LA COMMUNICATION AUTOUR DE L'INCERTITUDE DES STATISTIQUES PUBLIQUES

Ronan Le Saout[1]

[1] CREST-ENSAI. *E-mail : ronan.le.saout@ensae.fr*

Résumé. En tant que statisticien-économiste, il est souvent surprenant de constater que la précision des statistiques publiques ne soit pas publiée. Les raisons invoquées lors de discussions informelles sont multiples, le manque de moyens, les difficultés théoriques, l'absence de demande sociale, ou la difficulté de la pédagogie... Elles illustrent une absence de connaissance mais aussi une trop forte focalisation sur la précision au détriment d'une vision plus large de l'incertitude. L'objet de cet article est d'illustrer, à partir des statistiques diffusées par l'Insee, la typologie de Manski (2015) définissant l'incertitude selon trois composantes, permanente, transitoire et conceptuelle. Cette notion d'incertitude ne se limite ainsi pas à la précision et se rapproche de la notion de qualité. Ce caractère pluriel fait que les choix en termes de communication sont nécessairement multiples. Ils vont dépendre des produits statistiques (issus d'enquête, de sources administratives ou de sources multiples), de l'approche retenue (descriptive ou d'étude économique), et du public visé. Une évolution historique vers une plus grande transparence est mise en avant. L'émergence des données massives et ouvertes est enfin discutée à travers trois évolutions : la réappropriation de la notion de qualité des données par les chercheurs en sciences sociales, la délicate appréhension de l'incertitude lorsque la population est observée exhaustivement, le rôle de la data vizualisation dans la communication auprès du grand public.

Mots-clés. Incertitude, Qualité, Statistiques Publiques.

Je remercie Paul Champsaur, Jean-Michel Charpin, Éric Dubois, Dominique Ladiray et Olivier Sautory pour avoir accepté de répondre à mes questions concernant les choix effectués en termes de communication autour de l'incertitude. Je remercie grandement Pauline Givord pour l'orientation donnée à ce travail. Je remercie également Michel Armatte, Martin Chevalier, Laurent Davezies, Gaël de Peretti, Michel De Saboulin, Hélène Thélot ainsi que les participants au séminaire de socio-histoire des statistiques de la Sfds (2016) et des Jms (2018) pour leurs conseils et relectures. Ce qui est écrit dans cette communication n'engage bien sûr pas l'Insee.

"C’est une vérité, et il est malheureux que l’éducation française évite de la mettre en valeur, que l’activité humaine opère dans un monde incertain, mais qu’il faut bien distinguer incertitude et ignorance." Edmond Malinvaud, 1988.

1 Introduction

Manski (2015) fait le constat que les statistiques officielles, aux États-Unis mais également dans la majorité des pays, sont diffusées sans indicateurs de précision (intervalles de confiance par exemple). On dira par exemple que le taux de pauvreté est de 14 %, mais pas que l’intervalle de confiance est [13,7 %;14,3 %]. Le risque est alors de considérer que ces statistiques sont exemptes d’erreurs. Or les sources d’erreurs sont nombreuses : erreurs statistiques classiques liées à l’échantillonnage, mais également liées au mode de collecte ou de questionnement, ou aux concepts économiques sous-jacents.

L'objet de cette communication est d'étudier le cas des statistiques publiées par l'Insee et la communication de leur incertitude. Suivant Manski (2015), nous définissons l'incertitude à travers trois dimensions : permanente, transitoire et conceptuelle. Permanente, pour ce qui relève de l'erreur d'échantillonnage mais également de l'ensemble du processus d'enquête tel que la répondération, le calage sur marges ou l'imputation des données manquantes (et également de manière extensive l'erreur d'enquête totale et les effets de mode de collecte, dont on trouvera une présentation dans Razafindranovona 2015). C'est l'erreur de précision au sens classique de la théorie des sondages. Transitoire, car les statistiques publiques font souvent l'objet d'une publication précoce et sont ensuite révisées, ces modifications étant potentiellement invisibles pour les utilisateurs. L'objet ici n'est pas de savoir si ces corrections sont importantes ou non, mais si elles peuvent être connues aisément par les utilisateurs. Conceptuelle enfin, pour ce qui relève de la "boîte noire" des statisticiens telle que les méthodes de désaisonnalisation ou de correction de la qualité des indices. De même, on ne s'interrogera pas ici sur la pertinence des méthodes mises en oeuvre mais sur la disponibilité à la fois des indicateurs corrigés et non corrigés, les deux pouvant être utiles à un économiste. Cette vision extensive de l'incertitude rejoint ainsi plusieurs concepts de la notion de qualité, qui ne se limite pas à la précision des données diffusées.

Il convient ensuite d'identifier les vecteurs de communication de cette incertitude. Elle peut bien sûr s'établir à travers les tableaux et publications descriptives. Au-delà des chiffres, les commentaires peuvent préciser les évolutions par rapport à des chiffres provisoires, voire la précision de ces statistiques (liée à l’erreur d'échantillonnage et aux retraitement pour les sondages, à la modélisation pour la prévision). Les concepts et la méthodologie peuvent être synthétisés dans des encadrés ou courtes notes. Ce sont là les modes de publication “grand public”, qu'on peut qualifier de communication primaire. Mais elle peut aussi revêtir d'autres formes, à travers la transparence des méthodes mises en oeuvre (document de travail méthodologique, publication académique, ou séminaire) ou des caractéristiques des données (dites méta-données). C'est là une communication auprès d'un public spécialisé de statisticiens ou d'économistes. Enfin, le dernier mode de communication est celui des bases de données détaillées des enquêtes, et la possibilité pour les utilisateurs (chercheurs, chargés d'études, étudiants) d'appréhender l'incertitude des données avec les informations contenues dans la base de diffusion.

Peu de travaux en France ont été consacrés à cette question de la communication de l’incertitude. Les journées d’étude sur l’histoire de la statistique de 1976, et les contributions associées, n’ont pas abordé directement cette question, la communication de René Padieu sur *la diffusion de l’information statistique* (1987) traitant des publications et des domaines associés mais non de l’incertitude des statistiques. Cette question apparaît néanmoins dans plusieurs contributions de cadres de l’Insee, portant un regard critique sur le manque de communication autour de l’incertitude

des statistiques publiques (Malinvaud 1988; Domergue 1995; Dell *et al.* 2002). Armatte (2003), étudiant l'introduction des méthodes de sondage en France par Pierre Thionet sur la période 1938-1954, souligne également que ce débat n'est pas nouveau, citant Thionet (1954) « L'idéal serait de pouvoir publier côte à côte les résultats de l'enquête et des calculs d'erreur. Malheureusement ceci retarderait notablement la publication des résultats; et il a fallu y renoncer une fois pour toutes à l'I.N.S.E.E. Les calculs d'erreur sont donc effectués dans les temps morts… ». Cette communication vise donc à analyser les choix effectués, en identifiant les raisons à la fois théoriques, de coûts ou institutionnelles qui peuvent les expliquer. Elle s'appuie à la fois sur une étude bibliographique, la conduite d'entretiens et l'étude de quelques exemples (et non l'exhaustivité des situations rencontrées).

Cet article se concentre sur l'incertitude des statistiques dites descriptives, informations primaires du débat public. La première partie de cet article revient néanmoins sur les différences théoriques de l'incertitude pour les études économiques et les statistiques descriptives, l'Insee présentant la particularité d'effectuer les deux. La deuxième partie détaille quelques raisons aux choix effectués en termes de communication en les replaçant dans une perspective historique. La troisième suivante analyse plus particulièrement ces choix au regard des différences conceptuelles des statistiques issues d'enquêtes, de sources administratives exhaustives, établies à partir de diverses sources, ou résultant d'une modélisation pour la prévision économique. La dernière partie ouvre la discussion de la communication de l'incertitude des statistiques publiques à l'ère des *Big Data* et de l'*Open Data*.

Encadré : Méthodologie

Cette communication s'appuie sur une revue de littérature, une étude des informations publiques diffusées sur insee.fr et des entretiens réalisées avec des cadres de l'Insee. La revue de littérature a été effectuée à partir des mots clés « incertitude », « précision », « erreur », et « qualité » à partir des principales publications académiques de l'Insee (ou proche de l'Insee) *Annales de l'Insee* (devenue *Annales d'Économie et Statistique,* puis *Annals of Economics and Statistics*), *Économie et Statistique, Économie et Prévision, Courrier des Statistiques* ainsi que dans *Statistique et Société.*
Les entretiens ont visé à identifier des exemples où la question de l'incertitude des statistiques a pu peser sur le choix de diffuser ou non des statistiques, sur les raisons implicites ou explicites sur les modes de communication (ou la possibilité d'un non-choix), sur le rôle des institutions et des différences d'approches entre domaines (enquêtes, conjoncture, comptabilité nationale) ou entre producteurs de statistiques et chargés d'études économiques.

