinion au cours de trois mois consécutifs (analyse longitudinale)

  En particulier, ces travaux devraient compéter de premières constatations faites à l'occasion du lancement de CAMME en 1989 (cf. [5]) : il avait été mis en évidence en effet que la variation d'opinion d'un mois sur l'autre était assez fortement corrélée au changement de répondant. Il faudrait donc vérifier si ce constat tient toujours - c'est-à-dire analyser les profils des répondants et regarder la variabilité des réponses en fonction de l'identité du répondant - mais il est probable que oui, ce qui pourrait expliquer la relative difficulté à améliorer les résultats de l'enquête par calage ; car, par construction, les variables mises dans les équations de calage sont relatives au ménage (ou au « foyer téléphonique ») dans son ensemble : si donc on met en évidence une variabilité intra-ménage, celle-ci sera sans doute irréductible par les techniques de calage employées.

- D'autres travaux, fondés sur l'analyse des données, pourraient permettre de construire l'indicateur résumé synthétique le plus approprié, c'est-à-dire contenant le maximum d'information sur la variabilité de l'opinion (cf. [7]).

- Enfin, il conviendrait de pouvoir vérifier la validité de l'hypothèse forte avancée en matière d'échantillonnage, à savoir la neutralité du caractère incomplet de la base de sondage, qui n'atteint ni les seuls titulaires d'une ligne de téléphone mobile, ni ceux qui sont inscrits sur liste d'opposition.

  Des premières données issues de l'enquête Permanente sur les conditions de vie (PCV) de janvier 2005 devraient dans un premier temps apporter un éclairage, d'une part sur la substitution du téléphone mobile au filaire ou sur la multi-possession, en les analysant par catégorie, d'autre part sur les catégories les plus touchées par la « pénétration de la liste rouge ».

A terme, une véritable opération méthodologique serait utile mais sa programmation dépend des moyens disponibles.

L'idée serait de faire une enquête méthodologique sur CAMME. Il s'agirait de poser exactement les mêmes questions que CAMME, mais cette fois en face à face et sur un échantillon de logements et non plus de numéros de téléphone. Outre les questions d'opinion, quelques questions seraient posées sur la possession d'un téléphone fixe, portable et sur l'inscription sur une liste rouge ou orange. Cette enquête se ferait en même temps qu'une enquête CAMME habituelle.

L'intérêt serait alors de comparer, sur une période identique, les réponses aux questions d'opinion par deux protocoles de collecte et, surtout, de pouvoir vérifier si les réponses sont corrélées à l'attitude du ménage vis-à-vis de la téléphonie (mobile / filaire et liste d'opposition).

Naturellement, cette étude devrait prendre garde à plusieurs facteurs parasites :

- les dates d'enquêtes ne seraient pas rigoureusement identiques, la période de collecte de CAMME étant particulièrement courte (il faudrait garder la date d'enquête dans le fichier de l'enquête méthodologique).

- les protocoles de collecte seraient différents : capte-t-on de la même façon l'opinion au téléphone et en face à face ?

- si les deux échantillons des enquêtes sont bien disjoints, les réponses ne seront pas rigoureusement comparables puisque CAMME est en trois vagues, l'enquête méthodologique en une seule vraisemblablement.

Bibliographie

[1] CARON N. (1993) - Réflexions sur les erreurs de mesure, l'exemple de l'enquête de conjoncture auprès des ménages- *Document de travail n°9308 de la série des documents de travail de la Direction des Statistiques Démographiques et Sociales de l'INSEE.*

[2] CARON N., DEVILLE J.-C. et SAUTORY O. (1998) - Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE- *Document de travail n°9806 de la série Méthodologie statistique, INSEE.*

[3] CARON N., RAVALET P. et SAUTORY O. (1996) - Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises - *Document de travail n°9602 de la série Méthodologie statistique, INSEE.*

[4] CARON N. et RAVALET P. (2000) - Estimation dans les enquêtes répétées : application à l'Enquête Emploi en continu - *Document de travail n°0005 de la série Méthodologie statistique, INSEE.*

[5] CHRISTINE M. (1989), Enquête mensuelle de conjoncture auprès des ménages, rapport de fin d'expérimentation, 29 mai 1989.

[6] DEVILLE J.-C. (1998) - Estimation de variance pour des statistiques complexes : techniques des résidus et linéarisation - *Document de travail n°9802 de la série Méthodologie statistique, INSEE.*

[7] HILD F. (2002), « Les soldes d'opinion résument-ils au mieux les réponses aux enquêtes de conjoncture ? », Journées de Méthodologie statistique, Insee, décembre 2002.

