Donnees massives pour la statistique publique : enjeux, methodes et perspectives

Pauline GIVORD(*), Stéphanie COMBES(*)

(*)INSEE-DMCSI

Résumé

La prolifération exceptionnelle des données numériques ces dernières années laisse penser que de nombreuses disciplines, dont l'économie et la statistique, bénéficieront de ces nouveaux gisements d'information. Les technologies permettant de traiter de tels volumes de données se sont développées à un rythme impressionnant sur la période récente. L'utilisation de ce type de données pour la production d'indicateurs statistiques ou pour mener des analyses économiques se diffuse. Les applications les plus médiatiques ont porté sur le potentiel des requêtes internet pour fournir des indicateurs avancés des épidémies de grippe ("Google Flu") ou des indicateurs économiques [25, 13, 11]. Les instituts de statistique publiques s'intéressent également au potentiel de ces données, et plusieurs expérimentations sont en cours (utilisation des données de caisse pour l'indice des prix à la consommation par exemple). Les principaux apports identifiés seraient de pouvoir diffuser des indicateurs à un niveau de détail plus élevé (maille locale plus fine, sous-populations par exemple), et pour certaines d'entre elles de réduire les délais de publication grâce à un accès rapide aux flux de données. Le grand nombre d'observations peut également permettre d'améliorer la modélisation des processus économiques, en autorisant des non linéarités et/ou une meilleure prise en compte de l'hétérogénéité. Leur exploitation soulève néanmoins de nombreuses questions. L'utilisation des données massives est également un enjeu technique et statistique dont le praticien doit avoir une bonne compréhension pour faire des choix méthodologiques raisonnés. Nous proposons un premier aperçu de ces questions, sans viser à l'exhaustivité, ce qui serait illusoire compte tenu du rythme des innovations techniques et statistiques. Nous commençons par définir et présenter rapidement les caractéristiques de ces sources, et nous esquissons des pistes d'utilisation et les problèmes qui se posent pour les instituts de statistiques publiques (accès, confidentialité, qualité). Nous détaillons ensuite des aspects plus méthodologiques liés à l'exploitation de données de grande taille (analyse de données stockées en parallèle, méthodes de réduction de la dimension).

Introduction

Comme tout terme à la mode, le Big Data n'a pas de définition unique. De manière tautologique, le terme a été initialement développé pour désigner des données tellement volumineuses que les traitements classiques deviennent problématiques. A ce volume s'ajoute la fréquence élevée à laquelle ces données peuvent être collectées (vélocité), parfois en temps réel ce qui pose des questions spécifiques sur la gestion des flux de données. Enfin, leur complexité ou leur variété rendent caduques la plupart des outils de gestions de données usuels : il peut s'agir de données qui peuvent être semi voire non structurées (données issues d'internet par exemple), contenir des informations complexes à traiter (textes, images satellites...). Cette variété provient de la manière dont ces données sont générées : il s'agit de données brutes, issues de l'enregistrement d'activités (on les désignera parfois dans la suite de ce texte de données « passives »). Elles proviennent de capteurs (téléphones mobiles, GPS, satellites...) ou d'internet (comptage des requêtes, réseaux sociaux...), mais il peut aussi s'agir de données administratives (données hospitalières par exemple) ou de transaction (données de caisses ou de facturations...).

Les technologies permettant de traiter de tels volumes de données se sont développées à un rythme impressionnant sur la période récente. L'idée s'est aussi imposée que ces données recèlent des potentiels de création de valeur, conduisant à des investissements massifs dans les méthodes d'analyse de données. Des tentatives d'utiliser ce type de données pour la production d'indicateurs statistiques ou pour mener des analyses économiques ont été faites. Les plus médiatiques d'entre elles ont porté sur le potentiel des requêtes Internet pour fournir des indicateurs avancés des épidémies de grippe ("Google Flu") ou des indicateurs économiques [25, 13, 11]. Les instituts de statistique publiques s'intéressent également au potentiel de ces données, et plusieurs expérimentations sont en cours. L'Unece [1] et Eurostat ont mis en place des groupes de travail pour mener une réflexion commune aux INS sur les potentialités de ces nouvelles sources de données et les questions qu'elles soulèvent.