2 Deux théories associées à deux visions de l'incertitude: l'économétrie pour les études économiques, la théorie des sondages pour les enquêtes statistiques

L'Insee présente la particularité d'être à la fois producteur et utilisateur des données statistiques. Les théories utilisées dans les deux sphères sont néanmoins distinctes, théorie des sondages d'un coté, modélisation statistique et économétrie de l'autre.

Dans le cadre de l'estimation de la moyenne issue d'une enquête statistique, l'estimateur de Hájek est $\left(\sum_{i\epsilon s} 1/\pi_i\right)^{-1}\left(\sum_{i\epsilon s} Y_i/\pi_i\right)$ avec π_i la probabilité d'inclusion de l'individu *i* dans l'échantillon *s*. Un calcul de variance peut être conduit, tenant compte du plan de sondage, et des retraitements d'enquête. La formule complète est souvent longue et complexe. L'Insee se concentre sur une approche analytique de ce calcul[1], coûteuse d'un point de vue humain. Trouver un benchmark (ou

[1] Eurostat, pour des raisons pragmatiques, demande aux instituts nationaux de privilégier une approche par Bootstrap.

des gardes-fous permettant de juger si une variance est plausible ou non) reste délicat ("la légitimité est dans la longue note descriptive de ce calcul" pour reprendre les termes d'un méthodologue). Ce calcul reste donc complexe, et par ailleurs néglige les erreurs non liées à l'échantillonage telles que le questionnement, le mode de collecte, les interactions enquêtés-enquêteurs (Platek et Särndal, 2001).

Avec une perspective modèle, on retrouve ce même estimateur en estimant par les Moindres Carrés Ordinaire la constante $\hat{\alpha}$ du modèle (sans variables explicatives) $Y_i = \alpha + \varepsilon_i$, pondéré par l'inverse de la probalité d'inclusion et avec ε_i un terme d'erreur. Les logiciels permettent de plus le calcul d'intervalles de confiance de cet estimateur en tenant compte du caractère échantillonné des observations et d'hypothèses sur les résidus du modèle. Cet intervalle de confiance est en général différent de celui calculé par l'approche sondages.

Les notions d'aléa dans ces deux approches sont en effet différentes. En théorie des sondages, l'aléa est dans l'échantillonage. En économétrie, l'aléa est celui du modèle. Si on observe par exemple les comportements et caractéristiques de l'ensemble des habitants d'une petite ville française, le sondeur considérera que les statistiques de consommation ou de revenu sont exactes, leur variance nulle. Pour l'économètre, il restera du bruit lié à l'incertitude entourant la décision de consommation ou d'activités. Il fera l'hypothèse qu'il existe un modèle de "super-population", ce qui est observé n'en étant que des réalisations. Il y a donc également une différence conceptuelle sur la notion même de précision entre ces deux approches.

Ce simple exemple permet de comprendre que les deux approches permettent d'obtenir des indicateurs de précision mais aux concepts différents. Selon l'approche retenue, les choix en termes de communication de la précision seront différents. Dans le cas de statistiques descriptives, dont le cadre théorique est la théorie des sondages, il sera fait peu état du caractère imprécis de l'estimation dans les publications généralistes. Dans le cas d'un modèle, dont le cadre théorique est l'économétrie, elle sera assumée et plus communiquée. Ainsi, la publication *Insee Première* inclut ponctuellement des modélisations logistiques dont certains chiffres ne sont pas diffusés car non significatifs ("n.s. écart à la situation de référence non significatif à 5 %"), les intervalles de confiance restent néanmoins non diffusés. *A contrario* les évolutions descriptives même de faible ampleur de certains indicateurs restent commentées. Ces choix se retrouvent dans les publications plus détaillées telles que *Insee Références*.

Les raisons de ces choix *in fine* distincts ne sont pas explicites. Il y a en premier lieu une différence d'objectifs entre la description et l'explication, pour laquelle une preuve scientifique est attendue. Il reste que, lorsqu'un modèle est employé, l'objectif est souvent également descriptif. Il y a également la différence créée par les cursus de formation (ENSAI et ENSAE), dans lesquels peu de liens sont effectués entre les cours de statistiques descriptives, de théorie des sondages et d'économétrie. Les principales statistiques descriptives peuvent être représentées à l'aide d'un modèle économétrique, mais ce point n'est que rarement abordé. De même la majorité des cours d'économétrie négligent la question de la qualité des données et la prise en compte de leur construction. Il y a ensuite une partition entre les postes d'études et ceux de production statistique. Analysant les controverses entourant les études de Laroque et Salanié sur le chômage au début des années 2000, Coutrot et Exertier (2001) soulignent le non-dialogue entre ces deux sphères qui peut aboutir à des choix différents. Enfin, d'un point de vue plus sociologique, il apparaît qu'un modèle

---

Cette approche est plus simple mais ne fait pas consensus. Une des raisons en est que dès que le plan de sondage est complexe, les méthodes de bootstrap ne s'appliquent pas, ou alors au prix de simplifications du plan de sondage dont on ne peut pas mesurer les conséquences.

est associé à une étude des comportements nécessairement bruités alors qu'une statistique descriptive renvoie à une demande sociale d'exactitude, mesurée par une institution "indiscutable". Cette distinction de la demande sociale justifie ainsi une différence d'approche en termes de communication. Pour reprendre Desrosières 2003, "le statisticien scrupuleux peut parfois se trouver tiraillé entre, d'une part, une exigence professionnelle d'expliciter ses méthodes et ses conventions, en adoptant de fait une position quasi constructiviste, et, d'autre part, une demande sociale réaliste implicite de ne pas « encombrer le lecteur avec ces détails techniques », qui risquent de semer le doute ou de restreindre la portée des résultats présentés.".

La suite de l'article se concentre en premier lieu sur la communication autour de l'incertitude des statistiques dites descriptives, informations primaires du débat public.

3 Quelques raisons pour communiquer ou non l'incertitude

Plusieurs raisons peuvent être définies, qui relèvent d'un côté d'une non-disponibilité d'indicateurs d'incertitude que ce soit pour des raisons théoriques[2], informatiques (coût computationnel ou de stockage), ou humaines (manque de personnel), et de l'autre de choix stratégiques imposés ou non par les contraintes institutionnelles (européennes notamment, à travers la notion de qualité), la demande de transparence du public ou une volonté pédagogique de rendre les statistiques plus aisément compréhensibles. L'incertitude est dans ces cas connue mais non diffusée. Enfin, et notamment concernant les bases de données détaillées, il est possible que l'incertitude ne soit pas communiquée par absence de réflexions autour de ces questions, i.e. un non-choix, en l'absence de demande des utilisateurs.

3.1 Une communication à deux étages : le grand public versus un public spécialisé

Lors de leur communication aux JMS, Dell *et al.* (2002) écrivent en première phrase d'introduction "Une statistique n'a pas de valeur sans son écart-type." Cette communication aborde à la fois de manière théorique et pratique le calcul de la précision de statistiques descriptives simples ou complexes, en tenant compte de l'ensemble des traitements quantifiables d'une enquête (plan de sondages, non-réponse totale et partielle). Un ensemble de macros SAS est mis à la disposition des utilisateurs, pouvant permettre une estimation simplifiée de la variance du taux de pauvreté calculé à partir de l'*Enquête Revenus Fiscaux.* Les communications ultérieures du taux de pauvreté calculés par l'Insee n'inclurent cependant pas la précision de cette statistique (pour des raisons également techniques liées aux évolutions des enquêtes utilisées pour calculer ce taux de pauvreté).

