LE KIT DE DIFFUSION ESANE OU COMMENT DIFFUSER DES AGRÉGATS À DES NIVEAUX FINS TOUT EN ALERTANT L'UTILISATEUR SUR LA PERTINENCE DES CHIFFRES PUBLIÉS

Julien SENG (*)

(*) Insee, Direction des statistiques d'entreprises

Introduction

La réingénierie des statistiques sectorielles annuelles d'entreprises (le programme RESANE [1]), débutée en 2006, a abouti au système ESANE (Élaboration des Statistiques Annuelles d'Entreprises). Ce nouveau système, qui en 2012 en est à sa troisième campagne de production, met en jeu à la fois des données administratives (déclarations fiscales et sociales) et des données d'une enquête : l'ESA (Enquête Sectorielle Annuelle) pour les entreprises des secteurs non-industriels et des IAA (Industries Agricoles et Alimentaires), et l'EAP (Enquête Annuelle de Production) pour les entreprises du secteur industriel manufacturier.

La diffusion des résultats de la statistique structurelle d'entreprise se traduit par différents types de publications (chiffres clés, données détaillées, fiches sectorielles et Alisse[1]) sur le site insee.fr. Ces dernières peuvent se résumer à des agrégats au niveau sous-classe (quatre chiffres et une lettre) de la nomenclature d'activité française (Naf) et des agrégats au niveau groupe (trois chiffres) croisés avec des tranches d'effectifs (sur quatre tranches). C'est sur ces niveaux d'agrégation de la nomenclature que l'Insee s'est engagé en termes de qualité, et c'est donc sur ces niveaux qu'ont été optimisés les contrôles qualité opérés par les gestionnaires du centre de Nantes [2].

La méthodologie utilisée pour le calcul de ces agrégats est complexe. Il s'agit de la mise en œuvre d'un estimateur dit « composite » [3] qui combine les données de l'échantillon de l'enquête avec des données exhaustives provenant des liasses fiscales. Il donne des résultats différents de ceux obtenus par simple sommation des valeurs d'un fichier de données individuelles (ce qui était le cas jusqu'à présent pour les chiffres diffusés en statistique d'entreprises).

En outre, comme des utilisateurs internes ou externes au système statistique public souhaitent souvent obtenir des agrégats à des niveaux de nomenclature plus fins que ceux publiés en standard, il a été décidé de mettre à leur disposition un kit de diffusion. Ce kit, développé en SAS, utilise la méthodologie d'ESANE pour fournir des agrégats à la demande. Ainsi, en utilisant ce kit, l'utilisateur pourra retrouver les chiffres diffusés sur le site insee.fr et il pourra en calculer d'autres à des niveaux plus fins. Cependant, la significativité d'un résultat ne peut pas être assurée dans tous les cas car les calculs sont basés sur un échantillon. C'est pourquoi le kit accompagne chaque résultat d'indicateurs de qualité. Ceci permettra à l'utilisateur de juger de la fiabilité du chiffre qu'il a calculé.

Après avoir présenté les différentes fonctionnalités du kit et les choix méthodologiques qui ont été nécessaires pour sa mise en œuvre, les différents indicateurs qualité seront détaillés avec des exemples d'utilisation.

1. Fonctionnement et possibilités du kit

Le kit de diffusion permet de calculer des résultats plus fins que ceux publiés en standard. Pour cela, l'utilisateur du kit doit renseigner pas moins de cinq paramètres nécessaires à l'élaboration d'un résultat. Ces paramètres sont définis à l'aide d'une IHM (Interface Homme Machine) développée en SAS qui fait défiler différentes fenêtres présentant les critères à remplir (cf. Annexe 1).

[1] Alisse : Accès en Ligne aux Statistiques Structurelles d'Entreprises.

1.1. La population

Le premier paramètre à renseigner est la population sur laquelle les calculs vont s'effectuer. L'utilisateur a la possibilité de définir une liste de sections, divisions, groupes, classes ou sous-classes de la nomenclature d'activité française et/ou d'inscrire le nom d'une variable qui spécifie la population qu'il souhaite. Cette variable doit être calculée en amont de l'utilisation du kit et est à la charge de l'utilisateur. Il s'agit concrètement d'une indicatrice calculée pour l'ensemble des entreprises du champ d'ESANE et qui permet d'isoler les entreprises que l'on souhaite étudier (par exemple toutes les entreprises d'une catégorie juridique donnée).

Dans la suite de l'article, nous considérons l'exemple d'utilisation du kit suivant.
L'utilisateur cherche à obtenir les agrégats du chiffre d'affaires et des dépenses de carburant des secteurs des transporteurs terrestres de voyageurs ventilés par secteur d'activité (au niveau sous-classes de la NAF) et tranche de chiffre d'affaires.
Cet exemple servira de fil rouge pour illustrer chaque étape de la mise en œuvre du kit.
Dans le cadre de cet exemple, la population à renseigner est donc la liste de sous-classes suivante : 4910Z, 4931Z, 4932Z et 4939A.

1.2. Le domaine de diffusion

Après la population sur laquelle va reposer le calcul des agrégats, il convient de définir les domaines de diffusion. L'utilisateur peut choisir jusqu'à trois critères pour former les domaines de diffusion souhaités. Parmi ces critères figurent :

-Le type d'entreprise (société non financière ou entreprise individuelle)
-La catégorie juridique (sur 4 chiffres et en 258 modalités)
-Des tranches d'effectifs (en 4 modalités)
-Des tranches de chiffres d'affaires (en 8 modalités)
-La catégorie d'entreprise (Micro-entreprise, très petite, petite ou moyenne entreprise, entreprise de taille intermédiaire, grande entreprise et autre) [1]
-La région (en 27 modalités)
-Tous les niveaux de la Naf (Section, division, groupe, classe et sous-classe).

En plus de ces critères, l'utilisateur a la liberté de choisir un critère de son choix qui ne figure pas dans la liste ci-dessus mais qui existe dans la table de données individuelles (par exemple le code département).

Il peut également créer lui même une nouvelle variable pour définir ce domaine de diffusion (par exemple un regroupement de catégories juridiques). Cette création de variable se fait en amont du kit et reste à la charge de l'utilisateur à l'instar de ce qu'il est possible de faire pour la définition de la population décrite dans le paragraphe précédent.

[1] En attendant la prise en compte des entreprises profilées dans les statistiques d'ESANE, la catégorie d'entreprise est calculée à partir de l'effectif occupé, du chiffre d'affaires et du total du bilan des unités légales.

Dans le cadre de l'exemple présenté dans la partie précédente, l'utilisateur peut donc choisir les tranches de chiffre d'affaires et la sous-classe de la Naf comme critères de définition des domaines de diffusion. L'utilisateur obtiendra ainsi le tableau de résultats suivant :

Tranche de chiffre d'affaires	Secteur	Chiffre d'affaires	Dépense en carburant
Total[1]	4910Z		
De 0 k€ à 999 k€	4910Z		
De 1 000 k€ à 1 999 k€	4910Z		
De 2 000 k€ à 4 999 k€	4910Z		
De 5 000 k€ à 9 999 k€	4910Z		
De 10 000 k€ à 19 999 k€	4910Z		
De 20 000 k€ à 49 999 k€	4910Z		
De 50 000 k€ à 199 999 k€	4910Z		
Plus de 200 000 k€	4910Z		
Total[1]	4931Z		
De 0 k€ à 999 k€	4931Z		
…	…		

1.3. Le type de statistique

L'utilisateur choisit par la suite le type de statistique souhaitée :

- Des agrégats en niveau (en secteur et en branche),
- Des agrégats en évolution de l'année N sur N-1 (en secteur et en branche),
- Des distributions (centiles, indices),
- Des ratios en niveau et évolution.

