

# Projet Nosica

NOouvelles Sources pour les Indices de Chiffres d'Affaires

## Étude de faisabilité et d'opportunité

d'utilisation des données de caisse de la grande distribution alimentaire  
en remplacement de  
l'Enquête Mensuelle sur l'Activité des Grandes Surfaces Alimentaires

Réunion de concertation avec  
les enseignes de la grande distribution alimentaire

*10 février 2021*

Insee<sup>1</sup>



**Insee**

Institut national de la statistique  
et des études économiques

Mesurer pour comprendre

Direction des statistiques d'entreprises  
Département des indicateurs de court terme  
Division Indicateurs conjoncturels d'activité

Direction de la méthodologie et  
de la coordination statistique et internationale  
SSP<sup>2</sup> Lab

---

1 TIMBRE DG75-E320 - MONTROUGE - 88 avenue VERDIER - CS 70058 - 92541 MONTROUGE - FRANCE -  
www.insee.fr - Tél. standard : 01.87.69.50.00 - N° SIRET : 120 027 016 00019 - Code APE : 84.11Z

2 Service statistique public



## SOMMAIRE

Introduction.....	5
1 – L'enquête Emagsa.....	7
2 – Comparaisons entre sources.....	9
3 – Un processus en partie mutualisé.....	12
3.1 – Corrections et imputations.....	12
3.2 – Les outils de contrôle.....	12
3.3 – La mutualisation des processus.....	13
4 – Les apports des données de caisse.....	14
4.1 – Précocité des indicateurs.....	14
4.2 – Une meilleure couverture.....	15
4.3 – Des indices par produits.....	15
4.4 – Des indices par région.....	17
4.5 – Réduction de la charge pour les entreprises et l'Insee.....	18
5 – Les aspects techniques.....	18
5.1 – Les informations demandées.....	18
5.2 – Les conditions de transmission.....	19
5.3 – La fréquence et les délais de transmission.....	19
5.4 – Les espaces de conservation des données.....	19
5.5 – Les personnes autorisées à y accéder.....	19
5.6 – La durée de conservation.....	19
Conclusion.....	20
Liste des pièces jointes.....	21



# INTRODUCTION

---

La division Indicateurs conjoncturels d'activité (ICA) de l'Insee produit mensuellement des indices de chiffres d'affaires sur les secteurs de l'industrie, de la construction, du commerce et des services. Le calcul de ces indices répond à des impératifs nationaux et européens. Ils sont utilisés pour mesurer les variations mensuelles du chiffre d'affaires dans les secteurs marchands (non agricoles et non financiers). En tant que tels, ils constituent une information primaire pour suivre le cycle économique en France et sont utilisés, par exemple, par les comptes nationaux trimestriels français. La production des indices de chiffres d'affaires est régie, depuis le 1<sup>er</sup> janvier 2021, par le règlement European Business Statistics<sup>3</sup>.

La production des indices de chiffres d'affaires repose sur l'exploitation des données mensuelles de TVA (formulaire CA3). Cette source administrative, fournie par la DGFIP, permet de ne pas mettre à contribution les entreprises mais présente certains défauts : une couverture imparfaite puisque toutes les entreprises – en particulier les petites – ne sont pas placées sous le régime des déclarations mensuelles<sup>4</sup>, et des délais de publication relativement longs puisqu'il n'est pas possible de diffuser les indices avant la fin du second mois suivant le mois d'affaires. Ce délai de diffusion pose problème pour répondre à la demande européenne exprimée dans le règlement European Business Statistics concernant les statistiques de court terme qui demande aux États membres de l'Union européenne la production précoce (30 jours après la fin du mois) d'indicateurs d'activité sur le commerce de détail. La récente crise sanitaire a également rappelé, s'il en était besoin, l'intérêt de disposer de statistiques de qualité le plus rapidement possible.

Pour répondre à cette demande, l'Insee s'appuie sur un dispositif d'estimation mis en place à la fin des années 1990 et fondé principalement sur l'enquête mensuelle d'activité dans la grande distribution alimentaire (Emagsa), qui recueille des données auprès des entreprises non spécialisées du commerce de détail. L'enquête Emagsa permet de produire de façon précoce (30 jours après la fin du mois) des indices d'activité sur la grande distribution alimentaire et, en association avec d'autres sources et en s'appuyant sur des modèles économétriques, des estimations précoces sur l'ensemble du commerce de détail<sup>5</sup>. Cette source est également utilisée dans le cadre de l'élaboration du rapport sur les comptes commerciaux de la nation et de la construction des comptes nationaux sur le secteur du commerce. Plus précisément, l'enquête Emagsa fournit des informations statistiques sur le chiffre d'affaires des grandes surfaces alimentaires, ventilé par forme de vente (supermarchés / hypermarchés) et par grandes familles de produit (alimentaire / non alimentaire / carburants). Même si l'ampleur de l'enquête est modérée, la collecte de l'enquête Emagsa représente une charge pour les entreprises et mobilise des ressources à l'Insee.

Afin d'améliorer la qualité de ces indicateurs, notamment les premières estimations, l'Insee a lancé le projet Nosica (NOuvelles Sources pour les Indices de Chiffre d'Affaires) pour rechercher de nouvelles sources de données (telles que les données de paiements par carte bancaire ou les données de caisse des enseignes de la grande distribution à prédominance alimentaire) qui permettraient de pallier les défauts de la source TVA et de l'enquête Emagsa. Les objectifs du projet sont les suivants :

- réduire les délais de publication, améliorer la qualité de l'estimation anticipée de l'indice du volume du commerce de détail (30 jours après la fin du mois) ;
- produire une ventilation plus fine sur les produits, avec potentiellement des utilisations pour les comptes nationaux trimestriels ;

---

3 Règlement (UE) 2019/2152 du Parlement européen et du Conseil du 27 novembre 2019 relatif aux statistiques européennes sur les entreprises.

4 En France, les relevés de TVA sont effectués mensuellement, trimestriellement ou annuellement en fonction du chiffre d'affaires de l'entreprise. Les déclarations trimestrielles et annuelles ne sont donc pas utilisées dans le calcul des indices mensuels du chiffre d'affaires.

5 Le commerce de détail se compose du commerce de détail non spécialisé (notamment les grandes surfaces alimentaires), le commerce de détail en magasin spécialisé (différents secteurs selon les produits principaux vendus) et le commerce hors magasin (commerce en ligne notamment).

- réduire la charge pesant sur les entreprises par abandon de l'enquête Emagsa ;
- réduire les coûts pour l'Insee (toutefois, cet objectif ne sera atteint que si le coût de la collecte et du traitement de la nouvelle source de données est inférieur aux coûts correspondants de l'enquête Emagsa) ;
- élargir le champ des estimations précoces (actuellement sur le seul champ du commerce de détail) aux services fournis aux ménages (à partir de l'exploitation des données de paiement par cartes bancaires) ;
- enfin, d'autres objectifs peuvent également être explorés dans un second temps : ventilation géographique avec des indicateurs pouvant être utilisés pour le calcul du PIB régional et meilleure séparation des unités profilées<sup>6</sup> dans le commerce de détail.

Une première source s'est rapidement imposée dans la mesure où l'Insee en dispose déjà, celle des données de caisse de la grande distribution alimentaire, jusqu'alors réservée au calcul de l'indice des prix à la consommation (IPC). Depuis début 2020, l'IPC français est partiellement<sup>7</sup> calculé à partir des données de caisse.