Cet exemple reste illustratif de la communication de l'Insee depuis les années 1970 (et l'apparition de journaux tels que les *Annales de l'Insee*, ou *Économie et Statistique*). Une communication des méthodes est effectuée auprès de spécialistes à travers des publications académiques ou des séminaires, complétée par un calcul à usage interne de la précision, mais cette information n'est pas rendue aisément accessible au grand public. On retrouve ainsi des exemples analogues pour l'indice des prix (Ardilly et Guglielmetti 1993 ; Jaluzot et Sillard 2016), ou les statistiques de Patrimoine (Lamarche et Salembier 2015). Depuis la diffusion de fiches qualité qui incluent des indicateurs de

[2] L'erreur liée au mode de collecte, à la manière dont les questions sont posées et comprises, à l'ordre des questions... est rarement quantifiable. Les concepts établissant les comptes nationaux ou les indices peuvent être expliqués, mais mesurer l'influence des choix afférents (le classement des activités de R&D des entreprises en dépenses ou en investissements pour la comptabilité nationale, la prise en compte des substitutions entre produits pour l'indice des prix à la consommation) est délicat, si ce n'est impossible en production courante. Ces exemples visent à illustrer que les enjeux théoriques dépassent le calcul de la précision, qui peut lui aussi ne pas être possible (par exemple pour les indices de prix dits hédoniques pour lesquels la précision n'est en général pas calculée, ou l'absence de prise en compte des erreurs non liées à l'échantillonnage telles que les biais de collecte).

précision (cf. infra partie 2.3), la communication de la précision s'est néanmoins élargie.

Cette doctrine a également pu être résumée par l'éditorial de Philippe Domergue (1995) dans un numéro spécial d'*Économie et Statistique* sur la qualité de l'information statistique "Si les statisticiens ne manquent pas, du fait même de la difficulté de leur métier, d'être attentifs à la qualité de l'information qu'ils produisent, il n'est pourtant pas usuel en France qu'ils examinent "publiquement" les risques d'erreurs ou de biais, ou évaluent l'imprécision et les révisions de leurs estimations. Le plus souvent, ces débats très techniques ne dépassent pas un cénacle de spécialistes.". Une nécessité de transparence est ainsi soulignée, pour éviter toute mauvaise utilisation des statistiques mais également pour répondre à une demande accrue du public. Ce numéro spécial inclura ainsi des articles sur la précision de l'indice des prix, du PIB, ou des prévisions économiques. Pour autant, ce numéro spécial ne sera pas suivi par la suite d'une communication de ces incertitudes auprès du grand public.

Les contraintes théoriques, si elles existent, n'expliquent en général pas l'absence de diffusion de la précision des statistiques. Ces dernière années ont en effet été marquées par une forte implication des unités de l'Insee (UMS, puis DMS, et pôle PISE) sur le calcul de la précision, effectuée désormais pour presque toutes les enquêtes (et la mise à disposition d'outils informatiques). Au-delà de Dell *et al.* (2002), des documents de travail de l'Insee ont ainsi proposé des techniques de calculs de la précision (Roth 1989 pour la procédure PRECIS; Caron 1998 pour le logiciel POULPE). Gros et Moussallam (2015) ont détaillé le calcul de précision à partir des échantillons des enquêtes-ménages tirées dans le recencement de la population rotatif (qui a pu créer à court terme des contraintes théoriques). Les *Annales de l'Insee* (devenu *Annales d'Économie et Statistique,* puis *Annals of Economics and Statistics*) et *Économie et Statistique* ont été fondés en 1969 avec dès le départ des articles sur la précision des statistiques descriptives (Bodin 1969; Chartier 1969). Cette communication des méthodes s'est enfin renforcée par la création des Journées de Méthodologie Statistique en 1991 (Biau et de Peretti 2004) par trois cadres de l'Insee Jean-Claude Deville, Olivier Sautory et Dominique Ladiray. Elles intégreront des sessions spécifiques sur le calcul de la précision, ainsi qu'un ensemble de communications visant à ouvrir la "boîte noire" du statisticien et réduire ainsi l'incertitude conceptuelle. Elles seront complétées par les Journées sur la Correction de la Saisonnalité créées en 2008.

L'incertitude transitoire apparaît majoritairement communiquée à travers les commentaires des chiffres diffusés. Par exemple pour l'indice des prix des logements anciens, une section "révisions" précise pour le premier trimestre 2016 "L'indice des prix des logements anciens est révisé pour prendre en compte les transactions des périodes couvertes qui n'avaient pas encore été enregistrées lors de la publication précédente. Par rapport aux données publiées le 4 avril 2016, la variation des prix au quatrième trimestre 2015 est abaissée de 0,2 point au total, avec -0,2 point pour les maisons et -0,1 point pour les appartements. Elle s'établit à +0,2 % pour l'ensemble, au lieu de +0,4 % (chiffre actualisé au 4 avril), et +0,5 % estimé le 25 février. ". Cela n'a néanmoins pas toujours été le cas. L'*Informations Rapides* pour le premier trimestre 2009 n'incluait ainsi pas de section "révisions". On peut dater du milieu des années 2000 cet effort de communication de l'évolution des statistiques selon les chiffres provisoires, semi-définitifs ou définitifs (en parallèle d'une mise à disposition des données plus rapide), marquée par la diffusion d'une note méthodologique annuelle analysant les révisions des comptes nationaux. Auparavant, ce type d'informations était également diffusées à un public de spécialistes. Gallais (1995) détaille ainsi dans *Économie et Statistique* les révisions des comptes nationaux (PIB et ses composantes) pour la France et d'autres pays développés sur la période 1975-1990. Il reste que, s'il y a désormais communication sur ces actualisations, les bases de données n'incluent que les chiffres les plus actuels et non l'ensemble des millésimes diffusés (d'autres organismes que l'Insee ont créé de telles bases de données pour la France). Pourtant une demande extérieure existe, par exemple pour juger ex-post de la qualité des

prévisions (i.e. la comparaison entre la réalisation et la prévision), l'utilisation des seules données définitives n'est pas toujours suffisante.

Concernant l'incertitude conceptuelle, on peut noter la mise en œuvre d'une communication grand public et de vulgarisation scientifique centrée sur les statistiques phares de l'Insee à partir des années 2010 et de la rubrique *Insee en bref* du site Internet (Comprendre le PIB et la croissance, la mesure du chômage par l'Insee...). Néanmoins, si la communication sur les méthodes auprès d'un public plus spécialisé est riche, les chiffres "non transformés" (par exemple les prix moyens des logements, avant correction de la qualité, ou certaines données sectorielles non désaisonnalisées) ne sont pas toujours diffusés. Ils pourraient néanmoins être utiles à un public de chercheurs désirant mettre en œuvre des méthodologies plus spécifiques. Desrosières (2003), citant le cas de l'harmonisation des comptes nationaux (notamment entre les branches et les secteurs), souligne ainsi "dès lors que ces sources sont utilisées dans des contextes distincts, souvent sur des questions sectorielles ou locales, l'avantage éventuel de la coûteuse mise en cohérence peut être inférieur à la perte résultant des manipulations des sources de base uniquement destinées à cette mise en cohérence globale." Le coût lié à rendre disponibles ces informations pourrait être important, au-delà de la difficulté à communiquer sur ces incohérences.

3.2 Le rôle de l'évolution technologique

Depuis l'apparition des ordinateurs (et donc les années 1970-1980), les enjeux computationnels apparaissent mineurs quant à la possibilité de calculs de la précision, et des autres formes d'incertitude. Ces gains informatiques permettent néanmoins l'emploi de techniques nouvelles, par exemple l'utilisation du Bootstrap ou d'inférence Bayésienne pour les projections de population (Costemalle 2015 pour une présentation des travaux menés par l'ONU).

En parallèle, les coûts de stockage de l'information ont été réduits de manière drastique. Jusqu'aux années 1970, les cartes perforées et les bandes magnétiques étaient encore utilisées. Les premiers disques durs apparaissent dans les années 1960, la disquette dans les années 1970, le CD-Rom dans les années 1980. Les années 2000 sont marquées par l'arrivée des clés USB, et du Cloud. Alors qu'un Mo coûtait plusieurs milliers d'euros à stocker dans les années 1950, un disque dur d'1 To (1 million de fois plus grand) coûte moins de 100 euros en 2016. Ce coût impactait donc les choix en termes de stockage de l'information. Ils expliquent en partie qu'il n'était pas possible de stocker (et donc par la suite de communiquer) des informations complémentaires sur la méthodologie d'enquête (conserver les données imputées et originales dans une même base) ou l'incertitude (conserver l'ensemble des révisions d'une statistique) jusqu'à une période récente qu'on peut situer aux alentours des années 80. Ces choix historiques et contraints ont été associés à des chaînes de production spécifiques, dont la logique n'a pas forcément été modifiée malgré la baisse des coûts de stockage. Il y aurait par exemple un coût important de spécification et de vérification pour créer une base de données des différents millésimes diffusés de la comptabilité nationale. L'émergence des données massives montre que le coût de stockage de l'information reste une thématique d'actualité.

