Le *data editing* : Définition et principes généraux

Document de travail

N°M2025-06-Octobre 2025

Insee
Mesurer pour comprendre

Nathalie CARON

Institut national de la statistique et des études économiques

2025/06

Le *Data editing* : Définition et principes généraux

Nathalie CARON

Octobre 2025

Remerciements :

L'auteur remercie Pascal Rivière, Pascale Breuil et Philippe Zamora ainsi que les relecteurs au sein du département des méthodes statistiques pour leurs commentaires constructifs.

Direction de la méthodologie et de la coordination statistique et internationale
Timbre L001

88 Avenue Verdier - CS 70058 - 92541 Montrouge Cedex - France -
Tél. : 33 (1) 87 69 55 00 - E-mail : DG75-L001@insee.fr - Site Web Insee : http://www.insee.fr

Résumé

La traduction littérale de « *data editing* » donne « édition de données », expression peu parlante pour un statisticien français. Une traduction plus significative serait « vérification des données », mais elle resterait néanmoins imparfaite. En effet, le *data editing* correspond à l'ensemble des activités liées à la vérification des données individuelles qui visent à leur assurer la qualité requise en vue de leur exploitation en repérant et corrigeant les données individuelles tout au long du processus de production statistique. À vocation pédagogique, ce document de travail permet une première approche des principes généraux de mise en œuvre du *data editing* qui existent dans la littérature et qui reposent sur une succession de différents types de contrôles : les micro-contrôles et les macro-contrôles. Il est important de définir une stratégie globale de « nettoyage » des données qui commence dès la collecte avec l'utilisation mesurée de contrôles présents pendant le remplissage du questionnaire, puis se poursuit par des vérifications automatiques ou par des gestionnaires et se termine par des contrôles post-collecte. Cette stratégie à définir dépend de la qualité des données recueillies à la collecte et de la qualité des données attendue pour la diffusion en termes de précision, celle-ci devant idéalement être définie préalablement. Enfin, une fois le processus de production terminé, il est essentiel de conserver des informations détaillées sur la phase de *data editing* ainsi que sur le cycle de vie des données - depuis les données brutes renseignées par le répondant jusqu'aux données finales destinées pour la diffusion -. Cela permet d'évaluer a posteriori l'impact de cette phase et d'optimiser le processus de production concerné. Dans ce document, les méthodes de correction des données ou d'imputation de masse pour corriger les erreurs ne sont pas abordées, car les techniques utilisées pour imputer une valeur plausible sont les mêmes pour une valeur renseignée détectée inexacte et donc à modifier ou pour une valeur manquante en cas de non-réponse. Les références citées en bibliographie permettront au lecteur d'approfondir ses connaissances sur le *data editing*.

Mots-clés : *Data editing*, s*elective editing,* correction des données, micro-contrôles, macro-contrôles, GSDEM.

Codes JEL : C40, C80

Summary

Data editing refers to the overall process of checking data to ensure that they are of the quality required for analysis by identifying and correcting errors throughout the statistical process.
This document presents the basic principles of data editing as discussed in the literature with an educational purpose. The process involves a series of checks: micro-editing and macro-editing checks. We emphasise the importance of having a data "cleaning" strategy that starts with data collection, using built-in checks, continues with automated or manual checks, and ends with post-collection checks. The chosen strategy will depend on the quality of the data collected and the expected quality in terms of accuracy, which should be defined in advance.
The metadata accompanying the final dataset is also crucial. In this document, we do not cover methods for correcting data, such as imputation, because the techniques used to impute a plausible value are the same for an incorrect value that needs to be corrected or for a missing value in the case of non-response. We therefore refer the reader to the specific literature on this topic. The references cited in the bibliography will help the reader to find more information on data editing.

Table des matières

Introduction

Les données individuelles collectées par enquête ou issues de sources administratives sont rarement utilisables directement pour produire des indicateurs statistiques. En effet, elles contiennent généralement des erreurs et des valeurs manquantes. Ainsi, pour une question donnée, un répondant peut donner une réponse incorrecte de manière intentionnelle ou non, ou ne pas répondre du tout soit parce qu'il ne connaît pas la réponse, soit parce qu'il ne souhaite tout simplement pas répondre à la question.

On comprend aisément que la présence d'erreurs influe directement sur la qualité des statistiques élaborées à partir du fichier de données concerné. Ceci est d'autant plus problématique lorsque les erreurs sont nombreuses et/ou que les erreurs concernent des unités qui contribuent fortement dans le calcul des indicateurs. En ce qui concerne les non-réponses, celles-ci posent problème lorsque les non-répondants ont un comportement différent des répondants pour le thème de l'enquête et plus la différence est importante plus le problème l'est également avec des estimations qui sont biaisées. En général, la non-réponse est rarement le fruit du hasard et on se trouve dans la situation où les non-répondants sont différents des répondants. Les techniques de correction de la non-réponse par imputation ou par repondération ne sont pas détaillées dans ce document (voir par exemple Caron, 2005).

Dans ce document, on s'intéresse au processus global de vérification des données individuelles collectées qui vise à assurer la qualité requise pour les données. Ce processus global est appelé *data editing*. Fortement imbriqué dans le processus de production de la collecte brute des données jusqu'à la validation finale du fichier individuel, il concerne tout type de données : les enquêtes réalisées auprès des ménages ou des entreprises, les recensements de population ainsi que les sources de données administratives de nature entreprises ou ménages. Cependant, les enjeux de cette phase sont plus forts et spécifiques sur les données des entreprises. En effet, d'une part la population des entreprises est très hétérogène, certaines d'entre elles ayant un impact fort sur les résultats, et d'autre part les données collectées auprès d'elles sont souvent quantitatives, prenant des valeurs potentiellement très dispersées et rendant ainsi la vérification des données essentielle.

Une fois les données et/ou unités « douteuses » détectées par une succession de contrôles, une partie d'entre elles sont examinées par les gestionnaires d'enquêtes. Ils déterminent si c'est une erreur ou une donnée exacte mais atypique par rapport aux autres et ils peuvent corriger directement les valeurs en erreur grâce à leur expertise du domaine, l'historique des données ou dans certains cas en recontactant directement les entreprises. L'examen des données par les gestionnaires d'enquêtes est un processus très coûteux et long. Il pose ainsi nécessairement la question de l'arbitrage coût/qualité sachant que les gestionnaires ne peuvent traiter l'ensemble des unités douteuses pour un budget donné et dans un délai raisonnable. Il est donc nécessaire de hiérarchiser les unités à traiter et de déterminer celles qui le seront par les gestionnaires et celles qui le seront de manière automatique (*selective editing*). Adossé à cette pratique, il est également important de définir un critère d'arrêt des reprises manuelles, celui-ci dépendant du degré de finesse des agrégats publiés ainsi que du niveau de qualité attendu pour chacun d'entre eux. Dans l'idéal, la qualité serait appréhendée au sens de la précision en termes d'erreur quadratique moyenne. Celle-ci tiendrait compte non seulement des composantes classiques comme l'échantillonnage et la non-réponse mais également du "gain potentiel en qualité" apporté par les opérations de vérifications manuelles, c'est-à-dire l'écart entre l'agrégat estimé à tout moment du traitement de l'enquête et une valeur de référence finale de cet agrégat (obtenue par modélisation par exemple). Ainsi, dès que le niveau de qualité de l'agrégat atteint une valeur-cible que l'on considère comme satisfaisante, on arrête les contrôles manuels. Cet indicateur permettrait

également, au cas où pour des raisons de délai la production doit s'arrêter, d'avoir une estimation de la perte en qualité qui en résulte.

Les valeurs et/ou unités « douteuses » qualifiées comme des erreurs doivent être corrigées. En dehors des cas où ce sont les gestionnaires qui examinent les données et corrigent directement les erreurs, la correction des erreurs est automatisée. Les techniques utilisées pour imputer une valeur plausible sont similaires à celles utilisées pour une valeur manquante en cas de non-réponse, elles ne seront donc pas détaillées dans ce document. Toutefois, la phase d'imputation des erreurs peut être très imbriquée avec la phase de vérification statistique des données au sein du processus de production, et les méthodes d'imputation choisies le sont afin de permettre que les données imputées « satisfassent » les vérifications statistiques.

Après avoir défini ce que recouvre le *data editing* dans la première partie, nous détaillons ensuite les grands principes de cette phase en décrivant les différents types de contrôles : les micro-contrôles réalisés au niveau de chacune des unités (comme les contrôles internes au questionnaire) et les macro-contrôles permettant de repérer à partir d'agrégats définis les unités douteuses qui y contribuent le plus. La priorisation des unités à traiter via une procédure de *selective editing* ainsi que le critère d'arrêt des traitements par les gestionnaires sont abordés dans la partie 3. Nous insistons sur l'importance de définir une stratégie globale de « nettoyage » des données qui commence dès la collecte avec des contrôles pendant le remplissage du questionnaire, puis des vérifications par des gestionnaires ou de manière automatique, et qui se termine par des contrôles globaux après la collecte (partie 4). Cette stratégie dépend à la fois de la qualité des données en amont et de la qualité des données attendues qui doit par conséquent être définie préalablement. Il est également essentiel de conserver des informations sur la phase de *data editing* effectuée (contrôles mis en place, critère d'arrêt retenu, etc.) ainsi que de disposer de l'ensemble du cycle de vie des données, depuis les données brutes fournies par le répondant jusqu'aux données finales retenues pour la diffusion.

La dernière partie donne quelques éléments sur le cadre international, dont l'activité autour du *data editing* a été relativement dense depuis les années 1990 du fait essentiellement de spécialistes de la statistique publique. En effet, tout producteur de données d'enquêtes, et tout particulièrement les instituts nationaux statistiques (INS), est concerné par ce sujet qui mobilise une part non négligeable de ses ressources qu'il convient d'optimiser tout en maintenant la même qualité. Par conséquent, de nombreux travaux portent sur l'amélioration de l'efficacité de ce processus et investiguent de nouvelles méthodes basées sur l'intelligence artificielle (*machine learning)*.

Partie 1. Qu'est-ce que le *data editing* ?

Les données collectées sont rarement utilisables « directement », celles-ci peuvent d'une part être totalement manquantes pour une unité donnée ou incomplètes (cas de non-réponse totale et cas de non-réponse partielle) et d'autre part être présentes mais entachées d'erreurs.

Pour les rendre exploitables, il est donc nécessaire de détecter les erreurs puis de les corriger manuellement ou en réalisant un traitement de correction de la non-réponse.

1.1. Un ensemble d'activités contribuant à la qualité des données individuelles

De manière générale, le *data editing* concourt à rendre les données individuelles renseignées propres à leur exploitation et correspond à la phase qui permet de repérer les données susceptibles d'être inexactes, les expertiser et ainsi passer des données brutes aux données nettoyées. C'est une des étapes cruciales dans le processus de gestion des données qui vise à garantir la qualité des données de base avant leur utilisation (voir figure 1).

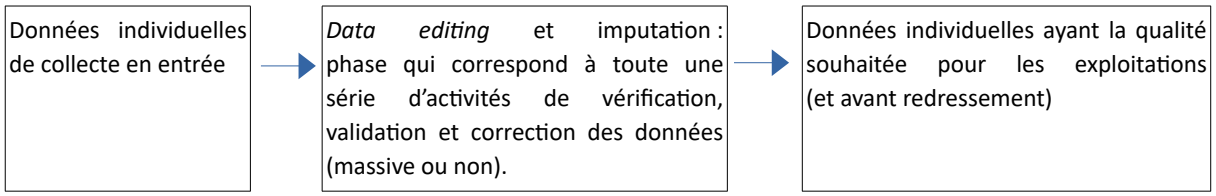


Figure 1 : La place du *data editing* dans le processus de production

Le *data editing* dépend directement de la qualité des données en amont et de la qualité des données attendue, et est une phase dont l'importance dépend de l'ampleur de la différence entre ces deux notions. Cette étape centrée sur les données individuelles se situe en amont des phases de redressement comme le calage et la correction de la non-réponse totale. Dans le cas des enquêtes entreprises, l'expertise des données et/ou unités identifiées comme douteuses ou au moins une partie d'entre elles est réalisée par des gestionnaires d'enquêtes qui prennent dans certains cas contact avec les entreprises concernées. Cette étape est donc très consommatrice de moyens de gestionnaires. Cependant, si les micro-données ne sont pas mises à disposition et que seuls des agrégats sont publiés, de nombreuses études indiquent qu'il n'est pas nécessaire de corriger toutes les erreurs et avoir une base finale « parfaite » pour publier des statistiques de qualité suffisante : seule la correction des erreurs des unités les plus « influentes » sur les statistiques est impérative.

1.2. Une définition variable selon les auteurs

Si on cherche dans la littérature une définition précise du *data editing,* on constate que ce concept n'est pas toujours défini de la même façon. Ainsi, par exemple d'après le glossaire des termes sur le *data editing* de la Commission économique pour l'Europe des Nations Unies (Unece) de 2000, le *data editing* est « l'activité visant à détecter et à corriger les erreurs (incohérences logiques) dans les données ». Avec cette définition, on se place donc dans le cadre de détection d'erreurs d'un certain type (les incohérences logiques) et la phase de correction de ces erreurs est incluse. En revanche, dans le « Handbook on Data Quality Assessment Methods and Tools » d'Eurostat de 2007, le *data editing* est défini comme « l'application de contrôles qui identifient les données manquantes, invalides ou incohérentes, ou qui signalent les enregistrements de données

susceptibles d'être erronés ». Dans cette définition, les données identifiées ne sont pas forcément en erreur et la phase de correction n'y est pas incluse.

Une des différences majeures entre ces deux définitions porte sur l'inclusion ou non de la phase de correction des données qui suit celle de détection de valeurs « hors normes » et fausses et qui lui est très liée. En général, si on inclut la phase d'imputation dans le *data editing*, on parle plutôt de « *data editing and imputation* », pour bien mettre en évidence que l'imputation est une dimension supplémentaire du processus. L'autre différence porte sur la nature des contrôles, la première définition étant plus restrictive en considérant uniquement des contrôles de premier niveau. Dans le cadre de ce document, nous excluons dès à présent la partie d'imputation de masse du *data editing* en considérant que c'est une étape postérieure, car les techniques utilisées pour imputer une valeur plausible sont les mêmes pour une valeur renseignée détectée inexacte et donc à modifier ou pour une valeur manquante en cas de non-réponse. En revanche, lors de leur expertise, les gestionnaires peuvent être amenés à corriger directement les données, cette phase de correction manuelle est dans la définition du *data editing*.

Le terme *data editing* issu de la littérature anglo-saxonne n'a pas d'équivalent satisfaisant en français. La traduction littérale donne « édition des données », ce qui n'est pas forcément explicatif. Cette traduction est pourtant utilisée à Statistique Canada au même titre que le terme *data editing* même si cet institut préfère séparer et nommer les différentes parties du processus de *data editing* avec par exemple les notions de validation des données, détection des valeurs aberrantes et imputation. En France, si on souhaitait traduire le terme de *data editing*, certes de manière imparfaite au vu des activités qu'il recouvre, ce pourrait être par « vérification des données » ou encore « contrôles et qualification des données ». Dans la suite de ce rapport, nous conserverons le terme de *data editing*.