[8] INSEE, Publications mensuelles des résultats de l'enquête*, Informations rapides*, site internet http://www.insee.fr/fr/indicateur/indic_conj/

[9] INSEE, Présentation de l'enquête de conjoncture mensuelle auprès des ménages (CAMME), site intranet de la Direction des Statistiques Démographiques et Sociales de l'INSEE.

[10] LOLLIVIER S. (1999) - Anticipations des ménages et environnement économique - *Economie et statistique N° 324-325, INSEE,aout 1999.*

[11] TILLE Y. (2001) - Échantillonnage et estimation en populations finies, *Dunod, Paris.*

[12] WOLTER K. M. (1985) - Introduction to Variance Estimation - *Springer-Verlag, New-York.*

Annexe 1 : Formulaire d'un plan de sondage en 2 phases : expression de l'estimateur d'un total, de sa variance et de sa variance estimée

Le total *Z* d'une variable *z* peut être estimé sans biais par :

$$\boxed{\hat{Z} = \sum_{k \in S_2} \frac{z_k}{\pi_{1k}\pi_{2k}}} \quad \text{où } \pi_{1k} = P(k \in S_1) \text{ et } \pi_{2k} = P(k \in S_2 / S_1)$$

Sa variance vaut :

$$\boxed{Var\left(\hat{Z}\right) = \sum_{k \in U}\sum_{l \in U} \frac{z_k z_l}{\pi_{1k}\pi_{1l}} \Delta_{1kl} + E\left(\sum_{k \in S_2}\sum_{l \in S_2} \frac{z_k z_l}{\pi_{1k}\pi_{1l}\pi_{2k}\pi_{2l}} \Delta_{2kl}\right)}$$

où $\Delta_{1kl} = \begin{cases} \pi_{1kl} - \pi_{1k}\pi_{1l} & k \neq l \\ \pi_{1k}\left(1 - \pi_{1k}\right) & k = l \end{cases}$ avec $\pi_{1kl} = P\left(k \in S_1, l \in S_1\right)$

et $\Delta_{2kl} = \begin{cases} \pi_{2kl} - \pi_{2k}\pi_{2l} & k \neq l \\ \pi_{2k}\left(1 - \pi_{2k}\right) & k = l \end{cases}$ avec $\pi_{2kl} = P(k \in S_2, l \in S_2 / S_1)$

Cette variance peut être estimée par :

$$\boxed{\hat{Var}\left(\hat{Z}\right) = \sum_{k \in S_2}\sum_{l \in S_2} \frac{z_k z_l}{\pi_{1k}\pi_{1l}} \frac{\Delta_{1kl}}{\pi_{1kl}\pi_{2kl}} + \sum_{k \in S_2}\sum_{l \in S_2} \frac{z_k z_l}{\pi_{1k}\pi_{1l}\pi_{2k}\pi_{2l}} \frac{\Delta_{2kl}}{\pi_{2kl}}}$$

Le principal problème réside dans le calcul pratique des probabilités d'inclusion doubles des tirages de 1ère et de 2ème phase.

Dans *<u>le cas particulier où la 2nde phase obéit à un plan poissonnien</u>*, les comportements de réponse individuels sont indépendants et on a :

$$\pi_{2kl} = \begin{cases} \pi_{2k}\pi_{2l} & k \neq l \\ \pi_{2k} & k = l \end{cases} \quad \text{et} \quad \Delta_{2kl} = \begin{cases} 0 & k \neq l \\ \pi_{2k}\left(1 - \pi_{2k}\right) & k = l \end{cases}$$

Ainsi :

$$\begin{aligned} \hat{Var}\left(\hat{Z}\right) &= \sum_{k \in S_2}\sum_{l \in S_2} \frac{z_k z_l}{\pi_{1k}\pi_{1l}} \frac{\Delta_{1kl}}{\pi_{1kl}\pi_{2kl}} + \sum_{k \in S_2}\left(\frac{z_k}{\pi_{1k}\pi_{2k}}\right)^2 \left(1 - \pi_{2k}\right) \\ &= \sum_{k \in S_2}\left(\frac{z_k}{\pi_{1k}}\right)^2 \frac{1 - \pi_{1k}}{\pi_{2k}} + \sum_{k \in S_2}\sum_{\substack{l \in S_2 \\ l \neq k}} \frac{z_k z_l}{\pi_{1k}\pi_{1l}} \frac{\Delta_{1kl}}{\pi_{1kl}} \frac{1}{\pi_{2k}\pi_{2l}} + \sum_{k \in S_2}\left(\frac{z_k}{\pi_{1k}\pi_{2k}}\right)^2 \left(1 - \pi_{2k}\right) \end{aligned}$$