Cependant pour la première année de production du kit, seuls les agrégats en niveau sont disponibles.

Dans notre exemple, les agrégats en niveau sont alors sélectionnés.

1.4. Les variables d'intérêt

L'utilisateur précise ensuite la liste des variables dont il souhaite les résultats. Pour des agrégats en niveau, l'utilisateur peut demander :

- soit des agrégats pour des variables fiscales (provenant des liasses fiscales) ou/et des variables de l'enquête sectorielle,
- soit demander des ventilations (en branches du chiffre d'affaires, en formes de ventes, en produits, etc).

Dans le cadre de l'exemple, il s'agit donc du chiffre d'affaires et des dépenses en carburant.

1.5. La sortie

Enfin, il est proposé à l'utilisateur une sortie en Excel, Html ou/et sas. Pour les deux premières sorties, un rappel des paramètres choisis est placé en en-tête du fichier.

La sortie correspondant à l'exemple présenté tout au long de cette partie figure en annexe 2.

[1] Ce total fait partie des chiffres standards publiés sur le site insee.fr

2. Les choix méthodologiques

Les entreprises du système ESANE sont réparties dans plusieurs champs [1]. Pour le calcul des agrégats, deux types de champs différents peuvent être distingués :
- le champ 1 qui correspond au champ de l'ESA. Sur ce champ, une estimation de l'ensemble des caractéristiques fiscales et d'enquêtes est possible car les réponses à l'enquête sont exploitables. En outre, sur ce champ il existe des contrôles et redressements supplémentaires qui sont réalisés sur les entreprises répondantes à l'échantillon de l'enquête (cf. [2]). D'un point de vue statistique, il convient donc d'inférer les corrections faites sur l'échantillon à l'ensemble du champ de l'enquête. C'est pourquoi cette population est considérée à part dans le calcul des agrégats.
- le champ dit « 2 à 4 » sur lequel seules les caractéristiques fiscales sont mobilisables ; aucun agrégat concernant des variables de l'enquête statistique n'est donc disponible sur ces entreprises.

A partir de cet ensemble de données, mêlant à la fois des données administratives (les liasses fiscales) et les données d'enquête (les questionnaires de l'ESA), le système ESANE fait appel à l'estimateur composite.

2.1. L'estimateur composite

Cet estimateur a été développé par l'UMS-E (Unité de Méthodologie Statistique - Entreprise) :

$$\hat{Y}_{Diff}^{X} = \sum_{i \in \text{exhaustif hors DOM}} Y_i^{Redi} \, 1\!\!1_{i \in champ} \, 1\!\!1_{APEenq=X}(i) +$$

$$\sum_{i \in \text{champ Esane 1 au lancement hors DOM et exhaustif couvert par } R \oplus HC} Y_i^{IEG} \, 1\!\!1_{APErep=X}(i) + \sum_{i \in R \oplus HC \text{ (hors DOM)}} w_i [\, Y_i^{Redi} \, 1\!\!1_{APEenq=X}(i) \, 1\!\!1_{i \in \text{champ Esane1}} - Y_i^{IEG} 1_{APErep=X}(i)]$$

$$+ \sum_{i \in \text{champ Esane 1 au lancement hors DOM et exhaustif non couvert par } R \oplus HC} Y_i^{IEG} \, 1\!\!1_{APErep=X}(i)$$

$$+$$

$$\sum_{i \in R \oplus HC \text{ (hors DOM)}} w_i \, Y_i^{Redi} \, 1\!\!1_{ape\_enq=X}(i) \, 1\!\!1_{i \in \text{champ Esane 2 à 4}} + \sum_{i \in \text{champ Esane 2 à 4}} Y_i^{IEG} \, 1\!\!1_{APErep=X}(i) + \sum_{i \in \text{champ ESANE 1 DOM}} Y_i^{IEG} \, 1\!\!1_{APErep=X}(i)$$

Avec :

- $\hat{Y}_{Diff}^{X}$: l'estimation de la variable Y avec l'estimateur composite par différence sur le secteur X,
- $\sum_{i \in \text{exhaustif hors DOM}} Y_i^{Redi} \, 1\!\!1_{i \in champ} \, 1\!\!1_{APEenq=X}(i)$: somme, pour toutes les unités i appartenant à la partie exhaustive du champ 1 hors dom et ayant pour ape à l'issue de l'enquête le secteur X, des valeurs REDI (REconciliation des Données Individuelles, cf. [2]).
- $R \oplus HC$: l'ensemble des unités répondantes et des unités hors champ (unités mises hors champ à l'issue de l'enquête pour cause de cessation ou d'ape hors champ et dont on dispose d'une liasse fiscale).
- $\sum_{i \in \text{champ Esane 1 au lancement hors DOM et exhaustif couvert par } R \oplus HC} Y_i^{IEG} \, 1\!\!1_{APErep=X}(i)$ somme, pour toutes les unités i appartenant à l'ensemble des unités issues du champ 1 hors DOM et exhaustif, couvrant l'ensemble $R \oplus HC$ et dont l'ape du répertoire Sirene vaut X, des valeurs IEG,
- $\sum_{i \in R \oplus HC \text{ (hors DOM)}} w_i \, Y_i^{Redi} \, 1\!\!1_{APEenq=X}(i) \, 1\!\!1_{i \in \text{champ Esane1}}$: somme pondérée, pour toutes les unités i appartenant à l'ensemble $R \oplus HC$ hors dom dont l'ape à l'issue de l'enquête vaut X, des valeurs REDI,

- $\sum_{i \in R \oplus HC \text{ (hors DOM)}} - w_i \; Y_i^{IEG} 1_{APErep=X}(i)$: somme pondérée, pour toutes les unités i appartenant à l'ensemble $R \oplus HC$ et dont l'ape issue du répertoire Sirene vaut X, des valeurs IEG (Informations Economiques Générales, ou valeur issue des liasses fiscales),

- $\sum_{i \in \text{champ Esane 1 au lancement hors DOM et exhaustif non couvert par } R \oplus HC} Y_i^{IEG} \; 1\!\!1_{APErep=X}(i)$: somme, pour toutes les unités i appartenant à l'ensemble des unités issues du champ 1 hors DOM et exhaustif, ne couvrant pas l'ensemble $R \oplus HC$ et dont l'ape du répertoire Sirene vaut X, des valeurs IEG,

- $\sum_{i \in R \oplus HC \text{ (hors DOM)}} w_i \; Y_i^{Redi} \; 1\!\!1_{ape\_enq=X}(i) \; 1\!\!1_{i \in \text{champ Esane 2 à 4}}$: somme pondérée pour toutes les unités i de $R \oplus HC$, qui à l'issue de l'enquête se retrouvent dans le champ 2 à 4 et dont l'ape à l'issue de l'enquête est X, des valeurs REDI,

- $\sum_{i \in \text{champ Esane 2 à 4}} Y_i^{IEG} \; 1\!\!1_{APErep=X}(i)$: somme pour toutes les unités appartenant au champ 2 à 4 et dont l'ape issue du répertoire Sirene vaut X, des valeurs IEG,

- $\sum_{i \in \text{champESANE1 DOM}} Y_i^{IEG} \; 1\!\!1_{APErep=X}(i)$: somme pour toutes les unités appartenant au champ 1 restreint aux unités provenant des DOM et dont l'ape issue du répertoire Sirene vaut X, des valeurs IEG.