1.3. Des activités très intégrées dans le processus de production

La principale raison pour laquelle le contour précis du concept de *data editing* varie selon les auteurs réside dans le fait que les activités de *data editing* sont très imbriquées dans le processus de production à différentes étapes et par conséquent il est difficile de les isoler. C'est particulièrement le cas lorsqu'une valeur en erreur est détectée et que celle-ci est imputée directement juste après sa détection.

Comme précisé dans la partie 1.1., le *data editing* est un ensemble d'activités qui visent à détecter les incohérences, les erreurs de saisie, les valeurs aberrantes et/ou influentes sur les résultats ainsi que d'autres anomalies qui pourraient compromettre la fiabilité des résultats obtenus à partir d'un jeu de données, que celles-ci soient issues d'enquêtes ou de sources administratives. Ces activités peuvent être réalisées à tous les stades du processus de production de données : dès la saisie des réponses aux questionnaires internet au moment de la collecte, lors de la disponibilité de chaque questionnaire pour en vérifier la cohérence globale ou encore en traitement centralisé après la collecte lorsque tous les questionnaires sont disponibles. Les vérifications ultimes permettant de s'assurer que les statistiques obtenues à partir du jeu de données concerné sont cohérentes par rapport à celles issues d'autres sources externes entrent également dans ce cadre dans la mesure où elles peuvent conduire à repérer des erreurs dans les données individuelles et à les corriger. Le *data editing* commence donc au moment de la collecte et se termine par la publication des données et/ou indicateurs finaux. Dans certains cas, un retour sur le processus d'enquête (en particulier sur le questionnaire) est réalisé afin de l'améliorer lorsque des sources d'erreurs systématiques sont identifiées. C'est d'ailleurs un des objectifs que donne Granquist (1995) au *data editing.*

De manière schématique, le *data editing* est composé d'opérations de vérification des réponses fournies par les enquêtés dans le cas des enquêtes. Ces opérations se font à l'aide de contrôles

automatiques qui permettent de repérer des unités qui ne « satisfont pas un certain nombre de critères pré-définis » et qui sont identifiées comme « douteuses », celles-ci pouvant alors soit être expertisées par des gestionnaires (c'est le cas pour les enquêtes entreprises, au moins pour une partie des unités « suspectes ») soit être traitées de manière automatique. Avec les éléments dont il dispose par son expertise acquise au fil du temps ou suite à une recherche sur Internet ou encore suite à un contact direct avec les entreprises enquêtées concernées, le gestionnaire décide soit de modifier la réponse initiale de l'entreprise soit de la confirmer soit encore de la mettre en données manquantes (et l'imputation se fera automatiquement par un programme comme dans le cas d'une non-réponse partielle). Dans certains processus, une valeur d'imputation peut être suggérée au gestionnaire lors de son expertise, valeur qu'il peut décider ou non de retenir.

Du fait de la forte imbrication du *data editing* dans le processus de production, les acteurs intervenant pour une opération donnée dans la phase de *data editing* sont nombreux :

- les enquêtés eux-mêmes ou les enquêteurs lors du recueil des réponses au questionnaire selon le mode de passation de l'enquête ;
- les gestionnaires d'enquêtes pour les enquêtes entreprises avec dans certains cas un nouveau contact avec les enquêtés et les divisions Enquêtes ménages dans les directions régionales pour les enquêtes ménages une fois la collecte de l'unité terminée ;
- et les responsables d'enquêtes pour des opérations centralisées ultimes de vérification ainsi que pour le redressement des données manquantes ou invalides par imputation.

1.4. Des activités essentielles en statistiques d'entreprises et consommatrices de moyens

Si on retrouve les grands principes du *data editing* dans tout type de recueil de données – les enquêtes réalisées auprès des ménages ou des entreprises, les recensements de population ou d'autres types ainsi que les traitements de données administratives de nature entreprises ou ménages –, le processus s'avère plus important, plus complexe et plus coûteux dans le cas de données entreprises pour les raisons suivantes :

– Les données relatives aux entreprises sont en général plutôt des données quantitatives qui sont liées entre elles, et peuvent prendre des valeurs très dispersées susceptibles d'avoir un impact fort sur les résultats. Celles-ci se prêtent donc plus à des contrôles et la phase de description des contrôles est relativement importante. Dans le domaine ménages, c'est moins le cas.

– La population des entreprises est très hétérogène et certaines d'entre elles peuvent avoir un impact considérable sur les agrégats. Cette caractéristique est prise en compte lors de la détermination des plans de sondage des enquêtes entreprises. Ce sont en général des plans de sondage stratifiés dans lesquels on retrouve les entreprises les plus importantes en termes d'effectifs et/ou de chiffre d'affaires dans une strate exhaustive. Cependant, il peut arriver que certaines entreprises qui ont des poids importants soient en réalité des entreprises de taille suffisante qui auraient dû être dans la strate exhaustive et qui ne l'ont pas été soit du fait d'erreurs de la base de sondage soit du fait d'un décalage trop important entre la date de constitution de la base de sondage et celle de la réalisation de l'enquête. C'est un cas classique qui conduit à des unités impactant fortement les agrégats. Dans les enquêtes ménages, il n'existe pas de strate exhaustive, chacun des ménages échantillonnés en représentant plusieurs milliers. Du fait de la grande hétérogénéité de la population des entreprises, les agrégats construits à partir de ces données sont très sensibles aux données influentes.

– Le plus souvent les échantillons incluent les mêmes unités au moins pour les grandes entreprises pour des périodes successives et l'utilisation de données historiques est donc possible.

Le fait que des gestionnaires d'enquêtes examinent des données et/ou unités douteuses rend le processus très coûteux et pose nécessairement la question de l'arbitrage coût/qualité sachant que les gestionnaires ne peuvent traiter l'ensemble des unités douteuses. L'arbitrage sur les moyens dépend fortement des « livrables » envisagés.

Sauf exception, dans le cas des enquêtes ménages et contrairement aux enquêtes entreprises, il n'y a pas de sélection d'unités à « examiner » par des gestionnaires avec re-contact éventuel de ces unités pour confirmer ou corriger leurs réponses. Les opérations réalisées sur les réponses recueillies auprès des enquêtés aux enquêtes ménages pour valider la collecte (vérification des concepts, correction des erreurs et des incohérences…) sont réalisées à deux niveaux : une partie des contrôles est réalisée dans ce qu'on appelle l' « apurement » de l'enquête ou la « validation de la collecte » par les divisions des enquêtes ménages dans les directions régionales avec l'aide dans certains cas des enquêteurs ; l'autre partie l'est par le responsable d'enquête à un niveau centralisé en particulier pour regarder la distribution de certaines variables clés de l'enquête. La validation de la collecte repose sur des premiers contrôles correspondant à des contrôles de validité et à certains contrôles de cohérence (voir partie 2.1.). Le travail réalisé pour examiner certaines variables par le responsable d'enquête est relativement proche de celui réalisé en statistique d'entreprises pour déterminer les données influentes, en particulier lorsque ces variables ont des distributions très asymétriques (par exemple le revenu ou le patrimoine).

Partie 2. Le *data editing* repose sur une succession de contrôles qui détectent des données douteuses …

Les contrôles permettent d'identifier les unités ou valeurs douteuses. On en distingue de deux types : les micro-contrôles et les macro-contrôles. Ces deux types de contrôles ont la même finalité, celle de détecter des données individuelles à valider/traiter, ils diffèrent sur la manière de les repérer.

2.1. Les micro-contrôles : des contrôles unité par unité directement au niveau des variables

D'après le glossaire des termes sur le *data editing* de l'Unece de 2000, les micro-contrôles consistent, pour une unité donnée, à détecter une erreur ou une suspicion d'erreur à partir des informations de cette seule unité. Ces contrôles s'effectuent par conséquent unité par unité sans tenir compte de l'impact potentiel d'une éventuelle erreur sur les statistiques diffusées.
Cependant, dans certains cas, la valeur de référence pour le contrôle peut être construite à partir de données courantes et/ou antérieures sur tout ou partie de la population concernée. Par exemple, détecter les unités dont la valeur de la variable *Y* est supérieure au quantile d'ordre 95 % (noté α ci-après) de la distribution empirique de cette même variable sur l'ensemble des unités observées est bien un exemple de micro-contrôle : en effet, le contrôle « $y_i > \alpha$ » ne porte que sur la valeur y_i de la variable *Y* pour l'unité considérée i, même si la valeur de référence α dépend de plusieurs unités.

Les contrôles des données peuvent être de simples vérifications de « bon sens » (contrôles de premier niveau) ou des vérifications plus complexes. Ils sont déterminés en tenant compte de l'expertise de spécialistes du sujet, de la structure et la complexité du questionnaire en termes de longueur et nombre de questions posées, de l'historique éventuel des données et de toute autre donnée auxiliaire disponible en lien avec les données.

Certains contrôles sont toutefois assez standardisés et leur construction est indépendante du thème de l'enquête comme le repérage des points extrêmes d'une distribution ou encore la vérification de la cohérence d'une donnée en fonction de son historique. Ce qui relève du spécialiste du domaine abordé dans l'enquête est le choix des variables à contrôler.

On distingue plusieurs natures de micro-contrôles :

- Les **contrôles de validité** s'assurent que les données respectent les contraintes définies (par exemple, pas de caractère dans un champ numérique), se situent dans l'étendue des valeurs autorisées et que les unités de mesure spécifiées ont été correctement utilisées. Une partie de ces contrôles sont souvent intégrés directement au système de collecte de données (questionnaire web). Un cas classique est l'erreur d'unité ou erreur de mesure flagrante. Par exemple, le chiffre d'affaires est demandé en euros, mais le répondant indique un montant en millions d'euros. La valeur collectée est alors fausse mais l'erreur en général se détecte directement et se corrige facilement de manière automatique. À noter que dans le cas où l'erreur de mesure est moins flagrante - erreur d'un facteur de 10 voire de 100 par exemple -, on se retrouve plutôt dans le cas des contrôles de vraisemblance (voir ci-dessous).
- Les **contrôles de cohérence** comparent différentes réponses au sein d'un même enregistrement pour s'assurer qu'elles sont cohérentes entre elles, et en particulier que les relations entre variables (économiques notamment) sont bien vérifiées. On distingue plusieurs sous-catégories de contrôles de cohérence :

– des contrôles d'égalité, par exemple CA=VA+CI où CA est le chiffre d'affaires, VA la valeur ajoutée et CI les consommations intermédiaires, ou encore la décomposition du chiffre d'affaires,

– des contrôles d'inégalité, par exemple CA ≥ 0,

– des contrôles de type « si… alors… », par exemple si une personne se déclare appartenir au groupe d'âge des 0 à 14 ans, alors elle ne peut pas être retraitée ou avoir son permis de conduire.

- Les **contrôles de vraisemblance** consistent à vérifier la « plausibilité » d'une ou plusieurs donnée(s). Par exemple, si on dispose d'un historique, on vérifie que la variable étudiée est vraisemblable par rapport à la même variable recueillie pour la même unité sur une période précédente. Ce type de contrôles est très présent dans le cadre des enquêtes conjoncturelles qui disposent par construction d'un historique profond en termes de données. La donnée étudiée peut également être comparée à une source auxiliaire. Par exemple, les enquêtes auprès des entreprises collectent généralement des données financières sur les entreprises. Les mêmes informations peuvent aussi être disponibles dans les déclarations fiscales de l'entreprise. Ainsi, les données fiscales peuvent être utilisées pour développer des contrôles de vraisemblance pour valider les données d'enquête. On vérifie également la plausibilité des ratios et variables calculées et tout écart trop important sera signalé. Les contrôles de vraisemblance peuvent aussi vérifier si la donnée est aberrante du point de vue de la distribution des données. Les contrôles statistiques de vraisemblance sont par exemple du type suivant :

  – la valeur y_i de la variable *Y* pour l'unité i sera jugée vraisemblable si elle est comprise dans l'intervalle [m-a*S ; m+b*S], où m est un indicateur de valeur moyenne de la variable *Y*, S un indicateur de sa dispersion et a et b deux valeurs égales ou non.

  – si l'on dispose pour une même unité i d'une variable auxiliaire x_i fiable ou déjà validée, la valeur y_i de la variable *Y* sera jugée vraisemblable si le ratio y_i/x_i (par exemple le chiffre d'affaires rapporté au nombre de salariés) est compris dans une fourchette prédéfinie.

Il faut aussi prévoir des contrôles permettant de se prémunir contre les répondants qui fournissent toujours la même réponse dans le cas d'une enquête répétée dans le temps que ce soit à un rythme mensuel, trimestriel ou annuel sans que la constance de la réponse soit justifiée sur longue période. La donnée peut tout à fait respecter les contrôles de validité, de cohérence ou de vraisemblance et par conséquent ne pas ressortir en donnée douteuse sans pour autant être exacte (c'est le problème des *inliers*, voir Granquist et Kovar, 1997). Une solution pourrait être de mettre un contrôle prenant en compte les évolutions pour cette variable d'unités similaires.

2.2. Le passage des résultats des micro-contrôles à la détection des données douteuses

Les micro-contrôles peuvent être très nombreux et la difficulté est qu'une même variable peut faire l'objet de plusieurs contrôles de différents types, certains d'entre eux n'indiquant aucune incohérence ni problème tandis que d'autres en signalent. Il reste alors à savoir comment se servir de toutes ces informations pour décider si la valeur de la donnée considérée est cohérente ou ne l'est pas.

Plusieurs méthodes peuvent être mises en place. La solution naturelle est de considérer que la variable est douteuse dès que l'un des contrôles qui implique la variable n'est pas satisfait. Cependant, cette solution conduit à avoir trop de variables douteuses et par conséquent à trop de vérifications.

Une autre solution est de calculer un « score » ou une « note de cohérence » pour chaque variable, ce score dépendant directement des contrôles qui sont satisfaits et de ceux qui ne le sont pas. Ce peut par exemple être la somme de notes qui sont attribuées à chaque contrôle pour caractériser la cohérence d'une variable, la note étant nulle si le contrôle ne signale aucun problème et non nulle s'il détecte un problème, la valeur de la note étant fonction de la « gravité » du problème. Ainsi, par exemple, dans le cas d'un contrôle statistique de vraisemblance basé sur un ratio, l'idée serait de fixer une note de 0 si le ratio est dans la fourchette prédéfinie, une note de 5 si le ratio est dans une fourchette un peu plus large que la fourchette initiale et une note de 10 si ce ratio est en dehors de ces deux fourchettes. Ainsi, si le score correspondant à la somme des notes est inférieur ou égal à un seuil alors la variable est jugée cohérente, si ce n'est pas le cas elle est jugée douteuse et doit donc être redressée et/ou examinée par un gestionnaire. Le nombre de contrôles concernant une même variable est aussi un critère qui devrait être pris en considération car plus il y a de contrôles sur une variable plus on a de chances de trouver des incohérences.