Ce document présente une brève synthèse de ces questions. Il commence par rappeler dans la partie 1 les opportunités offertes par ces données pour les instituts de statistiques publiques, mais aussi les problèmes qu'elles soulèvent. Ces problèmes sont de divers ordres. Ethique d'abord, puisqu'il s'agit de garantir la confidentialité de données individuelles a priori sensibles, légal aussi, puisque les conditions d'accès à des données parfois propriétés de partenaires privés ne sont pas toujours assurées. Comme pour l'utilisation de données administratives, la question de la qualité de données dont l'objectif initial est éloigné de la production d'indicateurs statistiques doit aussi être discutée. Enfin, traiter des données massives constitue également un enjeu technique. L'enregistrement sur des bases parallélisées est devenue une technique de stockage standard des données massives, qui peut nécessiter l'adaptation des traitements statistiques usuels. Au-delà du nombre d'observations, la grande dimension correspond souvent à de très nombreux régresseurs, et des techniques ont été apportées pour relever cet autre défi. Il s'agit, à la fois pour les choix technologiques et pour les méthodes statistiques, d'un domaine en pleine expansion. Nous nous contentons donc de présenter dans la partie 2 les technologies qui, si elles sont récentes, sont à l'heure actuelle devenues standards (avec le risque assumé de faire face à leur rapide obsolescence), et nous ne faisons qu'esquisser les questions statistiques et les

1. http://www1.unece.org/stat/platform/display/bigdata/Big+Data+Projects

pistes méthodologiques permettant d'y répondre.

1 Statistiques publiques et données massives

1.1 Big data : des opportunités pour le système statistique...

Sur les dernières années, les instituts de statistiques publiques ont commencé à s'intéresser à la possibilité d'intégrer certaines de ces sources dans leur production d'indicateurs économiques et statistiques. Plusieurs expérimentations ont été lancées dans ce but. Par exemple, l'Insee a commencé à explorer la possibilité d'utiliser des données de caisse (remontée des facturettes issues de la grande distribution) dans le calcul de l'indice des prix à la consommation [27, 23] et s'est intéressé à la possibilité d'utiliser les requêtes des utilisateurs pour prévoir la conjoncture économique [9] ; le potentiel des données de téléphonie mobile pour améliorer les statistiques de tourisme a été étudié dans le cadre d'un projet piloté par Eurostat [1] [2] ; Statistic Netherlands a construit un indicateur de confiance des ménages à partir des données de réseaux sociaux [14] ; l'institut de statistique australien s'intéresse à la prévision de productions agricoles à partir des données satellites [29]...

Pour les instituts statistiques, les opportunités identifiées seraient de plusieurs ordres (voir [18, 28, 15] pour une discussion générale). Tout d'abord, les données massives seraient susceptibles d'affiner certains champs. Le volume des enregistrements peut permettre de produire des indicateurs à une « granularité » plus fine, par exemple à l'échelle d'un territoire, sur des sous-populations ou certains marchés. Les données de caisse pourraient permettre d'améliorer la précision de l'indice des prix à la consommation, de fournir des prix moyens sur certains produits et permettre des comparaisons régionales [23]. S'il est peu vraisemblable que ces données passives puissent remplacer les données d'enquêtes, les enquêtes permettant des caractérisations socio-démographiques qui sont en général absentes des données massives, une utilisation complémentaire de ces sources voire un appariement présenterait de nombreux atouts. Il pourrait s'agir d'alléger la charge de réponse pour les enquêtés [26]. C'est particulièrement le cas pour des questions complexes, mobilisant un temps et/ou requérant des efforts de mémoire importants de la part des enquêtés (par exemple, la mesure des temps de transports, la description du budget de famille, les dépenses de santé). La réduction des coûts est également souvent évoquée, les données massives étant en principe issues d'enregistrements automatiques, mais ce gain devra être mis en regard du coût induit par le retraitement de données souvent complexes.