La source Emagsa et la source données de caisse permettent toutes deux d'estimer les ventes sur le commerce non spécialisé, selon une structuration toutefois différente (agrégation au niveau unité légale pour l'enquête Emagsa vs données disponibles par point de vente pour les données de caisse) et sous une forme bien moins riche pour l'enquête Emagsa actuelle puisque le détail par produits est réduit aux trois catégories alimentaire / non alimentaire / carburants. Les données de caisse offrent une couverture quasi exhaustive du champ des grandes surfaces alimentaires et sont reçues bien plus tôt que les données Emagsa (elles sont disponibles quelques jours après la fin du mois).

---

6 Le profilage consiste à analyser les groupes complexes pour identifier en leur sein, par delà leur organisation juridique en sociétés, des entreprises pertinentes en tant qu'acteurs économiques appelées « unités profilées ». Le profilage intègre aussi le fait d'identifier et de mettre en place au sein de ces groupes les modalités les plus efficaces pour la collecte de l'information statistique.

7 Les relevés de prix sur le terrain demeurent pour les types de points de vente non couverts par les données de caisse.

L'objectif de la présente étude de faisabilité et d'opportunité prévue au second alinéa de l'article 3 bis I de la loi n° 51-711 du 7 juin 1951 (cf. pièce jointe n°1) est :

- de s'assurer que les informations présentes dans les bases de données de caisse répondent aux objectifs de l'enquête Emagsa et améliorent la connaissance du secteur ;
- d'évaluer l'avantage présenté par la transmission par voie électronique par rapport à d'autres modes de collecte, notamment en termes de qualité et de précision, au regard de la charge pour les personnes morales de droit privé sollicitées et pour le service statistique public ;
- de proposer les conditions de transmission et de conservation de ces données à l'Institut national de la statistique et des études économiques afin de garantir la sécurité et la confidentialité des données.

Cette étude sera présentée et discutée lors de la concertation mentionnée au second alinéa du I de l'article 3 bis de la loi n° 51-711 du 7 juin 1951 rassemblant des représentants de l'administration et des personnes morales de droit privé concernées.

L'étude de faisabilité et d'opportunité, le rapport réalisé à la suite de la concertation ainsi que les observations des personnes morales de droit privé concernées seront transmis au Conseil national de l'information statistique pour avis.

En cas d'avis favorable du Cnis, un arrêté du ministre chargé de l'économie définit les personnes morales de droit privé concernées, les informations demandées et les conditions de leur transmission. Il détermine en particulier la fréquence, les délais de leur transmission et la forme qu'elle revêt. Il précise les espaces où sont conservées ces données, les personnes ayant l'autorisation d'y accéder, la durée de leur conservation qui ne peut excéder cinq ans à compter de leur transmission, ainsi que les modalités de leur destruction.

## 1 – L'ENQUÊTE EMAGSA

---

L'Enquête Mensuelle sur l'Activité des Grandes Surfaces Alimentaires (Emagsa) est réalisée par l'Insee auprès des entreprises qui gèrent des supermarchés et/ou des hypermarchés et permet de réaliser des estimations précoces sur ce champ (cf. *supra*)<sup>8</sup>. L'enquête a été présentée au CNIS et au comité du label<sup>9</sup> en 2019 pour le renouvellement de ses agréments. Le CNIS a renouvelé son avis d'opportunité favorable et le comité du label a attribué le label d'intérêt général et de qualité statistique à l'enquête. Les documents rendus à ces comités et les échanges qui y ont eu lieu faisaient mention des travaux initiés en vue d'un basculement vers de nouvelles sources plus efficaces et ne pesant pas sur les entreprises.

Environ 9 000 entreprises entrent dans le champ de l'enquête, hypermarchés, supermarchés et magasins multi-commerces qui constituent donc la base de sondage. L'échantillon est constitué d'environ 400 entreprises (unités légales) et est renouvelé chaque année. Il est exhaustif pour les entreprises ayant un chiffre d'affaires supérieur à 300 millions d'euros TTC annuel ou un effectif supérieur à 1 000 salariés. Dans les autres strates, un tirage coordonné positivement<sup>10</sup> avec les précédentes éditions est mis en œuvre pour

8 La première enquête a eu lieu en 1997 sous le nom d'enquête mensuelle des grandes surfaces (Emags).

9 Le Cnis vérifie l'opportunité de chaque opération en vérifiant qu'elle répond à un besoin d'intérêt général et ne fait pas double emploi avec d'autres sources existantes. Le Comité du label veille ensuite au respect des critères de qualité statistique.

10 L'objectif de la coordination d'échantillons consiste à prendre en compte, lors du tirage d'un échantillon donné, ceux des enquêtes précédentes, en général dans une optique de réduction de la charge statistique imposée aux

favoriser un recouvrement suffisant. La couverture géographique porte sur la France entière. Le réseau des enquêteurs entreprises de l'Insee accompagne les entreprises lors de leur entrée dans l'échantillon (cf. en pièces jointes n°2a la notice et n°2b le questionnaire de recrutement).

La collecte des données est réalisée mensuellement par Internet à partir du Portail de réponse aux enquêtes auprès des entreprises (<http://entreprises.insee.fr/>). Une notice explicative est accessible à partir du portail. Cette notice et un fac-similé du questionnaire sont disponibles en pièces jointes n°3a et n°3b. L'ouverture de chaque campagne mensuelle a lieu quelques jours après la fin du mois d'affaires enquêté. Les entreprises ayant transmis une adresse électronique sont informées par courriel de l'ouverture de la campagne. Les données sont traitées par le pôle Indicateurs Sectoriels d'Activité (ISA) de la direction régionale des Pays de la Loire qui gère aussi les relances des entreprises non répondantes.

Les questions posées portent sur le chiffre d'affaires TTC total de l'entreprise et sur la ventilation suivante :

- chiffre d'affaires réalisé en supermarchés dont :
  - vente de produits alimentaires ;
  - vente de produits non alimentaires ;
  - vente de carburants ;
- chiffre d'affaires réalisé en hypermarchés dont :
  - vente de produits alimentaires ;
  - vente de produits non alimentaires ;
  - vente de carburants.

Une partie du chiffre d'affaires total déclaré n'apparaît pas dans la ventilation. Il s'agit du chiffre d'affaires correspondant aux ventes relatives aux courses en ligne (retraits en drive, livraisons à domicile, retraits en magasin), aux ventes dans d'autres types de magasins gérés tels que les supérettes, les magasins de bricolage, de jardinage, de presse, les centres d'entretien et de réparation d'automobiles et aux prestations de services telles que la restauration en self-service, la location de véhicules automobiles, les services de lavage/nettoyage de véhicules proposés dans les stations essence.

Les indices sont diffusés sur le site internet de l'Insee, 30 jours après la fin du mois d'affaires. Ils sont diffusés par forme de vente (supermarchés et hypermarchés) et par grandes familles de produits (alimentaires, non alimentaires et carburants), soit 12 séries au total, permettant ainsi une analyse plus précise des mouvements conjoncturels. Les séries statistiques sont disponibles en valeurs brutes et corrigées des variations saisonnières et des effets de calendrier (correction des jours ouvrables). Un exemple de publication est disponible en pièce jointe n°4.