Cette méthodologie est par exemple mise en place dans le cadre de la production des statistiques structurelles de l'Insee dans le système Esane[1] (Beguin et Haag, 2017) mais également dès la fin des années 90 dans la 4ème génération des enquêtes annuelles d'entreprises (Riviere, 1997).

En fait, le sujet est encore plus complexe, car une valeur peut être jugée douteuse parce que les contrôles dans lesquels elle intervient mettent en jeu des variables qui sont elles-mêmes erronées, alors que si ces dernières avaient été redressées préalablement, le résultat des contrôles aurait été différent. Ceci renvoie à la question d'organiser les variables en groupes de variables et de traiter les contrôles et les redressements de façon hiérarchique groupe par groupe. Une fois que les variables d'un groupe sont jugées cohérentes, elles le resteront pour toute la suite du processus et leur validité ne sera donc plus remise en cause. C'est cette méthodologie qui a été choisie dans le cadre du traitement des enquêtes structurelles d'entreprises (Riviere, 1997). Le premier groupe rassemble les variables de base : effectif au 31/12, effectif moyen, frais de personnel, production, consommations intermédiaires, total des produits, total des charges. Les variables de ce groupe sont donc contrôlées et modifiées si besoin en premier et ne seront ensuite plus modifiées pour la vérification des variables des autres groupes.

2.3. Les macro-contrôles : des contrôles au niveau de données agrégées pour détecter des unités douteuses

Contrairement aux micro-contrôles, les macro-contrôles sont des contrôles effectués sur une statistique calculée à l'aide de données issues d'un ensemble d'unités (par exemple un agrégat ou un ratio de deux agrégats) au sein du processus de production des données. Ils permettent de détecter les unités ayant le plus d'influence sur l'(les) indicateur(s) en termes de contribution, puis de « redescendre » et d'examiner au niveau individuel ces unités. De nombreux articles (voir par exemple Unece, 1994 partie B ou Guggemos, 2010) portent sur la faible efficience de la phase de traitement des données, composée exclusivement de micro-contrôles et sur l'importance d'introduire des macro-contrôles dans les processus afin d'éviter « que les gestionnaires ne perdent du temps à corriger des données dont l'influence sur les agrégats à diffuser est totalement négligeable » avec en filigrane l'arbitrage coût/qualité.

Il existe également des macro-contrôles dits finaux (appelés *output editing*) appliqués à la fin de la chaîne de production sur un fichier complet de données et qui permettent de vérifier la cohérence globale des indicateurs soit par rapport à ceux issus d'autres sources, soit temporellement pour les

1 Esane repose sur un principe de réconciliation des données fiscales et des données d'enquêtes, qui, tout en réduisant la charge de réponse des entreprises, permet d'obtenir un estimateur plus précis de leur activité que les deux sources prises séparément.

dispositifs répétés dans le temps. Ceux-ci peuvent permettre de détecter certaines erreurs qui n'ont pas été décelées par les contrôles précédents, par exemple pour des variables qui n'ont pas été considérées comme des « variables cibles » lors de la mise en œuvre des contrôles. Pour cette phase, on peut également mettre à disposition le fichier de données finalisé ou les statistiques calculées auprès de premiers utilisateurs pour recueillir leur expertise métier. Les unités ayant un fort impact sur les statistiques peuvent également ressortir graphiquement (*graphical editing*).

Partie 3. … couplée avec un traitement de ces données réalisé de manière automatique ou par des gestionnaires

Une fois les contrôles réalisés, on se retrouve avec des unités qui ont des valeurs identifiées comme douteuses qu'il faut traiter.

3.1. En quoi consiste le traitement des données ?

Les données douteuses identifiées doivent être expertisées afin d'isoler celles jugées correctes qui sont à conserver pour l'exploitation des résultats de celles jugées en erreur qui doivent donc être modifiées et remplacées. Dans l'idéal, cette expertise est réalisée par des gestionnaires d'enquête. Cependant, il est impossible de faire vérifier l'ensemble des données douteuses par les gestionnaires dans un budget donné et dans un délai raisonnable, une partie le sera donc de façon automatisée (voir parties 3.2. à 3.5.).

Pour les données jugées en erreur, il faut définir comment les modifier (de façon automatique par un programme d'imputation de masse ou par le gestionnaire) et dans le cas des modifications faites directement par les gestionnaires, ceux-ci décident s'ils recontactent ou non les entreprises concernées pour définir les valeurs « de remplacement ». De nombreuses études indiquent qu'il n'est pas nécessaire de corriger toutes les erreurs pour avoir une base finale « parfaite » permettant de diffuser des statistiques de qualité suffisante. Granquist et Kovar (1997) suggèrent qu'il est possible d'éliminer jusqu'à 50 % de la phase de traitement des données douteuses par un gestionnaire sans effet significatif sur la qualité des données en priorisant de façon judicieuse les unités à vérifier et indiquent que « *la question n'est pas de savoir si nous pouvons nous permettre de réduire l'édition, mais plutôt si nous pouvons nous permettre de ne pas le faire* ». Ils soulignent également la nécessité de trouver le juste équilibre entre une phase de vérification excessive pouvant entraîner une perte précieuse d'informations en cas de correction erronée des données et une phase de vérification insuffisante pour garantir des données fiables. Ces deux situations extrêmes conduisent à des estimateurs qui ne reflètent plus la situation que l'on cherche à mesurer.

A noter que le fait de confirmer qu'une donnée douteuse est correcte et qu'elle reste donc identique avant la phase de *data editing* et après, fait partie de la phase de traitement des données au même titre que l'action qui consiste à changer les données jugées en erreur.

3.2. Le principe du *selective editing*

Afin de diminuer le coût de la phase de *data editing*, on combine deux types de vérification, les vérifications automatiques et les vérifications manuelles, pour vérifier les unités/données douteuses identifiées par les contrôles :

– la vérification automatique est réalisée sans intervention humaine et est en général couplée avec une procédure de traitement par imputation. Ces méthodes automatiques ont l'avantage d'être

reconductibles (et donc de donner le même résultat) ce qui n'est pas toujours le cas de l'expertise manuelle avec deux gestionnaires différents. La méthodologie la plus célèbre est celle de Fellegi et Holt (1976) qui s'appuie sur une liste prédéfinie de contrôles pour détecter les unités douteuses et sur une méthode visant à minimiser le nombre de variables à modifier et à imputer afin que le nouveau jeu de données « passe » l'intégralité de la chaîne de contrôles. De nombreux algorithmes ont été proposés pour localiser l'erreur à partir du paradigme de Fellegi et Holt, une des difficultés étant la définition, particulièrement chronophage, de l'ensemble des contrôles[2] à considérer dans le problème global de minimisation. Cette méthodologie a été étendue pour couvrir les données quantitatives continues (De Waal et Coutinho, 2005) et a été implémentée dans de nombreux logiciels utilisés par les instituts de statistiques. Par exemple, Statistique Canada utilise à ce jour le système généralisé BANFF qui offre des méthodes pour vérifier et imputer des données d'enquêtes sous la forme de neuf procédures statistiques SAS, la détection du nombre minimal de données à modifier reposant sur le paradigme de Fellegi et Holt (voir Kozak, 2005 pour la première version du système BANFF).

– la vérification manuelle met en jeu des gestionnaires d'enquête qui expertisent le questionnaire et contactent si nécessaire l'entreprise. Ils peuvent alors soit modifier directement la valeur initiale soit la confirmer en fonction des éléments dont ils disposent. En cas de doute, ils peuvent aussi mettre la valeur « à blanc » et la donnée sera imputée comme une donnée manquante dans un processus d'imputation générale des non-réponses. Une partie des modifications apportées par les gestionnaires sont liées aux « unités » observées, les entreprises, qui peuvent changer de contour, se déformer dans le temps, se « restructurer » par transfert d'activités. Les gestionnaires interviennent également dans le cadre de relances ciblées des unités non répondantes, afin d'avoir des données collectées les plus représentatives possibles.

Pour combiner les vérifications automatiques et les vérifications manuelles (démarche appelée *selective editing*), l'idée de base consiste à s'appuyer sur un classement des unités à traiter selon un certain critère, et à définir un seuil au-dessus duquel les unités seront contrôlées « à la main » par un gestionnaire (et dans certains cas recontactées) et au-dessous duquel les données feront l'objet d'un processus de vérification automatisé (voir parties 3.4. sur le seuil et 3.5. sur le critère d'arrêt).

3.3. Le classement des unités basé sur l'« importance » des erreurs potentielles

De façon générale, la hiérarchisation des unités repérées comme douteuses par les contrôles est liée à l'importance relative des données et/ou des unités les unes par rapport aux autres dans le ou les agrégats contrôlés. Plusieurs méthodes existent pour définir le critère de classement.

Une première méthode simpliste, pour les enquêtes entreprises, consiste à classer les unités selon leur taille en termes d'effectifs ou de chiffre d'affaires. Mais celle-ci ne permet pas de relier l'ordre du classement avec l'importance des erreurs potentielles sur la qualité obtenue. Pour en tenir compte, on peut par exemple regarder la contribution de chaque unité à une statistique donnée de type ratio en calculant la statistique qu'on obtiendrait avec l'ensemble de l'échantillon sans cette unité. Plus la statistique obtenue sans l'unité est différente de celle obtenue avec l'ensemble des unités, plus cette unité est considérée comme importante à contrôler.

Latouche et Berthelot (1992) proposent la fonction « DIFF » qui repose, pour une variable donnée, sur la différence pondérée entre la donnée brute fournie par l'unité et la valeur « attendue » : cette

2 L'ensemble des contrôles à prendre en compte sont les contrôles de départ appelés contrôles explicites ainsi que ceux dits implicites qui en sont dérivés. Illustrons ces concepts sur un exemple. Considérons trois variables, AGE avec 2 modalités (< 16 ans et ≥ 16 ans), STATUT avec 2 modalités (marié et non marié) et LIEN avec 3 modalités (époux, enfant, autres) ainsi que les deux contrôles explicites suivants « vérifier que si AGE < 16 ans alors STATUT = non marié » et « vérifier que si STATUT = non marié alors LIEN ≠ époux ». Ces deux contrôles explicites conduisent au contrôle implicite suivant : « vérifier que si AGE < 16 ans alors LIEN ≠ époux ».

différence sera utilisée comme estimation de l'impact attendu de la vérification de la variable sur un agrégat cible. Plus la valeur de la fonction « DIFF » est grande, plus cette unité est considérée comme importante à contrôler. Hesse (2005) fournit une formalisation théorique probabiliste de l'approche de Latouche et Berthelot (1992) ainsi qu'une généralisation de leur approche. Dans l'idéal, la variable « attendue » correspondrait à la « vraie » valeur qui peut être approximée par la valeur retenue par le gestionnaire. Comme celle-ci n'est bien évidemment pas connue à l'avance, on utilise un proxy, par exemple la valeur de la même variable à la précédente enquête (éventuellement multipliée par une évolution moyenne) ou la valeur médiane (ou moyenne) de la variable pour sa catégorie définie par exemple par le croisement de secteur d'activité et de tranche de taille à laquelle appartient l'unité. Lawrence et Mc Kenzie (2000) indiquent d'ailleurs que les contrôles retenus peuvent permettre de déterminer une valeur attendue cible (« expected amended value »). Par exemple, avec un contrôle de vraisemblance où on vérifie si une donnée est comprise entre les deux valeurs A et B, la moyenne de ces deux valeurs peut constituer une valeur attendue cible.

La fonction DIFF présentée ci-dessus repose sur une seule statistique donnée et donne pour chaque unité un score appelé « score local » pour cette statistique. Or, le travail de vérification par les gestionnaires se fait généralement sur l'ensemble du questionnaire. De plus, pour une enquête donnée, on diffuse plusieurs statistiques ou la même statistique calculée sur différents niveaux géographiques (national, régional, départemental par exemple) ou sur différents domaines. Bien évidemment, plus on diffuse à un niveau fin, plus on a d'unités à vérifier. Dans ce cas, se pose alors la question de savoir comment interclasser les unités sélectionnées par les divers scores locaux afin de pouvoir déterminer des priorités générales de traitement pour les gestionnaires d'enquête. L'idée est de combiner les scores locaux pour déterminer un score global à chaque questionnaire permettant ainsi de les **classer de manière unique** selon l'importance qu'on leur donne quant à la nécessité de les contrôler de façon approfondie. Pour obtenir ce score global, plusieurs combinaisons des scores locaux peuvent être envisagés : Latouche et Berthelot (1992) suggèrent la somme pondérée (ou non) des scores locaux selon l'importance subjective que l'on accorde aux variables, Hedlin (2003, 2008) indique le maximum des scores locaux et Farwell (2005) propose une distance euclidienne. Il est nécessaire de standardiser au préalable les scores locaux afin qu'ils soient invariants par changement d'unité des variables et donc comparables d'une variable à une autre.

La mise en œuvre de ces méthodes est loin d'être évidente, en particulier dans le cas d'unités présentant des données manquantes (Brilhault et Brion, 2008) ou encore pour classer judicieusement de manière unique. Des méthodes de *selective editing* pour donner des priorités dans le travail de vérification par les gestionnaires ont été introduites dans le cadre du dispositif d'Élaboration des statistiques annuelles d'entreprises (Esane) mis en oeuvre en 2014 à l'Insee (Beguin et Haag, 2017). A la mise en production, au fur et à mesure des réponses des entreprises, les scores choisis se sont révélés très instables en début de la collecte, ce qui a conduit à les modifier pour les rendre plus robustes. Par ailleurs, le processus de validation a été revu afin de porter sur une plus grande diversité de variables utilisées. Ainsi, de nouveaux indicateurs ont été mis au point pour détecter également des erreurs importantes sur des variables secondaires, non pris en compte dans les indicateurs initialement conçus.

Les méthodes présentées ci-dessus pour classer les unités identifiées comme douteuses ont une particularité commune, elles reposent sur une mesure de l'impact de l'unité sur la **valeur de l'agrégat**. Dans l'idéal, il serait plus pertinent d'estimer l'impact des modifications apportées sur une unité donnée sur la précision en termes d'erreur quadratique moyenne (qui correspond à la somme de la variance et du biais au carré) de l'agrégat considéré. Cette mesure de la précision devrait tenir compte des composantes classiques comme l'échantillonnage et la non-réponse mais également du "gain potentiel en qualité" apporté par les opérations de vérifications manuelles,

c'est-à-dire l'écart entre l'agrégat estimé avec les données disponibles (une partie d'entre elles ayant été vérifiées et traitées par les gestionnaires et l'autre partie pas encore) et une valeur de référence de cet agrégat. La difficulté de cette approche réside dans le fait que la précision est compliquée à estimer (Rivière, 2002b), cette précision devant pouvoir être estimée à tout moment du traitement de l'enquête. Cette approche nécessite également de définir une valeur de référence pour l'agrégat, cette valeur pouvant être choisie à partir d'un jeu de données sans non-réponse, jugées correctes ou corrigées exhaustivement par les gestionnaires ou encore par modélisation.