La complexité de l'estimateur tient, entre autre, au fait que ce dernier mobilise plusieurs sommes qui se basent à chaque fois sur des populations bien différentes. Par exemple, la première sous-somme est sur la partie exhaustive du champ 1, l'avant dernière sur le champ 2 à 4 etc.
De plus, sur chaque somme, il y a tantôt utilisation de la valeur des liasses fiscales et tantôt utilisation de la valeur issue de REDI.
L'implémentation de cet estimateur est donc difficile dans le sens où il faut établir de nombreux filtres pour séparer les entreprises pour qu'elles soient comptées dans la bonne sous-somme.

L'intérêt du kit réside dans le fait qu'il prend en compte cet estimateur. Ce dernier serait extrêmement difficile à mettre en œuvre pour un utilisateur extérieur à ESANE car il nécessite :
- d'une part de connaître les différentes sources prises en compte
- d'autre part de connaître la méthodologie mise en œuvre pour le contrôle de ces sources
- et enfin de connaître les différentes méthodes mises en œuvre pour le calcul des différents agrégats. Ces méthodes divergent en effet selon le type de variables et le niveau d'agrégation.

2.2. Les variables fiscales

Les agrégats des variables fiscales sont calculés de deux manières différentes. Au niveau groupe de la Naf, l'estimateur composite est utilisé. Cependant, cet estimateur composite atteint sa limite pour des niveaux plus fins (par exemple au niveau sous-classe). Vu sa formule, il est en effet possible d'obtenir des valeurs négatives pour des agrégats concernant des caractéristiques strictement positives. L'exemple le plus frappant concerne un secteur pour lequel l'ensemble des entreprises de l'échantillon aurait changé de secteur à la suite de la réponse à l'enquête, comme l'estimateur compte en négatif des unités qui changent se secteur, le résultat obtenu peut s'avérer négatif.
Ces problèmes sont d'autant plus fréquents que le domaine de diffusion est « petit ». C'est pourquoi, si ce problème n'est pas visible au niveau des groupes de la NAF, il commence à se produire plus fréquemment dès le niveau sous-classe (cf. [3]).

Ainsi, il a été décidé d'utiliser une méthodologie différente pour le niveau infra groupe. La solution adoptée consiste à appliquer aux agrégats du champ 1, calculés en composite au niveau groupe, la structure Horvitz Thompson qui est observée dans l'enquête :

$$Y_{\text{infra\_groupe}} = \hat{Y}_{groupe}^{composite} \frac{\sum_{i \in S} wi Y_i 1\!\!1_{i \in \text{infra\_groupe}}}{\sum_{i \in S} wi Y_i 1\!\!1_{i \in groupe}}$$

avec :

- $Y_{\text{infra\_groupe}}$: Le total de la variable Y estimée au niveau infra groupe,
- $\hat{Y}^{composite}_{groupe}$: Le total de la variable Y calculée au niveau groupe avec l'estimateur composite sur tout le champ 1,
- $\sum_{i \in s} wiY_i 1\mathrm{I}_{i \in \text{infra\_groupe}}$: L'estimation de Y au niveau infra groupe de l'enquête où $1\mathrm{I}_{i \in \text{infra\_groupe}} = \begin{cases} 1 \text{ si l'unité i appartient à l'infra\_groupe concerné} \\ 0 \text{ sinon} \end{cases}$,
- $\sum_{i \in s} wiY_i 1\mathrm{I}_{i \in groupe}$: L'estimation de Y au niveau groupe de l'enquête où $1\mathrm{I}_{i \in groupe} = \begin{cases} 1 \text{ si l'unité i appartient au groupe de la NAF concerné} \\ 0 \text{ sinon} \end{cases}$
- S : l'ensemble des unités de l'échantillon de l'enquête (ESA + EAP),

L'intérêt du kit est qu'il prend bien en compte cette méthodologie et donc qu'il permet de retrouver les agrégats publiés en standard.
De plus, le kit possède intrinsèquement les parties $\hat{Y}^{composite}_{groupe}$ et $\sum_{i \in s} wiY_i 1\mathrm{I}_{i \in groupe}$ ce qui améliore grandement le temps de calcul.

2.3. Les variables sectorielles de l'ESA

Comme précisé précédemment, les données de ces variables ne sont disponibles qu'au niveau du champ 1 (c'est à dire le champ de l'ESA).

Pour ces variables, l'estimateur Horvitz Thompson est utilisé. Il y a par la suite un recalage sur certaines variables (par exemple le chiffre d'affaires) sur le champ 1.

Le kit calcule donc une estimation sur le champ 1 et non sur le champ total (champ1 + champ 2 à 4). Cela a donc pour conséquence pour l'utilisateur, dans certains cas, de ne pas retrouver, via le kit, les agrégats sectoriels publiés sur le site. En effet, sur le site insee.fr les agrégats sectoriels (y compris ceux de l'ESA), ont été estimés sur l'ensemble du champ ESANE.
Ainsi, parmi les variables de l'ESA figurent des ventilations du chiffre d'affaires. La somme des ventilations du chiffre d'affaires par secteur correspond à l'agrégat du chiffre d'affaires sur le champ 1 uniquement. Or pour ce même secteur, le chiffre d'affaires publié sur le site insee.fr correspond à celui des champs 1 à 4.
Ceci n'est toutefois pas très pénalisant dans la mesure où la distinction des champs 1 et « 2 à 4 » est souvent sectorielle. Ainsi, une sous-classe de la Naf est soit entièrement dans le champ 1 soit entièrement dans le champ « 2 à 4 ».

Par ailleurs, il y a un calage de l'échantillon de l'ESA sur le chiffre d'affaires diffusé au niveau groupe sur le champ 1. De ce fait, il y a égalité de l'estimateur Horvitz Thompson calculé au niveau groupe et de l'estimateur composite au niveau groupe.

2.4. Le secret

Seul le secret primaire[1] est géré dans le kit et ce de façon incomplète. En effet, le kit dispose uniquement des cases à mettre en secret primaire calculées par le logiciel Tau argus dans le cadre de la diffusion standard sur le site insee.fr pour les deux niveaux suivants :

[1] Cette règle interdit la diffusion d'un résultat s'il y a moins de trois unités et/ou qu'il existe une unité représentant plus de 85% du résultat à diffuser.