Comme évoqué en introduction, et au-delà de l'objectif de réduction de la charge sur les entreprises, le projet vise à gagner en qualité par rapport à ce que permet d'obtenir l'enquête Emagsa : raccourcissement des délais de publication ; amélioration de la qualité au vu des benchmarks existants (mais disponibles plus tardivement) tels que les indices de chiffres d'affaires construits à partir de données fiscales (formulaires CA3) ou les statistiques structurelles annuelles (liasses fiscales) ; ventilations par produits et par région, qui intéressent d'autres services de l'Insee.



## 2 – COMPARAISONS ENTRE SOURCES

---

Plusieurs enseignes de la grande distribution alimentaire ont, par conventions, autorisé l'Insee à exploiter leurs données de caisse pour réaliser cette étude. Les données des supermarchés et des hypermarchés y sont intégrées, y compris les données des livraisons à partir des magasins et des drives, mais hors e-commerce. Ces enseignes représentent au total 44 % du chiffre d'affaires total de la grande distribution en 2019<sup>11</sup>. Aucune des quatre enseignes participantes (équivalent statistique de la notion de distributeur) ne dépasse les 2/3 du chiffre d'affaires étudié. Dans l'ensemble de l'analyse qui suit et des pièces jointes annexées, les tableaux et graphiques utilisent donc comme source les agrégations issues de ces différentes enseignes ayant signé une convention. Par la suite, la source n'est plus rappelée.

Les enregistrements utilisés couvrent la période allant du 1<sup>er</sup> janvier 2018 au 31 octobre 2020 et contiennent les informations suivantes :

- la date des ventes (quotidiennes) ;
- l'identifiant du magasin propre à l'enseigne ;
- l'identifiant de l'article vendu (ean) ;
- le libellé de l'article ;
- le chiffre d'affaires généré par les ventes de l'article dans la journée et le magasin.

Toutes les données issues des données de caisse ne sont pas intégrées dans l'indice des prix à la consommation pour des raisons méthodologiques. Par exemple, les prix des fruits et légumes en vrac, les viandes à la découpe ou la boulangerie continuent d'être relevés en magasins car ces produits ne sont pas documentés dans le référentiel de produits acheté par la division des prix à la consommation à une société d'études de marché. Il est donc impossible de vérifier, par exemple, que la variété de pommes est bien la même d'un mois sur l'autre et surtout, leur poids, qui varie à chaque vente, n'est pas connu. Au contraire, le projet de remplacement de l'enquête Emagsa nécessite l'utilisation de la totalité des remontées de caisse. Il a donc été nécessaire, pour les besoins de cette étude, de relire les fichiers originaux de données de caisse avec toutes les difficultés que cela comporte : volumétrie importante, quelques défauts parfois dans les fichiers, etc. Cette relecture a par ailleurs permis d'établir précisément les résultats sur les délais de réception des données et l'impact potentiel des doublons sur le chiffre d'affaires, doublons qui sont éliminés d'emblée lors du chargement des données dans les bases de l'IPC.

Les graphiques ci-dessous montrent les évolutions mensuelles et annuelles du chiffre d'affaires issu des données de caisse, de la source TVA et de l'enquête Emagsa.

Le chiffre d'affaires des données de caisse est celui des enseignes intégrées dans l'étude, supermarchés et hypermarchés confondus, y compris le chiffre d'affaires des drives. Le chiffre d'affaires de la source TVA est l'agrégation des chiffres d'affaires totaux des unités légales des secteurs 4711D (supermarchés) et 4711F (hypermarchés). Le chiffre d'affaires de l'enquête Emagsa est également le chiffre d'affaires total issu de cette enquête. Dans tous les cas, il s'agit de chiffres d'affaires bruts (non désaisonnalisés et sans correction des jours ouvrés), d'où des variations de fortes ampleurs.

---

11 Source LSA n°2587 du 23 janvier 2020

Figure 1 : comparaison des évolutions mensuelles (en%) de chiffre d'affaires entre les données de caisse, la source TVA et l'enquête Emagsa<sup>12</sup>

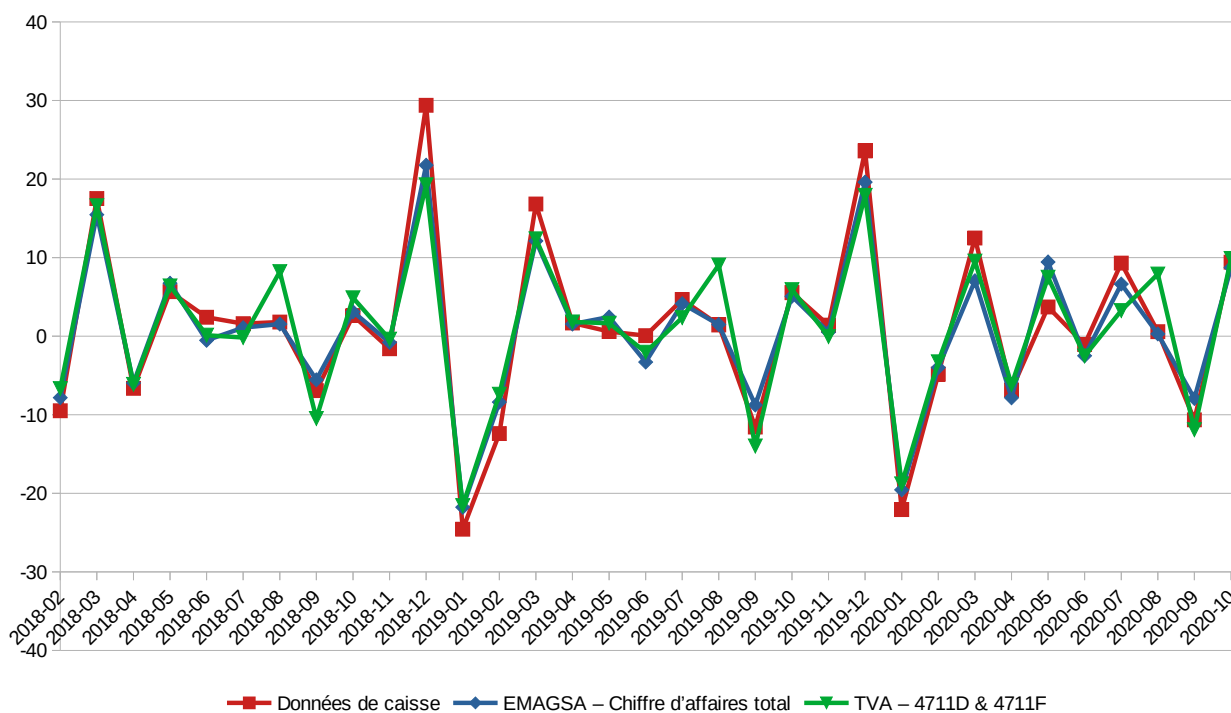
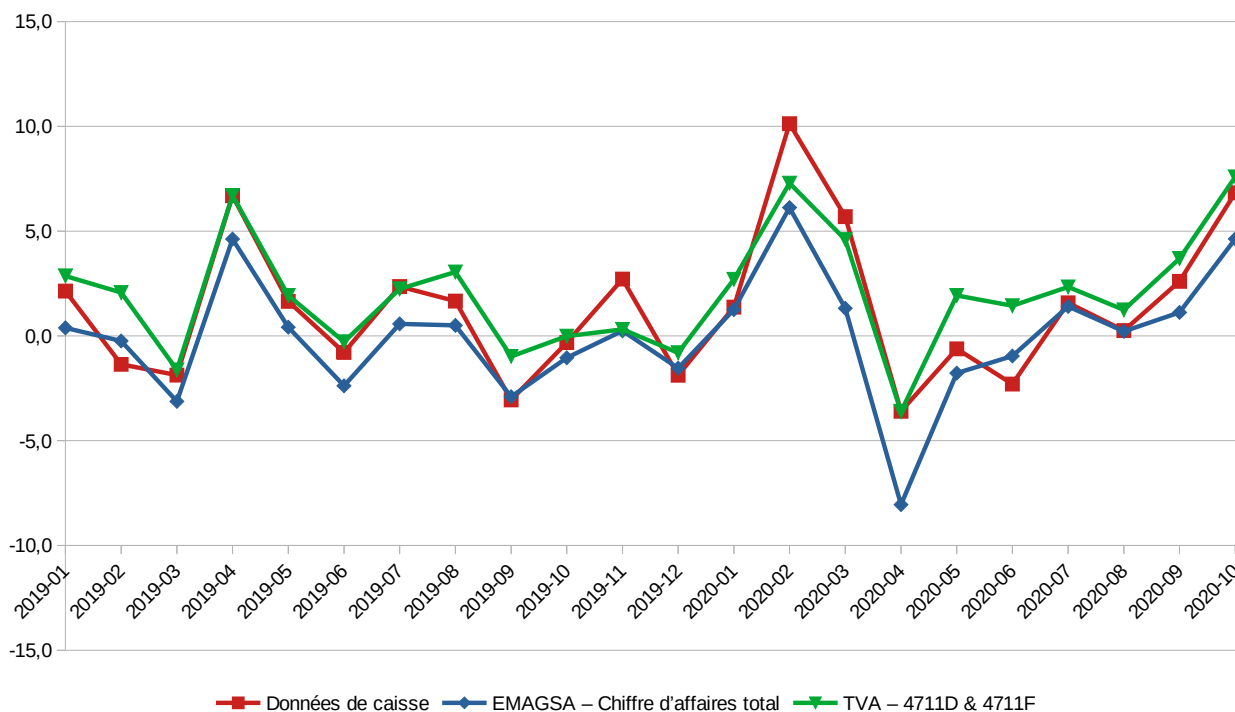