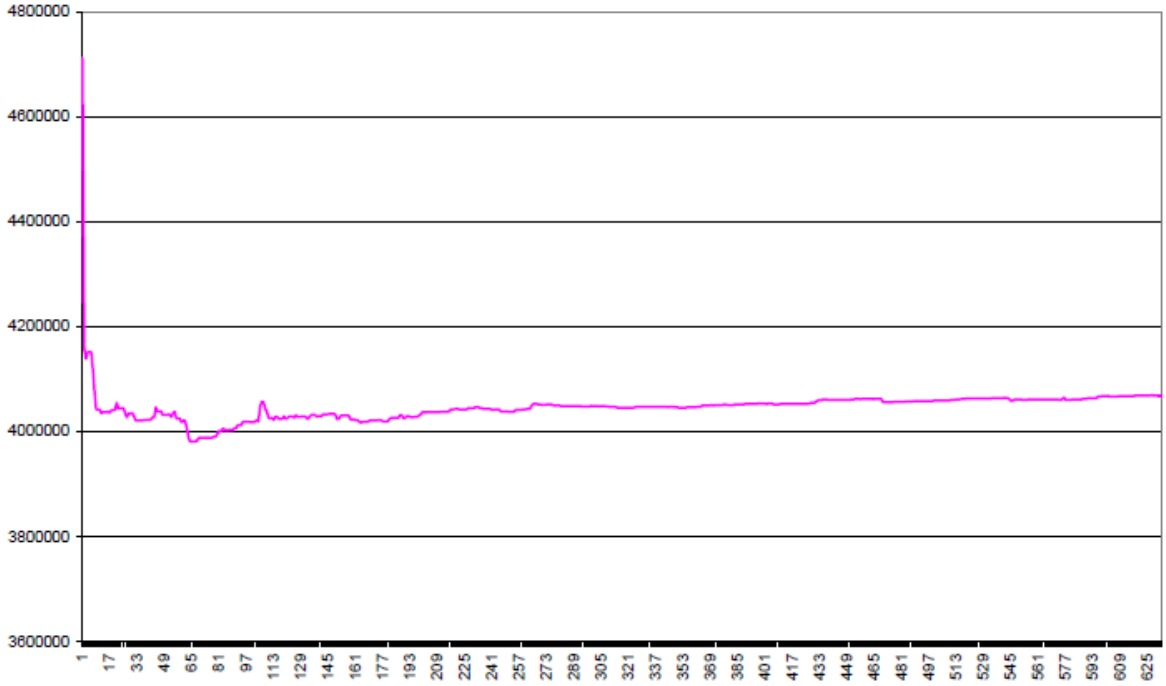
3.4. Le choix du seuil, jusqu'où pousser l'expertise manuelle ?

Une fois la priorisation réalisée par une des méthodes présentées dans la partie 3.3., il faut comme on l'a déjà signalé ci-dessus déterminer le seuil divisant l'ensemble des unités en deux parties : les unités pour lesquelles la vérification est automatique (et peut conduire soit à conserver la valeur brute, soit à la corriger par une procédure automatique d'imputation) et les unités pour lesquelles la vérification est manuelle. En théorie, le seuil doit être choisi afin d'éviter les traitements qui ne sont pas nécessaires au sens où les statistiques ou indices construits se stabilisent et ne varient plus en fonction des traitements manuels réalisés.

Les traitements devraient donc s'arrêter lorsque la qualité souhaitée des indicateurs est atteinte et les délais de production respectés. Dans le cadre du Handbook d'Eurostat concernant l'arbitrage entre précision et délais dans les enquêtes statistiques, Ph. Brion (2007) présente le résultat d'une simulation réalisée à partir des données de l'enquête structurelle annuelle. La figure 1 donne l'évolution de l'estimation du chiffre d'affaires du secteur du commerce de détail en fonction du nombre d'unités vérifiées manuellement, celles-ci étant classées en fonction de la valeur d'un score de priorité basé sur leur impact potentiel sur l'agrégat considéré. Pour les unités en vérification automatique, ce sont les données brutes qui sont utilisées et pour les données en vérification manuelle, ce sont les données modifiées ou confirmées par le gestionnaire qui sont utilisées. Par conséquent, sur la gauche de la figure, peu d'unités sont vérifiées manuellement et l'estimation est donc principalement basée sur des données brutes. À l'inverse, sur la droite, la plupart d'entre elles sont vérifiées manuellement. Plus précisément, le point en abscisse k sur le graphique a pour ordonnée la valeur de la statistique obtenue en la calculant avec les données vérifiées manuellement et modifiées par les gestionnaires pour les k premières unités (selon l'ordre donné par le score) et les données brutes pour toutes les autres unités. Cette figure montre que, dans ce cas, la vérification manuelle de moins de la moitié des unités suffit à produire une estimation robuste.

Figure 1

Example of a simulation conducted on data of the French annual enterprise survey for retail trade



Note de lecture : initialement, l'agrégat du chiffre d'affaires du commerce s'élève à 4,8 Mds€ quand il est calculé avec les données directement collectées. Les changements opérés sur les unités légales suite aux vérifications par les gestionnaires sont représentés par ordre d'influence sur l'agrégat. Les 65 premiers changements font baisser l'agrégat à 4,0 Mds€. Au bout de la 300ème modification, l'agrégat s'accroît jusqu'à 4,05 Mds€ et n'évolue quasiment plus par la suite.

3.4. Dans l'idéal, un seuil et un critère d'arrêt

Classer et hiérarchiser les vérifications et les unités à contrôler est une première étape, définir les unités qui nécessiteraient une vérification par les gestionnaires (le seuil) en est une seconde. En général, comme on l'a vu précédemment, le seuil dépend de l'impact des unités sur la valeur des agrégats considérés et le processus de *data editing* prend fin lorsque les gestionnaires ont terminé l'ensemble des vérifications qu'ils devaient faire.

Cette pratique risque cependant d'être trop coûteuse sans pour autant avoir un gain en qualité en rapport avec le coût – comme on l'a vu sur l'exemple précédent – ou d'avoir une durée trop longue par rapport à des impératifs de production comme pour les enquêtes réalisées dans le cadre de règlements européens avec de fortes contraintes de délais.

Dans une optique d'optimisation et d'efficience des moyens des gestionnaires et/ou de prise en compte des délais de l'enquête, il peut être intéressant de mettre en place un « critère d'arrêt » au processus de *data editing* manuel, c'est-à-dire un critère de décision qui serait objectif et calculable à tout moment de la phase de vérification, permettant de décider s'il est ou non nécessaire de continuer le travail de vérification par les gestionnaires. Définir un critère d'arrêt permet de réduire les coûts et les délais, mais il ne faut pas que ce soit au détriment de la qualité des indicateurs produits. L'idée générale est donc d'arrêter le processus de *data editing* manuel

dès lors que les unités qui n'auront pas été traitées par les gestionnaires ont un impact limité. Le critère naturel qui s'impose est un critère de précision estimant l'impact des modifications apportées à une unité donnée sur la précision en termes d'erreur quadratique moyenne de l'agrégat considéré (Riviere, 2002b). Ainsi, dès que la précision atteindrait une valeur-cible que l'on aurait définie préalablement, le processus de *data editing* serait terminé. S'il reste des questionnaires identifiés comme à traiter par les gestionnaires, ceux-ci peuvent basculer en reprise automatique ou rester identiques.

Le fait d'avoir une approche en termes d'erreur quadratique moyenne permet également de savoir mesurer le gain « perdu » en qualité s'il reste des données en reprise manuelle à traiter lorsque l'on est dans la nécessité d'arrêter le processus de *data editing* plus tôt que prévu pour des raisons de délai.

La priorisation des unités ainsi que la détermination d'un critère d'arrêt supposent que les objectifs de l'enquête en termes de qualité des données produites soient bien définies : quelles sont les variables cibles ? à quel niveau (région ? regroupement d'activités ?) seront-elles diffusées ? Et avec quelle précision souhaitée ? Les réponses à ces questions sont souvent très difficiles à obtenir. Il est évident que plus la finesse de diffusion est importante, plus le coût de vérification sera élevé : ainsi, une diffusion au niveau zone d'emploi plutôt qu'au niveau régional nécessite plus de reprises manuelles en *data editing*. Dans ce cadre, la définition du seuil et le choix du critère d'arrêt doivent tenir compte de tous les objectifs en termes de diffusion attendue. Il existe une abondante littérature sur l'optimisation de la phase de reprise manuelle, permettant ainsi de veiller à l'efficience des moyens de gestionnaires en définissant un critère d'arrêt des reprises manuelles en lien avec le degré de finesse des agrégats publiés et le niveau de qualité attendu (voir par exemple Rivière, 2002a, 2002c, Hesse, 2005).

Partie 4. Vers une stratégie globale de *data editing*

4.1. La mise en place d'une stratégie globale articulant les différents contrôles

Comme on l'a vu précédemment, un processus de *data editing* est basé sur une succession de contrôles et de vérifications de plusieurs types :

– des contrôles lors de la collecte des données ;

– des contrôles visant à vérifier la validité, la cohérence et la vraisemblance des enregistrements de données individuelles (les différents types de micro-contrôles) ;

– des contrôles visant à mesurer la contribution d'une unité sur un agrégat (macro-contrôles) au cours du processus de production ;

– des procédures qui ne ciblent que certaines unités ou variables qui seront vérifiées manuellement par des gestionnaires, celles-ci hiérarchisent le travail manuel et fixent des critères d'arrêt (*selective editing*) ;

– des vérifications (et des corrections) manuelles ou automatiques ;

– des ultimes vérifications des agrégats, soit par rapport à ceux d'autres sources, soit temporellement pour les dispositifs répétés dans le temps (*output editing*).

Définir une stratégie de *data editing* consiste pour une opération donnée à agencer ces différents types de contrôles et de vérifications en tenant compte du calendrier de production des statistiques, des produits (base individuelle ou indices) attendus et de la qualité de ces produits au sens de l'erreur quadratique moyenne, des moyens disponibles (et donc de l'arbitrage entre méthodes automatiques et manuelles).

Une modélisation du processus de *data editing* qui intègre également la partie imputation est présentée dans le manuel Edimbus de 2007 (voir figure 2 et partie 5.3.). Elle se décompose en plusieurs étapes. La première phase consiste à réaliser des premiers contrôles comme les contrôles de validité et ceux de cohérence et des premières corrections des données par imputation (phase « Initial E&I »). La seconde qui est axée sur la détection des valeurs influentes en calculant par exemple un score (phase « Influential Error ») sépare les unités qui seront examinées par des gestionnaires (phase « interactive E&I ») de celles qui seront traitées de manière automatique (phase « automatic E&I »). La phase suivante consiste à réaliser des macro-contrôles (phase « macro E&I ») et de repérer les unités ayant un impact important sur les agrégats qui sont alors renvoyées en expertise automatique ou manuelle. La dernière phase consiste à calculer les agrégats et à les comparer à ceux issus d'autres sources ou calculés lors de l'enquête précédente (phase « suspicious agregate »).

Dans le cas d'estimations suspectes, les unités sous-jacentes (les plus influentes) peuvent à nouveau être identifiées et vérifiées de manière interactive, généralement de manière sélective. Les données identifiées à ce stade correspondent à des cas qui n'ont pas été correctement identifiés ou traités lors des étapes précédentes du processus.

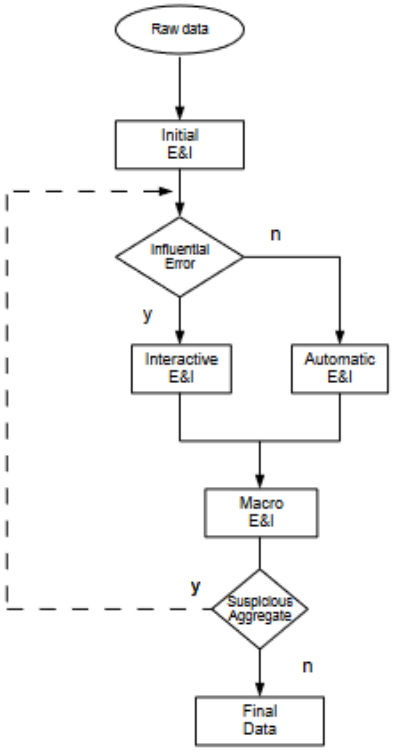


Figure 2 : *Exemple de processus général de data editing (Edimbus, 2007)*

***Note de lecture** : les différentes phases sont explicitées dans le paragraphe au-dessus de cette figure. Le terme « E&I » qui revient souvent correspond à « Edit and imputation » en anglais, soit « vérification des données et imputation » en français.*

Cette modélisation qui donne une vue générale du processus dans son ensemble nécessite d'être adaptée selon les enquêtes. Il est classique de distinguer les enquêtes ménages des enquêtes entreprises mais également de distinguer au sein des enquêtes entreprises les enquêtes conjoncturelles des enquêtes structurelles (voir annexe 1 pour d'autres exemples de processus). En effet, dans les enquêtes structurelles, le nombre de variables essentiellement quantitatives est élevé et on souhaite avoir une base individuelle de qualité. En revanche, pour les enquêtes conjoncturelles, ce sont en général des enquêtes avec peu de variables quantitatives, une forte profondeur historique, un processus de production très contraint avec des délais très courts et on souhaite disposer d'indices à des niveaux agrégés (donc avec moins d'exigence que pour une enquête structurelle sur la qualité de la base individuelle). Cependant, cette frontière est ténue car dans certains cas, il arrive que pour une enquête conjoncturelle on souhaite également avoir une base individuelle de qualité, en particulier lorsque celle-ci repose sur le traitement d'une source administrative qui est ensuite mise à disposition après traitement. Les enquêtes thématiques réalisées auprès des entreprises contiennent en général beaucoup de variables qualitatives (et donc sont plus proches des enquêtes ménages).

4.2. Conserver des informations sur le processus de *data editing* : une nécessité

Dans le fichier final au niveau « individu statistique[3] », il est très important de conserver les données brutes et de bien repérer l'ensemble du cycle de la donnée avec la valeur brute renseignée par le répondant à l'enquête, la valeur modifiée par le gestionnaire ainsi que la valeur

3 Par exemple, les établissements ou les entreprises pour les enquêtes économiques.

redressée et de savoir quelle méthode a été utilisée. Il est également très utile de disposer des données « avant / après » sur les unités qui sont traitées et redressées de façon automatique.