Soit :

$$\boxed{\hat{Var}\left(\hat{Z}\right) = \sum_{k \in S_2}\left(\frac{z_k}{\pi_{1k}\pi_{2k}}\right)^2 \left(1 - \pi_{1k}\pi_{2k}\right) + \sum_{k \in S_2}\sum_{\substack{l \in S_2 \\ l \neq k}} \frac{z_k z_l}{\pi_{1k}\pi_{1l}} \frac{1}{\pi_{2k}\pi_{2l}} \frac{\Delta_{1kl}}{\pi_{1kl}}}$$

Annexe 2 : Estimation de la précision de l'estimation du solde dans chacun des scénarii envisagés

1. *Scénario 1 : Plan simple sans remise de taille fixe de r répondants parmi M unités*

Nous considérons donc l'estimateur :

$$\hat{\bar{Y}}_{scenario\_1} = \frac{1}{r}\sum_{k\in S_2} y_k \ .$$

Sa variance peut s'estimer par :

$$\boxed{\hat{Var}\left(\hat{\bar{Y}}_{scenario\_1}\right) = \left(1-\frac{r}{M}\right)\frac{\hat{s}_y^2}{r}}$$

où $\hat{s}_y^2 = \frac{1}{r-1}\sum_{k\in S_2}\left(y_k - \hat{\bar{Y}}_{scenario\_1}\right)^2$ estime sans biais la dispersion $S_y^2 = \frac{1}{M-1}\sum_{k=1}^{M}\left(y_k - \bar{Y}\right)^2$ sous les hypothèses d'homogénéité des comportements admises précédemment, c'est-à-dire :

$$E\left(\frac{1}{r}\sum_{k\in S_2} y_k^2\right) = \frac{1}{M}\sum_{k=1}^{M} y_k^2 \ .$$

2. *Scénario 2 : Plan stratifié avec sondage aléatoire simple sans remise dans chaque strate h de taille fixe de r_h répondants parmi M_h unités*

L'estimateur du solde s'écrit :

$$\hat{\bar{Y}}_{scenario\_2} = \frac{1}{r}\sum_{h=1}\sum_{k\in S_{2h}} y_k = \sum_{h=1}\frac{r_h}{r}\hat{\bar{Y}}_h \qquad \text{où} \qquad \hat{\bar{Y}}_h = \frac{1}{r_h}\sum_{k\in S_{2h}} y_k \ .$$

Sa variance peut s'estimer avec :

$$\boxed{\hat{Var}\left(\hat{\bar{Y}}_{scenario\_2}\right) = \sum_{h=1}\left(\frac{r_h}{r}\right)^2\left(1-\frac{r_h}{M_h}\right)\frac{\hat{s}_{yh}^2}{r_h}}$$

où $\hat{s}_{yh}^2 = \frac{1}{r_h - 1}\sum_{k\in S_{2h}}\left(y_k - \hat{\bar{Y}}_h\right)^2$.

3. *Scénario 3 : Plan de 1ère phase stratifié avec sondage aléatoire simple sans remise dans chaque strate h de taille fixe n'$_h$ parmi M_h unités, suivi par un tirage de 2ème phase poissonnien avec, comme probabilité de réponse estimée, le ratio r/n' égal au nombre de répondants divisé par le nombre d'interrogés*

Nous pouvons ré-écrire le solde estimé comme :

$$\hat{\bar{Y}}_{scenario\_3} = \frac{1}{r}\sum_{k\in S_2} y_k = \frac{1}{r}\sum_{k\in S_2}\frac{y_k\pi_{1k}}{\pi_{1k}} = \sum_{k\in S_2}\frac{y_k\pi_{1k}\frac{1}{n'}}{\pi_{1k}\frac{r}{n'}} = \sum_{k\in S_2}\frac{z_k}{\pi_{1k}\pi_{2k}} = \sum_h\sum_{k\in S_{2h}}\frac{z_k}{\pi_{1k}\pi_{2k}}$$

où, pour $k \in U_h$,
$$\begin{cases} z_k = \dfrac{y_k\pi_{1k}}{n'} \\ \pi_{1k} = \dfrac{n'}{M} = \dfrac{n_h'}{M_h} = P(k\in S_1) \ \text{(allocations proportionnelles au nombre de ménages)} \\ \pi_{2k} = P(k\in S_2 / S_1) \ \text{que l'on modélise par } \pi_{2k} = \dfrac{r}{n'} \end{cases}$$