Figure 2 : comparaison des évolutions en glissements annuels (en%) de chiffre d'affaires entre les données de caisse, la source TVA et l'enquête Emagsa



12 N.B. : les indices issus de l'enquête EMAGSA et des déclarations de TVA n'étaient pas disponibles pour les mois de novembre et décembre 2020 au moment de la rédaction de l'étude.

Ces comparaisons mettent en évidence une forte cohérence dans les évolutions constatées d'une source à l'autre, et ce malgré des différences de champ importantes qui seront peu à peu neutralisées au fur et à mesure de l'avancement du projet. Ainsi les écarts s'expliquent par plusieurs raisons, dont certaines purement statistiques ne devraient pas apparaître en CVS-CJO :

- les données de caisse ne recouvrent dans cette étude (dans l'attente de l'intégration des autres enseignes) que 44 % du champ (secteurs 4711D et F) alors que les données issues de la source TVA comme de l'enquête Emagsa couvrent l'ensemble, soit par technique d'échantillonnage (Emagsa), soit par construction (TVA). De plus, la cession (ou la reprise) d'un magasin par une enseigne du champ de l'étude à une enseigne n'en faisant pas partie fait baisser (augmenter) le chiffre d'affaires des données de caisse dans cette étude. Des comparaisons à égalité de périmètre sur les trois sources – au niveau entreprise – ont été réalisées et donnaient de bons résultats mais ne peuvent être publiées sans risquer de trahir le secret statistique ;
- des différences de concept sur la mesure du chiffre d'affaires existent d'une source à l'autre :
  - les données de caisse permettent de mesurer le chiffre d'affaires des établissements classés en hypermarchés ou en supermarchés. La source Emagsa inclut en plus dans le chiffre d'affaires total des grandes enseignes certaines données en sus (supérettes et autres magasins de l'unité légale – bricolage, presse, parapharmacie, billetterie, services de location, services après-vente, restauration, ...) et les ventes de carburants. Enfin, la source TVA donne le chiffre d'affaires des unités légales, donc y compris activités secondaires n'appartenant pas forcément au secteur considéré (par exemple, le commerce de gros, le e-commerce) ;
  - les données des départements ultramarins ne sont pas comprises dans les données de caisse ;
  - le chiffre d'affaires issu de la source TVA comprend les chiffres d'affaires réalisés en France et à l'exportation, néanmoins ces derniers sont très faibles ;
  - la TVA est incluse dans les données de caisse et l'enquête Emagsa alors que les chiffres d'affaires issus de la source TVA sont hors taxe. Cependant, les taux de TVA sont inchangés en France depuis le 1<sup>er</sup> janvier 2014, il est donc aisé de corriger de cet effet ;
- la saisonnalité peut varier d'une source à l'autre, avec parfois des effets statistiques associés à une source en particulier. C'est le cas de la saisonnalité en juillet-août pour la source TVA : en raison de certaines facilités de déclarations durant cette période, la somme des déclarations sur juillet (transmises à l'organisme fiscal en août) peut sous-estimer les ventes réelles, et inversement en août où les ventes peuvent être surestimées. Cette particularité est neutralisée en CVS-CJO.

Pour plus d'informations sur les sources comparées, voir la pièce jointe n°5.

Tenant compte de ces différences de concepts entre les sources et vu la proximité des évolutions mensuelles et annuelles constatées, l'Insee souhaite abandonner l'enquête Emagsa et la remplacer par les données de caisse. Dans cette optique, une première esquisse du nouveau processus a été dessinée.

## 3 – UN PROCESSUS EN PARTIE MUTUALISÉ

---

### 3.1 – Corrections et imputations

Comme toute donnée, les données de caisse doivent être vérifiées et corrigées si nécessaire.

Le chiffre d'affaires peut être artificiellement gonflé par des enregistrements en double (souvent dus à des renvois de fichiers ou des problèmes de chargement en base) qui doivent être détectés et éliminés. La part de ces doublons chaque mois est présentée dans la pièce jointe n°6. Les doublons sont facilement détectables, il s'agit des enregistrements pour lesquels le magasin, le produit vendu, la date de vente, le chiffre d'affaires et la quantité sont identiques. Ils sont déjà détectés par la chaîne de traitement IPC lors de la mise en base des données.

A contrario, des problèmes techniques peuvent conduire certains magasins à transmettre tardivement leurs données ou ne pas les transmettre du tout pour certains jours, générant ainsi des valeurs manquantes. Il faut donc être en mesure de repérer et d'estimer les chiffres d'affaires manquants. Les indicateurs de détection de ces absences sont décrits dans le paragraphe suivant. Ils sont importants car ils ne doivent pas conduire à estimer le chiffre d'affaires d'un magasin ayant fermé et inversement, un chiffre d'affaires nul ne doit pas être imputé à un magasin qui est toujours en activité. L'utilisation des données de caisse étant moins exhaustive dans le calcul de l'IPC que dans celui des indices de chiffre d'affaires, les contrôles de la chaîne de traitements IPC vont devoir être renforcés. Ces contrôles renforcés profiteront au calcul de l'IPC. Une fois les valeurs manquantes détectées, les imputations pourront s'appuyer sur des techniques proches de celles utilisées par l'Insee pour les autres indicateurs d'activités et qui consistent à s'appuyer sur les observations d'unités proches.

