



Liberté • Égalité • Fraternité

RÉPUBLIQUE FRANÇAISE

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

RAPPORT D'ANALYSES ET DE PROPOSITIONS

Groupe de travail

*AUCHAN, CORA, MONOPRIX, SIMPLY MARKET,
SUPERMARCHES MATCH, SYSTÈME U, INSEE*

**Quel partenariat
pour permettre à la statistique publique
d'utiliser les données de caisse
pour le suivi des prix à la consommation ?**

Préambule du président

Le présent rapport d'analyses et de propositions est le fruit du travail d'un groupe, composé de représentants de grandes enseignes de la distribution alimentaire française et de représentants de l'Institut national de la statistique et des études économiques, et missionné pour instruire la faisabilité d'un projet de partenariat qui permettrait à la statistique publique d'utiliser les données de caisse pour le suivi des prix à la consommation.

Sans préjuger des suites que les enseignes souhaiteront ou non donner aux propositions formulées dans le rapport, je tiens à leur exprimer ma gratitude pour avoir à la fois accepté de participer au groupe de travail et permis à l'Insee d'accéder à des échantillons de leurs données de caisse.

Je remercie personnellement Messieurs Édouard DE CROZÉ, Philippe DUMÉNIL, Éric DUPRÉ, Fabien FOULON et Alain PARDON pour le temps qu'ils ont consacré à nos travaux et la qualité de leur contribution.

Merci également à Madame Isabelle VERGNAUD, représentante de la Fédération des Entreprises du Commerce et de la Distribution, qui a bien voulu assister à nos réunions, à Patrick MIDY qui a animé le groupe et rédigé le présent rapport ainsi qu'à Sébastien FAIVRE, Jean Marc ROUCH et Patrick SILLARD.

Jean-Louis LHÉRITIER

COMPOSITION DU GROUPE DE TRAVAIL

- Président du groupe

Monsieur **Jean-Louis Lhéritier**, chef du département des prix à la consommation, des ressources et des conditions de vie des ménages de l'Insee

- Rapporteur

Monsieur **Patrick Midy**, chargé du développement du partenariat entre les enseignes et l'Insee pour l'utilisation des données de caisse dans la production des indices de prix à la consommation

- Secrétariat

Assuré par la division des prix à la consommation

- Distribution

Monsieur Edouard de Crozé	Représentant de SIMPLY MARKET
Monsieur Philippe Duménil	Représentant d' AUCHAN
Monsieur Eric Dupré	Représentant de SYSTEME U
Monsieur Fabien Foulon	Représentant de MONOPRIX
Monsieur Alain Pardon	Représentant de CORA et Supermarchés MATCH
Madame Isabelle Vergnaud	Fédération des entreprises du commerce et de la distribution

- Administration(Insee)

Monsieur Sébastien Faivre	Chef de projet statistique Division des prix à la consommation
Monsieur Jean-Louis Lhéritier	Président
Monsieur Patrick Midy	Rapporteur
Monsieur Jean Marc Rouch	Responsable de secteur Division des prix à la consommation
Monsieur Patrick Sillard	Chef de la division des prix à la consommation

SYNTHESE DU RAPPORT

La production mensuelle de l'indice des prix à la consommation (IPC) figure parmi les missions majeures de l'Insee. Car il joue le triple rôle d'indicateur macroéconomique (mesure de l'inflation, déflateur d'agrégats), d'indicateur de référence pour l'indexation de contrats privés ou de dispositifs publics et d'indicateur fondamental de la politique monétaire. Dès lors, même si la statistique publique avait le souhait d'encore améliorer l'IPC - indice internationalement reconnu pour sa qualité et sa robustesse- ou d'enrichir le suivi des prix, son objectif prioritaire doit être de garantir la pérennité de la production de l'IPC, au moins à son niveau actuel de qualité.

Les données de caisse des grandes enseignes de la distribution alimentaire retracent, quotidiennement, pour chacun de leurs points de vente, l'exhaustivité des prix et des quantités de produits élémentaires - identifiés par leur code-barres - vendus au cours de la journée. Ces données permettraient, pour le calcul de l'IPC, d'appréhender quasiment en temps réel et au niveau le plus fin, la totalité des prix et des quantités de produits vendues. La taille de l'échantillon des produits suivis dans l'IPC pourrait être démultipliée, sans coût supplémentaire. Pour cela, l'Insee serait conduit à adapter ses méthodes et processus de production de l'IPC, ainsi que son organisation, à la réception et au traitement de données comportant à la fois un système d'identifiants plus complexes et volatils, et des volumétries très importantes.

Pour les enseignes de la grande distribution alimentaire, leurs données de caisse représentent une source d'information stratégique et hautement confidentielle. On peut y lire leurs stratégies de référencement des produits, leurs politiques de fixation des prix, leurs chiffres d'affaire détaillés et leurs parts de marché. Ces enseignes sont cependant parvenues à développer entre elles, moyennant le recours à des prestataires externes qui centralisent leurs données, un système d'information qui permet à chacune de connaître ses parts de marchés et, pour chaque article en vente dans un de ses propres magasins, les prix pratiqués dans les magasins de leurs concurrents.

Le groupe de travail Enseignes/Insee avait pour mission d'explorer les modalités techniques, méthodologiques, et juridiques, selon lesquelles il serait possible pour l'Insee d'avoir accès aux données de caisse, de les exploiter pour la production de l'IPC, de garantir la protection de leur confidentialité et de faire retour aux enseignes d'informations qu'elles auraient convenance à obtenir de sa part.

I. L'ETAT DES LIEUX

Du côté de l'Insee

L'objet général d'un indice de prix est de procéder au partage volume/prix des évolutions nominales des agrégats macroéconomiques. L'IPC français - strictement conforme aux prescriptions internationales - est calculé au moyen de 180 000 prix relevés chaque mois pour un panier fixe de biens représentatif de la consommation des ménages.

En cours d'année de l'ordre de 4 à 5% des produits élémentaires de l'échantillon disparaissent mensuellement des magasins dans lesquels ils donnaient lieu à relevés de prix. Un produit disparu est remplacé, dans l'échantillon, soit par un produit élémentaire équivalent -auquel cas le prix du remplaçant est directement comparé à celui du produit disparu, soit, si l'enquêteur ne parvient pas à trouver un produit équivalent, par un produit aussi proche que possible -auquel cas le traitement appliqué consiste à faire évoluer, pour le mois considéré, le prix du produit remplacé comme ceux des produits plus proches observés par ailleurs.

Le recours aux données de caisse permettrait : de tirer l'échantillon annuel dans une base exhaustive de couples [produits élémentaires | point de vente] ; de mieux appréhender les prix des produits vendus en promotion ; d'automatiser et d'enrichir la gestion des remplacements selon la technique du recouvrement. Accessoirement il serait possible de mener des recherches sur d'autres modes de calcul de l'IPC et sur des calculs d'indices alternatifs (par exemple des indices de prix moyens).

Du côté des enseignes

Le système d'information de chaque enseigne est structuré à partir d'identifiants de produits élémentaires (numéros à treize chiffres), transcrits en code-barres, attribués par les fabricants des produits, selon des standards et des pratiques définis par l'association internationale GS1.

S'agissant des produits frais à poids variable ou en libre service, c'est-à-dire ceux pour lesquels un conditionnement est réalisé sur place, les codes-barres sont apposés par l'enseigne.

Chaque enseigne gère son propre référentiel et sa propre classification des produits élémentaires

Les données de caisse de la quasi totalité des magasins des grandes enseignes de la distribution alimentaire donnent lieu à centralisation quotidienne, vers deux prestataires privés concurrents, qui les exploitent, notamment pour mettre à la disposition des enseignes des indicateurs relatifs à leurs parts de marché, d'une part, aux prix pratiqués au jour le jour, par leurs concurrents, d'autre part.

Ces exploitations sont rendues possible par l'usage, par ces prestataires de dictionnaires de codes à barres, qui permettent de bien les identifier individuellement et de leur attribuer une position dans une classification détaillée.

Les produits frais, en libre service ou à poids variable, qui ne sont pas correctement retracés dans ces répertoires et classifications échappent à ce système d'information fondé sur les données de caisse quotidiennes. En réalité il n'existe pas aujourd'hui de bonne classification de référence des produits frais, ni chez les enseignes, ni chez les opérateurs de leur SI, ni même au niveau international.

L'Insee, qui procède mensuellement à deux vagues de 40 000 relevés de prix de produits frais, pourrait le cas échéant, procéder à des exploitations particulières de ces données, selon des niveaux régionalisés à convenir avec les enseignes ou pour certaines formes de vente, afin de mettre à leur disposition des indicateurs des prix et de l'offre de produit frais.

II. L'INSTRUCTION DE LA FAISABILITÉ DU PROJET

Faisabilité technique et méthodologique

Il s'agissait de s'assurer qu'il serait effectivement possible d'utiliser les données de caisse pour le calcul d'indices de prix cohérents avec ceux qui résultent de la collecte traditionnelle.

A partir d'un échantillon de données de test, relatives à dix familles ou groupes de familles de produits, provenant des enseignes membres du groupe de travail et transmises par un des deux opérateurs de leur système d'information, il a été procédé :

- à un premier exercice de comparaison descriptive, portant sur près de 350 produits élémentaires, entre données issues de la collecte des prix en magasin par les enquêteurs de l'Insee et données de caisse ;
 - à un exercice de test systématique d'une méthode d'utilisation des données de caisse pour la détermination d'indices de prix relatifs aux seules familles de produits de l'échantillon de test.
- La comparaison des données a permis, pour 60% des relevés, de trouver une référence ayant des caractéristiques physiques compatibles avec celles relevées par les enquêteurs et dont le prix est identique à celui relevé. Pour 15% des relevés, il a été possible d'associer au moins une référence compatible et comportant un prix proche de celui du relevé (moins de 5% d'écart). Pour 25% des relevés, il n'a pas été possible d'identifier une référence compatible et ayant un prix proche du prix relevé.

Cet exercice a permis de constater que les caractéristiques techniques notées par les enquêteurs [dont l'objet n'est pas d'alimenter des bases de données mais de les guider dans leur démarche sur le terrain] sont nettement moins riches et moins précises que celles issues des données de caisse. Il peut même exceptionnellement arriver qu'elles soient erronées, si elles ne sont pas à jour et ne correspondent plus au produit donnant effectivement lieu à relevé, ce qui n'a aucun impact sur la qualité de l'IPC mais rend impossible tout appariement fondé sur la compatibilité des caractéristiques physiques. A contrario quand l'appariement des produits a été possible, et bien que les prix relatifs aux données de caisse utilisés ici soient des prix moyens hebdomadaires et non des prix relevés une journée donnée, l'identité ou la forte proximité des prix valide la qualité des données de prix.

- Le test d'une méthode d'utilisation des données de caisse pour le calcul de l'indice a consisté à transposer à un échantillon ou panier de [codes barre | point de vente] la méthode actuellement appliquée à un panier de [produit | point de vente].

Pour juger de la validité de la méthode, il a été procédé à 500 simulations différentes. Chacune comprend un tirage aléatoire, avec probabilités proportionnelles aux ventes, sur les données de décembre 2008, de trois paniers de 10, 20 et 50 000 couples [codes barre | point de vente]- soit des sondages de 1, 2 et 5%-conduisant à autant de séries mensuelles de décembre 2008 à décembre 2009 - moyennant remplacement automatique des codes-barres disparaissant chaque mois, selon une technique fondée sur la proximité, à l'intérieur de la famille des disparus, des numéros identifiants, avec application de la méthode du recouvrement.

Les résultats de ces simulations sur les huit familles de données de test font apparaître :

- qu'un taux de sondage à 2% garantit, pour chacune des huit familles ou postes étudiés, un indice de prix "données de caisse" précis à 1 point de pourcentage près, dans 98% des cas ;
- que l'indice de prix "données de caisse" de chacun des huit postes étudiés se situe dans l'intervalle de confiance à 95% de l'indice IPC.