Disposer dans une même base des différentes valeurs prises par une donnée depuis sa création lors du remplissage par le répondant à l'enquête jusqu'à la valeur utilisée et fixée pour la diffusion facilite les études méthodologiques sur le processus de *data editing.* Les études possibles sont très nombreuses, elles peuvent à titre d'exemples viser à déterminer l'impact sur la qualité des données/indicateurs diffusés de la phase de *data editing* dans son ensemble, de la phase de *data editing* concernant les reprises manuelles par les gestionnaires ou encore permettre de regarder sur un échantillon de données redressées automatiquement ce que donnerait une expertise par un gestionnaire par rapport à l'expertise automatique et vice versa. Cette base peut aussi permettre de repérer des erreurs trop systématiques et ainsi s'interroger sur la pertinence du libellé des questions ou de la place du contrôle les repérant dans la chaîne de production, ce qui contribue ainsi à une amélioration en continu du processus global de l'enquête.

En dehors de ce fichier final contenant le cycle de vie de toutes les données, il est également important de disposer systématiquement (après chaque production) d'informations sur la phase de *data editing* réalisée (contrôles mis en place, critère d'arrêt retenu, etc.) ainsi que d'indicateurs synthétiques qui permettent d'évaluer la qualité des statistiques produites, mais aussi d'optimiser le processus concerné (si celui-ci est amené à être reconduit).

On peut par exemple penser à trois types d'indicateurs synthétiques (voir Brion, 2006) :

- des indicateurs sur les données modifiées ou absentes. Par exemple, pour chaque variable : taux de non-réponse, proportion d'unités pour lesquelles la valeur finale est différente de la valeur brute, distribution avant et après modification des variables quantitatives, etc. et pour chaque agrégat, proportion d'unités modifiées ou imputées ;
- des indicateurs relatifs au processus de vérification des données. Par exemple, pourcentage d'unités avec au moins un contrôle invalidé, par variable : pourcentage d'unités traitées de façon automatique et de façon manuelle, pourcentage d'erreurs détectées (rapportées au nombre de données contrôlées) ;
- des indicateurs sur les moyens. Par exemple, le nombre de « personnes jours » réalisant des vérifications manuelles, la durée totale de la phase de production, celle du *data editing* ainsi que sa répartition dans le temps, le nombre de contacts avec les unités posant problème.

4.3. Distinction entre unités influentes et unités « en erreur »

Tout le processus de *data editing* basé sur un ensemble de contrôles consiste à détecter des données jugées douteuses. Mais les données douteuses ne sont pas nécessairement toutes des données « en erreur » qui doivent être obligatoirement corrigées parce qu'elles seraient fausses. Par exemple, si dans l'échantillon d'une enquête sur les revenus, on a sélectionné un footballeur professionnel, le montant de son revenu risque d'être identifié comme une donnée douteuse et pourtant il est exact.

Parmi les unités qui ont des valeurs douteuses ne passant pas les contrôles de vraisemblance, il est donc important de distinguer les unités « en erreur » avec des données fausses des unités non « en erreur ». Ces deux types d'unités ont un effet sur la qualité des estimations, cet effet étant d'autant plus important que les unités ont une forte contribution à un agrégat donné.

On définit les **unités influentes** (voir séminaire Insee de méthodologie statistique de 2013 et en particulier l'intervention de D. Haziza) appelées également par Chambers (1986) « unités aberrantes représentatives » comme des unités dont les valeurs collectées sont correctes mais

hors normes (donc ayant une forte contribution) et dont on ne peut supposer qu'elles sont uniques. En d'autres termes, il n'y a aucune raison valable de supposer qu'il n'existe pas d'autres valeurs similaires dans l'ensemble de la population concernée si cette unité a été échantillonnée car appartenant à la strate non exhaustive. Dans le cas de l'estimation d'un paramètre comme un total, ces unités ont une importance considérable dans l'estimation de celui-ci et on ne peut pas se permettre de leur mettre un poids « égal » à 1, car cela reviendrait à les considérer comme uniques. Les valeurs influentes sont problématiques car elles mènent généralement à des estimateurs instables, c'est-à-dire des estimateurs ayant une grande variance.

Ce problème survient particulièrement dans les enquêtes auprès des entreprises qui collectent des variables économiques dont les distributions sont fortement asymétriques. De plus, le plan de sondage est en général un plan de sondage stratifié selon la taille de l'entreprise et son secteur d'activité. Par conséquent, les poids des unités sont hétérogènes et tout décalage éventuel entre les données de la base de sondage utilisées pour construire les strates de tirage et les données recueillies lors de l'enquête (évolution à la hausse des effectifs et/ou changement de secteur d'activité ce qui fait que l'entreprise n'est pas classée dans sa « bonne strate » et peut avoir un poids trop élevé) est propice à l'apparition d'unités influentes dans l'échantillon. Même s'il est possible de minimiser leur effet au moyen d'un plan de sondage approprié en utilisant un plan stratifié comportant une strate exhaustive, il est généralement impossible d'éliminer complètement le problème des unités influentes. Ce problème peut également apparaître dans les enquêtes ménages portant sur les distributions de revenus ou de patrimoines des ménages. Ces unités influentes nécessitent donc des méthodes d'estimation robustes et adaptées afin de réduire leur impact, comme la winsorisation (voir encadré, séminaire Insee de méthodologie statistique de 2013 (intervention de D. Haziza), Deroyon, 2015, Favre-Martinoz, Beaumont, et Haziza, 2015). Le principe général de ces méthodes est de modifier les poids de sondage des unités influentes afin d'avoir des estimateurs plus stables. La modification des poids introduit du biais mais la précision globale obtenue des estimateurs en termes d'erreur quadratique moyenne (somme de la variance et du biais au carré) est significativement inférieure à celle des estimateurs que l'on a sans traitement particulier des données influentes. Les méthodes de winsorisation sont utilisées à l'Insee dans le cadre de l'Élaboration statistiques annuelles d'entreprises (Esane) et des enquêtes thématiques entreprises.

Les **unités « en erreur »,** qui sont les autres types d'unités douteuses, sont des unités qui contiennent une ou plusieurs donnée(s) incorrecte(s), ces données étant corrigées soit par des procédures classiques d'imputation de manière automatique soit par les gestionnaires d'enquêtes avec dans certains cas un contact avec l'enquêté (voir partie 3.). Comme indiqué dans la partie 1., les plans de sondage des enquêtes entreprises sont généralement des plans de sondage stratifiés avec les entreprises les plus importantes en termes d'effectifs et/ou de chiffre d'affaires dans une strate exhaustive. Il peut arriver dans la phase de *data editing* qu'il soit difficile de repérer les erreurs pour certaines unités et par conséquent de les corriger de façon adéquate. Dans les strates échantillonnées, une solution consiste à « traiter » ces unités dans la phase d'estimation en considérant qu'elles sont uniques et en leur diminuant arbitrairement leur poids à 1. Avec cette nouvelle pondération, l'impact des réponses de cette unité dans les agrégats devient faible.

Les méthodes de winsorisation

Si les valeurs déclarées pour une variable d'intérêt dépassent un certain seuil et qu'elles sont considérées comme influentes, celles-ci sont « rabotées » et on obtient des valeurs winsorisées. Le total de la variable d'intérêt est alors estimé comme la somme pondérée des valeurs winsorisées.

La winsorisation introduit un biais dans les estimateurs, en modifiant les réponses des entreprises, mais diminue la variance des estimations.

La méthode standard de winsorisation consiste à choisir un seuil K appelé seuil de winsorisation (K>0) et à définir pour chaque unité i de l'échantillon la valeur de la variable Y pour cette unité après winsorisation par :

$y_{iw} = y_i$ si $d_i \otimes y_i \leq K$ (où d_i = poids de sondage)

$= K / d_i$ si $d_i \otimes y_i > K$

L'estimateur retenu pour le total de la variable Y est donc $\hat{Y}_w = \sum_{i \in \text{échantillon}} \left(d_i \, y_{iw} \right)$

Le choix du seuil K est déterminant pour les propriétés de l'estimateur $\hat{Y}_w$. Il détermine si l'arbitrage biais-variance qu'introduit la winsorisation permet, au final, un gain réel en précision en termes d'erreur quadratique moyenne (correspond à la somme de la variance et du biais au carré). Ainsi, un mauvais choix du seuil K peut conduire à un estimateur moins précis que l'estimateur « naturel » d'Horvitz-Thompson.

Une méthode pour choisir K est de minimiser l'erreur quadratique moyenne estimée de l'estimateur robuste. On peut citer par exemple la méthode de Kokic et Bell (1994) dans le cas d'un sondage aléatoire stratifié à un degré. Ces méthodes nécessitent de l'information historique et/ou un modèle, elles sont basées sur des hypothèses simplificatrices et sont en général complexes à implémenter.

On peut également écrire l'estimateur sous une forme alternative faisant apparaître la winsorisation non pas comme une modification de la valeur des individus mais comme une modification de la pondération.

On a donc

$$\hat{Y}_w = \sum_{i \in \text{échantillon}} \left(d^*_i \, y_{iw} \right)$$

avec $d^*_i = d_i \times [\min(y_i ; K/d_i) / y_i]$

Avec cette écriture, on voit que seule la pondération des unités influentes est modifiée, celle des unités non influentes ne l'est pas. De plus, la modification conduit à des pondérations inférieures à 1 pour les unités concernées, ce qui peut être problématique car une unité se représente *a minima*. Pour pallier ce problème, il existe d'autres types de winsorisation comme celle de Dalèn-Tambay, appelée winsorisation de type 2. Les deux types de winsorisation donnent en général des résultats proches (voir Chambers, R. L., 1986)

Partie 5. Le contexte international, un cadre dynamique sur le *data editing* depuis le milieu des années 90

À partir du milieu des années 90, les travaux de recherche appliquée sur le *data editing* se développent au niveau international. La théorie et la pratique du *data editing* se sont ainsi formalisés grâce à des spécialistes émanant essentiellement du monde de la statistique publique. À ce jour, le *data editing* fait l'objet d'une littérature abondante et très complète (largement anglo-saxonne et très peu française) ainsi que de colloques internationaux réguliers sur le sujet.

Un des ouvrages de référence reste le « *Handbook of Statistical Data Editing and Imputation* » (De Waal et al., 2011) qui aborde de façon complète le *data editing* avec des points théoriques et des applications pratiques. Cet ouvrage décrit également de manière détaillée les différentes méthodes d'imputation que les auteurs estiment très liées à la phase de détection d'erreurs.

5.1. Un groupe international sur le *data editing* organisé par la Commission économique des Nations unies pour l'Europe (CEE-ONU) qui se réunit depuis plus de 30 ans

Le groupe de haut niveau pour la modernisation des statistiques officielles (HLG-MOS – High Level Group for the Modernisation of Official Statistics) de la Commission économique pour l'Europe des Nations unies (CEE-ONU et en anglais Unece pour *United Nations Economic Commission for Europe*) coordonne, comme son nom l'indique, les travaux internationaux relatifs à la modernisation des statistiques et encourage leur développement sur la base de standards de référence. Ainsi, sous l'égide de ce groupe, plusieurs référentiels ont été développés dont le modèle générique d'activité des organisations statistiques (GAMSO), le modèle générique de processus statistique (GSBPM), et le modèle générique d'information statistique (GSIM). Il existe aussi un modèle générique de *data editing* (GSDEM). Collectivement, ces modèles sont appelés les « modèles ModernStats ».

Ce groupe se réunit chaque année en workshop pour promouvoir de nouveaux développements, faciliter le partage d'expériences et d'idées de modernisation de la production statistique et identifier les opportunités de collaboration internationale pour l'année suivante.

Le *data editing* a été identifié par l'Unece comme un sujet de collaboration internationale il y a plus de 30 ans. Depuis les années 90, il existe toute une communauté internationale d'experts sur le *data editing* qui se réunissent 2 jours tous les 18 mois environ dans un pays différent (sauf les éditions de 2020 et 2022 qui étaient réalisées à distance). C'était dans le cadre d'une « work session on statistical data editing » dans les années 1990-2000 puis à compter de 2017 d'un « workshop on statistical data editing » et enfin depuis 2022 d'un « expert meeting on statistical data editing ». La session de 2014 s'est d'ailleurs déroulée à Paris du 28 au 30 avril et était pilotée par la Direction des statistiques d'entreprises (DSE) de l'Insee, la dernière s'est déroulée à Vienne du 7 au 9 octobre 2024. Ce groupe vise à identifier de nouvelles méthodes susceptibles d'améliorer la qualité et l'efficacité du *data editing* et à en faciliter le partage d'expériences. Environ une cinquantaine de personnes assistent à chaque réunion avec une vingtaine de personnes spécialistes de la statistique publique très actives et présentes régulièrement. Le fait qu'il n'y ait aucune session en parallèle facilite les échanges.

Les réunions de ce groupe donnent lieu à de nombreuses présentations de pays ainsi qu'à des documents accessibles en ligne (https://statswiki.unece.org/display/sde/Statistical+Data+Editing) sur le site de l'Unece. Sur les thèmes abordés, outre ceux relatifs aux améliorations des méthodes

traditionnelles de *data editing*, on y trouve également depuis quelques sessions des présentations d'exemples de mise en œuvre de nouvelles techniques comme l'utilisation du *Machine Learning* en intelligence artificielle pour le *data editing* et l'imputation.

Par ailleurs, Eurostat a lancé en 2024 un appel d'offre visant à créer un centre de ressources européen sur l'intelligence artificielle/le machine learning (IA/ML) pour les statistiques officielles. Dans celui-ci, un des groupes de travail porte spécifiquement sur les possibilités d'utilisation de ces méthodes en termes d'automatisation, d'efficacité et d'amélioration de la qualité du processus de *data editing* (voir partie 5.7.).

5.2. Des ouvrages de référence sur le *data editing* dès le milieu des années 90

Au début des années 90, la littérature scientifique est moins abondante sur la méthodologie des enquêtes réalisées auprès des entreprises que sur celle des enquêtes réalisées auprès des ménages. Les pratiques sont également très variables d'un pays à un autre. Ce constat partagé a conduit Statistique Canada et l'American Statistical Association (ASA) à organiser une conférence internationale spécifique appelée « International Conference on Establishment Surveys (ICES) » en juin 1993 à Buffalo qui avait les trois objectifs suivants : présenter des méthodes utilisées, présenter de nouvelles approches et permettre les échanges. Cette conférence a une caractéristique inhabituelle au sens où elle s'inscrit dans deux séries de conférences, l'une organisée par Statistique Canada (elle correspond au 10ème Symposium sur les questions de méthodologie d'enquête) et l'autre organisée par l'ASA (elle correspond à la 4ème Conférence sur les méthodes d'enquête). Cette conférence a donné lieu à un ouvrage appelé « Business Survey Methods » de plus de 700 pages (Cox, B.G. et al., 1995) regroupant des articles à contenu transversal en six parties, les cinq premières parties reprenant les principales étapes du processus de réalisation des enquêtes entreprises, la dernière portant sur une vision historique et prospective. Cet ouvrage contient des articles sur le *data editing* (en particulier des articles de Granquist, de Pierzchala et de Kovar), mais ceux-ci ne sont pas très nombreux.