### 3.2 – Les outils de contrôle

Le suivi quotidien du chiffre d'affaires d'une enseigne met facilement en évidence de gros défauts dans la transmission des données mais les petits défauts sont difficilement détectables ainsi car le chiffre d'affaires est très volatil au cours de la semaine et entre semaines même consécutives. En effet, il existe de très fortes différences dans le comportement d'achat du consommateur au cours de la semaine ; comportement qui a évolué avec le confinement du début d'année 2020. Par ailleurs, le chiffre d'affaires est fortement lié aux événements exceptionnels : jours fériés (à la hausse ou à la baisse), journées de promotion (soldes, black Friday, ...), etc. L'analyse doit donc être réalisée en comparant à la fois aux points proches et aux points équivalents de l'année précédente.

Un indicateur plus pertinent est le parc des magasins même s'il n'est pas exempt de défauts. Le nombre de magasins des données de caisse est un indicateur clé de la qualité des données. Cependant, il ne faut pas se limiter à compter les magasins (pour une entreprise), car l'ouverture d'un nouveau magasin pourrait éventuellement masquer l'absence de transmission de données par un autre magasin. Le suivi quotidien du chiffre d'affaires de chaque magasin permet d'identifier des ruptures dans les séries (absence ou apparition de données à partir d'une certaine date) qui seraient beaucoup moins visibles dans les données agrégées sur un mois entier, surtout si l'événement se produit vers la fin du mois. Lorsqu'aucune information n'est transmise pour un magasin, soit parce qu'elle a été oubliée, soit parce que le magasin a cessé ses activités, il est important de faire une distinction claire entre les deux ; l'examen des données jour par jour permet d'identifier toute omission occasionnelle de transmission. Symétriquement, l'examen des données jour par jour et magasin par magasin permet également d'identifier les changements de pratique, comme l'inclusion de l'activité de drive dans les chiffres transmis à partir d'une certaine date.

Pour les besoins du calcul de l'IPC, l'Insee a acquis (auprès d'une société d'études de marché privée) un référentiel des magasins régulièrement mis à jour, très utile pour identifier les créations et les fermetures de

magasins. À l'avenir, le référentiel des magasins pourrait éventuellement être construit en fusionnant d'une part le répertoire statistique des entreprises de l'INSEE, et d'autre part divers répertoires privés, ce qui nécessite toutefois que ces fichiers soient mis en correspondance soit sur un identifiant commun, soit sur leurs coordonnées géographiques. L'Insee va poursuivre ses travaux dans ce sens.

Le problème que ne résout pas ce référentiel est celui des ouvertures de magasins les dimanches et jours fériés. On compte une dizaine de jours fériés en France chaque année avec des autorisations législatives et une politique d'ouverture des enseignes différentes selon qu'il s'agit du 1<sup>er</sup> janvier, par exemple, jour pour lequel la presque totalité des points de vente sont fermés et le 11 novembre pour lequel ils sont en grande majorité ouverts. De même, les ouvertures le dimanche peuvent varier selon le mois de l'année et la saison. Certains hypermarchés du littoral ouvrent les dimanches de juillet et août ; une grande partie des magasins ouvrent les dimanches qui précèdent Noël.

Une étude doit encore être finalisée sur ce point afin d'identifier le mode de suivi exact : fréquence, périodes examinées, seuils, ... Un contrôle aussi fin n'existe pas actuellement dans la chaîne de calcul de l'indice des prix à la consommation.

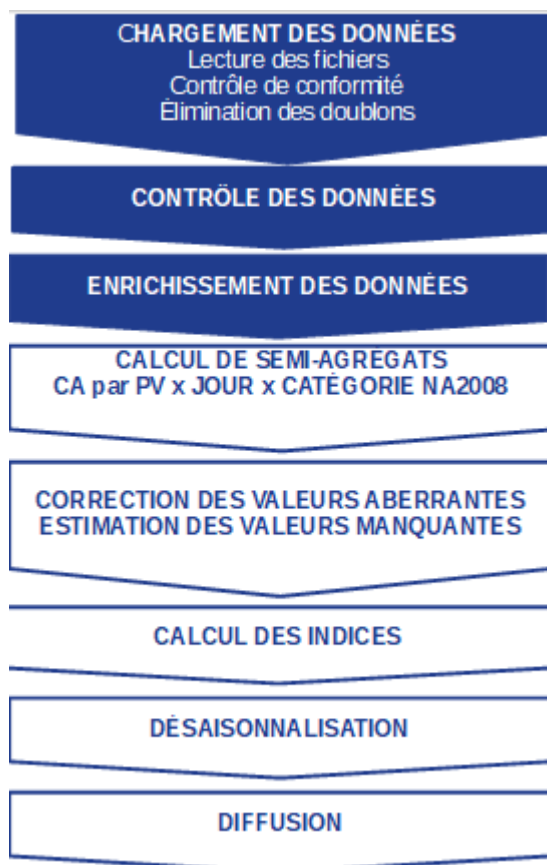
### **3.3 – La mutualisation des processus**

Plusieurs raisons poussent à mutualiser la partie réception / contrôle / enrichissement des données de caisse entre les deux chaînes de production – celle de l'IPC qui existe déjà et celle des ICA qui est à construire :

- les données de caisse sont volumineuses et les stocker à plusieurs endroits est coûteux, de même que les traiter plusieurs fois ;
- ces données sont sensibles et soumises à des règles strictes de conservation et de destruction. Leur stockage en un lieu unique est un garant de la bonne exécution de ces règles ;
- les améliorations dans le suivi de la qualité des données sera également bénéfique à l'indice des prix à la consommation ;
- l'ajout d'informations telles le classement des produits dans une nomenclature de production (en plus de la nomenclature en fonctions de consommation utilisée pour l'IPC), et l'attribution d'un identifiant Sirene aux magasins apporterait une information intéressante pour l'indice des prix à la consommation.

Pour la chaîne des traitements des indicateurs de chiffres d'affaires, le processus pourrait ressembler à celui dessiné ci-dessous. Les traitements sur fond sombre sont ceux communs avec la production de l'IPC ; les opérations sur fond clair sont propres à la chaîne de calcul des indices de chiffres d'affaires. Les traitements propres au calcul des indices de chiffres d'affaires débutent par le calcul de semi-agrégats : chiffres d'affaires quotidiens par magasin et catégorie NA2008.

Figure 3 : esquisse du futur processus



## 4 – LES APPORTS DES DONNÉES DE CAISSE

### 4.1 – Précocité des indicateurs

Les données de caisse sont pour la grande majorité des ventes reçues deux jours après la date des ventes considérée ; les données reçues très tardivement ne représentent qu'une infime part du chiffre d'affaires. Ainsi en fin de mois, 96 % du chiffre d'affaires est reçu en moyenne quand le minimum s'élève déjà à 94 % des données reçues (cf. tableau 1) ; deux jours après la fin du mois, il ne manque que 0,13 % du chiffre d'affaires en moyenne ; 0,02 % cinq jours après la fin du mois et 0,006 % dix jours après la fin du mois.