En 2003, l'Insee, alors dirigé par Jean-Michel Charpin, a décidé de diffuser gratuitement la majorité des statistiques produites. L'Insee a été un des premiers instituts nationaux de statistique à faire ce choix. La diffusion était historiquement payante car répondre à la demande du public était coûteux. L'Insee n'ayant pas les moyens de répondre à l'ensemble des demandes, ce caractère payant établissait un filtre. Avec l'arrivée d'Internet, le coût marginal de diffusion de l'information à un utilisateur supplémentaire est devenu nul. Il était donc légitime d'ouvrir cette diffusion à l'ensemble du public (Charpin 2010a) mais avec une contrainte, l'absence de demandes complémentaires. De ce point de vue, l'incertitude n'étant pas diffusée au préalable, il n'a pas été envisagé d'étendre la communication, pour des questions de coût. Mais il n'y a pas eu non plus par la suite de dialogue

entre la sphère méthodologique (qui calcule de nombreux indicateurs d'incertitude) et de la diffusion (qui met à disposition l'information statistique), pour savoir quelles informations complémentaires pourraient être ajoutées. Beaucoup d'indicateurs de précision sont désormais calculés mais non accessibles, sans que cela soit le résultat d'une volonté délibérée.

3.3 Un rôle moteur des institutions nationales et européennes

La notion d'incertitude ne se limite pas au calcul de la précision. L'incertitude permanente inclut par exemple le taux de non-réponse partielle (et donc d'imputation) de l'enquête. L'ajout de méta-données peut donc alerter les utilisateurs sur cette incertitude permanente. L'Autorité de la Statistique Publique s'est saisie de cette question dès sa création en 2009. Son premier rapport d'activité note ainsi "Les Bilans Qualité servent à rendre compte de la réalisation d'une enquête, à fournir des éléments quant à la précision des résultats et, dans le cas d'une enquête régulière, à donner des points de référence pour améliorer l'enquête suivante. Ces bilans existent déjà pour un grand nombre d'enquêtes « entreprises » et pour les enquêtes ménages liées à un règlement européen. L'objectif, dans un terme à préciser, est désormais la réalisation systématique de « fiches qualité ». Ces fiches sont une version synthétisée des bilans, portée à la connaissance des utilisateurs, pour toutes les enquêtes (domaines entreprises et ménages).". Le bilan annuel de l'ASP 2010 souligne néanmoins que "Les producteurs sont convaincus de l'utilité de ces fiches mais peinent à dégager du temps pour les réaliser, en particulier pour ce qui concerne les calculs de précision.". La partie "Sources et Méthodes" du site Internet de l'Insee met ainsi à disposition pour de nombreuses enquêtes et productions statistiques trois documents: une fiche descriptive, une documentation méthodologique et une fiche qualité. Pour les enquêtes entreprises, ces fiches qualité incluent les précisions des principaux indicateurs (jusqu'à plusieurs dizaines pour l'enquête ANTIPOL -Investissements dans l'industrie pour protéger l'environnement-), ainsi que les taux de non-réponse totale et partielle. Pour les enquêtes ménages, la diffusion des fiches qualité reste moins systématique, la précision n'étant diffusée que pour l'enquête CVS (cadre de vie et sécurité). La définition de ces fiches est par ailleurs reliée aux règles imposées par Eurostat, qui définissent des obligations de bilans et fiches qualité pour de nombreuses enquêtes[3]. Un projet européen d'harmonisation du contenu de ces fiches (projet RMéS) est en cours. Il reste que les bilans qualité, qui contiennent des calculs fins de précision, ne sont pas diffusés que ce soit au niveau national ou européen. Ce calcul de précision est alors utilisé comme un outil de validation interne des statistiques diffusées, non comme une information à usage du grand public. On retrouve ici la tension soulignée par Desrosières (2003) entre *qualité du produit*, celle qui intéresse l'utilisateur, et *qualité du processus de fabrication*, à laquelle s'attache l'organisateur de la production.

Cette présentation met ainsi en avant que les choix effectués ne sont pas indépendants des institutions nationales et européennes, notamment à travers la notion de qualité qui s'est renforcée au cours des années 2000. Avec la diffusion des fiches qualité, la communication de l'Insee sur la précision (et autres formes d'incertitude permanente) ne se limite plus aux publications spécialisées ou académiques. La communication de l'incertitude a largement bénéficié de l'adoption en 2005 d'un code des bonnes pratiques de la statistique européenne, née en partie de la crise des statistiques grecques des années 2000. La France joua un rôle central, le Directeur Général de l'Insee (J.-M. Charpin) présidant le groupe de travail en charge de la rédaction de ce code. L'approche extensive de l'incertitude proposée par Manski (2015) rejoint en effet plusieurs composantes de la qualité (pertinence, précision, délais, disponibilité, comparabilité (temporelle ou géographique), cohérence). Brion (2005), analysant la qualité des enquêtes, insiste également sur le rôle des institutions. Au niveau national, le Comité du Label de la Statistique Publique, créé en 1994, valide

[3] Un bouton "M" (pour "exploratary notes (metadata)"), donnant accès à ces fiches qualité, est associé à la diffusion des statistiques sur le site d'Eurostat, qui propose donc un accès plus large que l'Insee à ces données.

les choix techniques opérés pour la conduite des enquêtes. Des indicateurs cibles de précision peuvent ainsi être calculés dans ce cadre. Le CNIS (Conseil National de l'Information Statistique créé en 1984) juge au contraire de la pertinence des enquêtes et, en assurant la concertation entre les producteurs et les utilisateurs de la statistique publique, vise à identifier les nouveaux besoins. Il n'y a à ma connaissance eu que peu de demandes des utilisateurs relatives à l'incertitude au sein des instances du CNIS.

3.4 L'arbitrage entre transparence, pédagogie et défense de l'institution

La communication autour de l'incertitude par l'Insee a franchi des paliers, associée à plus de pédagogie autour des chiffres dans les médias. L'imperfection des sondages en particulier politiques (et non menés par l'Insee) a pu alerter le grand public sur cette problématique, à travers la diffusion de la précision de ces sondages d'opinion (et du nombre de personnes interrogées). Mais il convient également de ne pas créer de confusion entre par exemple la production d'instituts de sondages privés (à la méthodologie plus légère) et les chiffres publics. Moins communiquer sur l'incertitude peut aussi éviter de créer cette confusion, et apparaît ainsi comme une défense de l'institution, "Il ne faut pas couper la branche sur laquelle on est assis.". Même si son absence peut parfois amener à des conclusions fallacieuses, l'incertitude reste une information secondaire.

L'Insee est en effet un organisme public, mais il est également soumis depuis quelques années (et l'expansion de l'accès aux données privées issues d'internet) à une concurrence extérieure par d'autres organismes. Des indices de prix des logements sont par exemple diffusés de manière précoce par des agences immobilières, sans communication sur le champ couvert, les méthodes utilisées et leurs éventuelles imprécisions. L'objectif pour ces organismes est marchand. Il s'agit de s'assurer une visibilité et une publicité, et non de certifier la qualité d'un chiffre. L'Insee a créé des conditions de labellisation auxquelles peuvent se soumettre ces organismes, mais cette démarche reste bien sûr facultative (et peu suivie). Ce sont souvent les statistiques les plus précoces, et non les plus fiables, qui sont l'objet de l'attention médiatique.

La défense de l'institution apparaît ainsi nécessaire, car la crédibilité des statistiques est indispensable pour son acceptation par tous dans le débat public. Or l'acquisition de cette crédibilité et la définition de la communication associée restent délicates. Un sentiment diffus de doute sur la crédibilité des statistiques existe, difficile à expliquer, contrôler ou infléchir (Charpin 2010b). Mieux communiquer, ce n'est pas forcément en dire plus, Desrosières (2003) soulignant "comme c'est le cas pour de beaux monuments, l'exhibition de ces échafaudages de métadonnées peut, qu'on le veuille ou non, brouiller, sinon la beauté, du moins l'efficacité argumentative de l'évidence factuelle impliquée par la mise en avant d'un simple nombre, tout nu." L'indice des prix à la consommation, malgré une méthodologie éprouvée, a été ainsi l'objet de controverses liées à la distinction entre le ressenti de l'évolution des prix et ce qui était mesuré par l'indice. Ce fut le cas lors du passage à l'euro, lors d'une campagne publicitaire de Leclerc en 2004, et lors de la campagne présidentielle de 2007. L'Insee a largement communiqué sur ces incertitudes conceptuelles, sans lever complètement ces mises en causes. Au contraire, pour le recensement rénové de la population, Charpin (2010b) souligne que les critiques initiales disparurent grâce à la communication de l'Insee.