Ces résultats permettent au groupe de se prononcer positivement sur la faisabilité méthodologique et technique de l'utilisation des données de caisse pour le calcul de l'IPC.

Il y aura cependant lieu d'approfondir les deux questions suivantes, relatives au traitement des promotions et des remplacements.

- Dans la mesure où il serait possible de repérer systématiquement dans les données de caisse celles qui concernent les produits vendus en promotion selon quelle méthode conviendrait-il de les prendre en compte dans le calcul de l'indice des prix ? Faut-il considérer que ces conditions temporaires de vente, qui représentent de l'ordre de 12% du chiffre d'affaire des enseignes doivent faire l'objet d'un traitement particulier, comme s'il s'agissait de produits spécifiques analogues aux produits saisonniers ? ou convient il de les assimiler, au jour le jour, selon des calculs de prix moyens, à des produits courants ?
- Pour les remplacements, les données de caisse ouvrent la possibilité d'une mise en œuvre systématique de la méthode de recouvrement. Convient il de continuer à envisager les remplacements de façon individuelle, un produit disparu se voyant attribuer un remplaçant ? Est il envisageable de procéder à un remplacement mensuel global de l'ensemble des références disparues en faisant en sorte de maintenir une certaine représentativité de l'échantillon, par exemple au regard de l'âge des références selon leur cycle de vie ? Cette question complexe ne pourra être éludée, sauf à prendre le risque d'introduire des biais dans l'IPC.

Faisabilité juridique

Il est clair que les enseignes n'ont pas grand chose à gagner en entrant dans le partenariat qui leur est proposé, même si l'Insee est prêt à leur faire des retours, notamment sur des indices de prix de produits frais ou sous forme d'accès gratuits à une partie de son offre éditoriale. Elles ne peuvent se permettre de prendre le moindre risque concernant la protection de la confidentialité de leurs données.

Le dispositif juridique proposé par l'Insee, par voie contractuelle (conventions) et réglementaire (arrêté ministériel) garantira que les données transmises, couvertes par le secret statistique :

- ne seront utilisées que dans le cadre de l'élaboration de l'indice des prix à la consommation, à l'exclusion de tout autre usage,
- que seuls les agents de la division des prix à la consommation et ceux des services informatiques dédiés à la maintenance de l'IPC auront le droit d'utiliser ces données, à l'exclusion de tout autre agent,
- que les données de caisse transmises à la division des prix seront détruites sous un délai de 24 mois. Seules seront conservées les données de prix relatives à l'échantillon à partir duquel a été effectivement calculé l'IPC et les éléments de pondération généraux utilisés.

III DIAGNOSTIC PARTAGÉ DU GROUPE ET PROPOSITIONS OPÉRATIONNELLES

Sur la base des analyses et des éléments d’instruction du projet de partenariat développés précédemment et sans préjuger des décisions que chaque enseigne membre du groupe et l’Insee auront ou non convenance à prendre, l’ensemble des membres du groupe s’accorde sur le caractère faisable des sept propositions opérationnelles suivantes, dont certaines sont soit complémentaires soit exclusives les unes des autres, qui sont détaillées dans le rapport.

1. L’Insee proposera aux enseignes membres du groupe :
 - de leur faire retour d’indices de prix des produits frais, par forme de vente ou type d’agglomération, selon des modalités à convenir en commun ;
 - de leur fournir, à titre gracieux le fichier détail anonymisé, de l’enquête budget des familles ;
 - de leur ouvrir un accès, selon des modalités à définir en commun, à son offre éditoriale.

L’information des enseignes sur les magasins visités par les enquêteurs sera améliorée.

2. Les enseignes transmettront quotidiennement leurs données de caisse à l’Insee, à l’identique et selon les mêmes conditions techniques, que celles qui prévalent aujourd’hui pour les transmissions effectuées vers deux opérateurs centralisateurs de leurs données.
3. L’Insee sera autorisé à acquérir les données de caisse des enseignes membres du groupe auprès de l’un ou l’autre des deux opérateurs centralisateurs de leurs données.
4. L’Insee acquerra par abonnement auprès d’un prestataire privé, un dictionnaire de produits, à l’EAN, et sa classification détaillée, avec mises à jour mensuelles
5. L’Insee acquerra mensuellement, auprès d’un prestataire centralisateur des données de caisse un ensemble de relevés de prix issus des données de caisse, pour un échantillon de [code barre | point de vente] sous réserve de l’accord des enseignes concernées.
6. L’Insee acquerra auprès d’un prestataire privé des données de pondération pour le calcul de l’IPC
7. Un comité de coordination Enseignes/Insee se réunissant au moins une fois par an, sera chargé de suivre le bon déroulement du déploiement du partenariat.

SOMMAIRE

PRÉAMBULE.....	3
COMPOSITION DU GROUPE DE TRAVAIL.....	5
SYNTHÈSE DU RAPPORT.....	7
SOMMAIRE.....	13
INTRODUCTION.....	15
I ETAT DES LIEUX	19
I.A LE SUIVI DES PRIX À LA CONSOMMATION.....	19
I.A.1 <i>La problématique générale des indices de prix à la consommation dans la perspective de l'utilisation des données de caisse.</i>	19
I.A.2 <i>Description des modalités actuelles de confection et de production de l'indice mensuel des prix à la consommation</i>	21
I.A.3 <i>Les voies d'amélioration et d'enrichissement de l'existant</i>	23
I.B LE SYSTÈME D'INFORMATION DES ENSEIGNES	27
I.B.1 <i>Les identifiants des produits</i>	27
I.B.2 <i>Les données de caisse</i>	28
I.B.3 <i>Le système d'information fondé sur les données de caisse</i>	29
I.C LES DIFFÉRENTS MODES D'ACCÈS POSSIBLES AU SI DES ENSEIGNES POUR LA PRODUCTION DE L'IPC.....	31
II INSTRUCTION DE LA FAISABILITÉ DU PROJET.....	33
II.A UN TEST DE FAISABILITÉ D'UN INDICE DE PRIX À PARTIR DES DONNÉES DE CAISSE	33
II.A.1 <i>Un premier exercice de comparaison descriptive entre les données de caisse et les données de la collecte actuelle</i>	33
II.A.2 <i>Un exercice de test d'utilisation des données de caisse pour le calcul de l'IPC</i>	35
II.B LES QUESTIONS MÉTHODOLOGIQUES À TRAITER	47
II.B.1 <i>Quelles articulations entre les variétés de l'IPC actuel et les données de caisse ?</i>	47
II.B.2 <i>La question du traitement des promotions:</i>	47
II.B.3 <i>La question des remplacements</i>	48
II.C LA FAISABILITÉ JURIDIQUE	49
II.C.1 <i>Des conventions pluriannuelles de trois ans(2012, 2013, 2014) prorogables un an</i>	49
II.C.2 <i>Un arrêté du ministre de l'économie et des finances</i>	49
III DIAGNOSTIC DU GROUPE SUR LA FAISABILITÉ DU PROJET.....	51
IV PROPOSITIONS OPÉRATIONNELLES.....	53
 Annexe 1 Description de l'échantillon de données de test.....	 55
 Annexe 2 Comparaison entre la stabilité d'échantillons de [code barres point de vente] et d'échantillons de [produits élémentaires point de vente] : l'impact des promotions.....	 57

INTRODUCTION

Le suivi des prix à la consommation, à travers la production mensuelle de l'indice des prix à la consommation (IPC) figure parmi les missions majeures de l'Institut national de la statistique et des études économiques (Insee).

Car l'IPC joue un triple rôle.

- Indicateur économique, il permet de suivre mensuellement l'inflation d'une part, de calculer, à partir de leurs évolutions nominales, les évolutions "en volume" ou en termes "réels" de différents agrégats macroéconomiques -par exemple celle du pouvoir d'achat des ménages définie comme l'évolution réelle de leur revenu disponible- d'autre part.
- Publié au Journal Officiel chaque mois, il est utilisé pour la mise en œuvre des dispositifs d'indexation prévus dans différents cadres juridiques privés ou publics (indexation des pensions alimentaires ou des rentes viagères, actualisation annuelle des loyers, revalorisations du SMIC)
- Indicateur monétaire, il participe, conjointement avec les IPC nationaux des autres pays membres de l'Union européenne - moyennant une certaine harmonisation méthodologique coordonnée par la direction statistique de la Commission européenne (Eurostat)- à la mesure de l'inflation dans l'Union européenne, qui constitue la grandeur fondamentale sur la base de laquelle la Banque centrale européenne, élabore et arrête ses décisions de politique monétaire.

Strictement conforme aux recommandations et pratiques internationales telles que présentées dans le *Manuel de l'indice des prix à la consommation* publié sous l'égide collective de l'Organisation internationale du travail, de la Banque mondiale, du Fonds monétaire international, de l'OCDE et d'Eurostat, l'IPC français est internationalement reconnu pour sa robustesse et sa qualité¹.

Dans ces conditions l'objectif prioritaire de la statistique publique française, même si elle avait le souhait et la possibilité d'encore améliorer ou enrichir le suivi des prix à la consommation, doit être de garantir, au moins à son niveau actuel de qualité, la pérennité de la production de l'IPC.

*
* *

¹ Lors de l'audit conduit par le Fonds monétaire international en 2003, le rapport de la mission concluait à la bonne qualité de l'indice français et à sa conformité aux normes internationales : "*The compilation of price indices at Insee confirms France's reputation for a high degree of methodological soundness in price statistics and may serve as model for many countries.*"

Les données de caisse des grandes enseignes de la distribution alimentaire retracent, quotidiennement, par point de vente, pour chaque article référencé ou proposé à la vente, la quantité physique vendue, le prix de vente et la valeur nominale vendue.

L'utilisation de telles données pour le calcul de l'indice des prix à la consommation permettrait à l'Insee :

- d'appréhender les prix des produits au niveau le plus fin possible -celui de la référence élémentaire- et de façon plus exhaustive, soit une couverture plus large et plus précise des prix ;
- de mieux distinguer ou identifier les produits élémentaires constituant l'offre, grâce aux descriptifs individuels contenus dans des répertoires de références élémentaires ;
- de disposer, à travers les quantités vendues, pour le calcul de l'indice des prix, d'éléments de pondérations plus frais et plus fins ;
- d'augmenter, sans coût supplémentaire, la taille de l'échantillon sur la base duquel est calculé l'IPC.

Il faudrait pour cela que l'Insee adapte ses méthodes et processus de production de l'IPC et son organisation, à la réception et au traitement de données comportant un système d'identifiants individuels plus complexes et plus volatils que ceux des données actuellement utilisées pour le suivi des prix et des volumétries très importantes.

Accessoirement la richesse et la précision des données de caisse permettraient à la statistique publique française de mener des travaux de recherche ou d'exploration de certaines problématiques générales, qui ne sont pas spécifiques à notre pays, relatives à d'autres méthodes de calcul et de suivi des indices de prix à la consommation ou portant sur la conception d'indices alternatifs ayant une autre signification que celle de l'IPC.

*
* *

Pour les enseignes de la grande distribution alimentaire, leurs données de caisse représentent une source d'information stratégique et hautement confidentielle.

Car on peut y lire, par exemple,

- leurs stratégies de référencement des produits ;
- leurs politiques de fixation des prix;
- leurs résultats détaillés en termes de chiffres d'affaire et de parts de marché.

Alliées à d'autres sources elles pourraient être utilisées pour des analyses portant sur leurs marges.

Ces enseignes sont cependant parvenues à développer entre elles, moyennant le recours à des prestataires externes qui centralisent leurs données, un système d'information qui leur permet de connaître :

- leurs parts de marchés, selon un niveau d'agrégation suffisant pour rendre impossible l'identification des chiffres relatifs à un magasin ou à un produit particulier ;
- les prix pratiqués dans les magasins de leurs concurrents pour chaque article en vente dans un de leurs propres magasins².