- le niveau sous-classe et supérieur (c'est-à-dire classe, groupe, division et section),
- le niveau groupe croisé par tranche de taille et supérieur (division et section croisé par tranche de taille),

et les 6 variables :

-Total des amortissements sur immobilisations amortissables, à la fin de l'exercice,
- Total de l'actif,
- Effectif moyen du personnel,
- Total général des immobilisations,
- Chiffre d'affaires hors taxes,
- Nombre d'entreprises.

Le calcul du secret se fait de la façon suivante dans le kit : pour pouvoir obtenir le secret sur d'autres variables que ces six variables ci-dessus aux deux niveaux de nomenclature évoqués ci-dessus, il est associé à chacune d'elles une variable de cette liste. Cette dernière sert donc à générer le secret pour toutes les variables disponibles dans le kit. Par exemple, le secret pour le chiffre d'affaires hors taxes servira pour la définition du secret sur le total des ventes de marchandises.

Sur d'autres niveaux que les deux niveaux décris ci-dessus, le calcul se fait de la façon suivante :

- Si la sous-classe X est en secret alors tout résultat plus fin sur cette sous-classe X sera également mis en secret. Par exemple, si le nombre d'entreprises pour la sous-classe 0220Z est en secret, alors toutes les cases de cette sous-classe croisée avec n'importe quelle région seront en secret.
- Sinon, si le nombre d'unités de la case est inférieur à trois alors elle est mise en secret.

Dans les autres cas, la case est diffusée.

3. L'indicateur qualité

L'atout majeur du kit est qu'il met à disposition pour chaque résultat plusieurs indicateurs qualité qui permettent d'éclairer l'utilisateur quant à la diffusabilité de tel ou tel résultat. Ces indicateurs sont :

- Le coefficient de variation de l'estimation de la variable au niveau groupe de la Naf,
- L'intervalle de confiance pour l'estimation de la variable au niveau groupe de la Naf,
- Le coefficient de variation de l'estimation de la variable sur le domaine où est calculée la statistique,
- Le nombre d'unités provenant de l'échantillon pris en compte pour le calcul de l'agrégat,
- Le nombre (et le pourcentage) de liasses fiscales imputées prises en compte pour le calcul de l'agrégat,
- Un code qualité indiquant la pertinence des données (« Diffusable » si la qualité des données est suffisante, « Attention » si le résultat est peu fiable ou « Non diffusable »).

Les variables sectorielles de l'ESA ne peuvent bénéficier du nombre (et du pourcentage) de liasses fiscales imputées puisque ces variables sectorielles ne sont disponibles que dans l'enquête.

3.1. Le coefficient de variation sur le domaine

Le coefficient de variation est fourni par l'UMS-E (Unité de Méthodologie Statistique - Entreprise) sur 21 variables au niveau groupe, division et section. Par exemple :

- Effectif salarié en équivalent temps plein,
- Chiffre d'affaires hors taxes,
- Valeur ajoutée hors taxe - y compris autres produits et autres charges,
- Frais de personnel,
- Excédent brut d'exploitation,
- Résultat courant avant impôt,
- Achats.

La variance, calculée pour obtenir ces coefficients de variation, prend uniquement en compte la variance liée à l'échantillonnage et aux traitements statistiques (calage et correction de la non-réponse par la méthode des groupes de réponse homogène, cf. [4]) portant sur la partie échantillonnée (c'est-à-dire sur le champ de l'ESA). Le coefficient de variation se retrouve donc nul pour tout champ hors de l'ESA. Cette nullité du coefficient ne signifie donc pas que la précision est meilleure par rapport aux cases du champ de l'ESA.

Pour obtenir le coefficient de variation à des niveaux plus fins (par exemple : groupe croisé par tranche d'effectif), le coefficient de variation est calculé de la manière suivante (de façon « approximée ») :

$$CV(\hat{Y}_{\inf ra}) = \frac{EcartType(\hat{Y}_{Groupe})}{Y_{\inf ra}} \sqrt{\frac{n_G}{n_{\inf ra}}}$$

avec :

- $CV(\hat{Y}_{\inf ra})$: le coefficient de variation de la variable Y au niveau infra groupe (par exemple, au niveau groupe ventilé par tranche d'effectif),
- $EcartType(\hat{Y}_{Groupe})$: l'écart type de la variable Y au niveau groupe de la Naf,
- $Y_{\inf ra}$: l'estimation de la variable Y au niveau infra groupe,
- n_G : le nombre d'unités estimé au niveau groupe de la Naf,
- $n_{\inf ra}$: le nombre d'unités estimé au niveau infra groupe.

Enfin, le coefficient de variation sur d'autres variables que ces 21 variables est estimé, également de façon « approximée », par :

$$CV(\hat{Y}_{\inf ra}) = \frac{EcartType(\hat{Y}_{Groupe})}{\hat{Y}_{\inf ra}} \sqrt{\frac{n_G}{n_{\inf ra}}}$$

avec :

- $CV(\hat{Y}_{\inf ra})$: le coefficient de variation estimé d'une variable Y autre que les 21 variables ci-dessus,
- $EcartType(\hat{Y}_{Groupe})$: l'écart type d'une variable parmi les 21 variables ci-dessus au niveau groupe,
- $\hat{Y}_{\inf ra}$: l'estimation d'une variable parmi les 21 variables ci-dessus au niveau infra groupe.
- n_G : le nombre d'unités estimé au niveau groupe de la Naf,
- $n_{\inf ra}$: le nombre d'unités estimé au niveau infra groupe.

Pour chaque variable autre que les 21 ci-dessus, le principe revient à celui utilisé pour la pose du secret. Une variable de la liste est associée à la variable étudiée et elle servira à estimer le coefficient de variation.
Par exemple, les coefficients de variation du total de ventes de marchandises, du total de production vendue de biens et du total de production vendue de services au niveau infra groupe sont calculés à l'aide de l'écart type du chiffre d'affaires hors taxes et de son estimation au niveau infra groupe ; l'idée est d'obtenir un ordre de grandeur.

3.2. Le code qualité

Le code qualité est un indicateur synthétique permettant à l'utilisateur d'avoir un avis sur la diffusabilité d'une case. Afin qu'il soit simple et pratique, ce code est décomposé en trois modalités « Diffusable », « Attention » et « Non diffusable ». Il est calculé à partir :

- Du nombre d'unités provenant de l'échantillon pris en compte pour le calcul de l'agrégat (nb_unite_ech),
- Du coefficient de variation de la variable sur le domaine où est calculée la statistique (CV),
- Du coefficient de variation dit « seuil » qui correspond, pour un groupe de la Naf donné, au 9ème décile de la série formée par les coefficients de variation de la variable au niveau groupe de la Naf ventilé par tranche de taille et au niveau sous-classe de la Naf (CV_seuil).

Ce choix a été motivé par le fait que les résultats publiés sur le site insee.fr correspondent à ces deux niveaux (groupe croisé par tranche de taille et sous-classe). L'Insee est donc prêt à assumer le niveau de qualité obtenu à ces niveaux. Ainsi, toute case ayant un niveau de qualité supérieur peut être considéré comme diffusable.
- Du nombre d'entreprises estimé.