De plus, la pédagogie autour des notions statistiques reste délicate. Pour un lecteur non statisticien, un intervalle de confiance à 95 % sera perçu comme un intervalle de certitude, et non comme le fait qu'avec de nouveaux tirages d'échantillons, la statistique serait dans 5 % des cas en dehors de cet intervalle. Ces intervalles peuvent être relativement larges, et pourrait amener à ne pas commenter de nombreux chiffres, ce qui ne serait pas forcément un gain pour le débat public. Néanmoins, même sans diffuser un intervalle de confiance, la simple diffusion des statistiques peut donner une illusion de certitude. Malivaud (1988) souligne que la dernière décimale d'une statistique est

rarement significative. Il met en avant qu'une statistique même imprécise est souvent meilleure qu'une absence de statistique mais qu'il convient de reconnaître cette imprécision. "Je dois reconnaître avoir souvent été mal à l'aise à me dire que les statisticiens français faisaient des efforts peut-être insuffisants pour renseigner sur la précision de leurs résultats. [...] Dès lors qu'elle est objective, c'est-à-dire non intentionnellement biaisée, une information imprécise est meilleure que pas d'information du tout. L'utilisateur d'une statistique a ainsi intérêt à connaître le résultat obtenu tel quel, et ceci d'autant plus qu'il s'agit d'un utilisateur plus sérieux."

Il apparaît pour l'heure impossible de définir des règles générales, "la communication gardant une part de mystère, tant sont subtils les canaux qui vont influencer les utilisateurs dans leur jugement" (Charpin 2010b).

3.5 Une réflexion en devenir, le contenu des bases détaillées

Dernier point, celui du dernier niveau de diffusion, les bases de données détaillées. Des avancées récentes, le Comité du Secret et le Centre d'Accès Sécurisé aux Données (CASD), ont facilité l'accès (sous des conditions strictes) des chercheurs aux données individuelles d'enquêtes de la statistique publique et aux données administratives. Des données individuelles (anonymisées) sont même directement accessibles sur le site Internet de l'Insee. Le réseau Quetelet, qui archive également les données individuelles d'enquêtes, pousse à une mise à disposition extensive de l'information entourant chaque conduite d'enquête, mais sans formaliser le contenu des bases de diffusion qui reste du ressort des producteurs. Ce processus s'inscrit de manière plus générale dans le contexte d'une demande de reproductibilité des études économiques, de nombreux journaux académiques demandant l'accès aux données lorsqu'un article est accepté pour publication. Pour autant, une réflexion reste encore à mener sur le contenu de ces bases détaillées pour permettre une juste prise en compte de l'incertitude pour des utilisateurs extérieurs. Par exemple, les strates utilisées pour le plan de sondage ou les variables auxiliaires utilisées pour les redressements ne sont pas toujours connues (et renseignées dans la documentation d'enquête), alors même que ces informations sont nécessaires pour déterminer si un modèle doit être pondéré ou non ou calculer de manière adéquate la variance (Davezies et D'Haultfoeuille 2009). De même, si certaines bases de données précisent les observations imputées (à l'aide de variables dite "drapeaux"), ce n'est en rien systématique. Le risque est alors de laisser croire à tort à une grande qualité des données, la "théorie des sondages" et la méthologie d'enquête restant peu abordées dans les cursus classiques de statistiques. Le cours de sondages a par exemple été rendu facultatif à l'ENSAE pour les élèves non fonctionnaires (mais pas à l'ENSAI) au milieu des années 2000. Faute de connaissances de la part des utilisateurs, il ne faut donc pas attendre une demande de ces utilisateurs sur une meilleure communication des incertitudes entourant chaque processus d'enquête. La demande devra certainement être stimulée par l'Insee, le risque d'un point de vue de la reproductibilité des sciences serait de laisser croire que les données brutes sont celles diffusées après l'ensemble des traitements d'enquêtes. Des travaux sont en cours au sein de la Direction de la Méthodologie, par exemple sur l'économétrie des données d'enquête ou la fourniture de kits regroupant l'ensemble des données auxiliaires et des programmes nécessaires aux calculs de précision à partir des données individuelles.

Il n'y a donc pas de raison unique aux choix effectués en termes de communication autour de l'incertitude. Nous allons maintenant nous concentrer sur les raisons associées aux différences conceptuelles de production entre les statistiques issues d'enquêtes, de sources administratives, de modèles de conjoncture ou d'une construction économique.

4 Les différences conceptuelles de production

Abordant la notion de qualité statistique, Desrosières (2003) étudie chaque critère à l'aune d'une tension résultant du fait que les objets statistiques peuvent être vus « *à la fois* comme « réels » (ils existent antérieurement à leur mesure) et comme « construits à partir de conventions » (ils sont, d'une certaine manière, « créés » par ces conventions). ».

Quatre types de production statistique sont ici distingués, les statistiques issues d'enquêtes, de sources administratives exhaustives, celles fruits de diverses sources et d'une construction économique, et enfin la prévision économique. Toutes sont bien sûr construites à partir de conventions, mais la perception établie par les producteurs peuvent en être plus ou moins fortes.

Ainsi, dans le cas de statistiques issues d'enquêtes, un questionnaire et un corpus mathématique (la théorie des sondages) ancrent le calcul dans la réalité. Le calcul de la précision (et la communication autour de celle-ci) devient alors possible, voire demandé par le public. Pour le nouveau recensement, le bilan annuel 2012 de l'ASP faisait ainsi état d'un incident. Le journal Le Monde publia le 1er août 2012 un article mettant en cause la méthode du recensement, notamment la précision des résultats au niveau national et leur capacité à éclairer utilement au niveau local les communes dans l'exercice de leurs missions. L'Insee avait alors été obligé de communiquer sur la précision (0,02 % pour la population totale) et la cohérence des différents millésimes. Cette communication a été complétée depuis par un numéro spécial d'*Économie et Statistique* sur le recensement (Brilhault et Caron 2016 pour le calcul de précision). Autre indicateur phare, le taux de chômage calculé à partir de l'Enquête Emploi. Du milieu des années 1980 jusqu'en 2003, cette enquête était annuelle. L'incertitude n'était pas communiquée au grand public. Des publications spécialisées ont détaillé l'incertitude permanente et la précision (Deville et Roth 1986), ainsi que l'incertitude conceptuelle à travers la différence entre le nombre de chômeurs au sens du BIT estimé par l'enquête Emploi et le nombre de demandeurs inscrits à l'ANPE (Pôle Emploi actuellement) (Thélot 1987). Depuis 2003, l'enquête est devenue trimestrielle (Givord 2003) et la diffusion a été marquée par plusieurs polémiques. Depuis le milieu des années 80, les évolutions des taux de chômage calculées par l'Insee combinait deux sources : l'enquête Emploi pour le niveau annuel et les données de Pôle Emploi (à partir des inscriptions administratives) pour le profil mensuel (qui n'adoptent pas les mêmes définitions, Bureau International du Travail pour l'Insee, catégories de Pôle Emploi de l'autre). Pendant l'année en cours, l'enquête Emploi n'était pas disponible et l'estimation mensuelle était donc provisoire, fournie à partir de l'évolution d'une catégorie proche conceptuellement des chômeurs BIT inscrits à Pôle Emploi. On calait ensuite ces résultats sur l'enquête Emploi, et les divergences en évolution étant par ailleurs faibles. Ce ne fut plus le cas à partir de 2007, au moment même où une divergence d'évolution fut constatée entre ces deux sources. L'Insee a en effet rendu public le 8 mars 2007, comme annoncé de longue date, les résultats de l'enquête emploi du 4ème trimestre 2006, ainsi que ceux de l'année 2006. Mais il a simultanément annoncé qu'il décalait de six mois la mise en cohérence entre les deux sources pour se laisser le temps d'étudier l'importante divergence constatée, ce qui engendra des polémiques médiatiques sur la mesure du taux de chômage en France. Un rapport de l'Inspection générale des affaires sociales (Igas) et l'Inspection générale des finances (IGF) amena à modifier la méthode de calcul (utilisation unique de l'enquête Emploi trimestrielle) et à augmenter la taille de l'échantillon de l'enquête emploi. Il fut également décidé de préciser dans les *Informations Rapides* la précision par une note « Estimation à +/- 0,3 point près du niveau du taux de chômage et de son évolution d'un trimestre à l'autre », mais sans communiquer la précision pour l'ensemble des statistiques sur l'emploi (Gros et Moussallam 2016 pour une présentation du calcul de la précision de l'enquête emploi). L'analyse de ces défaillances s'est focalisée sur la précision, peut-être au détriment de l'incertitude conceptuelle et permanente non liée à l'échantillonnage. L'Insee a cependant décidé de