Sollicitées par l'Insee, six grandes enseignes de la distribution alimentaire

AUCHAN

CORA

MONOPRIX

SIMPLY MARKET

Supermarchés MATCH

SYSTÈME U

ont accepté de participer à un groupe de travail chargé d'explorer la faisabilité d'un projet de partenariat ayant pour objet de permettre à la statistique publique d'avoir accès à leurs données de caisse et de les utiliser pour le suivi des prix à la consommation.

Les travaux du groupe ont porté sur :

- la faisabilité méthodologique et technique de l'utilisation des données de caisse pour le suivi de l'IPC. A cette fin ont été menés des travaux d'expérimentation et de simulation à partir d'un échantillon de données de caisse, portant sur quelques familles de produit, que les enseignes membres du groupe ont permis à l'Insee d'obtenir ;
- la faisabilité juridique et contractuelle d'un dispositif propre à garantir une protection absolue de la confidentialité des données ;
- les modalités techniques de transmissions régulières à l'Insee des données de caisse ;
- les retours d'information que les enseignes auraient convenance à obtenir de la part de l'Insee.

² Une des principales grandes enseignes nationales ne participe pas à ce dispositif dit "de transparence".

Le présent document, qui constitue le rapport d'analyses et de propositions du groupe de travail, comporte :

I. Un état des lieux relatif :

- au suivi des prix à la consommation qui, à partir d'une illustration à portée générale des problématiques relatives au suivi des prix à la consommation, présente le processus d'élaboration de l'IPC puis envisage les voies d'amélioration qu'ouvrirait l'utilisation des données de caisse ;
- au système d'information des enseignes de la grande distribution.

II La présentation des travaux d'instruction relatifs à la faisabilité méthodologique, technique et juridique du projet de partenariat.

III. Un diagnostic partagé sur la faisabilité du projet

IV. Un ensemble de propositions opérationnelles

I ETAT DES LIEUX

I.A Le suivi des prix à la consommation.

Avant d'en venir, dans une partie I.A.2, à une description des modalités actuelles de suivi de l'indice national des prix à la consommation puis d'envisager les voies d'amélioration et d'enrichissement que permettrait le recours aux données de caisse (partie I.A.3), il a paru utile d'illustrer, autour de quelques concepts ou notions de base, la problématique générale des indices de prix à la consommation (partie I.A.1).

I.A.1 La problématique générale des indices de prix à la consommation dans la perspective de l'utilisation des données de caisse.

I.A.1.a Les indices de prix ont pour objet de procéder au partage "volume/prix" des évolutions nominales.

▪ Par exemple si un ménage a consommé au cours d'une période, par exemple un mois, les mêmes biens ou produits que le mois précédent, selon des volumes ou quantités physiques q_1 en progression uniforme de 3%, et si les prix p_1 de ces produits ont uniformément augmenté de 2% par rapport à ceux du mois précédent il est naturel de dire que la progression de sa dépense mensuelle de consommation de ces produits, soit approximativement 5%, se partage entre 2% de prix et 3% de volume.

$$\frac{\sum p_1 q_1}{\sum p_0 q_0} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_1 q_0}$$

Nominal (1,05) = Prix (1,02) × Volume (1,03)

Dans ce cas simple et purement théorique :

- l'indice de volume désigne, sans ambiguïté, une véritable augmentation de volumes au sens physique ;

- l'indice de prix, rapport entre la moyenne des prix de la période finale et celle des prix de la période initiale, toutes deux pondérées par les quantités consommées à la première période, peut s'interpréter comme l'augmentation nominale du budget du consommateur nécessaire pour acheter en deuxième période exactement les mêmes quantités de biens que celles achetées en première période. Soit une conception de l'indice des prix comme l'augmentation nominale du budget nécessaire pour consommer les mêmes quantités physiques ;

- l'indice des prix est le même que les pondérations par les quantités consommées soient celles de la première (indice de Laspeyres) ou de la seconde période (indice de Paasche).

L'indice national des prix à la consommation est un indice de Laspeyres. Il s'interprète comme l'augmentation budgétaire nécessaire au maintien du volume de la consommation des ménages.

I.A.1.b. En réalité :

- D'une période à la suivante, ni les quantités consommées ni les prix n'évoluent de façon uniforme.

Les indices de Laspeyres et de Paasche prennent des valeurs différentes.

Le consommateur a la possibilité, à budget constant, compte tenu des changements de prix relatifs de l'offre de biens, d'opérer certaines substitutions entre les quantités de biens qu'il achète pour maintenir au meilleur niveau possible l'utilité (ou la satisfaction) qu'il retire de sa consommation.

L'indice de Laspeyres - calculé de façon stricte - ne rendrait pas compte de ces possibilités de substitution et n'exprimerait donc pas exactement l'augmentation de budget nécessaire au maintien de l'utilité que les ménages retirent ou attribuent à leur consommation.

Pour rapprocher l'indice national des prix à la consommation, d'*un indice de prix à utilité constante*, une méthode de calcul par moyenne géométrique des prix des produits élémentaires a été adopté, à l'intérieur des familles de produits non substituables, conformément à la théorie et aux règles fixées au niveau européen et international³.

- Aujourd'hui l'utilité ou la satisfaction retirée par un ménage de sa consommation n'est plus uniquement liée à la quantité de chaque bien qu'il achète ou consomme mais également à sa capacité budgétaire à accéder sans trop attendre aux nouveaux produits qui apparaissent de façon continue. C'est-à-dire à sa faculté de se conformer à la norme de consommation à laquelle il s'identifie, sous peine de se sentir déclassé.

Cette problématique relève d'une autre approche de calcul, par exemple à travers celle de calculs *d'indices de prix moyens* d'exercice des fonctions de consommation, soit une notion *d'indices du coût de la vie*⁴. Ces calculs, qui nécessitent la connaissance en temps réel, à titre de pondération, des quantités de produits élémentaires consommés, seraient rendus possibles par les données de caisse.

Une telle approche, outre qu'elle relève d'une conception un peu révolutionnaire des indices de prix à la consommation serait difficile à mettre opérationnellement en œuvre. Car elle suppose une répartition des produits élémentaires en familles de produits substituables et participant à l'exercice de fonctions de consommation, répartition qui pourrait soit s'avérer impossible à finaliser soit comporter plusieurs hypothèses concurrentes. En d'autres termes, cette démarche pourrait soit ne pas aboutir soit déboucher sur le calcul de plusieurs indices de prix moyens concurrents.

En tout état de cause on ne voit pas très bien l'usage concret qui pourrait être fait de tels indices de prix, lesquels ne feront pas consensus, et alors même qu'existera par ailleurs un indice reconnu et conforme aux prescriptions internationales.

Pour les raisons évoquées précédemment, le cœur du projet "données de caisse" consiste en la transposition aux données de caisse, dans toute la mesure du possible, des méthodes actuelles de suivi de l'IPC.

Ce qui n'interdira nullement de mener, à partir des données de caisse, quelques travaux d'expérimentation méthodologique relatifs à des indices de prix complémentaires.

³ *Manuel de l'indice des prix à la consommation* publié sous l'égide collective de l'Organisation internationale du travail, de la Banque mondiale, du Fonds monétaire international, de l'OCDE et d'Eurostat

⁴ Proposition extraite du rapport au Conseil d'analyse économique "Mesurer le pouvoir d'achat" de MM. Moati et Rochefort, 2008

I.A.2 Description des modalités actuelles de confection et de production de l'indice mensuel des prix à la consommation

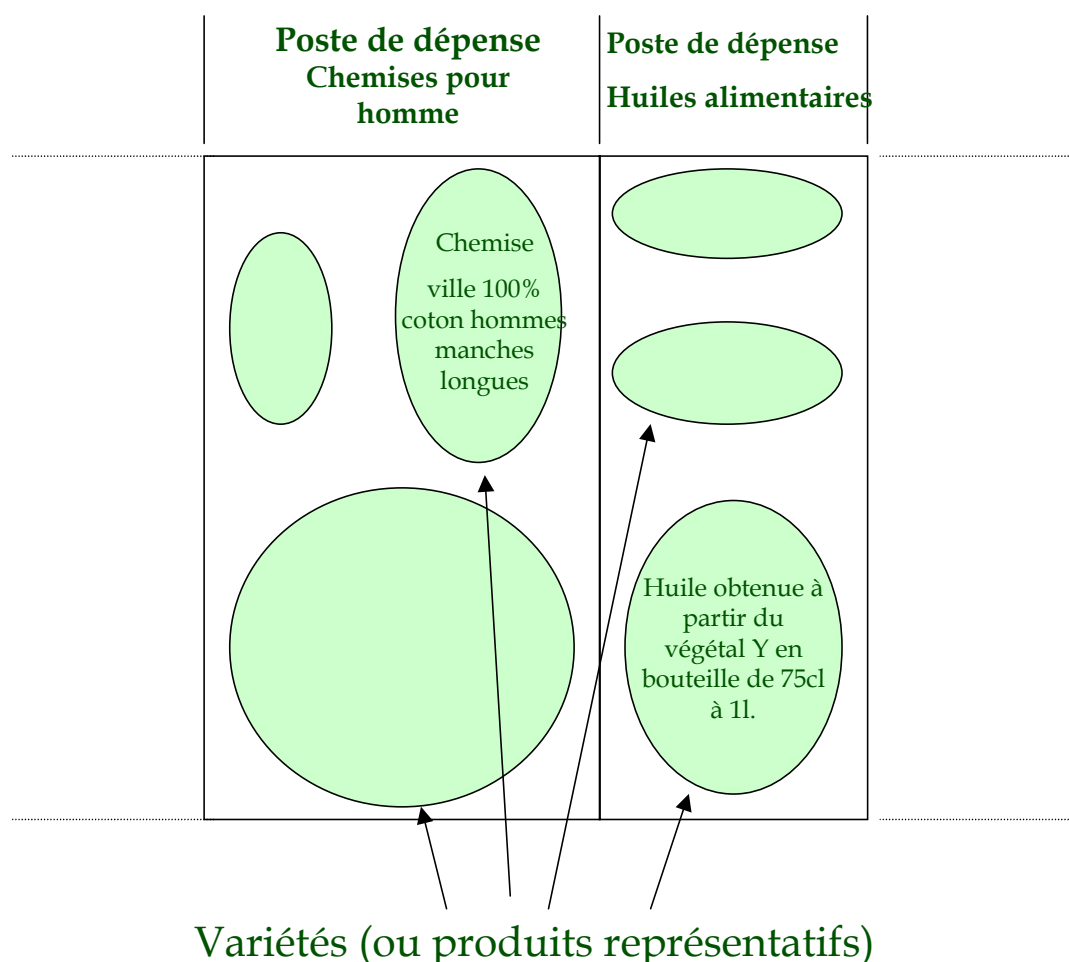
Tel que défini par une résolution de l'Organisation internationale du travail, « l'objectif d'un indice des prix à la consommation est la mesure des variations au cours du temps du niveau général des prix des biens et services acquis, utilisés ou payés par la population de référence pour sa consommation. Cet indice est estimé à partir d'une suite de mesures synthétiques des variations relatives, d'une période à l'autre, des prix d'un ensemble fixe de biens et de services de consommation constants en quantité et par leurs caractéristiques... ».

Conformément à ce principe général, l'indice des prix à la consommation est calculé en France par l'Insee au moyen de 180 000 prix relevés chaque mois pour un panier fixe de biens représentatif de l'ensemble des biens et services consommés par les ménages. Ce panier est figé durant une année. Il est composé d'environ 110 000 produits élémentaires dont les prix sont relevés chaque mois, principalement par des enquêteurs de l'Insee dans 27 000 points de vente et, pour quelques dizaines d'entre eux, de façon centralisée (tarifs du gaz, de l'électricité, des assurances, des services financiers ...).

I.A.2.a La constitution de l'échantillon

Ont été définies des familles de produits élémentaires, appelées variétés, sous-ensembles jugés représentatifs des postes de consommation des ménages, selon le schéma ci dessous.