Avec ces quatre variables, le code qualité est calculé de la manière suivante :

nb_unite_ech	0<CV<CV_Seuil	CV_seuil<CV<CV_seuil*2	CV_seuil*2 < CV	CV=. ou CV=0
0-5	Non diffusable	Non diffusable	Non diffusable	Non diffusable
6-20	Attention	Attention	Attention	Non diffusable
>20	Diffusable	Attention	Non diffusable	Non diffusable

Enfin, à la suite de cette codification, si le rapport du nombre d'unités provenant de l'échantillon et le nombre d'entreprises estimé est supérieur à 0.5 alors la case est jugée « Diffusable ». Ceci permet de diffuser des cases ayant un effectif faible a priori et pour lesquelles on dispose d'au moins 50% de répondant à l'ESA.

Les différentes tranches de nombre d'unités provenant de l'échantillon ont été choisies arbitrairement. Il apparaît donc qu'une case est jugée « Diffusable » à partir du moment où :

- La case a été construite avec au moins 20 unités provenant de l'échantillon et que le coefficient de variation de la case est assez faible,
- Ou alors que le nombre d'unités provenant de l'échantillon est assez proche du nombre d'entreprises estimé.

Ce code synthétique de qualité se « dégrade » par la suite avec les modalités « Attention » et « Non diffusable » au fur et à mesure que le nombre d'unités provenant de l'échantillon diminue et/ou que le coefficient de variation de la case augmente. Des statistiques de diffusabilité sont données dans la partie suivante.

Par ailleurs, afin que ce code soit cohérent avec les agrégats publiés sur le site de l'Insee, il a été jugé diffusable tout résultat qui peut se retrouver sur internet (à savoir le niveau groupe de la Naf croisé par tranche d'effectif et le niveau sous-classe). C'est l'objet de la prise en compte du coefficient de variation seuil.

Il est à noter que la notion de diffusabilité est calculée indépendamment du secret de la case. Il est alors tout à fait possible qu'une case jugée « diffusable » soit en secret. Le code n'est représentatif que de la qualité du résultat alors que le secret s'attarde sur l'anonymat des entreprises.

3.3. Le coefficient de variation et intervalle de confiance au niveau groupe de la Naf

Le kit met également à disposition le coefficient de variation au niveau groupe (il y a donc le coefficient de variation au niveau groupe de la Naf et au niveau du domaine de diffusion) et l'intervalle de confiance de la variable au niveau groupe.
Tout comme le coefficient de variation au niveau du domaine de diffusion, une variable de la liste des 21 variables est associée à la variable étudiée et servira à estimer le coefficient de variation ainsi que l'intervalle de confiance au niveau groupe.
Par exemple, le coefficient de variation et intervalle de confiance au niveau groupe du total des ventes de marchandises seront identiques au coefficient de variation et intervalle de confiance du chiffre d'affaires hors taxes au niveau groupe.
L'ensemble de ces indicateurs qualité figurent dans l'exemple de sortie de l'annexe 2.

4. Quelques statistiques

4.1. Le code qualité

Le code qualité est donc intrinsèquement lié au coefficient de variation de la case et au nombre d'unités provenant de l'échantillon servant au calcul de la case en question. La diffusabilité d'une case se dégrade dès lors que ce dernier diminue et que le coefficient de variation augmente. C'est-à-dire que plus il y a de critères de définition des domaines de diffusion, plus le nombre de cases diffusables diminue. Ceci est bien entendu logique car on arrive au point où le nombre d'unités provenant de l'échantillon qui permet de calculer les agrégats devient trop faible pour que l'on puisse juger la loi des grands nombres comme applicable.

Prenons l'exemple de la répartition de ce code qualité suivant le nombre de critères servant à définir le domaine de diffusion pour le nombre d'entreprises au niveau sous-classe de la Naf :

critère	Critère n°1	Catégorie d'entreprise	Catégorie d'entreprise	Catégorie d'entreprise	Catégorie d'entreprise
	Critère n°2		Tranche de chiffre d'affaires	Tranche de chiffre d'affaires	Tranche de chiffre d'affaires
	Critère n°3			Catégorie juridique	Type d'entreprise (SNF/EI[1])
code qualité	Diffusable	74%	64%	51%	58%
	Attention	6%	9%	6%	10%
	Non diffusable	20%	27%	43%	32%

Le nombre de critères pour le domaine de diffusion tend à augmenter le nombre de case « Non diffusables ». Cependant, ce nombre dépend également du type de critère. En effet, la demande de résultat par type d'entreprise conduit à subdiviser chaque résultat suivant 2 modalités alors que le critère catégorie juridique subdivise chaque résultat en 27. En résumé, tout dépend du niveau de détail de la variable critère.

4.2. Le temps de calcul

Le temps de calcul nécessaire au kit dépend forcément de la configuration de l'ordinateur utilisé[2].
Avec l'ordinateur utilisé pour cet article :

- Demander des agrégats fiscaux nécessite 6 à 10 minutes (le nombre de variables et la population affectent peu le temps de calcul),
- Demander des agrégats sectoriels de l'ESA demande plus de temps (environ 10-15 minutes).

[1] Sociétés Non Financières (SNF), Entreprises Individuelles (EI)
[2] L'ordinateur utilisé pour obtenir les temps est un AMD Athlon(tm) 64 x2 Dual Core Processor 4400+ à 2.29 Ghz muni de 2.75 Go de RAM.

Conclusion

Le kit de diffusion des données ESANE est donc un outil novateur dans le sens où il peut répondre aux besoins propres à tel ou tel utilisateur tout en l'informant de la pertinence du chiffre calculé. C'est en cela que réside le principal apport du kit par rapport aux diffusions précédentes des statistiques structurelles de l'Insee qui se basaient sur un fichier complet (FICUS). Ces diffusions passaient sous silence l'utilisation de l'échantillon pour le calcul de l'Activité Principale Exercée (APE) ce qui pouvait « leurrer » l'utilisateur qui pensait travailler sur des données exhaustives. L'estimateur composite cherche à « extrapoler » les résultats obtenus sur l'échantillon à l'ensemble de la population. Et c'est donc ce que propose de faire le kit. Toutefois en agissant de la sorte, l'utilisateur se confronte aux limites liées à la taille de l'échantillon et il est important qu'il en soit prévenu. Libre ensuite à l'utilisateur d'utiliser le résultat même s'il a été jugé comme étant non diffusable dans la mesure où le chiffre n'engage que son auteur.