Tableau 1 : part du CA mensuel reçu à J+...de la fin du mois

	0 jours	2 jours	5 jours	10 jours
min	94,1344	98,0589	99,7160	99,9717
max	98,2262	99,9992	100,0000	100,0000
moyenne	96,0127	99,8712	99,9793	99,9936

Les délais par mois sont présentés dans la pièce jointe n°7.

Ces délais permettent d'envisager une publication avancée en milieu de mois suivant, soit nettement plus rapidement qu'aujourd'hui, et une estimation révisée en fin de mois ou le mois d'après au moment de la publication des indices du mois suivant fondés sur les données TVA. Le gain potentiel en termes de délai apparaît encore plus pertinent à la lumière du besoin d'informations qu'a fait ressurgir la crise sanitaire depuis le début de l'année 2020.

## 4.2 – Une meilleure couverture

Les calculs des indices sont actuellement fondés sur le recueil de données auprès d'un échantillon d'entreprises. Le recours aux données de caisse permettrait une couverture exhaustive qui ouvre la possibilité de produire des indices détaillés par produit (notamment pour des besoins interne) ou par région.

L'abandon de l'enquête éliminerait également les problèmes liés au renouvellement annuel de l'échantillon, en particulier les risques de mauvaise compréhension des agrégats attendus même si ces risques sont limités grâce à l'accompagnement des entreprises nouvellement incluses par le réseau d'enquêteurs de l'Insee. Ce phénomène a déjà été constaté, même pour des entreprises pérennes dans l'échantillon, lors d'un changement de correspondant. Avec les données de caisse, l'Insee maîtriserait, de manière uniforme pour toutes les entreprises, la constitution des agrégats.

Les données de caisse des départements ultramarins ne sont encore que rarement transmises à l'Insee. Les deux divisions des prix à la consommation et Indicateurs conjoncturels d'activité sont intéressées par ces données puisque la couverture géographique des deux indicateurs est la même (France métropolitaine et départements d'Outremer). Elles ont initié des contacts dans ce sens.

De même, pour les données de vente des carburants qui ne figurent pas dans les données de caisse, une source alternative est recherchée (source macroéconomique du Comité professionnel du pétrole utilisée par exemple par les comptes trimestriels ou données issues des paiements par carte bancaire, qui font l'objet d'un projet parallèle).

## 4.3 – Des indices par produits

L'objectif est de produire des indices au niveau A129 de la NA2008<sup>13</sup> pour les besoins internes (Cf. pièce jointe n°8), soit environ 50 indices puisque les produits vendus dans les grandes surfaces alimentaires ne couvrent pas l'ensemble des catégories, notamment sur les services. La nomenclature et le niveau retenu résultent d'un compromis entre les besoins utilisateurs, en particulier les comptables nationaux, et les possibilités de classement automatique. En effet, le référentiel de produits acheté par la division des prix à la consommation pour le calcul de l'IPC ne couvre pas l'ensemble des produits vendus dans les grandes surfaces. Les produits bien décrits par ce référentiel sont les produits industriels de grande consommation (produits alimentaires industriels, produits d'entretien de la maison et produits d'hygiène-beauté). N'y figurent pas les produits frais en vrac ou à la découpe (légumes, fruits, viande, poisson, fromages, ...) qui représentent une part non négligeable du chiffre d'affaires. L'enjeu est donc, pour produire des indices par grandes catégories, d'être capable de classer tous les produits dans la NA2008 à partir de leur libellé disponible dans les données de caisse. Pour cela, des outils s'appuyant sur l'analyse textuelle et des algorithmes de « machine learning » sont en cours d'expérimentation ; les premiers résultats sont brièvement présentés ci-après. La liste définitive des niveaux donnant lieu à publication sera établie en tenant compte du respect du secret statistique tel que défini par l'Insee – à savoir des indices calculés à partir des données d'au moins trois entreprises et dont aucune d'entre elles ne représentent plus de 85 % du chiffre d'affaires – et de leur intérêt en termes d'information conjoncturelle. En ce sens, les niveaux publiés seront plus agrégés que les indices utilisés pour des besoins internes comme les comptes nationaux. Les

---

13 Cette nomenclature agrégée permet de répondre aux besoins de la production de données de synthèse pour l'analyse économique et la diffusion et de faciliter les comparaisons internationales.

niveaux trop fins en termes de chiffre d'affaires et/ou de précision des indices ou ne respectant pas le secret statistique seront regroupés.

L'algorithme retenu est le module supervisé<sup>14</sup> de la librairie fastText sous python (<https://fasttext.cc/>)<sup>15</sup>. FastText est un réseau de neurones optimisé pour être entraîné en utilisant peu de ressources informatiques ; Bojanowski et al. (2016)<sup>16</sup> ont montré que sur un même jeu de données, l'entraînement d'un modèle fastText se fait en une dizaine de minutes contre plusieurs jours pour d'autres modèles pour des performances comparables. Ce point est important car la volumétrie des données de caisse est grande et les traitements quotidiens. Ce modèle de « deep learning » prend en entrée des unités de vocabulaire définies de façon arbitraire (tokens). Au niveau le plus simple, cela peut être un mot mais, selon l'application utilisée, un token peut être un groupe (n-grams) de mots, un groupe de caractères, ou simplement, une lettre ou un chiffre<sup>17</sup>. L'utilisation des n-grams de caractères est très utile pour développer un algorithme robuste aux fautes d'orthographe, aux fautes de frappes et aux abréviations qui sont très courantes dans les données de caisse. Pour plus d'informations concernant le traitement naturel du langage, le lecteur peut se référer à Jurafsky and Martin (2019)<sup>18</sup>.

Chaque libellé passe par une phase de pré-traitement constituée d'une quarantaine de règles déterminées après expertise des données et qui sont trop précises pour être exposées en détails ici. On peut citer par exemple :

- la suppression des chaînes de caractères parasites telles que « INCONNU », « FIELD COLLECTED », « IMAGE RECEIVED » ou bien de résidus de codes comme « &AMP » ;
- la suppression des accents et la mise en lettres capitales ;
- le remplacement des unités de poids par un symbole #POIDS ;
- le remplacement des tailles par un symbole #TAILLE ;
- le remplacement des unités de volume par #VOLUME ;
- la suppression des chiffres isolés (certains produits ont un EAN a la place de la description) ;
- la suppression des libellés vides.

Le but de ce processus est d'une part d'uniformiser les libellés, et d'autre part de supprimer les informations faiblement pertinentes mais coûteuses en mémoire. En particulier, l'algorithme va enregistrer autant de tokens que de chiffres dans l'échantillon d'apprentissage, ce qui augmente considérablement la taille du dictionnaire. Ainsi par exemple :

- côte de bœuf 250 g deviendra COTE DE BŒUF #POIDS ;
- t-shirt homme M deviendra T-SHIRT HOMME #TAILLE ;
- biscottes 12 x 3 deviendra BISCOTTES #LOT.

---

14 Une tâche d'apprentissage supervisée fait référence à une tâche pour laquelle la bonne réponse est connue (ici, la catégorie correcte de chaque libellé de produit est connue), en opposition à une tâche dite non-supervisée pour laquelle il n'y a pas de « bonne » réponse prédéfinie.

15 Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 427-431, Valencia, Spain. Association for Computational Linguistics.

16 Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2016). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5.

17 Un n-gram de mot est un token composé de n mots. Ainsi, le libelle BABA AU RHUM est composé des bi-grams BABA AU et AU RHUM. Un n-gram de caractères est un token composé de n caractères.