A la même époque, l'Unece rassemble dans deux ouvrages des articles de référence sur le *data editing* dans le cadre du projet sur le *data editing* lancé par la Conférence des statisticiens européens, ces deux ouvrages contenant explicitement le terme *data editing* dans leur titre (Conference of European Statisticians Statistical Standards and Studies – no. 44, Statistical Data Editing). Le premier volume (Unece, 1994) insiste particulièrement sur l'intérêt du *macro editing* pour rationaliser et optimiser le processus de contrôle des données ainsi que sur sa mise en œuvre. Le second volume (Unece, 1997) détaille les techniques efficaces pour le *data editing* avec en particulier le *selective editing* et aborde la question de l'évaluation de cette phase. Ces deux ouvrages contiennent une bibliographie très fournie.

Quelques années plus tard, en 2011, Wiley édite le « *Handbook of Statistical Data Editing and Imputation* » qui aborde de façon complète le *data editing* avec des points théoriques et des applications pratiques ainsi que les principales méthodes d'imputation. Cet ouvrage reste à ce jour un des ouvrages de référence sur le sujet.

5.3. Des préconisations harmonisées sur le *data editing* dès 2007

En 2007, un manuel sur le *data editing* pour les enquêtes entreprises (ISTAT, CBS and SFSO, 2007) a été réalisé dans le cadre du projet soutenu par Eurostat « Pratiques recommandées pour l'édition et l'imputation dans les enquêtes transversales auprès des entreprises » appelé EDIMBUS. Ce manuel a été coordonné par l'Institut national italien de statistique (ISTAT) avec la participation du Centraal Bureau voor de Statistiek Netherlands (CBS) et de l'Office fédéral suisse de statistique (OFS).

Ce manuel constitue une grande avancée dans la stratégie du *data editing* en fournissant un cadre de travail, assorti de recommandations d'ordre général sur les bonnes pratiques à mettre en œuvre, y compris sur les méthodes d'imputation, aussi bien dans le domaine des statistiques structurelles que dans celui des statistiques conjoncturelles. De manière générale, il contribue à améliorer le rapport coût-efficacité du processus de *data editing* et les délais de production tout en maintenant la qualité attendue. Cependant, les auteurs insistent dès l'introduction sur le fait que les recommandations qui y sont formulées ne s'appliquent pas toutes de la même manière à tous les contextes d'enquête ; ainsi, leur pertinence et leur applicabilité réelle doivent être soigneusement évaluées par des experts, en tenant compte des objectifs, de l'organisation et des contraintes de l'enquête.

Des pays comme l'Espagne se sont emparés de ce manuel dans le cadre de leurs enquêtes et ont ainsi enrichi la stratégie générale, par exemple en faisant mieux apparaître la phase de contrôles à la collecte, en identifiant les types de contrôles réalisés (Rama, S. et Salgado, D., 2014).

5.4. Dans le Generic Statistical Business Process Model – GSBPM –, la phase de *data editing* est imparfaitement identifiée

Le GSBPM (Generic Statistical Business Process Model) est un modèle générique de description de processus de production statistique que les instituts de statistique suivent pour produire des statistiques officielles. Celui-ci est conceptualisé en 8 phases (définition des besoins, conception, élaboration, collecte, traitement, analyse, diffusion, évaluation), chacune de ces phases étant composée de sous-processus. Il est indépendant de la source de données et peut être utilisé quelle que soit la nature de l'opération, enquête et/ou source administrative.

Développé en 2008 par un groupe de travail commun Unece/Eurostat/OCDE, le modèle est revu périodiquement depuis et la dernière version du GSBPM date de 2019[4].

Dans le GSBPM, on retrouve le processus de *data editing* dans la phase 5 de traitement des données et plus spécifiquement dans le sous-processus 5.3 « Review and validate » (en français « examen et validation » d'après le document en français de l'Unece sur le le GSBPM) ainsi que dans le sous-processus 5.4 « Edit and impute » (en français « édition et imputation des données »). Il est également présent dans la phase 6 d'analyse et plus spécifiquement dans le sous-processus 6.2 « valider les résultats ».

De manière plus détaillée :

- C'est dans le cadre du sous-processus 5.3 que les données sont examinées pour identifier les erreurs potentielles et les problèmes, tels que les valeurs aberrantes, les non-réponses aux questions et les erreurs de codage. Cette étape de validation des données s'effectue avec une phase de contrôles, les données douteuses étant examinées par des gestionnaires ou de manière automatique. Les activités de correction qui modifient les

4 Une nouvelle version du GSBPM (version 5.2) a été adoptée lors de la conférence des statisticiens européens de l'Unece à Genève en juin 2025.

données avec une variété de méthodes pour y parvenir sont répertoriées dans le cadre du sous-processus 5.4.

- Le sous-processus 5.4 comprend la décision de modifier ou non les données identifiées comme douteuses *via* le sous-processus 5.3, la méthode utilisée en cas de modification, l'ajout de la nouvelle valeur dans l'ensemble de données et le « marquage » de cette donnée comme modifiée ainsi que les métadonnées associées.
- Le sous-processus 6.2 concerne les activités pour valider les résultats/agrégats produits. Celles-ci comprennent en particulier :
  - la comparaison des statistiques obtenues dans le temps lorsque c'est possible ;
  - la vérification de la cohérence globale des indicateurs par rapport à ceux issus d'autres sources lorsqu'elles existent ;
  - l'utilisation de *macro editing* pour détecter (avant correction éventuelle ou validation) les unités ayant un impact important sur les agrégats.

Cependant, même si le *data editing* est identifié clairement dans les sous-processus du GSBPM cités ci-dessus, c'est un processus à part entière imbriqué dans l'ensemble du processus de production d'une enquête et qui est donc susceptible d'affecter l'ensemble des phases du GSBPM. En effet, certaines règles de validation peuvent être directement dans les outils de collecte afin de limiter les erreurs ou se produire parallèlement aux activités de collecte (par exemple sous-processus 4.3 « réaliser la collecte » et 4.4 « finaliser la collecte »).

5.5. En 2015, création du référentiel spécifique le Generic Statistical Data Editing Model – GSDEM

Lors de la réunion Unece en 2014 (work session on statistical *data editing*, Paris, 28-30 avril 2014), les experts de *data editing* ont éprouvé le besoin de disposer d'un cadre commun élaboré dans un contexte international qui isole spécifiquement le *data editing*. Le compte-rendu de cette réunion indique qu'un des points à réaliser pour la prochaine réunion est :

> *« Development of a common, generic process framework for statistical data editing. This could be done by a task team under the High-Level Group for the Modernisation of Statistical Production and Services, and presented at the next Work Session. »*

Par conséquent, suite à cette réunion, un groupe réunissant des experts de *data editing* de Finlande, d'Italie, de Norvège, des Pays-bas et de la France a été mis en place sous l'égide de l'ONU dans le cadre du groupe de haut niveau sur la modernisation de la production et des services statistiques. La première version du référentiel pour le *data editing* appelé GSDEM (Generic Statistical Data Editing Model) a été diffusée en 2015, celui-ci a été actualisé en 2019.

En donnant des standards, le GSDEM qui est cohérent avec les autres standards internationaux - en particulier le GSBPM - vise à faciliter la compréhension, la communication, la pratique et le développement dans le domaine *data editing*.

Les processus de *data editing* y sont décomposés en une succession d'étapes élémentaires qui peuvent être combinées différemment selon les enquêtes. Chacune de ces étapes a une des 3 finalités ci-dessous et pour chaque finalité on décrit ce que l'on souhaite faire (le pourquoi) et la méthode utilisée (le comment) :

– l'examen, les variables à regarder : on définit ce que l'on souhaite faire et la méthode pour y parvenir, il s'agit de l'étape de description du contrôle envisagé ;

– la sélection des unités : on souhaite sélectionner les unités nécessitant un traitement ultérieur et caractériser le traitement en fonction du problème identifié sur les données (erreur ou valeur aberrante par exemple, validation automatique ou manuelle), et on décrit la méthode pour réaliser ces deux opérations ;

– le traitement, par exemple par imputation et on décrit le type d'imputation.

Cependant, l'application de ce cadre conceptuel est limitée par le fait qu'en pratique, les méthodes mises en œuvre dans un processus de production ne font souvent pas la distinction entre ces trois étapes. En effet, il arrive qu'une méthode combine les fonctions des trois phases en une seule opération. C'est par exemple le cas avec une règle du type IF-THEN : la partie IF contient l'examen sous la forme de l'évaluation d'un contrôle, la sélection est dans la décision que cette règle doit provoquer le traitement d'une ou plusieurs variables (celles spécifiées dans la partie THEN) et le traitement est spécifié par la méthode qui fournit une nouvelle valeur.

Dans la dernière partie du document sur le référentiel GSDEM, on trouve plusieurs exemples de représentation de processus de *data editing* (voir annexe 1) en tenant compte des unités interrogées (entreprises ou ménages), de la nature des variables (quantitatives ou qualitatives), des types de sources (enquêtes et fusions de sources). Les représentations graphiques des processus reprennent les grands principes de celle déjà présente dans Edimbus (ISTAT, CBS and SFSO, 2007) et reprise en figure 2 dans ce document en l'adaptant et en la complétant. La représentation la plus proche de celle dans Edimbus est relative au processus des enquêtes structurelles réalisées auprès des entreprises (A.1). On trouve en outre un exemple pour les enquêtes conjoncturelles d'entreprises (A.2) caractérisées le plus souvent par un faible nombre de variables, des délais de production très réduits et un fort attendu sur la qualité des agrégats. L'enjeu essentiel concerne donc la vérification et le traitement des erreurs ayant un fort impact sur les agrégats. Le troisième exemple (A.3) concerne une enquête ménage dans laquelle on a des données ménages et individus avec deux sous-processus de vérification des données séquentiels le premier au niveau ménage et le second au niveau individus et dans laquelle il n'y a que des données quantitatives. Ce modèle est plus complexe lorsqu'il y a des variables quantitatives et qualitatives. Le dernier exemple (A.4) concerne une opération de fusion de sources où il existe un processus de *data editing* au sein de chacune des sources puis un processus de *data editing* une fois la fusion des sources réalisées.

5.6. Le *data editing* identifié dans le cadre d'assurance qualité européen dès 2011

En parallèle à la mise en place de référentiels par l'Unece, la Commission européenne adopte en 2005 le 1er code des bonnes pratiques de la statistique européenne et le complète en 2009 avec le règlement n°223/2009, relatif aux statistiques européennes.

Au sein du code de bonnes pratiques, la qualité des processus et l'harmonisation des normes et des méthodes est un des principes-clés. La phase de *data editing* est surtout couverte par le 8ème principe qui veille à la mise en place de procédures statistiques adaptées, mises en œuvre tout au long des processus statistiques. Le cadre d'assurance qualité (Quality Assurance Framework of the European Statistical System) qui précise le code des bonnes pratiques aborde explicitement la notion de *data editing*. Dès 2011, celui-ci préconise la « conformité aux normes des techniques d'*editing*, d'imputation et de contrôle de la communication statistique. Les techniques d'*editing*, d'imputation et de contrôle de la communication statistique respectent les règles méthodologiques et les bonnes pratiques, et sont documentées ».

Le règlement européen n°223/2009 impose aux autorités statistiques nationales de transmettre avec les données envoyées à Eurostat des métadonnées ainsi que des rapports qualité[5]. Le contenu de ces rapports s'est progressivement enrichi depuis 2009. En 2021, dans son rapport sur la qualité des statistiques européennes, la Cour des Comptes européenne recommande que les rapports qualité nationaux soient standardisés sur une norme commune et qu'Eurostat procède à des évaluations plus rigoureuses. La Commission a ainsi officiellement recommandé (à travers la recommandation 2023/397) aux INS en 2023 de respecter le cadre de référence SIMS (Single Integrated Metadata Structure Guidelines). La phase de *data editing* est couverte par 2 des 19 sections qui composent le cadre de référence SIMS. On y retrouve des éléments généraux sur les principales opérations de contrôle et validation. Néanmoins, cette documentation reste de nature qualitative, ne comporte que quelques indicateurs de qualité et ne constitue donc pas une base suffisante permettant de comparer les processus entre les pays.

5.7. De nouvelles approches du *data editing*

Comme indiqué dans la partie 5.1., dans les réunions internationales et plus spécifiquement dans celles du groupe sur le *data editing* de l'Unece, on voit apparaître depuis quelques sessions, à côté de présentations d'amélioration des méthodes traditionnelles du *data editing,* des présentations de mise en œuvre de techniques d'intelligence artificielle, en particulier des méthodes de *machine learning* appliquées dans le cadre des vérifications des données. De nombreux INS investissent sur ce sujet.

Plusieurs motivations sont à l'origine des investissements des INS sur de nouvelles approches :

– **Produire des indicateurs avancés rapidement avant la diffusion des indicateurs définitifs.** Dans ce cas, l'idée est de réduire au minimum voire de supprimer la phase d'expertise à la main pour se reposer sur des méthodes de détection et de correction des anomalies totalement automatisées afin de publier très rapidement des indicateurs provisoires. La chaîne « traditionnelle » reposant sur du *selective editing* continue d'exister pour produire les indicateurs définitifs dans les délais habituels. Les méthodes de *machine learning* utilisées peuvent soit être des méthodes non supervisées soit des méthodes supervisées, la différence entre ces deux types reposant sur le fait que dans le cadre des méthodes supervisées on dispose d'un jeu de données « étiquetées » (le plus souvent un échantillon) - c'est-à-dire d'un jeu de données brutes ainsi que des données vérifiées et corrigées par les gestionnaires -, ce jeu étant appelé jeu d'apprentissage.

– **Moderniser les processus actuels en optimisant la recherche des unités à expertiser par les gestionnaires.** Cette motivation n'est pas nouvelle, la validation manuelle étant identifiée par tous les INS comme une étape très coûteuse en ressources sur laquelle des investissements sont régulièrement réalisés pour l'optimiser. Les méthodes de *machine learning* offre de nouvelles perspectives en permettant par exemple de repérer les données problématiques les plus influentes sur le résultat et définir des priorités pour le travail des gestionnaires dès le début de la collecte lorsque les données disponibles ne sont pas suffisamment nombreuses pour mettre en place des macro-contrôles.

Enfin, les INS utilisent de plus en plus des sources administratives pour leur production, celles-ci étant plus volumineuses et plus exhaustives que les sources d'enquêtes. Par conséquent, leur traitement nécessite une automatisation plus accrue du *data editing* pour aider à la détection des valeurs douteuses et à leur traitement. De plus, recontacter les unités dont les gestionnaires traitent les données apparaît moins évident même si on dispose de données de contact. En effet, la source administrative sera très certainement relative à une période de référence plus lointaine que celle d'une enquête dont le traitement par les gestionnaires sera réalisé juste après la collecte.