Postes de dépense et variétés



On compte 1 100 variétés⁵, qui couvrent plus de 95% de la consommation des ménages. La liste des variétés et leurs périmètres individuels sont actualisés annuellement pour tenir compte de l'évolution et du renouvellement de l'offre (en général, chaque année, trois ou quatre variétés sont supprimées et remplacées, les périmètres de 25 à 30 variétés sont ajustés)⁶.

Les produits élémentaires retenus dans l'échantillon sont sélectionnés, **à l'intérieur de chaque variété**, de façon à être bien représentatifs de la diversité des formes de vente, des niveaux de qualité ou de standing de leurs marques.

Les 27 000 points de vente retenus dans l'échantillon sont sélectionnés dans 106 agglomérations de plus de 2 000 habitants représentatives de l'ensemble du territoire de France métropolitaine et des départements d'outre-mer. Les différentes formes de ventes - notamment hypermarchés, supermarchés, maxi-discounts, superettes, magasins populaires, petits magasins traditionnels et marchés- sont représentées conformément à leur part de marché.

En réalité les prix relevés sont ceux de couples [produit élémentaire | point de vente] car, pour un même produit élémentaire, la localisation et l'organisation physique du point de vente sont des facteurs de différenciation de l'offre.

A titre d'exemple, pour une variété donnée, il pourra y avoir mensuellement 150 relevés de prix de [produit élémentaire | point de vente] effectués dans 24 agglomérations différentes selon la répartition croisée suivante entre agglomérations et formes de vente.

AGGLO	HYPER - MARCHÉ	SUPER - MARCHÉ	Maxidiscounte	Sup é rette	Magasin Populaire	AUTRE			MARCHE	TOTAL
1	2	2	1	0	0	1	6
2	1	2	1	0	0	0	4
3	1	1	1	1	1	0	5
....
....
15	8	6	3	1	1	0	...	1	1	21
....
....
24	2	1	1	0	0				0	4
TOTAL	60	52	20	5	5	0	...	4	4	150

N.B. : Le choix du [produit élémentaire | point de vente] donnant lieu à relevés de prix est effectivement réalisé sur le terrain par les enquêteurs sur la base de plans d'échantillonnage de ce type

⁵ 1000 pour la France métropolitaine et 100 pour les départements d'outre mer.

⁶ A l'occasion de ce renouvellement, l'évolution de la qualité des produits est neutralisée par une double collecte des prix en décembre de l'année précédente pour l'ancien et le nouvel échantillon, ces deux échantillons n'étant donc pas directement comparés.

I.A.2.b La gestion des remplacements

Chaque relevé consiste, comme le font de temps en temps la plupart des consommateurs lors de leurs achats, à suivre le prix d'un même produit (de même marque, de même conditionnement, ...) dans le même point de vente tout au long de l'année.

En cours d'année de l'ordre de 4 à 5% des [produit élémentaire|point de vente] de l'échantillon disparaissent chaque mois, et doivent être remplacés.

A cette fin l'enquêteur chargé des relevés des prix du couple disparu recherche en priorité dans le même point de vente un produit élémentaire équivalent, c'est-à-dire appartenant à la même variété et ayant les mêmes caractéristiques techniques.

Dans la moitié des cas il est ainsi procédé à des remplacements par des produits élémentaires équivalents. Les prix des deux produits sont alors directement comparés (remplacement dit en équivalent).

Si au contraire il n'est pas possible de trouver un produit équivalent, l'enquêteur retiendra un produit de remplacement aussi proche que possible du produit disparu.

Une comparaison directe des prix des deux produits n'est alors pas possible. Le traitement généralement appliqué à ces cas consiste à faire évoluer, pour le mois considéré, les prix des produits disparus comme ceux des produits les plus proches observés par ailleurs. Les mois suivants, le prix du nouveau produit retenu est suivi.

I.A.2.c Le calcul de l'indice des prix à la consommation

A partir des 180 000 relevés effectués chaque mois, des évolutions de prix sont calculées au niveau des variétés en tenant compte du poids démographique des différentes agglomérations ; puis à des niveaux de plus en plus agrégés jusqu'à l'indice d'ensemble en tenant compte du poids des différents produits dans la consommation des ménages. Les pondérations utilisées sont mises à jour au début de chaque année et publiées au niveau des 305 postes et de leurs regroupements. Pour éviter tout risque de manipulation, la liste des variétés est, elle, gardée confidentielle.

I.A.3 Les voies d'amélioration et d'enrichissement de l'existant

Par rapport aux données relevées par des enquêteurs les données de caisse comportent plusieurs avantages évidents qui contribueraient à une amélioration qualitative de l'IPC :

- absence d'erreurs de saisie ;
- coût marginal d'augmentation de la taille de l'échantillon faible, voire nul ;
- bénéfice de répertoires de produits ou références élémentaires, exhaustifs et comprenant des descriptifs physiques généralement plus riches que ceux-nécessairement sommaires - relevés par les enquêteurs.

Plus structurellement le recours aux données de caisse ouvrirait au moins trois voies d'amélioration de la qualité de l'IPC.

I.A.3. a Tirage de l'échantillon annuel dans une base exhaustive

Actuellement la base de sondage dans laquelle est tiré l'échantillon annuel est limitée, tant sur le plan géographique (106 agglomérations de plus de 2 000 habitants) que sur celui des produits (ceux qui font partie des 1 100 variétés).

Il est procédé de cette façon car, d'une part, une couverture plus exhaustive du territoire national imposerait des coûts de collecte plus élevés, et, d'autre part, il est nécessaire, grâce à la définition de chaque variété, de guider les enquêteurs dans la mise en œuvre des procédures annuelles de sélection des produits élémentaires à retenir dans l'échantillon.

A partir de données de caisse qui porteraient de façon exhaustive sur l'ensemble des points de vente et des références de produits, il serait possible de procéder par tirage aléatoire dans une base exhaustive de [produit élémentaire | point de vente] ce qui, toutes choses égales par ailleurs, améliorerait la qualité de l'estimation de l'IPC.

I.A.3.b.le suivi des promotions

En l'absence d'informations exogènes systématiques sur l'offre promotionnelle proposée dans les points de vente donnant lieu à relevés de prix il n'est pas toujours aisé pour les enquêteurs de repérer les éventuelles promotions concernant des produits de l'échantillon.

En tout état de cause les promotions de courte durée ne peuvent être appréhendées que si elles sont effectives le jour de relevé.

Le recours aux données de caisse, à condition de disposer d'indications sur le statut promotionnel éventuel des références des produits permettrait de rationaliser le suivi des prix des produits vendus en promotion. En résulterait une amélioration de la qualité du suivi des prix de l'offre et donc de l'IPC.

Par ailleurs, compte tenu de l'importance prise par les ventes promotionnelles dans les chiffres d'affaires des enseignes de la grande distribution et donc dans les achats de biens alimentaires des ménages, il pourrait être utile d'envisager les modalités d'une éventuelle actualisation des méthodes de prise en compte des prix des promotions dans les indices de prix à la consommation.

I.A.3.c la gestion des remplacements en dissemblable

Actuellement quand un [produit élémentaire | point de vente] disparaît de l'échantillon, dans le cas où son remplaçant n'est pas équivalent, l'estimation de la différence de qualité entre l'élément disparu et son remplaçant n'est pas automatique et peut s'avérer délicate.

Le recours aux données de caisse permettrait, au moins pour les cas de cette nature, de mettre en œuvre la technique du recouvrement, qui consiste à comparer directement le prix du produit remplaçant à son propre prix le mois précédent. Soit une prise en compte implicite et automatique de l'effet qualité. Cette technique- visée et approuvée dans les recommandations internationales relatives au suivi des indices de prix à la consommation -

ne peut être utilisée dans la collecte traditionnelle car le prix du remplaçant au cours du mois précédent le remplacement, n'a, par définition, pas été relevé et n'est donc pas disponible.

Soit un facteur d'amélioration de la qualité de l'IPC.

I.A.3.d Les enrichissements rendus possibles par le recours aux données de caisse

- La rotation des produits en cours d'année est devenue importante. Les évolutions mensuelles des prix des produits nouveaux, qui sont souvent marquées, ne peuvent pas être parfaitement prises en compte au cours de l'année de leur apparition dans le cadre d'un dispositif d'échantillon annuel. D'où une interrogation sur le choix de la bonne périodicité de renouvellement de l'échantillon. Faut-il continuer de chaîner les échantillons annuellement? ou les chaîner trimestriellement, mensuellement, hebdomadairement, voire quotidiennement. L'accès quotidien et exhaustif aux données de caisse rendrait possible la mise en œuvre de ces différents scénarii. La difficulté serait de choisir entre ces méthodes qui conduisent à des indices de prix présentant des évolutions très différenciées et qui, en tout état de cause, perdraient leur comparabilité internationale ou européenne
- Moyennant une augmentation appropriée de la taille de l'échantillon, pourraient être suivis des **indices de prix à la consommation régionalisés**.
- L'exploitation des données de caisse permettrait à l'Insee de procéder à des calculs expérimentaux de prix moyens pour certaines familles de produits, dans le cadre de **travaux de recherche relatifs à la faisabilité d'indices de prix moyens**, et, en particulier, de répondre à une demande de la direction générale de la Commission européenne (Eurostat) qui, dans le cadre d'un projet de recherche sur ce sujet, souhaite avoir communication de résultats de calculs de prix moyens relatifs à certaines familles de produits.

I.B Le système d'information des enseignes

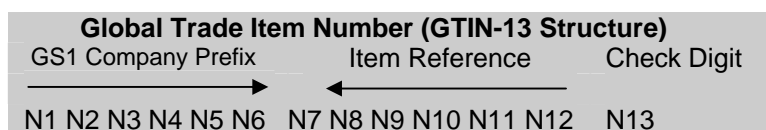
I.B.1 Les identifiants des produits

L'identification, lors de leur passage en caisse, des produits vendus dans les grandes enseignes de la distribution alimentaire s'effectue à l'aide des codes à barres qui figurent sur chaque article. A chaque code-barres correspond une séquence de 13 chiffres formant l'EAN (European Article Numbering)⁷ qui constitue l'identifiant individuel des données de caisse.

En France ce système de numérotation des articles est piloté par la société GS1 France, membre de l'association internationale GS1 spécialisée notamment dans la conception et l'harmonisation des conditions de mise en œuvre de standards d'identifiants des produits.

Toute entreprise qui adhère en France au système d'identifiants par code barres -EAN se voit attribuer par GS1 France, une plage d'EAN qu'elle peut utiliser librement pour identifier ses produits, dans le respect des règles d'usage prescrites par GS1.

La composition du numéro est décrit dans le schéma ci dessous proposé par GS1.



- Les trois premiers chiffres de l'EAN indiquent le pays dans lequel l'entreprise a adhéré au système EAN. Par exemple les trois premiers chiffres des numéros attribués par l'intermédiaire de GS1-France sont compris entre 300 et 379 (ceux attribués par GS1 Germany entre 400 et 440 et par GS1 China entre 690 à 695).
- Le code formé par les six premiers chiffres -ou préfixe du code barres constitue le numéro mondial d'identification de l'entreprise adhérente de GS1. Ce code barre permet ainsi l'identification individuelle du fabricant du produit correspondant.
- Le code formé par les douze premiers chiffres est un identifiant du produit.
- Le treizième chiffre est une clé de contrôle.

Ces numéros sont utilisés comme identifiants des produits dans les catalogues des fabricants, dans les transactions entre fabricants et distributeurs, dans les référentiels de produits des enseignes, et dans les bases centralisant les données de caisse.

Il convient de noter que les produits frais à poids variable et, de façon générale, les produits ou articles qui donnent lieu à conditionnement dans le point de vente - c'est-à-dire ceux qui ne relèvent pas, pour les distributeurs, d'une activité de pur négoce - sont étiquetés de "faux

⁷ Dénommé GTIN (Global Trade Item Number) par GS1

EAN ", qui en permettent l'identification lors du passage en caisse, au moyen de référentiels et des classifications propres au point de vente ou à l'enseigne.