18 Jurafsky, D. and Martin, J. H. (2019). Speech and Language Processing (3rd ed. Draft). Prentice Hall PTR, USA, 3rd edition.



Pour entraîner le modèle et mesurer son efficacité, il est nécessaire de disposer de données préalablement classées. Le référentiel d'articles, qui contient plus de 8 millions de produits classés en plus de 700 grandes familles de consommation, a permis, de classer, relativement rapidement, une partie des données de caisse dans la nomenclature NA2008 en affectant directement une catégorie NA2008 à l'ensemble des produits d'une famille du référentiel. Cependant, ce référentiel ne contient pas la totalité des produits commercialisés dans la grande distribution alimentaire. De plus, le classement de l'ensemble des familles du référentiel dans la nomenclature NA2008 n'a pas été possible en raison de familles très hétéroclites au regard de la NA2008 et dont il faudrait classer les articles un par un.

Les articles ont ensuite été répartis aléatoirement entre l'échantillon d'entraînement du modèle et l'échantillon de test qui permet de comparer la catégorie prédite par le modèle avec celle attribuée grâce au référentiel. Au sein de chaque famille, 80 % des articles ont été assignés à l'échantillon d'entraînement.

Il est ressorti de ce test que l'algorithme est performant, dans la mesure où, pour 98 % des articles, il prédit la bonne catégorie. À noter que les catégories où l'algorithme est le moins efficace sont aussi les catégories contenant le moins d'articles (e.g. production de films, produits à base de tabac). Les résultats détaillés sont présentés dans la pièce jointe n°9. Ce taux de réussite porte uniquement sur les produits présents dans le référentiel.

Les travaux sont en cours d'extension à l'ensemble du champ, c'est-à-dire à la fois les données sur lesquelles a été mis en place l'apprentissage et les données non incluses dans le référentiel (et donc non classées a priori). À défaut de référentiel permettant de calculer une précision de l'algorithme sur ce champ, une vérification ex-post sur un échantillon significatif a été réalisée. Il en ressort un taux d'articles bien classés de l'ordre de 85 % et qui devrait être bien supérieur en part de CA (calcul en cours). Sur la partie alimentaire, le taux de bon classement (dans une dizaine de classes cibles : légumes frais, poisson, viande, produits laitiers, ...) est quant à lui compris entre 95 % et 100 %.

Ce résultat devrait par ailleurs être nettement amélioré dans les mois qui viennent. En effet, une partie des mauvais classements trouve son origine dans des produits mal rangés dans le référentiel et, par transition, mal classés dans la nomenclature cible, et des produits pour lesquels la correspondance entre la nomenclature du référentiel et la nomenclature cible est erronée (bien souvent car les produits d'une famille du référentiel n'appartiennent pas tous à la même catégorie de la nomenclature cible). Des travaux de vérification du référentiel sont en cours, ce qui devrait permettre de réduire très fortement cette source de divergence. Un taux de bons classement cible de l'ordre de 95 % paraît atteignable.

L'amélioration du classement passe donc à présent par une meilleure classification des articles de l'échantillon d'entraînement (correction d'erreurs de classement par des experts en nomenclatures, meilleure classification des articles appartenant à des familles du référentiel hétéroclites entre autres, amélioration de la classification des articles hors référentiel). Ce travail sera mené progressivement avant le passage en production du nouveau dispositif puis en continu. Cette amélioration est nécessaire pour déterminer la liste des indices qui seront produits puisqu'elle s'appuiera à la fois sur le poids de la catégorie dans l'ensemble des ventes et sur la précision du classement par catégorie. Des regroupements seront réalisés pour les catégories trop petites ou dont la précision sera trop faible.

## **4.4 – Des indices par région**

Sans être dans les priorités du projet, la possibilité de pouvoir segmenter des indicateurs d'activité à un niveau infra-national a assez vite été soulevée. De nouveau, la crise actuelle, avec son impact parfois hétérogène selon les régions, rehausse l'intérêt de ce sujet.

Sur ce point, un référentiel des points de vente, acheté ou issu du répertoire des entreprises géré par l'Insee, est l'élément clé puisqu'il permet d'affecter une région à chacun des points de vente. Un travail d'appariement entre ce référentiel d'un côté, et le référentiel statistique de l'Insee de l'autre côté, sera mené durant l'année, en s'appuyant notamment sur la géolocalisation des établissements.

Les enseignes n'étant pas également réparties sur le territoire, la publication d'indices régionaux devra régulièrement prendre en compte les modifications dans les parts de marché régionales pour respecter le secret statistique. Le secret statistique qui sera appliqué est celui défini par l'Insee : au moins trois entreprises dont aucune ne couvre plus de 85 % du chiffre d'affaires de la région. Il ne sera d'ailleurs a priori pas possible pour cette raison de publier des statistiques locales avec un même niveau de finesse sur les produits (cf. section précédente).

## 4.5 – Réduction de la charge pour les entreprises et l'Insee

Même si la charge mensuelle pour les entreprises est faible (estimée à 20 minutes chaque mois – temps de collecte des informations, de connexion au site et de remplissage du questionnaire), elle contribue tout de même à alourdir la charge statistique globale pesant sur les entreprises que l'Insee essaie de réduire. À noter que pour les entreprises des groupements qui se trouvent dans la partie renouvelable de l'échantillon, le coût d'entrée est significatif (questionnaire portant sur les 12 derniers mois) et doit être ajouté à la charge mensuelle.

La charge de collecte – apurement représente pour l'Insee environ 1 ETP<sup>19</sup>. Dans ces conditions, il n'est pas certain de pouvoir la réduire fortement même en cherchant à mutualiser au maximum les processus avec la division des prix à la consommation. L'idée est plutôt de « faire mieux » – en produisant des indices plus détaillés – à coûts constants.

# 5 – LES ASPECTS TECHNIQUES

---

Dans le but d'alléger la charge des entreprises concernées et sa propre charge, l'Insee souhaite mutualiser la réception et le stockage des données de caisse utilisées par la division des prix à la consommation et la division Indicateurs conjoncturels d'activité. Les points ci-dessous reprennent donc les éléments de l'arrêté du 13 avril 2017 rendant obligatoire la transmission de données par voie électronique à des fins de statistique publique (cf. pièce jointe n°10).

## 5.1 – Les informations demandées

Ces informations sont demandées à toutes personnes morales dont un des établissements a une activité de commerce de détail en magasin non spécialisé à prédominance alimentaire de plus de 400 m<sup>2</sup>.

Les informations nécessaires au calcul des indicateurs de chiffres d'affaires et au contrôle de ce calcul sont, pour chaque article vendu au cours d'un jour donné dans chacun des points de vente des personnes morales désignées ci-dessus :

- le chiffre d'affaires (ou à défaut le prix de vente et la quantité d'articles vendus) par date de vente, point de vente et produit vendu ;
- la date de la vente ;
- a minima un identifiant du point de vente pouvant être relié à des informations sur le point de vente (localisation et type) via un référentiel acheté par l'Insee (celui de la société IRI actuellement). L'Insee lui affectera un identifiant (SIRET) de son répertoire Sirene ;
- l'identifiant du produit (l'ean) ;
- le ou les libellés descriptifs de l'article tels qu'ils apparaissent dans les données de caisse.