5 Article 12, alinéa 3 ; article 17bis, alinéa 4 du règlement européen 223.

En 2024, afin de développer les connaissances et les cas d'utilisation de solutions basées sur l'intelligence artificielle/le machine learning (IA/ML), Eurostat a lancé un appel d'offre visant à créer un centre de ressources européen sur l'IA/ML pour les statistiques officielles (*one-stop-shop for artificial intelligence/machine learning for official statistics - AIML4OS*). Un groupe de travail porte explicitement sur le *data editing*. La durée du projet est de 4 ans, du 1er avril 2024 au 31 mars 2028, et rassemble un consortium de 14 pays dont la France (13 États membres et la Norvège).

Les principaux résultats attendus sont les suivants :

– la mise en place d'un cadre pour le développement de solutions d'IA/ML à utiliser dans le contexte de la production de statistiques officielles et européennes ;

– l'accès des pays membres à des solutions/ressources d'IA/ML établies et éprouvées dans le contexte de la production de statistiques officielles ;

– l'encouragement à s'engager dans l'IA/ML et l'accélération du passage à la production réelle.

Le projet global est divisé en 13 groupes avec 6 groupes de nature transversale et 7 groupes axés sur des cas d'utilisation de l'IA/ML dans les statistiques officielles, parmi lesquels le *data editing* et l'imputation constituent deux axes spécifiques. Plus précisément, le groupe sur le *data editing,* piloté par l'INS allemand Destatis, a pour objectif de développer, tester et mettre en œuvre des solutions basées sur l'IA/ML conduisant à la mise en œuvre de standards et de méthodologies. Il se décompose en trois sous-objectifs :

– découvrir les possibilités d'utilisation de l'IA/ML en termes d'automatisation, d'efficacité et d'amélioration de la qualité du processus de *data editing* et imputation (en mettant l'accent sur le *data editing*) ;

– explorer ces possibilités à l'aide d'exemples pratiques, sans oublier les normes de qualité des statistiques officielles ;

– étudier les effets sur les charges de travail correspondantes de *data editing* dans les instituts de statistique.

Une présentation globale du projet ainsi que des deux groupes *data editing* et imputation ont fait l'objet d'une session lors de la réunion du groupe « Expert Meeting on Statistical Data Editing » d'octobre 2024 à Vienne.

Bibliographie

Beguin, J.-M., Haag, O., 2017, Méthodologie de la statistique annuelle d'entreprises : Description du système « Ésane », Insee méthodes

Brilhault, G., Brion, Ph., 2008, Vérification sélective des données pour le futur système français de statistiques structurelles d'entreprises, Methodes de sondage, pages 21-26, sous la direction de Philippe Guilbert, David Haziza, Anne Ruiz-Gazen, Yves Tillé, Dunod

Brion, Ph., 2007, Guidelines for Finding a Balance Between Accuracy and Delays in the Statistical Surveys / Arbitrages entre délais et précision dans les enquêtes statistiques, document de travail DSE E2007/18, Insee et consultable également à l'adresse suivante https://ec.europa.eu/eurostat/documents/64157/4374310/28-GUIDELINES-FOR-BALANCE-BETWEEN-ACCURACY-AND-DELAYS-2007.pdf/cad273c7-8534-4c4d-8662-10adfd411b0f

Caron, N., 2005, La correction de la non-réponse par repondération et par imputation, document de travail M0502, Insee série méthodologie statistique, https://www.bnsp.insee.fr/ark:/12148/bc6p06zrh1j.texteImage

Chambers, R. L., 1986, Outlier robust finite population estimation, Journal of the American Statistical Association, Vol 81, N°396, pp 1063-1069

Cox, B.G., Binder, D.A. and Co, 1995, Business Survey Methods, Wiley.

Deroyon, T., 2015, Traitement des valeurs atypiques d'une enquête par winsorization - Application aux Enquêtes Sectorielles Annuelles, Actes des Journées de Méthodologie Statistique

De Waal, T., Coutinho, W., 2005, Automatic editing for business surveys : an assessment of selected algorithms, International Statistical Review, vol 73, n°1

De Waal, T., Pannekoek, J., Scholtus, S., 2011, Handbook of statistical data editing and imputation, Wiley.

Farwell, K., 2005, Significance Editing for a Variety of Survey Situations. Paper presented at the 55th session of the International Statistical Institute, Sydney, 5-12 April.

Favre-Martinoz, C., Beaumont, J.-F., Haziza, D., 2015, Une méthode de détermination du seuil pour la winsorisation avec application à l'estimation pour des domaines, Techniques d'enquête, Vol. 41, N°1

Favre-Martinoz, C., Deroyon, T., 2018, Traitement des valeurs influentes dans les enquêtes, fiche méthodologique n°10, https://www.insee.fr/fr/information/2838097

Fellegi, I.P., Holt, D. , 1976, A systematic approach to automatic edit and imputation, Journal of the American Statistical Association, Vol. 71, Number 253

Guggemos F., 2010, Rapport du groupe de travail sur les procédures de contrôles et redressements dans le cadre du programme Premice, document de travail DSE E2010/09, Insee https://intranet.insee.fr/jcms/1388184_DBFileDocument/fr/guggemos-gt-contr-redr-pgm-premice-2010-09

Granquist, L., 1995, Improving the Traditional Editing Process. In Business Survey Methods, eds. B. Cox, D. Binder, N. Chinappa, A. Christianson, M. Colledge, and P. Kott, New York, Wiley, 385-401.

Granquist, L., Kovar, J. G., 1997, Editing of survey data : How much is enough ?, Survey measurement and process quality, Wiley

Hedlin, D., 2003, Score Functions to Reduce Business Survey Editing at the UK Office for National Statistics, Journal of Official Statistics, vol. 19, pages 177-199

Hedlin, D., 2008, Local and global score functions in selective editing, Unece, work session on statistical data editing, Vienna

Hesse, C., 2005, Vérification sélective de données quantitatives, document de travail DSE 2005/04, Insee, https://www.bnsp.insee.fr/ark:/12148/bc6p06zqwmj/f1.pdf

ISTAT, CBS and SFSO, 2007, Recommended practices for editing and imputation in cross-sectional business surveys, manuel mis au point dans le cadre du projet Edimbus d'Eurostat disponible a l'adresse suivante : https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf

Kokic, P. N., Bell, P. A.,1994, Optimal Winsorizing cutoffs for a stratified finite population estimator, Journal of Official Statistics, Vol. 10, pages 419-435

Kozak, R., 2005, The BANFF system for automated editing and imputation, congrès annuel de la Société Statistique du Canada.

Latouche, M., Berthelot, J.-M., 1992, Use of a score function to prioritize and limit recontacts in editing business surveys, Journal of Official Statistics, Vol. 8, n°3

Lawrence, D., McKenzie, R., 2000, The general application of significance editing, Journal of Official Statistics, Vol. 16, n°3, pages 243-253.

Rama, S., Salgado, D., 2014, Standardising the editing phase at Statistics Spain : a little step beyond Edimbus, document de travail 05/2014, INE, Espagne

Riviere, P., 1996, enquêtes annuelles d'entreprises : à la rencontre du 4ème type, Courrier de statistiques, N°78, Insee

Rivière P., 2002a, General principles for data editing in business surveys and how to optimise it, Document de travail n° 0203, Insee série méthodologie statistique, et contribution à l'Unece Work Session on Statistical Data Editing mai 2002, https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2002/05/sde/16.e.pdf

Rivière P., 2002b, Estimating total survey error in business survey, Document de travail E2002/9, Insee série Direction des statistiques d'entreprises.

Rivière P., 2002c, Impact on mean squared error as a score to handle data editing, Document de travail E2002/10, Insee série Direction des statistiques d'entreprises, https://www.bnsp.insee.fr/ark:/12148/bc6p06zqwxn/f1.pdf.

Rivière P., 2005, « Optimiser la production statistique : lien entre coût minimal de vérification et finesse de diffusion », Actes des Journées de Méthodologie Statistique, https://journees-

methodologie-statistique.insee.net/optimiser-la-production-en-statistique-dentreprise-lien-entre-cout-minimal-de-verification-et-finesse-de-diffusion/.

Séminaire de Méthodologie Statistique du département des méthodes statistiques, 2013, Le traitement des unités influentes dans les enquêtes, https://www.insee.fr/fr/information/2387481.

Statistique Canada, 2003, Méthodes et pratiques d'enquêtes, https://www150.statcan.gc.ca/n1/fr/catalogue/12-587-X

Statistique Canada, 2021, « les statistiques : le pouvoir des données », module de formation en ligne, https://www150.statcan.gc.ca/n1/edu/power-pouvoir/toc-tdm/5214718-fra.htm

Unece, 1994, Conference of european statisticians statistical standards and studies - no. 44, statistical data editing volume no. 1, Methods and techniques , United nations, Geneva

Unece, 1997, Conference of european statisticians statistical standards and studies - no. 48, statistical data editing volume no. 2, Methods and techniques , United nations, Geneva

Unece, 2000, *Glossary of Terms on Statistical Data Editing*, United Nations, Geneva, https://unece.org/fileadmin/DAM/stats/publications/editingglossary.pdf

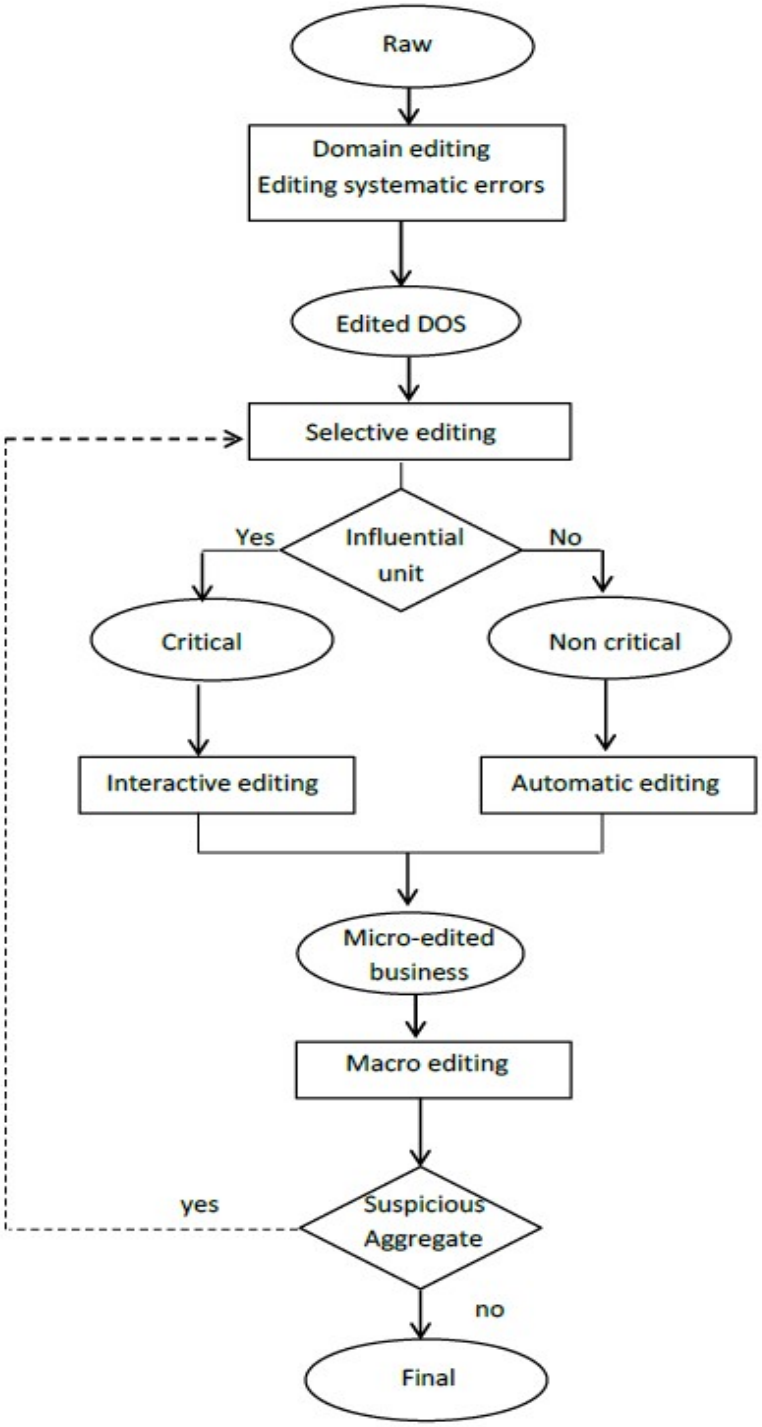
Unece, 2014, Modèle générique du processus de production statistique (version en français) https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2014/1-Mod%C3%A8le_g%C3%A9n%C3%A9rique_du_processus_de_production_statistique.pdf

Unece, 2019, Generic Statistical Data Editing Model (GSDEM), référentiel data editing https://statswiki.unece.org/display/sde/5+SDE+Flow+Models

Documentation relative aux "work sessions" des Nations unies consacrées au data editing, disponible à l'adresse suivante https://statswiki.unece.org/display/sde/Statistical+Data+Editing