publier les statistiques du “halo” du chômage à cette période, permettant ainsi aux utilisateurs d’appréhender ces incertitudes conceptuelles. Les deux sources vivent maintenant leurs vies indépendamment. Des divergences plus ponctuelles ont continué à exister depuis lors et suscitent périodiquement l'interrogation du public (et de l'ASP concernant les chiffres diffusés par Pôle Emploi) sur la qualité des statistiques, par exemple en 2013 à l'occasion d'un changement mineur du questionnaire de l’enquête emploi. Les chiffres de l'Insee peuvent se trouver discrédités. Ces polémiques montrent ainsi que communiquer sur l'incertitude est délicat, les arguments techniques trouvant peu d'écho dans le débat public. De plus, l'incertitude dépasse bien souvent la notion de précision, et renvoie à une incertitude plus subjective associée au processus de construction d'une statistique (qui inclut de très nombreuses sources d'erreurs, autres que celles d’échantillonnage) et aux garde-fous que peuvent s'imposer les statisticiens pour valider une statistique. Dans ces exemples, on note ainsi une nécessité de communiquer autour des différentes formes d'incertitudes, permanentes (précision, modes de collecte), provisoires notamment à travers les différents millésimes du recensement, et conceptuelles à travers les différences de mesures du taux de chômage.

La réflexion autour des statistiques issues de sources administratives est plus récente. Ces sources sont exhaustives et normalement exactes. Pour autant, ces sources restent imprécises car elles ont pour usage premier une utilisation administrative et non statistique. Il y a donc conversion d'une « donnée » en un concept statistique. Or les règles administratives peuvent évoluer dans le temps, source d'incertitude conceptuelle. Dans l'exemple des données de demandeurs d'emploi, les règles d'enregistrement des différentes catégories par Pôle Emploi peuvent changer. De même, les chiffres de la délinquance évolueront avec les règles régissant l'enregistrement des plaintes ou les logiciels d'enregistrement utilisés (et la codification afférente). Il n'y a pas non plus de règle unique régissant les comptes des collectivités locales, rendant délicate l'estimation de ceux-ci.

D'autres statistiques sont par contre issues de sources multiples, voire le fruit d'une théorie économique. Elles ne découlent pas directement d'un questionnaire mais n'en sont que le résultat indirect (à l'instar également des sources administratives évoquées ci-dessus). Les conventions adoptées sont ainsi plus explicites, mais la notion de précision en devient aussi plus difficile voire impossible à définir. Il apparaît ainsi délicat de définir le sens statistique que pourrait revêtir un intervalle de confiance pour le PIB et ses composantes. L'incertitude sera alors principalement abordée à travers les révisions et les concepts. Les comptes nationaux en sont l'exemple le plus connu, qui découlent directement d'une construction économique (le prix Nobel d'économie 1984 Richard Stone a ainsi été récompensé pour ses travaux sur la création d'un système de comptabilité nationale). Ses révisions faisaient historiquement l'objet d'une communication ponctuelle à travers des publications spécialisées (Gallais 1995). Depuis le milieu des années 2000, l'analyse des révisions inclut à la fois une comparaison du chiffre diffusé à celui de la précédente période, complétée par une note méthodologique analysant les révisions des comptes annuels sur plus longue période. Pour les comptes trimestriels, pour lesquels 3 estimations sont réalisées (à 1, 2 et 3 mois après la fin du trimestre), ces révisions sont faibles (0,06 point en moyenne entre 1991 et 2013). Mais elles cachent des révisions ultérieures, car à la différence des comptes annuels, les comptes trimestriels peuvent être continuellement révisés (notamment par calage sur les révisions des comptes annuels, ou des corrections de variations saisonnières). Ces révisions, plus importantes (0,24 point au bout de 3 ans sur la période 1991-2013), sont précisées dans la note méthodologique accompagnant la diffusion des comptes trimestriels. Pour les comptes annuels, il y a 3 évaluations du PIB de l’année n, le compte dit définitif étant obtenu en n + 3. Les comptes annuels ne faisaient pas l'objet d'une communication spécifique mais simultanée aux comptes trimestriels. Ce n'est plus le cas depuis 2016, où pour des raisons de délai de publication, les comptes annuels font l'objet d'une diffusion distincte. Or la révision de la croissance pour l'année 2014 (du chiffre provisoire au chiffre semi-définitif) a été forte mais non exceptionnelle, passant de 0,2 % à 0,6 %, ce qui a créé

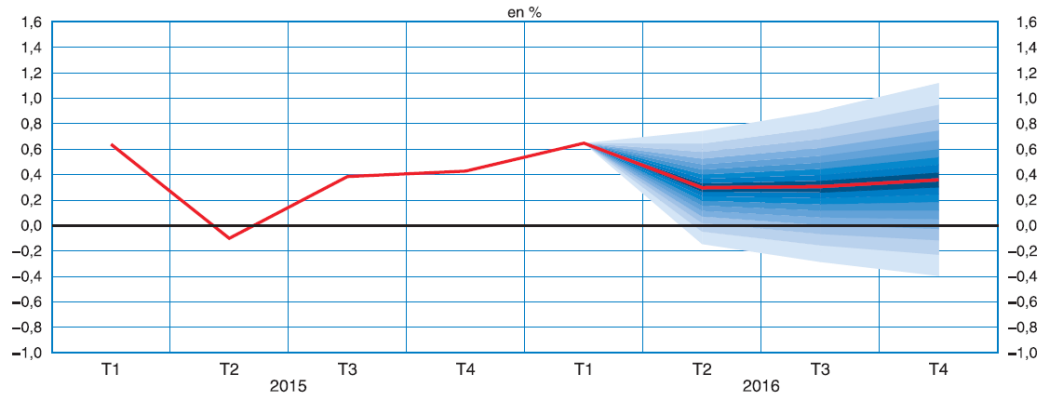
des débats publics. Cet exemple récent nous montre qu'un changement de communication peut à la fois mieux informer le public car les chiffres gagnent alors en visibilité et transparence, mais que la pédagogie autour de ces révisions et donc de l'incertitude reste délicate alors même qu'une note méthodologique analysant les révisions annuelles des comptes est diffusée depuis le milieu des années 2000 (ce qui souligne également que cet effort restait moins visible). De plus, de fortes révisions ne sont pas toujours le signe d'une mauvaise qualité des chiffres initiaux mais d'informations complémentaires utilisées. Il reste que dans le cadre d'une communication de second niveau (chercheurs, chargés d'études...), trouver l'information historique sur les révisions reste délicat, en l'absence de base de données comprenant à la fois les chiffres provisioires et définitifs.

Il n'y a par ailleurs pas de communication de l'Insee sur les différences statistiques ("statistical discrepancy") obtenues selon l'approche retenue (revenu, dépense, ou production). D'un point de vue conceptuel, seul est comptabilisé ce qui est visible, Zucman (2013) souligne ainsi que les actifs possédés par les ménages à l'étranger sont sous-estimés car ils ne tiennent pas compte des actifs possédés dans les paradis fiscaux. La prise en compte de l'économie numérique (substitution des taxis par Uber, des hôtels par AirBnb, des agences de voyage par les agences en ligne...) peut engendrer des difficultés conceptuelles et méthodologiques d'après le rapport Bean (2016) sur les statistiques publiques anglaises. Par ailleurs, les aspects institutionnels jouent ici un rôle également, le code des bonnes pratiques de la statistique européenne poussant par exemple à la diffusion simultanée de séries corrigées et non corrigées (des variations saisonnières et des jours ouvrables) et à une meilleure communication autour des méthodes. Pour conclure et montrer la difficulté à communiquer sur ces séries temporelles, mélant construction statistique et théorie économique, on peut citer Desrosières (2003) "L'interprétation des séries temporelles a été à l'origine de techniques statistiques très spécifiques (calcul de moyennes mobiles, correction des variations saisonnières, modèles linéaires stochastiques) visant à démêler des réalités d'ordres différents, depuis des « variations aléatoires » de court terme jusqu'à des « tendances longues séculaires », en passant par diverses formes de cycles. L'articulation des rhétoriques réaliste et conventionnaliste est, dans le cas des séries temporelles, très originale, et mériterait une étude spécifique".