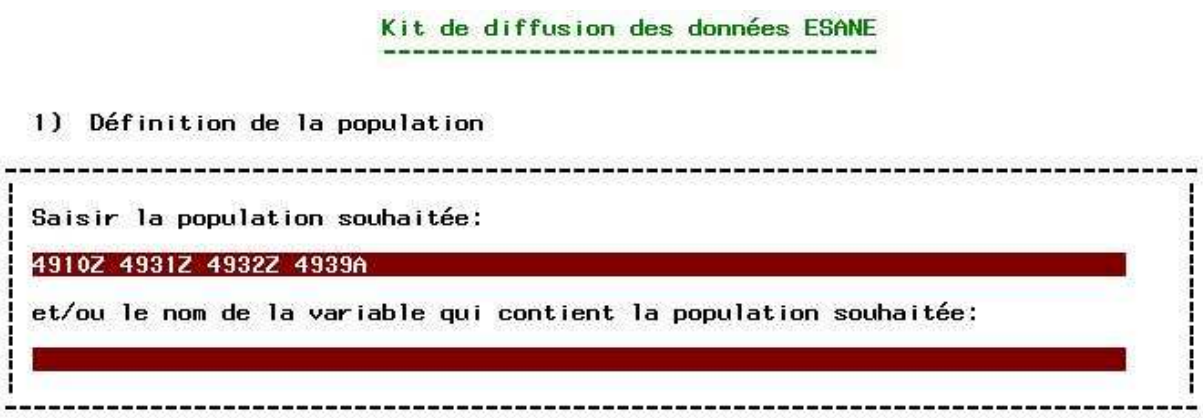
Pour cette première année de mise à disposition du kit, seuls les agrégats en niveaux sont disponibles. Cependant, il est prévu d'y ajouter le calcul de ratios, de distributions et même la décomposition des agrégats en évolution entre deux années N et N-1 [5]. Ce kit évoluera également en fonction des remarques faites par les utilisateurs.

Annexe 1 :

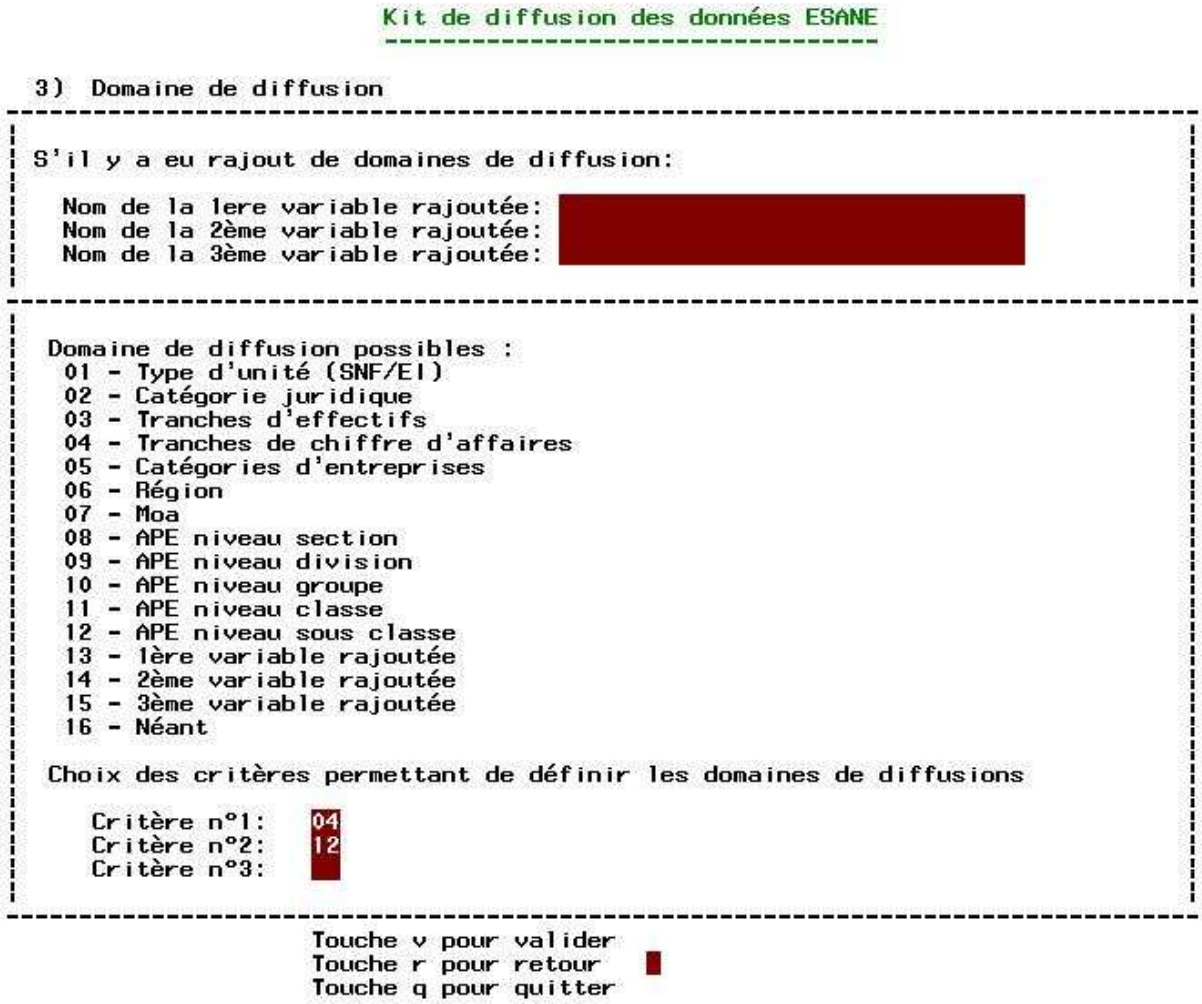
Comme précisé dans la première partie, le kit fait défiler une succession de fenêtres qui ont pour but de renseigner les divers paramètres nécessaires à l'élaboration d'un résultat.

Cet annexe présente les différentes fenêtres à remplir pour l'obtention du chiffre d'affaires et des dépenses en carburant pour le secteur des transporteurs terrestres de voyageurs au niveau sous-classe croisé par tranche de chiffre d'affaires. Le résultat sera présenté dans un fichier Excel.

Pour le choix de la population, il s'agit donc de définir les sous-classes 4910Z, 4931Z, 4932Z et 4939A:

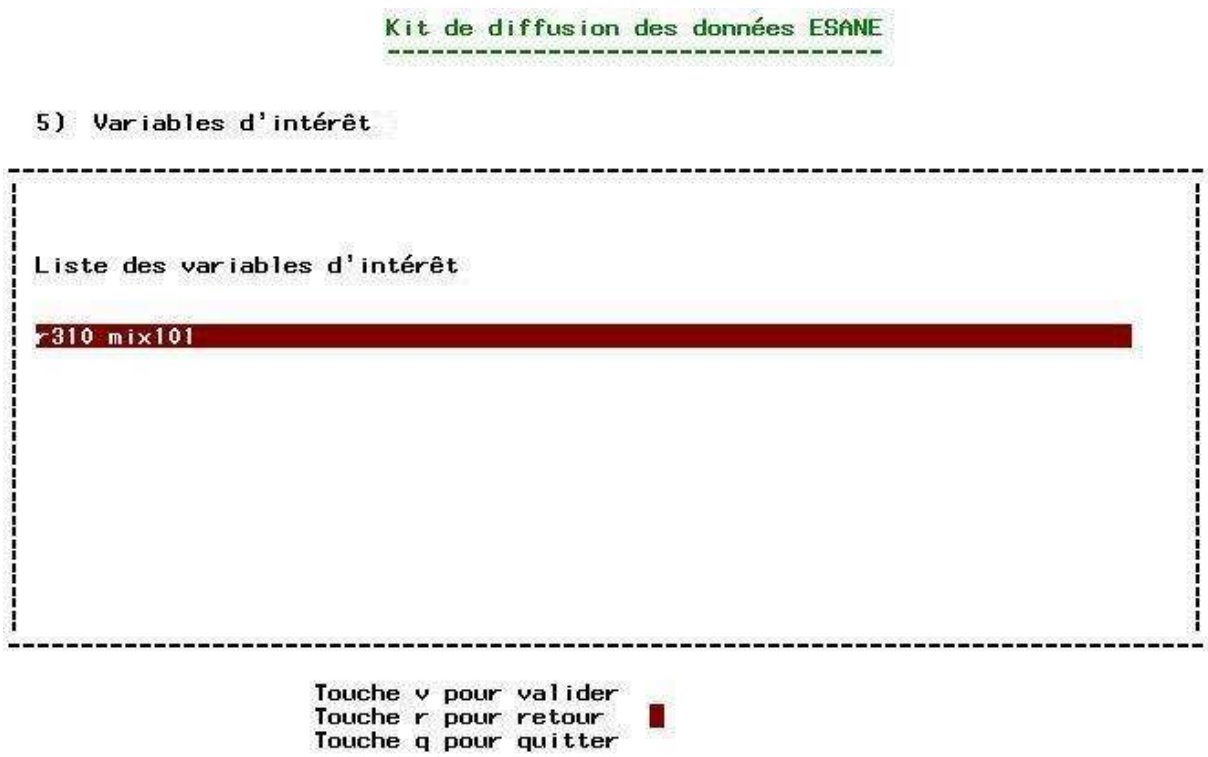


Il faut ensuite définir le domaine de diffusion : les sous-classes croisées par tranches de chiffre d'affaires :

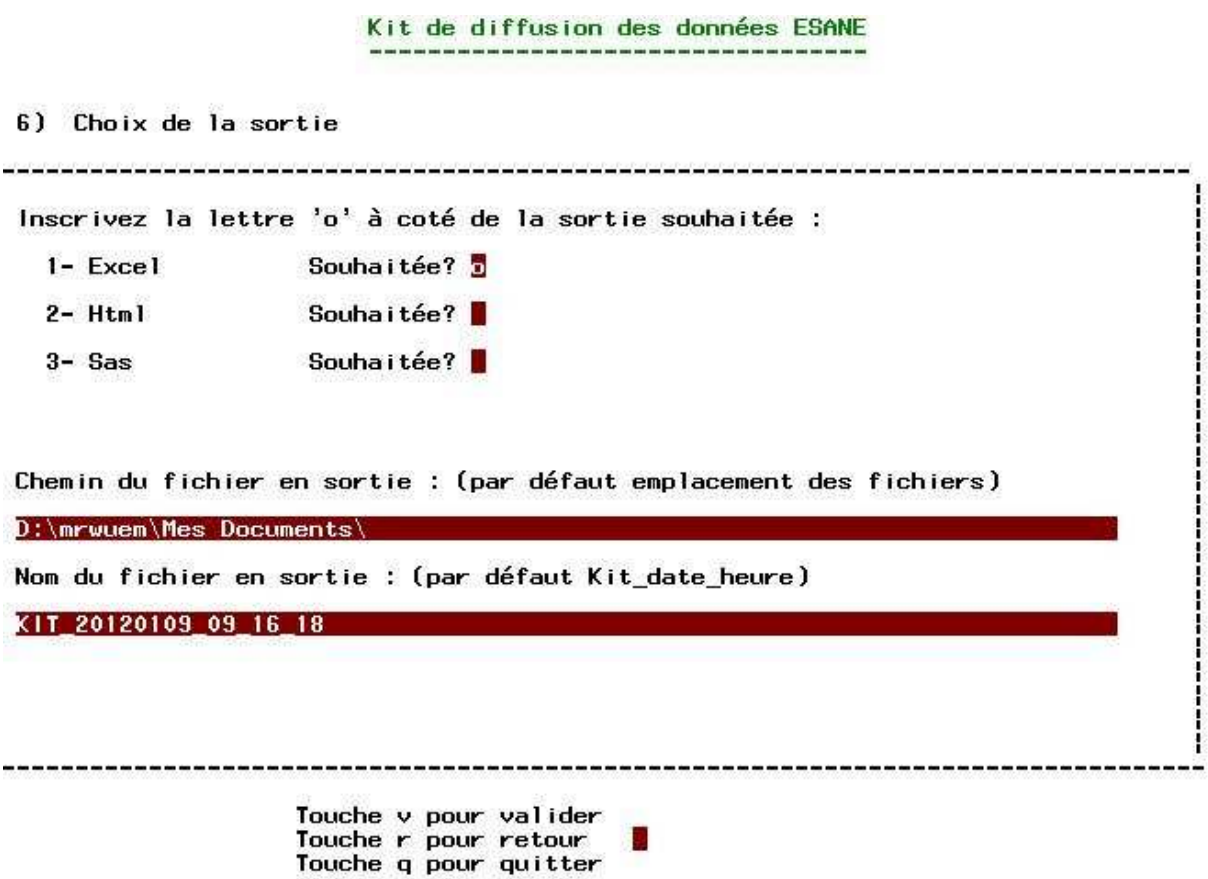


Ensuite le type de statistique :

Viennent ensuite les variables d'intérêt (la variable nombre d'entreprises est calculée automatiquement) :