D'après le formulaire rappelé en annexe 1 et adapté au scénario envisagé, nous obtenons pour variance estimée :

$$\hat{Var}\left(\hat{\bar{Y}}_{scenario\_3}\right)=\sum_h\sum_{k\in S_{2h}}\left(\frac{z_k}{\pi_{1k}\pi_{2k}}\right)^2\left(1-\pi_{1k}\pi_{2k}\right)+\sum_{k\in S_{2h}}\sum_{\substack{l\in S_{2h}\\ l\neq k}}\frac{z_k z_l}{\pi_{1k}\pi_{1l}}\frac{1}{\pi_{2k}\pi_{2l}}\frac{\Delta_{1kl}}{\pi_{1kl}}$$

avec $$\pi_{1kl}=P\left(k\in S_1,l\in S_1\right)=\begin{cases}\dfrac{n_h^{'}}{M_h}\cdot\dfrac{n_{h'}^{'}}{M_{h'}} & k\in U_h,l\in U_{h'},h\neq h'\\ \dfrac{n_h^{'}}{M_h} & k=l,k\in U_h\\ \dfrac{n_h^{'}}{M_h}\dfrac{n_h^{'}-1}{M_h-1} & k,l\in U_h,k\neq l\end{cases}$$

et $\Delta_{1kl}=\pi_{1kl}-\pi_{1k}\pi_{1l}$.

4. ***Scénario 4 : Plan de 1ère phase stratifié avec sondage aléatoire simple sans remise dans chaque strate h de taille fixe n'h parmi Mh unités, suivi par un tirage de 2ème phase poissonnien stratifié avec, comme probabilité de réponse estimée dans une strate h quelconque, le ratio rh/n'h égal au nombre de répondants divisé par le nombre d'interrogés de la strate.***

Le solde estimé peut s'écrire ainsi :

$$\hat{\bar{Y}}_{scenario\_4}=\frac{1}{r}\sum_{k\in S_2}y_k=\frac{1}{r}\sum_h\sum_{k\in S_{2h}}\frac{y_k\pi_{1k}}{\pi_{1k}}\frac{\frac{r_h}{n_h^{'}}}{\frac{r_h}{n_h^{'}}}=\frac{1}{\sum_{k\in S_2}1}\times\sum_h\left(\sum_{k\in S_{2h}}\frac{z_k}{\pi_{1k}\pi_{2k}}\times\sum_{k\in S_{2h}}1\right)$$

avec, pour $k\in U_h$, $$\begin{cases}z_k=\dfrac{y_k\pi_{1k}}{n_h^{'}}\\ \pi_{1k}=\dfrac{n'}{M}=\dfrac{n_h^{'}}{M_h}=P\left(k\in S_{1h}\right)\\ \pi_{2k}=P(k\in S_{2h}/S_{1h})\ \text{que l'on modélise par}\ \pi_{2k}=\dfrac{r_h}{n_h^{'}}\end{cases}$$

Ré-écrivons le solde estimé comme :

$$\hat{\bar{Y}}_{scenario\_4}=\frac{\sum_h\hat{A}_h\hat{B}_h}{\sum_h\hat{B}_h}=\frac{\sum_h\hat{A}_h\hat{B}_h}{\hat{B}}$$

avec $\hat{A}_h=\sum_{k\in S_{2h}}\frac{z_k}{\pi_{1k}\pi_{2k}}$; $\hat{B}_h=r_h=\sum_{k\in S_{2h}}1=\sum_{k\in S_{2h}}\frac{\pi_{1k}\pi_{2k}}{\pi_{1k}\pi_{2k}}$ et $\hat{B}=r=\sum_{k\in S_2}1=\sum_h\hat{B}_h$.

La variance de $\hat{\bar{Y}}_{scenario\_4}$ peut s'approcher d'après la linéarisée d'un ratio, c'est-à-dire :

$$\hat{Var}\left(\hat{\bar{Y}}_{scenario\_4}\right)=\frac{1}{\hat{B}^2}\left[\hat{Var}\left(\sum_h\hat{A}_h\hat{B}_h\right)-2\hat{\bar{Y}}\hat{Cov}\left(\sum_h\hat{A}_h\hat{B}_h,\hat{B}\right)+\hat{\bar{Y}}^2\hat{Var}\left(\hat{B}\right)\right]$$

avec $$\begin{cases}\hat{Var}\left(\sum_h\hat{A}_h\hat{B}_h\right)=\sum_h\hat{Var}\left(\hat{A}_h\hat{B}_h\right)\\ \hat{Var}\left(\hat{B}\right)=\sum_h\hat{Var}\left(\hat{B}_h\right)\end{cases}$$