I.B.2 Les données de caisse

I.B.2.a Les référentiels des enseignes

L'entrée d'un nouveau produit ou article dans la base de références d'une enseigne, est initiée par l'intermédiaire d'une fiche produit émise par son service acheteur, chaque fois qu'il contracte avec un fabricant ou producteur.

Les référentiels des enseignes sont quotidiennement actualisés et enrichis des nouvelles références, et bénéficient de vérifications et compléments - apportés aux données issues des fiches produits - en interne ou par l'intermédiaire de prestataires externes, par exemple des sociétés gestionnaires de catalogues de produits.

Chaque ligne de référence comprend notamment outre son code interne éventuel, son EAN, l'indication de sa position dans la classification de produits propre à l'enseigne, une désignation de l'article selon un libellé textuel qui donne les principales caractéristiques de son conditionnement, par exemple le volume ou le poids, le nombre d'unités.

Ces référentiels ne comportent pas de description détaillée des caractéristiques physiques des produits.

I.B.2.b Les classifications de produits

Propres à chaque enseigne et comportant jusqu'à sept ou huit niveaux, elles s'appuient principalement sur :

- au niveau le plus élevé, une répartition en rayons. Selon un ordre de grandeur d'une dizaine pour les produits de grande consommation et produits frais industriels (PGC et PFI). Par exemple : "épicerie salée", "épicerie sucrée", "crèmerie", "surgelés" ;
- à un niveau intermédiaire, une répartition en familles et sous familles de produits. Selon un ordre de grandeur d'un millier de familles ou sous familles pour les PGC et PFI. Par exemple pour une enseigne, le poste "céréales enfants" est une famille alors que pour une autre il s'agit d'une sous famille de la famille "céréales".
- au niveau le plus fin l'unité de besoin, qui ne semble guère utilisée par elle-même, mais constitue le poste de base de regroupement des références élémentaires.

I.B.2.c Description des données de caisse

Les données de caisse comportent, quotidiennement, par point de vente, et par référence de produit vendue au moins une fois au cours de la journée : le nombre d'unités vendues, le prix, le chiffre d'affaire correspondant, ainsi que le libellé désignant le produit.

L'identifiant des références est soit l'EAN, pour les produits de "pur négoce", y compris les marques de distributeurs, soit le "faux" EAN propre à l'enseigne.

Ces données, centralisées dans des bases de données de ventes, permettent aux enseignes de suivre, au niveau le plus fin de leurs classifications, les volumes physiques et nominaux de leurs ventes.

I.B.3 Le système d'information fondé sur les données de caisse

A l'exception des sociétés allemandes de hard discount implantées en France et de quelques magasins d'un grand distributeur français, l'ensemble des enseignes de la grande distribution alimentaire transmet, quotidiennement, l'intégralité de ses données de caisse, telles que décrites plus haut, à deux prestataires privés [IRI France et AC Nielsen] opérateurs du système d'information de la grande distribution alimentaire.

I.B.3.a Les sorties du système d'information (SI)

En retour des transmissions de données de caisse, effectuées par elles à titre gratuit, les enseignes ont accès, auprès de ces prestataires, à des données relatives à leurs parts de marché, selon un niveau d'agrégation suffisant pour rendre impossible l'identification des chiffres relatifs à un magasin ou à un produit particulier.

Les enseignes ayant adhéré à un dispositif contractuel dit "de transparence", géré par un prestataire, ont également accès aux prix pratiqués dans les magasins de leurs concurrents pour chaque article en vente dans un de leurs propres magasins⁸. Plus précisément toute enseigne adhérente a la possibilité d'acheter au prestataire les "relevés" de prix électroniques, dans tout magasin d'une autre enseigne adhérente, de produits sélectionnés par elle, parmi ceux qu'elle vend dans ses propres magasins.

I.B.3.b Les répertoires centraux d'EAN

Afin de pouvoir exploiter les données qu'ils centralisent, les opérateurs du SI gèrent des dictionnaires et systèmes de classification des produits tels qu'identifiés par leur codes barres ou EAN.

Ces répertoires constitués à partir de différentes sources, (catalogues de produits des fabricants, enquêtes menées directement sur les points de vente pour visualiser les articles et en repérer les caractéristiques, prospectus publicitaires distribués par les enseignes notamment à l'occasion d'opérations de promotion), comportent, notamment, outre un libellé textuel désignant le produit, sa description à travers plusieurs dizaines de caractéristiques physiques, et sa position dans une classification propre à l'opérateur.

L'organisation de ces classifications, est analogue à celles des enseignes, avec lesquelles il est aisé de faire le lien.

S'agissant des produits dotés de "vrais " EAN c'est-à-dire des produits de grande consommation et des produits frais industriels, la précision et la richesse des répertoires et

⁸ Une des principales enseignes françaises de la grande distribution alimentaire n'a pas adhéré à ce mécanisme. En conséquence elle n'a pas accès, dans ce cadre, aux prix pratiqués dans les magasins des autres enseignes pas plus que les autres enseignes n'ont accès aux prix pratiqués dans ses propres magasins.

classifications permettent aux enseignes de procéder ou faire procéder- sous réserve du respect de leur confidentialité - à toutes les analyses qu'elles jugent utiles.

Ce n'est pas le cas pour les produits frais traditionnels.

I.B.3.c Les difficultés de suivi des produits frais traditionnels

L'offre de produits frais, tels que les fruits et légumes ou les poissons, présente la particularité d'être en renouvellement permanent - selon les saisons, mais aussi d'une année à la suivante -et d'être composée d'objets individuellement différents les uns des autres car constitués de matière vivante - laquelle est elle-même susceptible d'évoluer en rayon.

Il n'existe pas aujourd'hui de bonne classification de référence des produits frais, ni chez les enseignes, ni chez les opérateurs de leur SI, ni même au niveau international.

En l'absence d'un minimum d'harmonisation entre les classifications - internes aux enseignes - de ces produits, il est impossible d'exploiter valablement les données de caisse portant sur cette catégorie de produits.

A titre d'exemple si dans la classification des produits frais d'une enseigne le niveau le plus fin de classement des poissons est un poste intitulé "Poissons", aucune exploitation centrale des données relatives à chacune des variétés de poisson n'est possible⁹.

En réalité les produits frais traditionnels échappent au SI des enseignes.

L'Insee effectue, mensuellement, deux vagues de 40 000 relevés de prix, portant, pour chacune des variétés observées dans les points de vente visités par les enquêteurs, sur l'exhaustivité des produits proposés à la vente. Soit un suivi de l'offre de produits frais autant que de leurs prix.

Cette source de données sur les produits frais pourrait, le cas échéant, donner lieu à des exploitations particulières, par exemple à des niveaux régionalisés, ou pour certaines formes de vente, afin de mettre à la disposition des enseignes quelques indicateurs des prix et de l'offre de produits frais.

⁹ Car la seule information disponible sur la nature précise du produit est qu'il s'agit de poisson

I.C Les différents modes d'accès possibles au SI des enseignes pour la production de l'IPC

Pour produire l'IPC à partir d'un échantillon ou panier de [codes barres | point de vente], il est nécessaire de disposer

- de données de prix à l'EAN ;
- d'un répertoire de produits à l'EAN qui permette notamment de les classer dans les postes ad hoc retenus pour le calcul de l'IPC et de procéder aux remplacements des références qui, en cours d'année, disparaissent de l'échantillon.
- d'éléments détaillés sur le poids de ces produits en terme de quantités vendues, autant pour le tirage d'un échantillon représentatif que pour le calcul de l'indice.

Le tableau suivant croise ces catégories de données avec les différentes sources possibles.

	IRI	Nielsen	Enseignes	Autres
Données de caisse	Oui	Oui	Oui	
• Prix	Oui	Oui	Non	Oui (collecte Insee)
• Pondérations	Oui	Oui	Non	Oui(enquête conso)
Référentiel produits(EAN)	Oui	Oui	Non(insuffisant)	Oui(Insee mais lourd)

A priori il est envisageable soit ① d'obtenir directement les données de caisse, soit ② d'obtenir séparément des données de prix et des données de pondération.

① Les données de caisse permettent directement de constituer un échantillon annuel ou panier de [codes barres | point de vente] et de calculer des indices mensuels et annuels de prix à condition d'avoir par ailleurs accès à un répertoire d'EAN approprié.

A priori les données de caisse peuvent techniquement être obtenues auprès de chacun des deux opérateurs nationaux du SI des enseignes ou directement, auprès des enseignes.

Le référentiel de produit peut, a priori, être obtenu auprès d'un des deux opérateurs, ou être produit en régie directe par l'Insee, en prenant notamment appui sur les répertoires propres aux enseignes, mais s'agissant d'un développement lourd, il se justifierait mal dans la seule perspective de la production de l'IPC.

② Les éléments relatifs aux prix d'un panier prédéterminé de [codes barres | point de vente] et ceux relatifs aux pondérations pourraient être obtenus séparément, auprès de l'un ou l'autre des deux opérateurs. Cette disjonction entre les sources, qui conduirait à se priver de la richesse de données croisées [prix x quantités] et, partant, en limiterait les possibilités d'exploitation, semble devoir être écartée.

Alternativement l'Insee pourrait continuer de procéder à ses relevés de prix et obtenir les éléments de pondérations par ses enquêtes sur la consommation des ménages.

La question de l'accès à un référentiel de produits(EAN) se pose dans les mêmes termes qu'au ①.

II INSTRUCTION DE LA FAISABILITÉ DU PROJET

II.A Un test de faisabilité d'un indice de prix à partir des données de caisse

Afin de vérifier qu'il serait effectivement possible d'utiliser les données de caisse pour le calcul d'indices de prix cohérents avec ceux qui résultent de la collecte traditionnelle, il a été procédé, à partir d'un échantillon de données de test provenant des enseignes membres du groupe de travail,

- à un premier exercice de comparaison descriptive, portant sur près de 350 produits élémentaires, entre données issues de la collecte des prix en magasin par les enquêteurs de l'Insee et données de caisse ;
- à un exercice de test systématique d'une méthode d'utilisation des données de caisse pour la détermination d'indices de prix relatifs aux seules familles de produits de l'échantillon.

L'échantillon de données de test, présenté en détail à l'annexe 1, comprend, pour dix familles ou groupes de familles de produits, trois années de données hebdomadaires de caisse, relatives à mille points de vente des enseignes participantes.

II.A.1 **Un premier exercice de comparaison descriptive entre les données de caisse et les données de la collecte actuelle**

Le but de cet exercice était, à travers un essai d'appariement entre les données issues des deux sources, de bien appréhender leurs différences concrètes de nature, à l'exclusion de toute considération relative à la qualité des données.

Parmi les 1000 points de vente de l'échantillon ont été sélectionnés ceux dans lesquels des relevés de prix ont été effectués en décembre 2009¹⁰.

Parmi les produits élémentaires donnant lieu à relevés ont été sélectionnés ceux qui appartiennent à l'une des dix familles ou groupes de familles de produits de l'échantillon de test, soit 345 produits.

D'où un ensemble de couples [point de vente | produit élémentaire] pour chacun desquels les données provenant des deux sources sont disponibles.

C'est à partir de ces données, décrites de façon synthétique dans le tableau ci-après, que l'on a cherché à repérer, au mieux, dans les données de caisse, les références des produits ayant donné lieu aux relevés physiques de prix par les enquêteurs de l'Insee.

¹⁰ Par commodité on s'est limité aux points de vente situés dans les dix régions françaises dans lesquelles les enseignes participantes sont les mieux implantées (Nord et Est de la France).

Dans la mesure où :

- les familles de références élémentaires ne recouvrent pas exactement les variétés ;
- les descriptifs individuels des objets sont organisés de façon sensiblement différentes entre les deux sources ;
- les prix retenus dans l'échantillon sont des moyennes hebdomadaires ;

le but de l'exercice était d'apprécier les termes et les difficultés de la construction d'une passerelle entre les deux sources et non de parvenir à un appariement complet entre la collecte effectuée par les enquêteurs IPC et les données de caisses.