Dernier point, celui des chiffres de prévision, qui ne relèvent pas dans la majorité des pays des statistiques calculées par les instituts de statistique[4]. Néanmoins, cet exemple apparaît dans Manski (2015), comme exemple de bonnes pratiques pour la Grande-Bretagne. Il reste ainsi important de souligner que l'Insee a adopté des pratiques similaires en reprenant le graphique des risques diffusé par la Banque Centrale d'Angleterre (et non par l'Institut de statistiques) et qui illustre graphiquement la précision et les intervalles de confiance des prévisions. La communication des erreurs de prévision s'établissait au préalable à travers des publications spécialisées (Borowski et al. 1991, Bouthevillain et Mathis 1995). Le premier graphique des risques (exemple ci-dessous pour juin 2016) apparaît dans la note de conjoncture de juin 2008 avec un encadré méthodologique (à la fois pour les prévisions annuelles et trimestrielles, limité ensuite aux prévisions trimestrielles) rédigé par Éric Dubois alors chef du Département de la conjoncture à l'Insee, et complété à partir de décembre 2008 d'un retour sur la précédente prévision. L'un des objectifs en termes de communication était de contribuer à acclimater l'idée que les prévisions ne sont pas de même nature que les statistiques, parce qu'elles ne s'appuient pas uniquement sur des données mais aussi sur une modélisation, plus ou moins explicite, des comportements économiques des agents. Elles ne sont pas la « vérité révélée ». C'est aussi la nature particulière de ces chiffres qui a poussé à mieux mettre en avant leur incertitude. Le second objectif était d'éviter une appréhension de l'incertitude par des scénarios, qui peuvent eux-mêmes donner une illusion de précision. Le choix a par contre été fait de

[4] Les prévisions sont notifiées à la Commission Européenne par le Ministre de l'Économie, alors que les comptes nationaux le sont par le Directeur Général de l'INSEE.

ne pas établir d'intervalles pour le passé (contrairement à la Banque Centrale d'Angleterre), au vu des difficultés conceptuelles de la notion de précision de la croissance du PIB.



Lecture : le graphique des risques retrace, autour de la prévision centrale (en trait rouge), 90 % des scénarios probables. La première bande, la plus foncée, décrit les scénarios les plus probables autour du scénario central, l'ensemble de ces scénarios ayant une probabilité totale de 10 %. La deuxième, un peu moins foncée, se décline en deux sous-bandes immédiatement au-dessus et immédiatement en dessous de la bande centrale ; elle contient les scénarios qui suivent en termes de probabilité, de telle sorte que la probabilité totale des deux premières bandes atteigne 20 % ; et ainsi de suite, en allant du centre vers l'extérieur et de la bande la plus foncée à la bande la plus claire, jusqu'à atteindre 90 % (cf. la *Note de conjoncture* de l'Insee de juin 2008, pages 15 à 18). On peut alors estimer que la première estimation qui sera publiée dans les comptes trimestriels pour le deuxième trimestre 2016 a 50 % de chances d'être comprise entre +0,1 % (bas de la cinquième bande en partant du bas) et +0,5 % (limite supérieure de la cinquième bande en partant du haut) et 90 % de chances d'être compris entre –0,1 % et +0,7 %.

Graphique 1 : Exemple de graphique des risques diffusé dans la note de conjoncture de Juin 2016

5 L'incertitude des statistiques publiques à l'ère des *Big Data* et de l'*Open Data*

Les nouvelles sources de données, issues d'internet, de capteurs, des réseaux sociaux… communément appelées Big Data sont sources de recherche et de nouveaux challenges pour la statistique publique. Trois évolutions sont ici mises en avant et discutées: la réappropriation de la notion de qualité des données par les chercheurs en sciences sociales, la délicate appréhension de l'incertitude lorsque la population est observée exhaustivement, et le rôle de la data vizualisation dans la communication auprès du grand public.

L'accès aux données administratives (mais également aux données massives privées) crée pour la recherche de nouvelles opportunités et une nécessité de réfléchir à la qualité de ces données (Einav et Levin 2014). L'un des exemples le plus connu est le « Billion Prices Project (BPP) » qui vise à définir des indices de prix quotidiens et mondiaux à partir des données disponibles sur Internet. Ce projet est développé au MIT depuis 2008. De manière concomitante, la recherche appliquée en sciences sociales traverse une crise de reproductibilité avec la nécessité d'une plus grande transparence sur les méthodes et les données lors de la publication d'un article scientifique. L'émergence d'une offre concurrente à la statistique publique et cette demande accrue de transparence de la recherche publique poussent ainsi les instituts nationaux à mieux communiquer sur leurs méthodes, concepts et données.

Concernant la question de l'incertitude lorsque la population est observée exhaustivement, une demande sociale visible à travers les recommandations de l'Autorité de la Statistique Publique (2014) a émergé pour juger ou non significatives les évolutions des statistiques mensuelles de demandeurs d'emploi. Une recommandation était d'établir des seuils en dessous desquels les variations ont une faible signification. Ce questionnement est proche de celui de la signification des tests statistiques lorsqu'on dispose de données massives (plusieurs millions d'observations). Il n'y a pas actuellement de cadre théorique permettant de répondre simplement à la question de l'incertitude de telles statistiques. Des aides à la compréhension, non prévues dans les lignes directrices européennes, doivent donc être construites. Cela peut s'appuyer alors sur les distributions empiriques passées (pour les évolutions des chiffres du chômage) ou des hypothèses de

modélisation (par exemple la modélisation de la délinquance à l'aide d'une loi de Poisson). Mais la théorie générale reste encore à écrire.

Enfin, l'ensemble des arguments présentés jusqu'à présent a trait à la vision du producteur, celui qui émet la statistique. La connaissance de la compréhension des notions d'incertitude (ou d'intervalles de confiance) par le grand public, celui qui reçoit cette information, mériterait d'être approfondie. Tak *et al.* (2014) ont ainsi effectué des expérimentations en visualisation des données où plusieurs représentations graphiques des intervalles de confiance étaient proposées. Avec une présentation graphique adaptée et testée, la notion d'intervalle de confiance est comprise par un public non statisticien et peut même améliorer la compréhension des statistiques diffusées. Des recherches en visualisation des données semblent donc nécessaires pour faire parler l'incertitude des statistiques publiques au-devant de tous les publics.

6 Conclusions

La communication autour de l'incertitude diffère selon les produits statistiques et les publics visés. Elle se limite à des indicateurs phares (précision du taux de chômage, révision des comptes nationaux) dans les publications à diffusion large et pour la vulgarisation scientifique des méthodes. Elle est beaucoup plus précise et détaillée dans des publications académiques ou spécialisées. Avec la diffusion des fiches qualité depuis les années 2010, la communication de l'Insee sur la précision (et autres formes d'incertitude permanente) s'est élargie et ne se limite plus aux publications spécialisées ou académiques. L'incertitude entourant les statistiques issues de sources administratives et exhaustives reste encore à définir. Il convient également de bien distinguer les statistiques des prévisions. L'effort fourni à partir de 2008 pour mieux communiquer les incertitudes entourant les prévisions de croissance (graphique des risques) visait aussi à alerter le public sur la distinction entre ces deux productions, une prévision étant par nature incertaine.

D'un point de vue historique, les évolutions vers une plus grande communication ont souvent été associées à des crises combinées à une évolution continue de l'informatique (coût de stockage moins élevé, gains computationnels, diffusion par internet). Crise institutionnelle liée aux statistiques grecques et à la gouvernance d'Eurostat au début des années 2000 qui a débouché sur le code de bonnes pratiques de la statistique européenne et l'accent mis sur la notion de qualité. Problèmes rencontrés sur les chiffres du chômage et de l'emploi. Crise économique de 2008 alertant sur l'incertitude des prévisions et légitimant le recours à un graphique des risques pensé avant cette crise économique. La crise de reproductibilité des sciences sociales actuellement en cours pourrait déboucher sur une meilleure information sur l'incertitude entourant les données détaillées accessibles aux chercheurs (par exemple les variables utiles pour la construction du plan de sondage ou le retraitement des données). L'accès aux données administratives (mais également aux données massives privées) crée aussi pour la recherche de nouvelles opportunités et une nécessité de réfléchir à la qualité de ces données (Einav et Levin 2014). La communication se verra renforcée dans les années à venir dans ce sens. Il est ainsi prévu dans le cadre d'Insee 2025 d'expérimenter la mise à disposition publique de données et méthodes statistiques développées par la statistique publique (modèles, algorithmes, méthodes d'estimations), et la parution de documentations exhaustives sur les méthodes et outils statistiques, les politiques de révision des indicateurs et d'informations synthétiques sur l'accès aux micro-données via le CASD, le réseau Quetelet ou le site insee.fr.