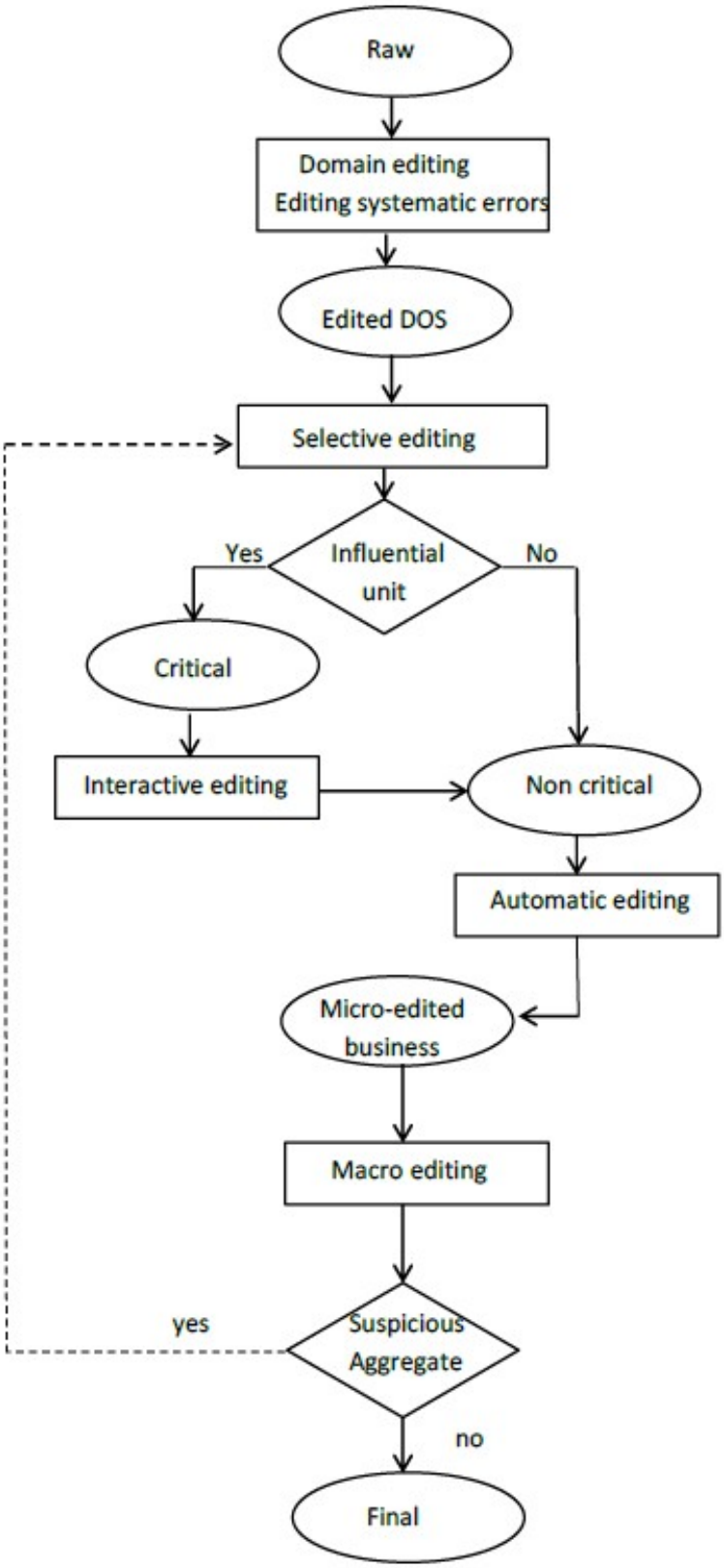
Annexe 1 : Exemples de processus de *data editing* présentés dans le GSDEM[6]

A.1 - pour les données issues d'enquêtes structurelles d'entreprises

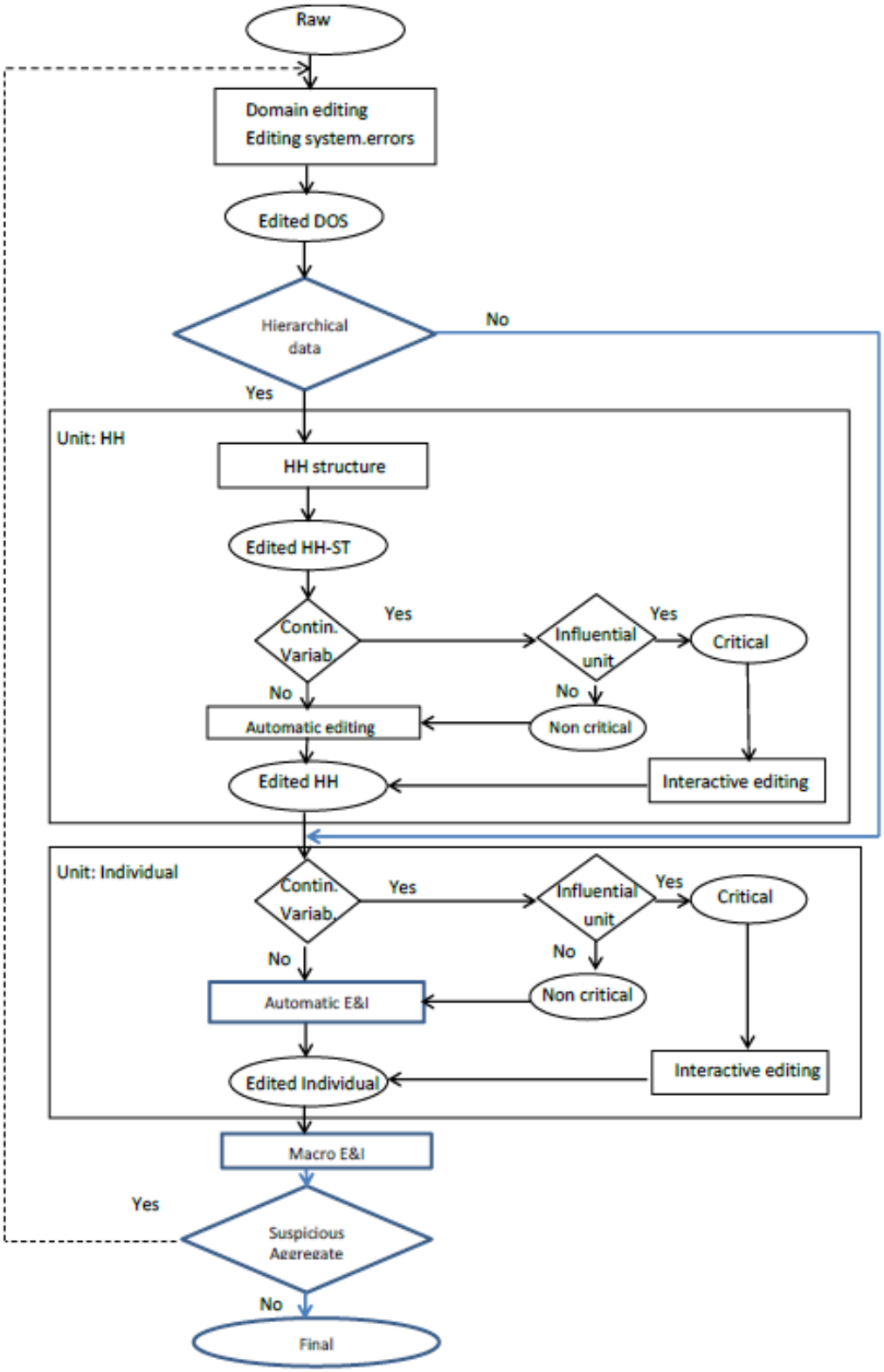


6 Se réferer au GSDEM (https://statswiki.unece.org/display/sde/5+SDE+Flow+Models) pour la signification des cases.

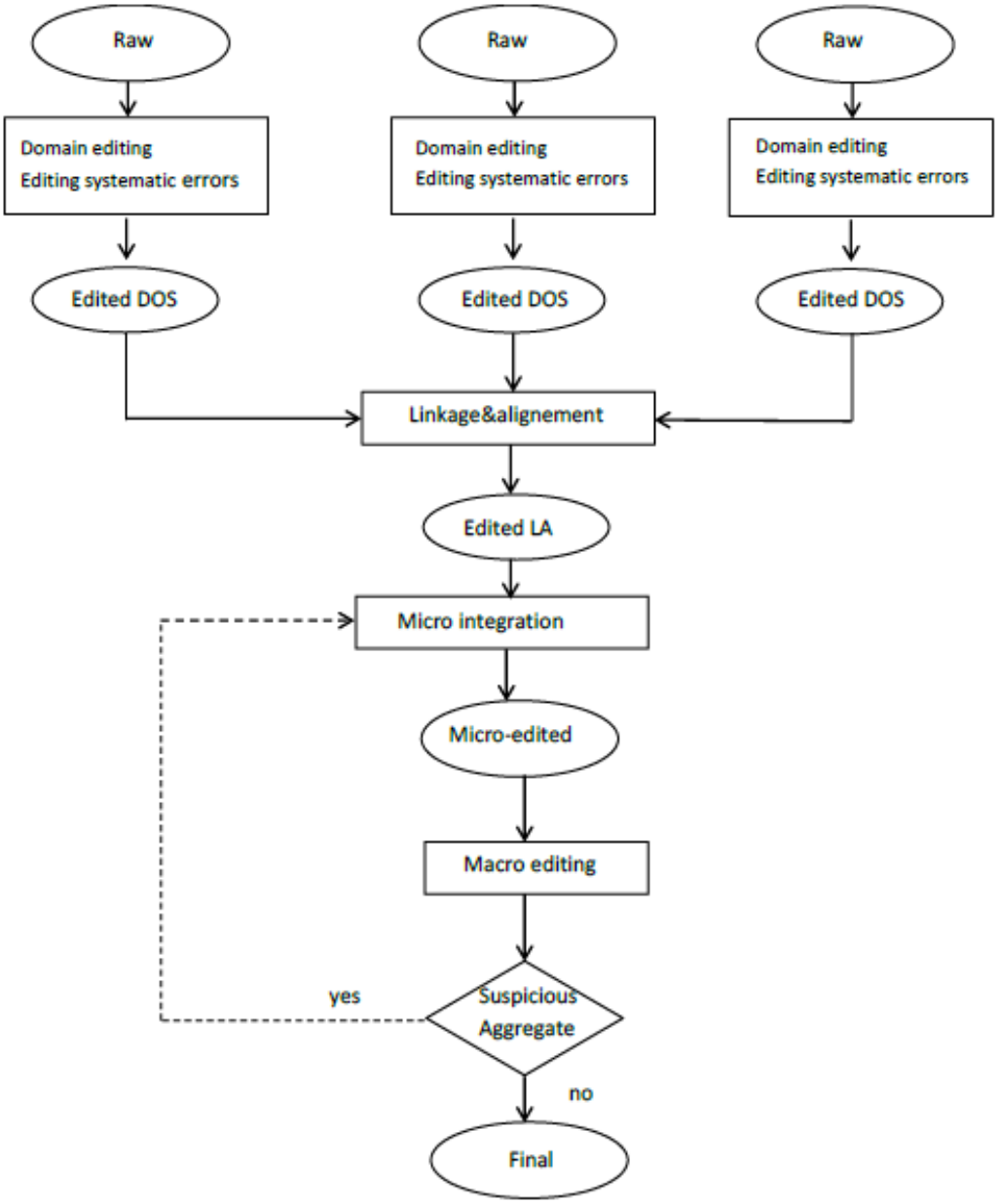
A.2 - pour les données issues d'enquêtes conjoncturelles d'entreprises



A.3 - pour les données issues d'enquêtes ménages



A.4 - pour les données issues de trois sources administratives qui sont « fusionnées »



Série des Documents de Travail « Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

0502 : Correction de la non-réponse par répondération et par imputation
N. CARON

0503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD,
X. D'HAULTFOEUILLE

M2014/01 : La microsimulation dynamique : principes généraux et exemples en langage R
D. BLANCHET

M2015/01 : la collecte multimode et le paradigme de l'erreur d'enquête totale
T. RAZAFINDROVONA

M2015/02 : Les méthodes de Pseudo-Panel
M. GUILLERM

M2015/03 : Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse
E. GROS
K. MOUSSALAM

M2016/01 : Le modèle Logit Théorie et application.
C. AFSA

M2016/02 : Les méthodes d'estimation de la précision de l'Enquête Emploi en Continu
E. GROS
K.MOUSSALAM

M2016/03 : Exploitation de l'enquête expérimentale Vols, violence et sécurité.
T. RAZAFINDROVONA

M2016/04 : Savoir compter, savoir coder. Bonnes pratiques du statisticien en programmation.
E. L'HOUR
R. LE SAOUT
B. ROUPPERT

M2016/05 : Les modèles multiniveaux
P. GIVORD
M. GUILLERM

M2016/06 : Econométrie spatiale : une introduction pratique
P. GIVORD
R. LE SAOUT

M2016/07 : La gestion de la confidentialité pour les données individuelles
M. BERGEAT

M2016/08 : Exploitation de l'enquête expérimentale Logement internet-papier
T. RAZAFINDROVONA

M2017/01: Exploitation de l'enquête expérimentale Qualité de vie au travail
T. RAZAFINDROVONA

M2018/01: Estimation avec le score de propension sous R
S. QUANTIN

M2018/02: Modèles semi-paramétriques de survie en temps continu sous R
S. QUANTIN

M2019/01 : Les méthodes de décomposition appliquées à l'analyse des inégalités
B. BOUTCHENIK
E. COUDIN
S. MAILLARD

M2020/01 : L'économétrie en grande dimension
J. L'HOUR

M2021/01 : R Tools for JDemetra+ - Seasonal adjustment made easier
A. SMYK
A. TCHANG

M2021/02 : Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman
L. CASTELL
P. SILLARD

M2021/03 : Conception de questionnaires auto-administrés
H. KOUMARIANOS
A. SCHREIBER

M2022/01 : Introduction à la géomatique pour le statisticien : quelques concepts et outils innovants de gestion, traitement et diffusion de l'information spatiale
F. SEMECURBE
E. COUDIN

M2022/02 : Le zonage en unites urbaines 2020
V. COSTEMALLE
S. OUJIA
C. GUILLO
A. CHAUVET

M2023/01 : Les réseaux de neurones appliqués à la statistique publique : méthodes et cas d'usages
D. BABET
Q. DELTOUR
T. FARIA
S. HIMPENS

M2023/02 : Redressements de la première vague de l'enquête epicov : un exemple de correction des effets de sélection dans les enquêtes multimodes
L.CASTELL
C. FAVRE-MARTINOZ
N. PALIOD
P. SILLARD

M2023/03 : Appariements de données individuelles : concepts, méthodes, conseils
L.MALHERBE

M2023/04 : Victimations déclarées et effets de mode : enseignements de l'expérimentation panel multimode de l'enquête cadre de vie et sécurité
L. CASTELL
M. CLERC
D. CROZE
S. LEGLEYE
A. NOUGARET

M2024/01 : Estimation en temps réel de la tendance-cycle : apport de l'utilisation des filtres asymétriques dans la détection des points de retournement
A.QUARTIER-LA-TENTE

M2024/02 : La disponibilité des coordonnées de contact dans fidéli-nautile - quels enseignements pour les protocoles de collecte ?
G. CHARRANCE (INED)

M2024/03 : Discuter l'existence d'un effet de sélection dans un cadre multimode grâce à une analyse de sensibilité - Application aux enquêtes annuelles de recensement
L. COURT
S. QUANTIN

M2024/04 : Vers une désaisonnalisation des séries temporelles infra-mensuelles avec JDemetra+
A. SMYK
K. WEBEL

M2025/01 : Les estimations par capture-recapture ou par système multiple : quelques éléments théoriques
P. ARDILLY
H. KOUMARIANOS

M2025/02 : Tests cognitifs pour les enquêtes auto-administrées : quelques éléments de méthode
D. GUILLEMOT
J. DIRAND
C. FLUXA

M2025/03 : Statistiques fondées sur des données administratives - esquisse d'un cadre général
H. KOUMARIANOS
P. RIVIÈRE

M2025/04 : Peut-on estimer un effet de mesure sur une enquête à partir d'un essai croisé ab/ba : la question de la non-réponse non ignorable dans l'enquête test emploi du temps
Loreline COURT
Simon QUANTIN

M2025/05 : L'apport des technologies cloude pour industrialiser le processus d'innovation statistique
R. AVOUAC
T. FARIA
F. COMTE

M2025/06 : Le data editing : Définition et principes généraux
N.CARON