Principales caractéristiques des relevés IPC et des données de caisse utilisées

	RELEVES DE PRIX INSEE	DONNEES DE CAISSE
Numéro d'identification dans un répertoire partagé	Non	Oui pour les produits de pur négoce : EAN fabricant (ou distributeur pour les MDD) Non pour les produits frais (faux EAN attribués en interne par les distributeurs)
Date d'application du prix	Le jour	Le jour (en réalité la semaine dans l'échantillon utilisé)
Classification	Variété d'appartenance (les variétés sont définies selon une nomenclature mise à jour annuellement)	Famille d'appartenance' (pour l'échantillon) ou position dans une classification partagée pour les produits de pur négoce. Position dans une classification propre au distributeur pour les produits frais
Caractéristiques techniques du produit élémentaire	Un ensemble, spécifique à chaque variété - et précisé dans sa définition - de quelques caractéristiques physiques du produit élémentaire objet du relevé de prix, <i>saisies de façon obligatoire par l'enquêteur</i> . [Par exemple la quantité physique du produit, le type de marque, la marque.] Un ensemble non formalisé de zéro à deux ou trois caractéristiques du produit <i>notées à l'initiative de l'enquêteur et choisies par lui</i> , qui lui permettent de bien identifier le produit pour en faciliter le relevé du prix [par exemple la couleur du conditionnement pour un paquet de pâtes]	Pour les produits de pur négoce : un ensemble bien formalisé de quelques dizaines de caractéristiques physiques, variables selon la famille d'appartenance de la référence élémentaire. Pour les produits frais <i>absence d'homogénéité des descriptifs physiques</i> des références élémentaires

Il a été procédé selon deux critères. En premier rang ont été recherchés les EAN dont les caractéristiques physiques étaient compatibles avec les caractéristiques techniques notées à l'appui des relevés des enquêteurs de l'Insee. En second rang le prix des références a été comparé avec le prix relevé par les enquêteurs de l'Insee.

Le constat résultant de cet exercice est le suivant :

- Pour 60% des relevés des enquêteurs a été trouvée une référence(EAN) ayant des caractéristiques physiques compatibles avec les caractéristiques techniques du relevé de l'enquêteur, et dont le prix est identique au prix relevé ;
- Pour 15% des relevés il a été possible d'associer au relevé au moins une référence compatible avec les caractéristiques techniques et comportant un prix proche de celui du relevé (moins de 5% d'écart) ;
- Pour 25% des relevés il n'a pas été possible d'identifier un EAN compatible et ayant un prix proche du prix relevé.

Parmi les facteurs explicatifs des difficultés d'appariement entre les articles physiques objet des relevés de prix par les enquêteurs de l'Insee et les données de caisse :

- les différences de périmètre entre les variétés de l'IPC et les familles d'EAN auxquelles se sont limitées les recherches d'EAN susceptibles de correspondre aux relevés. Certains relevés portaient sur des articles dont les références ne figuraient pas parmi les familles retenues dans l'échantillon de données de caisse. Ce point devra être approfondi, en vue d'une parfaite articulation entre les familles d'EAN et les variétés de l'IPC.
- les incompatibilités entre les descriptifs physiques des données de caisse et ceux des relevés qui ne sont pas formalisés selon les mêmes termes. D'une part la partie facultative notée par les enquêteurs, si elle joue bien son rôle pour la mise en œuvre de la collecte, est difficile à exploiter lors d'un croisement avec d'autres sources. D'autre part les mises à jour des caractéristiques obligatoires sont parfois effectuées avec retard, c'est-à-dire qu'elles ne correspondent plus à l'article donnant lieu à relevé de prix -ce qui ne présente aucun inconvénient du point de vue de la qualité de l'IPC- mais interdit tout rapprochement avec les données de caisse. Il pourrait être jugé utile, à condition de ne pas alourdir le travail des enquêteurs, de chercher les voies d'amélioration des descriptifs physiques des articles donnant lieu à relevés.

II.A.2 Un exercice de test d'utilisation des données de caisse pour le calcul de l'IPC

II.A.2 ① La méthode et les modalités de simulation de sa mise en œuvre

L'exercice a consisté à transposer à un échantillon ou panier de [code barres | point de vente] la méthode actuelle de calcul de l'IPC, laquelle s'applique à un panier de [produit élémentaire | point de vente].

Afin de tirer des enseignements statistiques sur la validité de la méthode, 500 simulations différentes ont été réalisées. Chacune consiste en un tirage aléatoire - effectué sur les données du mois de décembre 2008 - de trois paniers de 10000, 20000 et 50000 couples [code barres | point de vente] conduisant à autant de séries mensuelles - de décembre 2008 à décembre 2009. Les produits suivis qui disparaissent un mois donné sont remplacés automatiquement(voir infra).

Procéder par simulation permet d'estimer la variabilité des résultats obtenus.

Les tirages n'ont porté que sur les codes barres appartenant aux huit familles ou groupes de familles de références correspondant précisément aux postes "café", "chocolat en tablette",

"huile", "riz", "yaourt", "fromage à pâte molle", "œufs", "jus de fruits" de la consommation des ménages, qui donnent lieu à publication de huit sous-indices de l'IPC, auxquels les huit sous-indices de prix "données de caisse" résultant des simulations peuvent donc être comparés.

II.A.2 ②.a Le mode de tirage de l'échantillon

Dans l'IPC actuel, l'objectif associé à la constitution du panier annuel de référence est d'obtenir un échantillon de produits représentatif de la diversité de la consommation des ménages d'une part, de sélectionner des produits « biens suivis, biens vendus » dans le point de vente où ils sont vendus de manière à pouvoir observer leur prix chaque mois sur une longue période de temps d'autre part. De façon non accessoire, cette manière de procéder conduit à un nombre de remplacements mensuels inférieur à celui qui résulterait d'un tirage purement aléatoire dans l'ensemble des produits offerts à la vente.

S'agissant de constituer un panier de codes-barres, la meilleure représentativité est assurée par un tirage aléatoire avec probabilité de sélection proportionnelle aux quantités vendues. Il se trouve que les produits les plus vendus sont également les plus suivis, c'est-à-dire ceux dont les taux de remplacements annuels sont les plus faibles, ainsi que le montre l'étude préalable sur la stabilité des codes barres présentée ci-dessous.

Ont été déterminées parmi toutes les séries de la base de sondage (mois de décembre 2008) celles qui étaient stables au cours de l'année 2009, c'est-à-dire ceux dont les ventes n'étaient nulles pour aucun des mois de l'année 2009. On obtient par famille de produits IRI les résultats suivants :

Numéro Famille (IRI)	Description Famille IRI	Nombre de séries (croisement code- barres et point de vente) au mois de base décembre 2008	Vente totales 2009 des séries présentes au mois de base (en milliers d'euros)	Taux de stabilité des séries sur l'année 2009
-------------------------------------	--	---	--	--

La sélection des échantillons de [codes-barres|point de ventes] a été faite au mois de décembre 2008, au sein de chaque famille IRI, par tirage aléatoire avec probabilités proportionnelles aux ventes de décembre 2008¹¹.

II.A.2 ②.b La taille de l'échantillon

Actuellement dans le champ des familles de produits de l'échantillon de données de test, pour les points de vente des enseignes participant au test le taux de sondage de la collecte, IPC est voisin de 0,1% (de l'ordre de 1 500 relevés effectués pour un total de 1,5 millions de couples [codes- barres | point de ventes] présents dans les données de test).

Pour mener les simulations ont été tirés, dans les seules huit familles ou groupes de familles retenues pour le test, des échantillons de 10 000, 20 000 et 50 000 couples [codes-barres | point de vente] ce qui correspond sensiblement à des taux de sondage respectifs de 1, 2 et 5% sur les 1,1 millions de séries présentes en décembre 2008 dans les 1000 points de ventes sélectionnés pour le test.

II.A.2 ②.c La gestion des remplacements

Dans le cadre des simulations réalisées, une procédure de remplacement automatique simple a été adoptée, fondée sur le fabricant - ou la marque- tel qu'il est identifié par le préfixe du code barres (EAN).

Lorsqu'un code-barres de l'échantillon disparaît dans un point de vente -c'est-à-dire lorsque ses ventes mensuelles sont nulles - on le remplace par un code-barres vendu dans le même point de vente, stable [au sens de l'étude préalable présentée page 24], dont les 12 premiers chiffres forment un nombre le plus proche possible de celui du code barre disparu et dont les 6 premiers (formant le préfixe) sont identiques.

Si un tel remplaçant est trouvé, on impute à la série mensuelle du produit disparu l'évolution du prix du produit remplaçant par rapport au mois précédent (méthode de recouvrement).

S'il n'existe dans le point de vente aucun code-barres stable ayant les 6 mêmes premières positions que le code-barres disparu, on impute à la série l'évolution moyenne des prix de la famille concernée dans le point de vente, calculée comme la moyenne géométrique des évolutions de prix mensuelles des codes-barres stables ou pour lesquels un remplaçant a été trouvé.

¹¹ Par famille, enseigne et marque c'est-à-dire que le nombre de séries par enseigne dans l'échantillon pour une famille IRI est proportionnel à la part des ventes de l'enseigne pour cette famille IRI, et de même pour le nombre de séries par marque

Ce procédé de choix des remplaçants est volontairement frustré à ce stade, car il présente l'avantage d'être automatique. Ce mode de remplacement permet également d'éviter la démultiplication des remplacements successifs dans une même série mensuelle. Il conviendra naturellement de l'améliorer pour une utilisation des données de caisse dans l'IPC.

Les remplacements ayant été ainsi effectués on dispose d'une table donnant un prix mensuel pour chacune des 10 000, 20 000 ou 50 000 séries que comporte chacune des 500 simulations.

II.A.2 ①.d Le mode de calcul de l'indice des prix élémentaire d'une famille IRI

Compte tenu du mode de tirage des séries dans une base de sondage globale pour chaque famille IRI, on calcule ici directement un indice global par famille IRI.

Comme il s'agit de familles relativement larges et donc constituées de produits hétérogènes, on calcule un indice par moyenne géométrique des évolutions de prix de l'ensemble des produits de l'échantillon tiré dans la famille.

II.A.2 ② Les résultats des simulations effectuées

II.A.2 ② .a Les résultats obtenus au niveau « poste » de l'indice des prix

Afin d'obtenir des résultats significatifs, les indices de prix ne sont calculés que pour des familles IRI ou groupes de familles IRI correspondant à un « poste » dans la nomenclature de diffusion de l'indice des prix¹² (soit un ensemble de variétés de produits représentant une part suffisamment importante de la consommation des ménages pour que le nombre de relevés de prix effectué permette de calculer un indice faisant sens). On obtient ainsi des indices de prix « données de caisses » pour huit postes de l'indice des prix.

① Précision des indices obtenus

Les simulations montrent une précision satisfaisante des indices de postes obtenus (indices significatifs à 1% près sur le champ des enseignes participant au test¹³) pour les taux de sondage sur données de caisse retenus.

A titre d'exemple, on donne ici les résultats observés pour différentes tailles d'échantillons pour le poste étudié le moins important en terme de consommation (riz) et celui qui est le plus important (yaourts)

Riz (famille IRI 4206) :

¹² Le niveau poste est le niveau le plus fin dans la nomenclature de diffusion de l'indice des prix.

¹³ C'est-à-dire que si l'indice calculé pour un poste donné sur le champ des enseignes participant au test est de +3%, on peut dire qu'il est compris entre +2% et +4%.

Lecture du tableau : pour une taille d'échantillon de 10 000 séries, 5% des échantillons tirés donnent un indice d'évolution annuel des prix du riz en 2009 inférieur ou égal à **-3,0%**.

On constate ainsi que pour un échantillon de global de 10 000 codes-barres (soit un taux de sondage de 1% des codes-barres), on obtient dans 90% des cas un indice des prix du riz, sur les 6 enseignes considérées, significatif à 1% près (puisque l'indice des prix moyen est de -2,1% et que 90% des échantillons donnent un indice compris entre -3,0% et -1,1%).