On calcule, d'après l'annexe 1,

$$\begin{cases} \hat{V}ar\left(\hat{A}_h\right) = \sum_h \left[\sum_{k\in S_{2h}} \left(\frac{z_k}{\pi_{1k}\pi_{2k}} \right)^2 \left(1-\pi_{1k}\pi_{2k}\right) + \sum_{k\in S_{2h}} \sum_{\substack{l\in S_{2h} \\ l\neq k}} \frac{z_k z_l}{\pi_{1k}\pi_{1l}} \frac{1}{\pi_{2k}\pi_{2l}} \frac{\Delta_{1kl}}{\pi_{1kl}} \right] \\ \hat{V}ar\left(\hat{B}_h\right) = \sum_{k\in S_{2h}} \left(1-\pi_{1k}\pi_{2k}\right) + \sum_{k\in S_{2h}} \sum_{\substack{l\in S_{2h} \\ l\neq k}} \frac{\Delta_{1kl}}{\pi_{1kl}} \end{cases}$$

dont on déduit par sommation :

$$\hat{V}ar\left(\hat{B}\right) = \sum_h \hat{V}ar\left(\hat{B}_h\right) = \sum_h \left[\sum_{k\in S_{2h}} \left(1-\pi_{1k}\pi_{2k}\right) + \sum_{k\in S_{2h}} \sum_{\substack{l\in S_{2h} \\ l\neq k}} \frac{\Delta_{1kl}}{\pi_{1kl}} \right]$$

La variance de $\hat{A}_h \hat{B}_h$ peut s'obtenir en linéarisant un produit :

$$\hat{V}ar\left(\hat{A}_h \hat{B}_h\right) = \hat{A}_h^2 \hat{V}ar\left(\hat{B}_h\right) + \hat{B}_h^2 \hat{V}ar\left(\hat{A}_h\right) + 2\hat{A}_h \hat{B}_h \hat{C}ov\left(\hat{A}_h, \hat{B}_h\right)$$

avec : $\hat{C}ov\left(\hat{A}_h, \hat{B}_h\right) = \sum_{k\in S_{2h}} \frac{z_k}{\pi_{1k}\pi_{2k}} \left(1-\pi_{1k}\pi_{2k}\right) + \sum_{k\in S_{2h}} \sum_{\substack{l\in S_{2h} \\ l\neq k}} \frac{z_k}{\pi_{1k}\pi_{2k}} \frac{\Delta_{1kl}}{\pi_{1kl}}$

Nous avons donc : $\hat{V}ar\left(\sum_h \left(\hat{A}_h \hat{B}_h\right) \right) = \sum_h \hat{V}ar\left(\hat{A}_h \hat{B}_h\right).$

Enfin, nous calculons: $\hat{C}ov\left(\sum_h \hat{A}_h \hat{B}_h, \hat{B} \right) = \sum_h \hat{C}ov\left(\hat{A}_h \hat{B}_h, \hat{B}_h\right) = \hat{A}_h \hat{V}ar\left(\hat{B}_h\right) + \hat{B}_h \hat{C}ov\left(\hat{A}_h, \hat{B}_h\right).$

Finalement, nous obtenons donc :

$$\boxed{\hat{V}ar\left(\hat{\bar{Y}}_{scenario\_4}\right) = \frac{1}{\hat{B}^2} \sum_h \left[\left(\hat{A}_h - \hat{\bar{Y}}\right)^2 \hat{V}ar\left(\hat{B}_h\right) + \hat{B}_h^2 \hat{V}ar\left(\hat{A}_h\right) + 2\hat{B}_h \left(\hat{A}_h - \hat{\bar{Y}}\right) \hat{C}ov\left(\hat{A}_h, \hat{B}_h\right) \right]}$$

que l'on calcule avec $\pi_{1kl} = \frac{n_h'}{M_h} \frac{n_h' - 1}{M_h - 1}$ pour $k, l \in S_{2h}$ et $\forall h = 1, ..., H$.

5. Scénario 5 : Plan de 1ère phase stratifié avec un tirage systématique à probabilités inégales dans chaque strate, suivi par un tirage poissonnien en 2ème phase avec, comme probabilité de réponse estimée, le ratio r/n' égal au nombre de répondants divisé par le nombre d'interrogés.