---

19 Équivalent Temps Plein

Pour des raisons pratiques liées à la mutualisation des données de caisse entre les services des prix à la consommation et des indicateurs conjoncturels d'activité, ce dernier souhaite un accès complet aux données de caisse fournis à l'Insee ; ce qui signifie un accès également aux :

- nombres d'articles vendus au cours de la journée et/ou le prix unitaire de l'article pour cette journée ;
- codes de classement des articles dans la nomenclature utilisée par l'enseigne.

Par ailleurs, devront être transmises, à un rythme au moins annuel et aussi souvent que nécessaire, les données suivantes permettant la lecture des données quotidiennes : dessin et format des fichiers sous dénomination normalisée, répertoire des points de vente concernés.

## **5.2 – Les conditions de transmission**

Les données sont transmises électroniquement sous forme cryptée, et via un mode de transmission du type ftpssl. Elles peuvent être transmises via un intermédiaire désigné par l'Insee.

## **5.3 – La fréquence et les délais de transmission**

Les données décrites dans l'annexe sont transmises quotidiennement, de préférence le lendemain des ventes. Elles devront avoir été transmises dans un délai de deux jours pour au moins 90 % d'entre elles, dans un délai d'au plus un mois pour le reliquat.

## **5.4 – Les espaces de conservation des données**

Les données transmises par les personnes morales désignées au paragraphe 5.1 sont stockées dans un espace électronique sécurisé créé sur le réseau électronique de l'Insee.

## **5.5 – Les personnes autorisées à y accéder**

Une décision du directeur général de l'Insee désigne les personnes habilitées à accéder à cet espace électronique.

La division Indicateurs conjoncturels d'activité souhaite que les responsables de la production des indices de chiffres d'affaires dans le commerce de détail, à la direction générale comme dans les directions régionales (Pays de la Loire actuellement), et les informaticiens responsables des applications de production de ces indicateurs soient autorisés à accéder aux données de caisse.

## **5.6 – La durée de conservation**

Les données transmises seront détruites au plus tard le 31 décembre de la troisième année suivante la collecte. Seules pourront être conservées, au-delà de cette date, des données agrégées de chiffres d'affaires ayant servi à la production des indices de chiffres d'affaires qui garantissent l'anonymat des personnes morales de droit privé ayant participé à l'enquête.

# CONCLUSION

---

L'Insee calcule et publie chaque mois (+30 jours après la fin du mois) des indices de chiffre d'affaires des grandes surfaces alimentaires à partir de l'Enquête mensuelle sur l'activité des grandes surfaces alimentaires (Emagsa) réalisée par internet auprès d'un échantillon d'entreprises qui gèrent des supermarchés et/ou des hypermarchés. L'enquête répond à une forte demande française d'analyse conjoncturelle et européenne que ne peuvent complètement satisfaire les sources fiscales en raison de leur disponibilité tardive. L'utilisation des données de caisse, source alternative déjà présente à l'Insee pour le calcul de l'indice des prix à la consommation, permettrait de réduire les délais de publication en publiant les premières estimations au milieu du mois suivant le mois d'affaires ; d'alléger la charge statistique pesant sur les entreprises en supprimant l'enquête Emagsa ; d'améliorer la qualité des indicateurs en produisant des indices par produits grâce à l'analyse textuelle des libellés des données de caisse et par région, à moyens constants pour l'Insee.

Les premières comparaisons réalisées sur un échantillon de données provenant d'enseignes qui ont accepté de participer à cette expérimentation montre des évolutions cohérentes, malgré les différences de concepts entre les données de caisse et les sources auxquelles elles sont comparées. Ces évolutions sont suffisamment proches pour conduire l'Insee à étudier plus en détail le processus de remplacement de l'enquête Emagsa par les données de caisse.

Le processus de calcul des indices de chiffres d'affaires serait mutualisé avec celui de calcul de l'indice des prix à la consommation pour la partie réception / contrôle / enrichissement des données de manière à limiter la volumétrie et garantir une bonne conservation des données. Des informations et contrôles seraient ajoutés à ce système d'information pour les besoins du calcul des indices de chiffres d'affaires (SIRET des magasins, classement des produits en nomenclature de production, détection des valeurs manquantes via le parc des magasins) rendant cette mutualisation bénéfique pour les deux acteurs.

Les données auxquelles la division Indicateurs conjoncturels d'activité demande l'accès sont exactement celles actuellement transmises à l'Insee pour le calcul de l'indice des prix à la consommation. Les conditions de transmission et de conservation seront identiques à celles définies dans l'arrêté propre à l'indice des prix à la consommation. Seule la liste des personnes autorisées à traiter ces données sera élargie aux statisticiens et informaticiens chargés du calcul des indices de chiffres d'affaires.

Les données de caisse sont la source sur laquelle le projet est le plus avancé en grande partie grâce aux contacts noués depuis une dizaine d'années avec les propriétaires des données et la bonne connaissance qu'a pu acquérir l'Insee de ces données. Le projet prévoit également l'exploitation d'une autre source prometteuse, celles des données de paiement par carte bancaire. Cette source – disponible au plus tard le 15 du mois suivant le mois d'affaires – permettrait de construire des indices d'évolution de l'activité (approximée par la valeur des transactions CB par établissement) dans différents sous-secteurs où la part des paiements par carte bancaire est importante : notamment le commerce de détail mais aussi les secteurs des services à la personne, arts & spectacles, hébergement & restauration et transports. Des contacts ont été établis avec les propriétaires des données et, dans les faits, ces données sont déjà utilisées depuis la crise sanitaire pour l'estimation à +30 jours sur la partie non alimentaire du commerce de détail. D'autres pistes avec d'autres fournisseurs pourront le cas échéant être explorées, chaque source ayant ses avantages et ses inconvénients. Toutes ces sources ont pour but de se compléter pour améliorer la qualité ou servir de point de comparaison. Elles ne peuvent pas remplacer néanmoins en l'état la source TVA qui par son exhaustivité permet de publier des indicateurs de qualité sur un champ très étendu, mais de façon plus tardive.

# LISTE DES PIÈCES JOINTES

---

- Pièce jointe n°1 : Décret n°2017-463 du 31 mars 2017 portant application de l'article 3 bis de la loi du 7 juin 1951 relative à l'obligation, la coordination et le secret en matière de statistique
- Pièce jointe n°2a : Enquête mensuelle sur l'activité des grandes surfaces alimentaires (emagsa) – Enquête de recrutement – Indications pour remplir le questionnaire initial
- Pièce jointe n°2b : Enquête mensuelle sur l'activité des grandes surfaces alimentaires (EMAGSA) – Questionnaire initial
- Pièce jointe n°3a : Enquête mensuelle sur l'activité des grandes surfaces alimentaires (EMAGSA) – Indications pour remplir les questionnaires mensuels
- Pièce jointe n°4 : Insee Informations Rapides N°2020-334
- Pièce jointe n°5 : Comparaison des sources fiscale (TVA), Emagsa et données de caisse
- Pièce jointe n°6 : Poids des doublons dans le chiffre d'affaires
- Pièce jointe n°7 : Délais de réception des données de caisse
- Pièce jointe n°8 : Nomenclature NA2008
- Pièce jointe n°9 : Précision de l'algorithme de classement des produits
- Pièce jointe n°10 : Arrêté du 13 avril 2017 rendant obligatoire la transmission de données par voie électronique à des fins de statistique publique