Lorsqu'on double la taille de l'échantillon, l'indice est significatif à 1% dans 98% des cas.

Yaourt (famille IRI 5701) :

Dans le cas de la famille Yaourt, un taux de sondage de 1% suffit à obtenir un indice significatif à 1% dans 98% des cas.

Les résultats complets pour l'ensemble des postes sont donnés en annexe 1.

② Comparaison avec les indices issus de la collecte IPC

Les indices calculés au niveau de chaque poste à partir des données de caisses sont comparés avec les indices du poste suivants issus de l'IPC :

- indice poste IPC calculé dans le cadre de la collecte actuelle (toutes enseignes et toutes formes de ventes, sur la base d'indices élémentaires calculés pour chaque [variété | agglomération])
- indice « grande distribution » calculé toutes enseignes confondues sur les formes de ventes étudiées (hypermarchés, supermarchés et magasins populaires), comme la moyenne géométrique des évolutions de prix pour l'ensemble des produits IPC du champ appartenant au poste
- indice « grande distribution/enseignes du test » calculé sur les enseignes participant au test et les formes de ventes étudiées, comme la moyenne géométrique des évolutions de prix pour l'ensemble des produits IPC du champ appartenant au poste

On obtient les résultats suivants :

Riz

Huiles alimentaires

Jus de fruits

Chocolat en tablette

Yaourts

Cafés

Œufs

Fromages à pâte molle

On constate ainsi une bonne concordance entre les indices Données de Caisses et les indices IPC grande distribution toutes enseignes confondues (écart limité à 1%). Les divergences avec l'indice IPC calculé sur le seul champ des enseignes participant au test viennent avant tout, de l'incertitude des indices correspondants fondés sur la collecte IPC, compte tenu du champ très restreint sur lequel portent les indices et de la faiblesse du nombre de relevés de prix associés.

II.A.2 ② .b Calculs d'indices de prix globaux sur les 8 postes étudiés

① Indice global portant sur les seules enseignes participant au test

On étudie l'indice d'évolution globale des prix sur les 8 postes étudiés et sur les 6 enseignes participant au test. On compare l'indice global issu des données de caisses (moyenne pondérée des indices de postes « données de caisses » calculés au paragraphe précédent) et l'indice issu de la collecte IPC (moyenne pondérée des indices de poste « IPC grandes distribution sur les enseignes du test »). Les pondérations sont les ventes totales annuelles 2008 des enseignes participant au test (estimées à partir des données de test). On obtient les résultats suivants :

La proximité de l'indice Données de Caisses (-1,4%) et l'indice fondé sur les données de l'IPC (-2,0%) est satisfaisante, compte tenu du champ restreint sur lequel ils portent (environ 3% des ventes de la grande distribution) et donc de l'importance de l'erreur d'échantillonnage des indices calculés à partir des données de la collecte IPC.

② Indice global portant sur l'ensemble de la grande distribution

On calcule ici des indices portant sur l'ensemble du champ de la grande distribution. On compare un indice « purement IPC » calculé à partir des seuls relevés IPC avec un indice « mixte » calculé à partir des données de caisses pour les enseignes participant au test et à partir des relevés IPC pour les enseignes non participantes.

Méthode de calcul des indices « purement IPC » et des indices mixtes

Les données de ventes annuelles 2008 « toutes enseignes confondues » sont issues des fichiers de consommation annuelle fournis à l'INSEE par la société Nielsen.

L'indice IPC « toutes enseignes confondues » est calculé comme la moyenne pondérée par les ventes totales annuelles 2008 « toutes enseignes confondues » des indices de poste « toutes enseignes confondues » calculés au paragraphe précédent.

Pour calculer l'indice mixte « Données de Caisse/IPC » global, on calcule d'abord un indice mixte au niveau de chaque poste, comme une moyenne pondérée de l'indice données de caisses pour les enseignes participant au test et de l'indice IPC calculé sur les enseignes ne participant pas au test.

Le poids de l'indice « données de caisses » dans l'indice mixte correspond à la part des enseignes participant au test dans les ventes totales de la grande distribution pour le poste considéré, calculée

comme le rapport entre les ventes totales annuelles 2008 des enseignes participant au test (estimées à partir des données de test) et les ventes totales annuelles 2008 de la grande distribution (issues des données Nielsen)

L'indice mixte global « Données de Caisses/IPC » est ensuite calculé comme la moyenne des indices mixtes au niveau poste pondérée par les ventes annuelles 2008 « toutes enseignes confondues ».

On obtient les résultats suivants :

On constate une bonne proximité entre l'indice IPC « toutes enseignes confondues » (-1,5%) et l'indice mixte (-1,3%), compte-tenu du fait que les 8 postes suivis couvrent de l'ordre de 10% des ventes de la grande distribution.

Annexe : inflation annuelle au niveau poste observée pour différentes tailles d'échantillon

On obtient sur les 500 simulations de tirage de séries les résultats suivants sur l'inflation annuelle 2009 :

Riz (famille IRI 4206) :

On constate ainsi que pour un échantillon global de 10 000 codes-barres (soit un taux de sondage de 1% des codes-barres), on obtient dans 90% des cas un indice des prix du riz sur les 6 enseignes considérées significatif à 1% près.

Lorsqu'on double la taille de l'échantillon, l'indice est significatif à 1% dans 98% des cas.

Huiles alimentaires (familles IRI 3202 et 3203) :

On constate qu'un taux de sondage de 1% des codes-barres (taille globale de 10 000 codes-barres) suffit pour avoir un indice des prix significatif à 1% dans 98% des cas.

Jus de fruits (famille IRI 8002) :

On constate qu'un taux de sondage de 1% des codes-barres (taille globale de 10 000 codes-barres) suffit pour avoir un indice des prix significatif à 1% dans 98% des cas.

Chocolat en tablette (famille IRI 7904) :

On constate également qu'un taux de sondage de 1% des codes-barres (taille globale de 10 000 codes-barres) suffit pour avoir un indice des prix significatif à 1% dans 98% des cas.

Yaourt (famille IRI 5701) :

Café (famille IRI 0205, 0701 et 0703) :

Pour le poste « cafés », il est nécessaire d'avoir un taux de sondage de 2% des codes-barres pour avoir un indice des prix significatif à 1% dans 98% des cas.

Œufs (famille IRI 6402) :

Pour le poste œufs également, il est nécessaire d'avoir un taux de sondage de 2% des codes-barres pour avoir un indice des prix significatif à 1% dans 98% des cas.

Fromages à pâte molle (familles IRI 8271 à 8274) :

Dans le cas des fromages à pâte molle, on obtient une très bonne précision avec un taux de sondage de 1%, puisque l'amplitude maximale est de + ou - 0,7%.

II.B Les questions méthodologiques à traiter

II.B.1 Quelles articulations entre les variétés de l'IPC actuel et les données de caisse ?

D'une part les données de caisse permettront de connaître de façon fine et en temps réel les produits effectivement achetés par les ménages. Ceci devrait faciliter les actualisations annuelles de la liste des variétés et les ajustements annuels de leurs périmètres et engendrer une meilleure réactivité dans la prise en compte de l'évolution de l'offre, facteur d'amélioration de la qualité de l'IPC.

D'autre part, dans l'autre sens, le classement - à effectuer le plus possible en temps réel- des EAN apparaissant en cours d'exercice, selon les variétés de l'IPC, permettra de disposer à tout moment pour chaque variété de la liste des EAN qui la composent. Il en résultera, sans doute, une possibilité d'amélioration et de rationalisation de la gestion des remplacements effectués mensuellement dans la collecte traditionnelle résiduelle.

II.B.2 La question du traitement des promotions:

Elle comprend deux volets.

- Comment repérer dans les données de caisse les articles vendus en promotion ?

Deux situations doivent être distingués :

- ou bien le produit fait l'objet d'un conditionnement particulier de la part du fabricant ce qui implique un EAN spécifique, et la mention "special pack"(par opposition à "regular pack") dans sa documentation soit un repérage aisé du caractère promotionnel. ;
 - ou bien le produit bénéficie d'une promotion spécifique aux magasins et à l'enseigne qui le proposent à la vente, sans changement de code barre. Dans ce cas le repérage du statut promotionnel peut passer, par exemple, par des enquêtes auprès d'un échantillon national de magasins (visites dans les points de vente, collecte et exploitation des prospectus publicitaires émis à l'occasion des campagnes de promotion) et par le recours à des modèles de propagation à l'ensemble des points de vente des grandes enseignes, des indications issues des enquêtes de terrain¹⁴.
- Selon quelle méthode prendre en compte dans le calcul de l'IPC les prix des produits vendus en promotion ?. Faut il considérer que ces conditions temporaires de vente, qui représentent de l'ordre de 12% du chiffre d'affaire des enseignes doivent faire l'objet d'un traitement particulier, comme s'il s'agissait de produits spécifiques analogues aux produits saisonniers ? Ou bien convient il de les assimiler, au jour le jour, selon des calculs de prix moyens, à des produits courants ?

¹⁴ Méthode utilisée par la société IRI

II.B.3 La question des remplacements.

Le taux de remplacement mensuel d'un échantillon d'EAN sera supérieur à celui de l'échantillon de produits de l'IPC actuel. Mais le traitement de ces remplacements sera facilité par l'application de la méthode du recouvrement, à laquelle les données de la collecte traditionnelle ne permettent pas de recourir. Sa mise en œuvre pourra être largement automatisée.

Dans la mesure où la méthode du recouvrement traite implicitement les effets qualité se pose la question suivante. Convient-il de continuer à envisager les remplacements de façon individuelle, un produit disparu se voyant attribuer un remplaçant ? Est-il envisageable de procéder à un remplacement mensuel global de l'ensemble des références disparues en faisant en sorte de maintenir la représentativité de l'échantillon, par exemple au regard de l'âge des références selon leur cycle de vie ? Cette question complexe ne pourra être éludée, sauf à prendre le risque d'introduire des biais dans l'IPC.

II.C La faisabilité juridique

Afin d'apporter aux enseignes une garantie de respect de la confidentialité de leurs données le dispositif suivant leur sera proposé.

II.C.1 Des conventions pluriannuelles de trois ans (2012, 2013, 2014) prorogables un an

Ces conventions préciseront les modalités techniques selon lesquelles l'Insee pourra effectivement accéder aux données de caisse [par exemple selon des transmissions quotidiennes directes identiques à celles dont bénéficient actuellement IRI et Nielsen] et stipuleront expressément que l'Insee s'engage :

- i. **à n'utiliser ces données que dans le cadre de l'élaboration de l'indice des prix à la consommation, à l'exclusion de tout autre usage** [les enseignes signataires qui y auront convenance pourront convenir qu'elles autorisent l'Insee à utiliser leurs données pour des calculs d'IPC régionalisés, selon des zonages à préciser, et/ou à mener des travaux de recherche portant exclusivement sur d'autres modalités de calcul de l'IPC ou sur le calcul d'indices de prix alternatifs tels que des indices de prix moyens à la condition que d'éventuelles publications des résultats de ces recherches ne permettent pas l'identification individuelles d'un point de vente, d'une enseigne, ou d'un produit élémentaire ;]
- ii. **à ce que seuls les agents de la division des prix à la consommation et ceux des services informatiques dédiés à la maintenance de l'IPC auront le droit d'utiliser ces données, à l'exclusion de tout autre agent.**
- iii. **que les données de caisse transmises à la division des prix à la consommation seront détruites sous un délai maximum de 24 mois.** Seules seront conservées les données de prix relatives à l'échantillon à partir duquel a été effectivement calculé l'IPC et les éléments de pondération généraux utilisés.

Ces conventions viseront expressément l'arrêté du ministre de l'économie et des finances mentionné ci dessous, qui sera annexé à chaque convention.

Il sera stipulé que toute modification apportée à cet arrêté devra être notifiée par l'Insee au cosignataire de la convention et que faute d'avoir été actée par avenant, avant sa date d'entrée en vigueur, une telle modification, entrainera annulation de la convention.