De manière analogue au scénario 3, nous avons ré-écrit le solde estimé comme :

$$\hat{\bar{Y}}_{scenario\_5} = \frac{1}{r} \sum_{k\in S_2} y_k = \frac{1}{r} \sum_{k\in S_2} \frac{y_k \pi_{1k}}{\pi_{1k}} = \sum_{k\in S_2} \frac{y_k \pi_{1k} \frac{1}{n'}}{\pi_{1k} \frac{r}{n'}} = \sum_{k\in S_2} \frac{z_k}{\pi_{1k}\pi_{2k}} = \sum_h \sum_{k\in S_{2h}} \frac{z_k}{\pi_{1k}\pi_{2k}}$$

avec ici : $\begin{cases} z_k = \frac{y_k \pi_{1k}}{n} \\ \pi_{1k} = \frac{n}{N} \frac{1}{1-\tau_k} = P(k \in S_1) \quad \text{où } \tau_k \text{ est le taux de pénétration de la liste rouge là où habite k} \\ \pi_{2k} = P(k \in S_2 / S_1) \text{ que l'on modélise par } \pi_{2k} = \frac{r}{n} \end{cases}$

Ce qui nous permet d'écrire comme précédemment :

$$\hat{Var}\left(\hat{\bar{Y}}_{scenario\_5}\right)=\sum_h\left[\sum_{k\in S_{2h}}\left(\frac{z_k}{\pi_{1k}\pi_{2k}}\right)^2\left(1-\pi_{1k}\pi_{2k}\right)+\sum_{k\in S_{2h}}\sum_{\substack{l\in S_{2h}\\ l\neq k}}\frac{z_k z_l}{\pi_{1k}\pi_{1l}}\frac{1}{\pi_{2k}\pi_{2l}}\frac{\Delta_{1kl}}{\pi_{1kl}}\right] \quad (1)$$

Le calcul de cette quantité nécessite la connaissance des probabilités d'inclusion d'ordre 2 du tirage de 1ère phase, c'est-à-dire π_{1kl}. Or, ces probabilités dépendent de l'ordre de rangement de la population que nous ne maîtrisons pas.

Lorsque la population est triée aléatoirement dans chaque strate *h*, on montre qu'elles peuvent être approchées par : $\pi_{1kl}\cong\left(1-\frac{1}{n'_h}\right)\pi_{1k}\pi_{1l}+\frac{1}{n'_h}\left(\pi_{1k}^2\pi_{1l}+\pi_{1k}\pi_{1l}^2\right) \qquad \forall k\neq l\in U_h, h=1,...,H$

Bien que le tirage des échantillons soit pratiqué sur un fichier trié selon les probabilités d'inclusion (ce qui génère une stratification implicite donc un gain de précision), nous avons considéré cette approximation dans une première approche afin d'obtenir une estimation de la variance qui devrait a priori surestimer sa vraie valeur.

Nous avons donc implémenté l'expression (1) avec : $\frac{\Delta_{1kl}}{\pi_{1kl}}=\frac{\pi_{1k}\pi_{1l}}{n'_h}\left(\pi_{1k}+\pi_{1l}-1\right)$ pour $k\in S_2$.

6. ***Scénario 6 : Plan de 1ère phase stratifié avec un tirage systématique à probabilités inégales dans chaque strate, suivi par un tirage poissonnien en 2ème phase stratifié et de probabilité de réponse estimée dans une strate h quelconque le ratio r_h /n'_h égal au nombre de répondants divisé par le nombre d'interrogés de la strate(scénario 6).***

De manière analogue au scénario 4, le solde estimé peut s'écrire ainsi :

$$\hat{\bar{Y}}_{scenario\_6}=\frac{1}{r}\sum_{k\in S_2}y_k=\frac{1}{r}\sum_{h=1}\sum_{k\in S_{2h}}\frac{y_k\pi_{1k}}{\pi_{1k}}\frac{\frac{r_h}{n'_h}}{\frac{r_h}{n'_h}}=\frac{1}{\sum_{k\in S_2}1}\times\sum_{h=1}\left(\sum_{k\in S_{2h}}\frac{z_k}{\pi_{1k}\pi_{1k}}\times\sum_{k\in S_{2h}}1\right)$$

avec ici, pour $k\in U_h$, $\begin{cases} z_k=\frac{y_k\pi_{1k}}{n'_h} \\ \pi_{1k}=\frac{n'}{M}\frac{1}{1-\tau_k}=\frac{n'_h}{M_h}\frac{1}{1-\tau_k}=P(k\in S_{1h}) \\ \pi_{2k}=P(k\in S_{2h}/S_{1h}) \quad \text{que l'on modélise par } \pi_{2k}=\frac{r_h}{n'_h} \end{cases}$

que l'on écrit de la même manière : $\hat{\bar{Y}}_{scenario\_6}=\frac{\sum_h\hat{A}_h\hat{B}_h}{\sum_h\hat{B}_h}=\frac{\sum_h\hat{A}_h\hat{B}_h}{\hat{B}}$

avec $\hat{A}_h=\sum_{k\in S_{2h}}\frac{z_k}{\pi_{1k}\pi_{2k}}$; $\hat{B}_h=r_h=\sum_{k\in S_{2h}}1=\sum_{k\in S_{2h}}\frac{\pi_{1k}\pi_{2k}}{\pi_{1k}\pi_{2k}}$ et $\hat{B}=r=\sum_{k\in S_2}1=\sum_h\hat{B}_h$