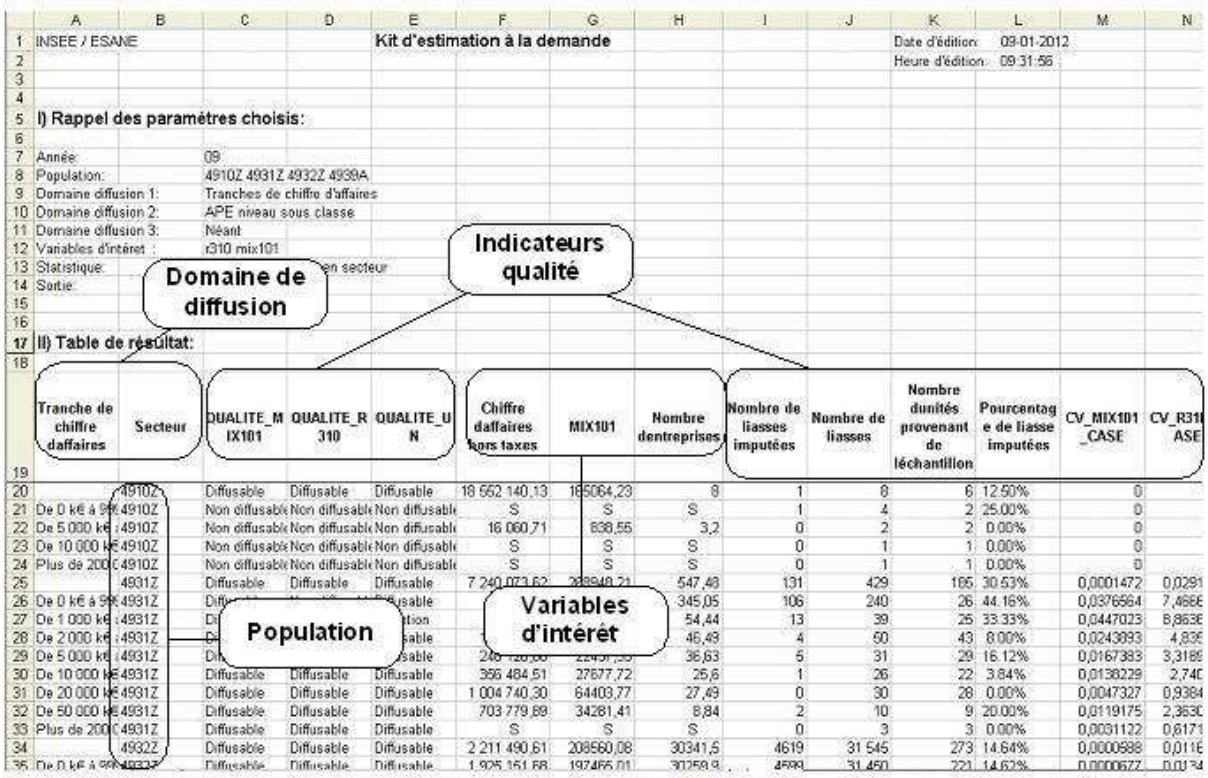


Et enfin le type de sortie et éventuellement le nom du fichier ainsi que sa localisation sur l'ordinateur :



Annexe 2 :

L'exemple d'utilisation du kit en partie 1 conduit à ce résultat :



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	INSEE / ESANE				Kit d'estimation à la demande						Date d'édition:	09-01-2012		
2											Heure d'édition:	09:31:56		
3														
4														
5	I) Rappel des paramétres choisis:													
6														
7	Année:		09											
8	Population:		4910Z 4931Z 4932Z 4939A											
9	Domaine diffusion 1:		Tranches de chiffre d'affaires											
10	Domaine diffusion 2:		APE niveau sous classe											
11	Domaine diffusion 3:		Néant											
12	Variables d'intérêt :		r310 mix101											
13	Statistique:													
14	Sortie:													
15														
16														
17	II) Table de résultat:													
18														
19	Tranche de chiffre daffaires	Secteur	QUALITE_MIX101	QUALITE_R310	QUALITE_UN	Chiffre daffaires hors taxes	MIX101	Nombre dentreprises	Nombre de liasses imputées	Nombre de liasses	Nombre dunités provenant de léchantillon	Pourcentage de liasse imputées	CV_MIX101_CASE	CV_R31 ASE
20		4910Z	Diffusable	Diffusable	Diffusable	18 552 140,13	185064,23	8	1	8	6	12,50%	0	
21	De 0 k€ à 99	4910Z	Non diffusable	Non diffusable	Non diffusable	S	S	S	1	4	2	25,00%	0	
22	De 5 000 k€	4910Z	Non diffusable	Non diffusable	Non diffusable	16 060,71	638,55	3,2	0	2	2	0,00%	0	
23	De 10 000 k€	4910Z	Non diffusable	Non diffusable	Non diffusable	S	S	S	0	1	1	0,00%	0	
24	Plus de 200	4910Z	Non diffusable	Non diffusable	Non diffusable	S	S	S	0	1	1	0,00%	0	
25		4931Z	Diffusable	Diffusable	Diffusable	7 240 073,62	238948,21	547,48	131	429	185	30,53%	0,0001472	0,0291
26	De 0 k€ à 99	4931Z	Diffusable	Diffusable	Diffusable	1	1	345,05	106	240	26	44,16%	0,0376564	7,4668
27	De 1 000 k€	4931Z	Diffusable	Diffusable	Diffusable	1	1	54,44	13	39	25	33,33%	0,0447023	8,8638
28	De 2 000 k€	4931Z	Diffusable	Diffusable	Diffusable	1	1	46,49	4	50	43	8,00%	0,0243893	4,835
29	De 5 000 k€	4931Z	Diffusable	Diffusable	Diffusable	246 120,66	22437,55	36,63	5	31	29	16,12%	0,0167383	3,3189
30	De 10 000 k€	4931Z	Diffusable	Diffusable	Diffusable	356 484,51	27677,72	25,6	1	26	22	3,84%	0,0138229	2,740
31	De 20 000 k€	4931Z	Diffusable	Diffusable	Diffusable	1 004 740,30	64403,77	27,49	0	30	28	0,00%	0,0047327	0,9384
32	De 50 000 k€	4931Z	Diffusable	Diffusable	Diffusable	703 779,89	34281,41	8,84	2	10	9	20,00%	0,0119175	2,3630
33	Plus de 200	4931Z	Diffusable	Diffusable	Diffusable	S	S	S	0	3	3	0,00%	0,0031122	0,6171
34		4932Z	Diffusable	Diffusable	Diffusable	2 211 490,61	208560,08	30341,5	4619	31 545	273	14,64%	0,0000588	0,0116
35	De 0 k€ à 99	4932Z	Diffusable	Diffusable	Diffusable	1 925 151,68	197466,01	30259,9	4599	31 450	221	14,62%	0,0000677	0,0134

Et avec en feuille 2, le coefficient de variation au niveau groupe et l'intervalle de confiance :

	A	B	C	D	E	F	G	H	I	J
1	III) Coefficients de variation et intervalles de confiance au niveau de référence									
2										
3	SECTEUR	CV_MIX101_GROUPE_ESTIME	IC_MIX101_BAS_GROUPE_ESTIME	IC_MIX101_HAUT_GROUPE_ESTIME	CV_R310_GROUPE	IC_R310_BAS_GROUPE	IC_R310_HAUT_GROUPE	CV_UN_GROUPE	IC_UN_BAS_GROUPE	IC_UN_HAUT_GROUPE
4	493	0,00	36 447,09	36 957,25	0,00	16 509 646,88	16 610 801,42	0,00	36 447,09	36 957,25
5										

Références :

[1] ***Esane, le dispositif rénové de production des statistiques structurelles d'entreprises,***
P. Brion, Courrier des statistiques n°130, 2011

[2] **Esane : À la recherche d'une cohérence maximale des données multi-sources sur les entreprises par le biais de micro et macro contrôles**, Olivier Haag*, XIème Journées de Méthodologie Statistique*, janvier 2012.

[3**] Esane ou les malheurs de l'estimation composite : comment gérer les valeurs négatives d'estimateurs par différence ?**, Emmanuel Gros*, XIème Journées de Méthodologie Statistique*, janvier 2012.

[4] **Le plan de sondage de l'ESA (enquête sectorelle annuelle du futur dispositif Esane)**, Perrine Bauer, Gwennaëlle Brilhault, Emmanuel Gros, Xème Journées de Méthodologie Statistique, janvier 2009

[5] **Diffuser des taux d'évolution entre l'année N et N-1, c'est bien, décomposer cette évolution en différents facteurs explicatifs, c'est mieux !**,Heïdi Koumarianos*, XIème Journées de Méthodologie Statistique*, janvier 2012.