II.C.2 Un arrêté du ministre de l'économie et des finances

Le projet d'exploitation des données de caisse sera, préalablement à la signature des conventions, soumis à l'avis des formations compétentes du Conseil National de l'Information Statistique, ce qui formalisera la couverture des données par le secret statistique prévu par la loi n°51-711 du 7 juin 1951 sur l'obligation, la coordination et le secret en matière de statistiques.

Il sera approuvé par un **arrêté du ministre chargé de l'économie qui reprendra formellement les stipulations i, ii, iii mentionnées ci dessus.**

III DIAGNOSTIC DU GROUPE SUR LA FAISABILITÉ DU PROJET

1. L'utilisation des données de caisse pour le suivi des prix à la consommation serait très positive pour la statistique publique.
2. La faisabilité méthodologique et technique de l'utilisation des données de caisse pour le calcul de l'indice des prix est avérée, grâce aux travaux de simulation menés sur l'échantillon de test (hors produits frais).
3. Pour réaliser tout le gain qualitatif potentiel du recours aux données de caisse il sera nécessaire d'approfondir la méthodologie générale de prise en compte des promotions dans l'indice des prix, et de développer une méthode spécifique de remplacement des références qui disparaissent, en cours d'exercice, des échantillons.
4. Du côté de l'Insee la faisabilité informatique , telle que testée à partir de données quotidiennes en grandeur réelle transmises par quelques enseignes, est attestée par le service compétent, sous réserve de l'affectation des moyens humains nécessaires à la mise en œuvre du projet.
5. Il est clair que les enseignes n'ont pas grand chose à gagner en entrant dans ce partenariat, même si l'Insee est prêt à leur faire des retours, notamment sur des indices de prix des produits frais, ou sous forme d'accès gratuits à une partie de son offre éditoriale.
6. Elles ne peuvent se permettre de prendre le moindre risque concernant la protection de la confidentialité de leur données.
7. Le dispositif juridique proposé par l'Insee, par voie contractuelle (conventions) et réglementaire (arrêté ministériel) garantira que les données transmises, couvertes par le secret statistique,
 - o ne seront utilisées que dans le cadre de l'élaboration de l'indice des prix à la consommation, à l'exclusion de tout autre usage ;
 - o que seuls les agents de la division des prix à la consommation et ceux des services informatiques dédiés à la maintenance de l'IPC auront le droit d'utiliser ces données, à l'exclusion de tout autre agent ;
 - o que les données de caisse transmises à la division des prix à la consommation seront détruites sous un délai maximum de 24 mois.
8. Sur cette base, sans préjuger des décisions que chaque enseigne et l'Insee auront ou non convenance à prendre le groupe s'accorde sur le caractère faisable d'un ensemble de propositions opérationnelles.

IV PROPOSITIONS OPÉRATIONNELLES

Sur la base des analyses et des éléments d’instruction du projet de partenariat développés précédemment et sans préjuger des décisions que chaque enseigne membre du groupe et l’Insee auront ou non convenance à prendre, l’ensemble des membres du groupe s’accorde sur le caractère faisable des sept propositions opérationnelles suivantes, dont certaines sont soit complémentaires soit exclusives les unes des autres.

1. L’Insee proposera aux enseignes membres du groupe :
 - de leur faire retour d’indices de prix des produits frais, par forme de vente ou type d’agglomération, selon des modalités à convenir en commun ;
 - de leur fournir, à titre gracieux le fichier détail anonymisé, de l’enquête budget des familles ;
 - de leur ouvrir un accès, selon des modalités à définir en commun, à son offre éditoriale.

L’information des enseignes sur les magasins visités par les enquêteurs sera améliorée.

2. Les enseignes transmettront quotidiennement leurs données de caisse à l’Insee, à l’identique et selon les mêmes conditions techniques, que celles qui prévalent aujourd’hui pour les transmissions effectuées vers deux opérateurs centralisateurs de leurs données.
3. L’Insee sera autorisé à acquérir les données de caisse des enseignes membres du groupe auprès de l’un ou l’autre des deux opérateurs centralisateurs de leurs données.
4. L’Insee acquerra par abonnement auprès d’un prestataire privé, un dictionnaire de produits, à l’EAN, et sa classification détaillée, avec mises à jour mensuelle
5. L’Insee acquerra mensuellement, auprès d’un prestataire centralisateur des données de caisse un ensemble de relevés de prix issus des données de caisse, pour un échantillon de [code barre | point de vente] sous réserve de l’accord des enseignes concernées.
6. L’Insee acquerra auprès d’un prestataire privé des données de pondération pour le calcul de l’IPC
7. Un comité de coordination Enseignes/Insee se réunissant au moins une fois par an, sera chargé de suivre le bon déroulement du déploiement du partenariat.

PREMIER TEST DE LA FAISABILITÉ DU PROJET DONNÉES DE CAISSE.
Description de l'extrait des données de caisse des hypermarchés et supermarchés demandé par l'Insee.

- Période étudiée : données hebdomadaires enregistrées lors des passages en caisse au cours des trois années 2007 à 2009, soit 156 semaines
- Produits suivis : données relatives aux gencodes correspondant à des produits appartenant à l'une des 10 familles ou regroupements de segments de marché suivants.

1	Famille	Café Torréfié
2	Famille	Chocolat en tablette
3	Famille	Jus de fruits
4	Famille	Œufs
5	Famille	Papier hygiénique
6	Famille	Yaourt
7	Famille	Huiles alimentaires
8	Famille	Riz
9	3 segments de la famille « Fromage »	Fromages à pâte molle croûtes fleuries Fromages à pâte molle croûtes lavées Fromages à pâte persillée
10	1 segment de la famille "Entrées, pizza, fast food surgelé"	Tartes salées surgelées

Globalement ces produits représentent, en moyenne, de l'ordre de 1000 gencodes par point de vente (hypermarchés et supermarchés).

- Points de vente : 1000 points de vente (hypermarchés et supermarchés) choisis parmi les 7000 existants de la façon suivante.
 - 700 tirés de manière purement aléatoire ;
 - de l'ordre de 120 sélectionnés parmi ceux dans lesquels les enquêteurs de l'Insee procèdent effectivement, pour les produits retenus pour le test, à un nombre de relevés supérieur à un seuil minimal (il s'agit d'assurer une bonne comparabilité entre les données de caisse et les données IPC) ;
 - de l'ordre de 180 correspondant à l'ensemble des points de vente (hypermarchés et supermarchés) situés dans quelques départements, dont plusieurs agglomérations sont effectivement couvertes par les enquêteurs de l'Insee.

- Informations demandées, pour chaque triplet (semaine, point de vente, code-barres) :

Identifiant de la semaine, identifiant du point de vente, identifiant code-barres, montant des ventes totales hebdomadaires, quantité ou nombre d'unités vendues dans la semaine, prix de vente électronique hebdomadaire (rapport entre le montant total des ventes hebdomadaires et des quantités totales hebdomadaires vendues), et, au mieux de ce qu'il sera possible d'obtenir, des variables de documentation du code-barres (nature du produit, marque, caractéristiques techniques : forme, composition, emballage, conditionnement, ... présence ou non de promotions).

ANNEXE 2 COMPARAISON ENTRE LA STABILITÉ D'ÉCHANTILLONS DE [CODE BARRES | POINT DE VENTE] ET D'ÉCHANTILLONS DE [PRODUITS ÉLÉMENTAIRES | POINT DE VENTE] : L'IMPACT DES PROMOTIONS

I Stabilité des échantillons de séries observés dans les simulations

On a simulé le tirage d'échantillon de 10000 séries dans la base de sondage annuelle 2009, réparties au sein des 17 familles IRI proportionnellement au volume des ventes en 2009 de chaque famille¹⁵.

500 tirages d'échantillon ont été effectués dans le cadre des simulations, selon une méthode de tirage permettant d'assurer une bonne représentativité de l'échantillon de séries en termes d'enseigne et de marque¹⁶.

On constate en moyenne sur les 500 échantillons tirés qu'environ 72% des séries sélectionnées sont stables¹⁷, soit un taux bien supérieur au taux de stabilité des séries dans la base de sondage (55%).

La proportion de remplacements à effectuer dans les échantillons simulés est donc de 28% en moyenne, contre 45% si on avait sélectionné les séries indépendamment des ventes.

II Comparaison avec la collecte IPC

Une étude du taux de remplacements effectués en 2009 dans le cadre de la collecte enquêteurs actuelle sur les produits correspondant aux 10 familles suivies dans les données de test a été effectuée.

On constate ainsi en 2009 **un taux de remplacement de 17%** pour les produits suivis dans le cadre de la collecte **enquêteurs actuelle**, à comparer avec un **taux de remplacement de 28%** pour les échantillons annuels de séries issus des **données de caisses**.

Les produits à vie courte sont mieux pris en compte dans les échantillons de séries issus des données de caisses

Une première explication de la différence sur les taux de remplacements vient de la prise en compte ou non des promotions « fabricant ». On entend par promotions « fabricant » des produits différents en termes de conditionnement et de packaging du produit habituellement vendu, et dont le prix au litre ou au kg est plus bas que pour le produit standard : par exemple un lot de deux bouteilles avec 15% gratuit au lieu de la bouteille à l'unité vendue habituellement (par opposition aux promotions « magasins » dans lesquelles le magasin baisse le prix du produit habituellement en rayon)

En effet, ces promotions « fabricant » correspondent à des produits (et des codes-barres) temporaires, qui donneront lieu dans la très grande majorité des cas à un remplacement. Par exemple, si on a une bouteille de soda d'1,5 litre suivie dans le panier IPC annuel, et qu'elle coexiste un mois donné avec une bouteille de 2 litres du même soda avec « 15%

¹⁵ Ventes totales par famille calculées à partir des séries présentes dans la base de sondage 2009.

¹⁶ L'introduction de critères de représentativité supplémentaires a été testée mais conduisait à des temps de traitements informatiques trop importants. Avec des deux critères de représentativité introduits ici, le temps total nécessaire aux simulations a été de 66 heures.

¹⁷ Ce taux variant peu d'un échantillon à l'autre puisqu'il est toujours compris entre 71% et 74%.

gratuits », l'enquêteur saisira uniquement le prix de la bouteille d'1,5 litres habituellement suivie (on notera que la bouteille de soda deux litres avec « 15% gratuit » correspond dans les données de caisses à un code-barres différent de la bouteille de soda habituelle d'1,5 litres, et donc à une série différente dans la base de sondage).

En revanche, dans les simulations d'échantillons de codes-barres, les promotions « fabricant » sont traitées comme les autres codes-barres, ce qui fait qu'on retrouve en moyenne dans les échantillons tirés la même proportion de codes-barres en promotions que dans la base de sondage.

On a les résultats suivants sur l'importance des codes-barres associés à une promotion « fabricant » dans la base de sondage :

On constate ainsi que les promotions « fabricant » représentent 7% des codes-barres, pour un peu moins de 10% des ventes.

En supposant que toutes les promotions « fabricant » correspondent à des codes-barres instables (durée de vie du produit inférieure à un an), on retrouve des ordres de grandeur cohérents sur les remplacements, puisque le taux de remplacement annuel global dans les données de caisses (28%) est proche de la somme du taux de remplacements dans l'IPC (17%) où les produits suivis excluent dans la très grande majorité des cas les promotions « fabricant » et du taux de produits correspondant à des promotions « fabricant » (7%).

La différence restante (4%) pourrait venir du fait que l'IPC suit uniquement des produits « biens suivis biens vendus », tandis que les échantillons de séries contiennent aussi des codes-barres correspondant à des produits moins vendus et plus instables, même si c'est en faible proportion compte-tenu du montant limité de leurs ventes.

Il ressort donc de cette étude que l'augmentation des remplacements lors du passage de la collecte enquêteurs actuelle à un échantillon de codes-barres reste limitée (taux passant de 17% à 28%) et que cette hausse peut s'expliquer par une meilleure couverture des produits à durée de vie courte, d'une part les promotions « fabricant » et d'autre part des produits plus atypiques avec un faible chiffre d'affaires.