Finalement, comme pour le scénario 4, nous obtenons donc :

$$\hat{Var}\left(\hat{\bar{Y}}_{scenario\_6}\right)=\frac{1}{\hat{B}^2}\sum_h\left[\left(\hat{A}_h-\hat{\bar{Y}}\right)^2\hat{Var}\left(\hat{B}_h\right)+\hat{B}_h^2\hat{Var}\left(\hat{A}_h\right)+2\hat{B}_h\left(\hat{A}_h-\hat{\bar{Y}}\right)\hat{Cov}\left(\hat{A}_h,\hat{B}_h\right)\right]$$

que l'on calcule ici avec :

$$\pi_{1kl}\cong\left(1-\frac{1}{n'_h}\right)\pi_{1k}\pi_{1l}+\frac{1}{n'_h}\left(\pi_{1k}^2\pi_{1l}+\pi_{1k}\pi_{1l}^2\right) \quad \text{pour } k,l\in S_{2h} \text{ et } \forall h=1,...,H.$$

Annexe 3 : résultats numériques sur la précision de quelques soldes publiés en mars, juillet ou octobre 2004

Tableau 2 : précision obtenue pour quelques soldes publiés en mars, juillet ou octobre 2004

Indicateur	*Mois d'enquête 2004*	*Solde estimé*	*Intervalle de confiance à 95% pour chaque scénario*					
			1	*2*	*3*	*4*	*5*	*6*
Evolution passée du niveau de vie en France	Mars	**-0,5600**	± 0,02585	± 0,02587	± 0,02931	± 0,02933	± 0,03522	**± 0,03533**
	Juillet	**-0,5073**	± 0,02679	± 0,02676	± 0,03043	± 0,03054	± 0,0352	**± 0,03543**
	Octobre	**-0,4985**	± 0,02768	± 0,0277	± 0,03092	± 0,03096	± 0,03546	**± 0,03559**
Perspectives d'évolution du niveau de vie en France	Mars	**-0,2798**	± 0,02753	± 0,02756	± 0,02839	± 0,02843	± 0,03001	**± 0,03013**
	Juillet	**-0,2506**	± 0,02812	± 0,02811	± 0,02901	± 0,02912	± 0,03029	**± 0,03056**
	Octobre	**-0,2906**	± 0,0277	± 0,02765	± 0,02878	± 0,0289	± 0,03056	**± 0,03092**
Evolution passée de la situation financière des ménages	Mars	**-0,1634**	± 0,02575	± 0,02579	± 0,02609	± 0,02611	± 0,02668	**± 0,02675**
	Juillet	**-0,1468**	± 0,02605	± 0,02601	± 0,02637	± 0,02651	± 0,02687	**± 0,0272**
	Octobre	**-0,1584**	± 0,02674	± 0,02676	± 0,02711	± 0,02717	± 0,02765	**± 0,02782**
Perspectives d'évolution de la situation financière des ménages	Mars	**0,0178**	± 0,02395	± 0,02392	± 0,02393	± 0,02403	± 0,02396	**± 0,02424**
	Juillet	**-0,0359**	± 0,02559	± 0,02556	± 0,02559	± 0,02571	± 0,02564	**± 0,02594**
	Octobre	**-0,0059**	± 0,02532	± 0,02525	± 0,02528	± 0,02543	± 0,02532	**± 0,02573**
Opportunité de faire des achats importants	Mars	**-0,1333**	± 0,02839	± 0,02842	± 0,0286	± 0,02863	± 0,02895	**± 0,02907**
	Juillet	**-0,0506**	± 0,02934	± 0,0294	± 0,02941	± 0,02943	± 0,02942	**± 0,02948**
	Octobre	**-0,1327**	± 0,02902	± 0,02902	± 0,02925	± 0,02935	± 0,0296	**± 0,02983**
Perspectives d'évolution de la situation économique générale	Mars	**-0,1616**	± 0,03065	± 0,03067	± 0,03091	± 0,03096	± 0,03141	**± 0,03157**
	Juillet	**-0,1403**	± 0,03282	± 0,0327	± 0,03298	± 0,03323	± 0,03342	**± 0,03405**
	Octobre	**-0,2678**	± 0,03206	± 0,03199	± 0,03284	± 0,03299	± 0,0342	**± 0,03464**
Perspectives d'évolution du nombre de chômeurs	Mars	**-0,4738**	± 0,03072	± 0,03067	± 0,03281	± 0,0329	± 0,03679	**± 0,0371**
	Juillet	**-0,5473**	± 0,02959	± 0,02958	± 0,03343	± 0,03351	± 0,03849	**± 0,0387**
	Octobre	**-0,5762**	± 0,02855	± 0,02856	± 0,03269	± 0,03275	± 0,03837	**± 0,03853**