L'incertitude recouvre plusieurs formes, quantifiables ou non, permanente, transitoire ou conceptuelle. Communiquer sur l'incertitude, c'est faire un choix, qui peut lui-même donner l'illusion d'exactitude. Des révisions très faibles ne nous disent pas forcément que les statistiques sont de grande qualité, mais peut aussi signifier que peu d'informations complémentaires ont été

prises en compte. Cette communication interfère aussi avec le jugement de la crédibilité des statistiques, indispensable pour leur acceptation dans le débat public. C'est pourquoi l'arbitrage entre transparence, pédagogie et défense de l'institution demeure délicat.

Cette analyse de la communication de l'incertitude s'appuie dans une très large mesure sur la vision du producteur, celui qui émet la statistique. Elle montre que, même s'il n'y a pas toujours communication, l'incertitude entourant le processus de construction des données est un fait dont les statisticiens tiennent compte de manière centrale dans leur travail. De ce point de vue, il y a certainement des recherches à effectuer pour comprendre ce que peut et désire recevoir celui qui reçoit cette information, en distinguant également deux publics. La visualisation des données pourrait ainsi éclairer la diffusion de l'incertitude auprès du grand public, mais il n'y a pas forcément de demande forte d'une communication élargie. Pour le public de chercheurs, l'absence de demandes sur l'incertitude des données détaillées (données imputées ou non, poids intermédiaires, variables de calages...) peut cacher une mauvaise compréhension du processus de construction des données. Dans une perspective de reproductibilité des sciences, le risque est alors de considérer que les données brutes sont celles diffusées après l'ensemble des traitements effectués par les statisticiens. Il apparaît dans ce cas que les instituts de statistiques doivent avoir un rôle moteur sur la communication de l'incertitude des bases de données détaillées.

Bibliographie

[1] Ardilly P., et Guglielmetti F. (1993) La précision de l'indice des prix : mesure et optimisation, *Économie et statistique*, 267, 13-29.
[2] Armatte M. (2003) Pierre Thionet et l'introduction des méthodes de sondage en France 1938-1954, *Journal de la société française de Statistique*, 144(1-2), 227-255.
[3] Biau, O. et de Peretti G. (2004), Les Journées de Méthodologie Statistique: « cheval de Troie » des méthodologues à l'Insee ? Un regard historique porté par des étudiants de l'ENSAE, *Courrier des Statistiques*, 111, 39-45.
[4] Bodin, J.-L. (1969), Une Comparaison Expérimentale de la Précision Obtenue par Divers Tirages Systematiques de Grappes, *Annales de l'Insee*, 2, 47-71.
[5] Bean, C. (2016), Independent review of UK economic statistics: final report, Rapport au Ministre des finances du Royaume-Uni.
[6] Borowski D., Bouthevillain C., Doz C., Malgrange P., et Morin P. (1991) Vingt ans de prévisions macroéconomiques: une évaluation sur données françaises, *Économie et prévision*, 99(3), 43-65.
[7] Bouthevillain K., et Mathis A. (1995) Prévisions : mesures, erreurs et principaux résultats, *Économie et statistique*, 285-286, 89-100.
[8] Brion, P. (2005), La prise en compte de la qualité dans les enquêtes auprès des entreprises, *Courrier des Statistiques*, 115, 41-48.
[9] Brilhault, G., et Caron N. (2016) Le passage à une collecte par sondage : quel impact sur la précision du recensement?, *Économie et statistique*, 483-484-485, 23-40.
[10] Caron, N. (1998), Le Logiciel Poulpe: Aspects Méthodologiques, Actes des Journées de Méthodologie Statistique de l'Insee; Insee Méthodes n°84-85-86, 173-200.
[11] Charpin, J.-M. (2010a) L'information statistique en perspective: six grands changements, *Revue d'Economie Financière*, 98/99, 15-26.
[12] Charpin, J.-M. (2010b) Statistiques: les voies de la confiance, *Revue Economique*, 61(3), 371-393.
[13] Chartier, F. (1969), Erreurs d'Échantillonnage pour Divers Plans de Sondage de Logement, *Annales de l'Insee*, 2, 3-45.
[14] Costemalle, V. (2015), Projections de populations: l'ONU adopte une méthode bayésienne, *Statistique et Société*, 3(3), 9-14.
[15] Courtot, T., et G. Exertier. (2001), La loi des grands nombres ou quand le « non-emploi » efface le chômage…, L'année de la régulation, Economie, Institution, Pouvoirs, Presses de Sciences-Po.
[16] Davezies, L., et X. D'Haultfoeuille. (2009), Faut-il pondérer ? … Ou l'éternelle question de l'économètre confronté à des données d'enquête, Document de travail de l'Insee n°G2009/06.
[17] Dell, F., d'Haultfoeuille X., Février P. et et Massé E. (2002), Mise en Oeuvre du Calcul de Variance par Linéarisation, Actes des Journées de Méthodologie Statistique de l'Insee, 73-104.
[18] Desrosières, A. (1993), La politique des grands nombres – Histoire de la raison statistique, Ed. La Découverte.
[19] Desrosières, A. (2003), Les qualités des quantités, *Courrier des Statistiques*, 105-106, 59-66.
[20] Deville J.-C., et Roth N. (1986) La précision des enquêtes sur l'emploi, *Économie et statistique*, 193-194, 127-134.
[21] Domergue P. (1995) Présentation générale, *Économie et statistique*, 285-286, 3-8.
[22] Einav L., et Levin J. (2014) Economics in the age of big data, Science, 346(6210), 715-722.
[23] Gallais, A. (1995), Révisions et précision des comptes nationaux français, *Economie et Statistique*, 285-286, 59-80.
[24] Givord, P. (2003) Une nouvelle enquête emploi, *Économie et statistique*, 362, 127-134.
[25] Gros, E., et Moussallam K. (2015), Les méthodes d'estimation de la précision pour les enquêtes ménages de l'Insee tirées dans Octopusse, Document de travail de l'Insee n°M2015/03.
[26] Gros, E., et Moussallam K. (2016), Les méthodes d'estimation de la précision de l'Enquête

Emploi en Continu, Document de travail de l'Insee n°M2016/02.
[27] Jaluzot, L., et Sillard P. (2016), Echantillonnage des agglomérations de l'IPC pour la base 2015, Document de travail de l'Insee n°F1601.
[28] Lamarche, P., et Salembier L. (2015), Précision de l'Enquête Patrimoine 2010, Document de travail de l'Insee n°F1503.
[29] Malinvaud, E. (1988), Le service public de la statistique en occident : le point actuel, *Journal de la société de statistique de Paris*, 129(4), 227-243.
[30] Manski, C. F. (2015), Communicating Uncertainty in Official Economic Statistics: An Appraisal Fifty Years after Morgenstern, *Journal of Economic Literature*, 53(3), 631-53.
[31] Padieu, R. (1987) La Diffusion de l'Information Statistique in Pour une histoire de la statistique, Economica-Insee, vol. 2, ed. par J. Affichard, *Matériaux*.
[32] Platek R., et Särndal C.-E. (2001), Can a Statistician Deliver ?, Journal of Official Statistics, 17(1), 1-20.
[33] Razafindranovona, T. (2015), La Collecte Multimode et le Paradigme de l'Erreur d'Enquête Totale, Document de travail de l'Insee n°M2015/03.
[34] Roth, N. (1989), Une procédure d'estimation de précision dans les enquêtes auprès des ménages: PRECIS, *Courrier des Statistiques*, 49, 73.
[35] Tak S., Toet A., et van Erp J. (2014), The Perception of Visual UncertaintyRepresentation by Non-Experts, *IEEE Transactions on Visualization and Computer Graphics*, 20(6), 935-943.
[36] Thélot C. (1987) La mesure de l'évolution récente du chômage. *Économie et statistique*, 205, 39-48.
[37] Zucman, G. (2013) The Missing Wealth of Nations, Are Europe and the U.S. net Debtors or net Creditors?, *Quarterly Journal of Economics*, 128(3), 1321-1364.