Annexe 4 : résultats numériques sur la précision de l'indicateur du moral des Français » de mars, juillet et octobre 2004

Tableau 3 : précision obtenue pour le « moral des Français » de mars, juillet ou octobre 2004

Indicateur	*Mois d'enquête 2004*	*Solde estimé*	*Intervalle de confiance à 95% pour chaque scénario*					
			1	*2*	*3*	*4*	*5*	*6*
Le moral des Français	Mars	**-0,2237**	± 0,01619	± 0,01621	± 0,0171	± 0,0171	± 0,0188	**± 0,0188**
	Juillet	**-0,1983**	± 0,01675	± 0,01672	± 0,0177	± 0,0178	± 0,019	**± 0,0192**
	Octobre	**-0,2172**	± 0,01672	± 0,01671	± 0,0178	± 0,0178	± 0,0193	**± 0,0194**

Annexe 5 : résultats numériques sur la précision des évolutions de quelques indicateurs entre mars et juillet 2004 et entre mars et octobre 2004

Tableau 4 : précision d'évolutions entre mars et juillet 2004 et entre mars et octobre 2004

Indicateur	*Périodes de 2004*	*Différence entre les soldes*	*Intervalle de confiance à 95% pour chaque scénario*					
			1	*2*	*3*	*4*	*5*	*6*
Evolution passée du niveau de vie en France	de mars à juil.	**0,0527**	± 0,03723	± 0,03722	± 0,04225	± 0,04234	± 0,04979	**± 0,05003**
	de mars à oct.	**0,0615**	± 0,03787	± 0,0379	± 0,0426	± 0,04265	± 0,04998	**± 0,05015**
Perspectives d'évolution du niveau de vie en France	de mars à juil.	**0,0292**	± 0,03935	± 0,03937	± 0,04059	± 0,0407	± 0,04264	**± 0,04292**
	de mars à oct.	**-0,0108**	± 0,03905	± 0,03904	± 0,04043	± 0,04054	± 0,04283	**± 0,04317**
Evolution passée de la situation financière des ménages	de mars à juil.	**0,0166**	± 0,0366	± 0,0366	± 0,0370	± 0,0372	± 0,0378	**± 0,0381**
	de mars à oct.	**0,005**	± 0,0371	± 0,0372	± 0,0376	± 0,0376	± 0,0384	**± 0,0385**
Perspectives d'évolution de la situation financière des ménages	de mars à juil.	**-0,0537**	± 0,03505	± 0,03501	± 0,03504	± 0,03519	± 0,03509	**± 0,0355**
	de mars à oct.	**-0,0237**	± 0,03485	± 0,03478	± 0,03481	± 0,03499	± 0,03486	**± 0,03535**
Opportunité de faire des achats importants	de mars à juil.	**0,0827**	± 0,04083	± 0,04089	± 0,04102	± 0,04106	± 0,04128	**± 0,0414**
	de mars à oct.	**0,0006**	± 0,0406	± 0,04062	± 0,04091	± 0,041	± 0,0414	**± 0,04165**
Perspectives d'évolution de la situation économique générale	de mars à juil.	**0,0213**	± 0,0449	± 0,0448	± 0,0452	± 0,0454	± 0,0458	**± 0,0464**
	de mars à oct.	**-0,1062**	± 0,0444	± 0,0443	± 0,0450	± 0,0452	± 0,0464	**± 0,0468**
Perspectives d'évolution du nombre de chômeurs	de mars à juil.	**-0,0735**	± 0,0427	± 0,0426	± 0,0468	± 0,0469	± 0,0532	**± 0,0536**
	de mars à oct.	**-0,1024**	± 0,0419	± 0,0419	± 0,0463	± 0,0464	± 0,0531	**± 0,0534**
Le moral des Français	de mars à juil.	**0,0255**	± 0,0233	± 0,02329	± 0,0246	± 0,0247	± 0,0267	**± 0,0269**
	de mars à oct.	**0,0065**	± 0,02327	± 0,02328	± 0,0247	± 0,0247	± 0,0269	**± 0,027**