1.2 ...mais aussi de nombreux défis

1.2.1 Confidentialité

Collecter et utiliser des données individuelles à partir des achats, des recherches internet, des déplacements soulèvent des questions sur les garanties qui peuvent être données en termes de respect de la vie privée. La tentation d'apparier les différentes sources de données rend les possibilités de réidentifications élevées (on parle d'effet mosaïque). Cette question dépasse les instituts statistiques, mais peut légitimer leur intervention dans la

2. L'ensemble des rapports de cette étude de faisabilité menée par un consortium sont disponibles à l'adresse : `http://mobfs.positium.ee/`

définition et l'application de règles autour de l'utilisation et de l'appariement de ces nouvelles sources de données. Elle nécessite un investissement conséquent dans les travaux de recherche sur les techniques permettant de garantir la confidentialité.

1.2.2 Accès aux données

La plupart des nouvelles sources sont produites par des opérateurs privés, dont elles sont donc la propriété. Au-delà de la nécessité et des possibilités d'encadrer les utilisations à des fins commerciales sans le consentement explicite des consommateurs, cela soulève plusieurs questions pour la statistique publique.

Trivialement, une question d'accès à ces sources de données, qui n'est pas garanti actuellement par un cadre législatif. La question est d'autant plus sensible que des opérateurs privés peuvent être amenés à commercialiser certaines de leurs données. Cela peut rendre moins audible la production des instituts de statistiques publiques, et doit conduire à une réflexion sur la délimitation de leur champ d'intervention légitime.

Au-delà des aspects légaux ou financiers d'accès aux données, se pose la question de la pérennité et de l'indépendance de cet accès. S'appuyer sur des données privées (ou administratives) pour produire des indicateurs statistiques nécessite de garantir que l'accès ne puisse être remis en cause, et les données ne puissent être manipulées. Or la stabilité même de certaines de ces données n'est pas toujours assurée : c'est en particulier le cas pour les données de réseaux sociaux, dont l'utilisation est volatile et soumise aux effets de mode. Rien n'assure par exemple que Twitter, dont la création date de 2006, sera encore un réseau social utilisé dans quelques années. Ce constat est aussi fait par [9], qui s'intéressent à la possibilité d'utiliser l'outil Google Trend pour améliorer les prévisions conjoncturelles. La méthodologie de cet outil, qui fournit des chroniques correspondant à l'évolution de l'intérêt des internautes pour une requête ou une classe de requêtes, est parfois modifiée. Cela se traduit par une variabilité au cours du temps des chroniques disponibles pour l'utilisateur et pourrait potentiellement entraîner celle de la chronique elle-même pour une requête précise. Par conséquent, une vérification fréquente de la méthodologie serait nécessaire pour garantir la qualité d'une production statistique qui serait fondée sur ce type de données.

1.2.3 Qualité des données

La statistique publique s'est construite historiquement sur des dispositifs de collecte, recensements et sondages, spécifiques, construits de manière à mesurer des indicateurs définis a priori. Les enquêtes par sondage s'appuient sur une théorie bien établie. De nombreux outils et méthodes statistiques dédiés ont été développés dans ce cadre. Utiliser des données « passives » qui résultent simplement de l'enregistrement d'activités oblige à changer de perspectives. Même si ces questions ne sont pas nouvelles pour la statistique publique, dont la production s'appuie depuis plusieurs décennies sur l'utilisation de sources administratives, elles prennent une nouvelle dimension avec des données qui peuvent être complexes.

La première question est celle de la représentativité des données massives. Celles-ci sont en général générées par une fraction de la population uniquement (les abonnés d'un opérateur de téléphonie mobile, les utilisateurs de réseaux sociaux...). Cette population

est a priori singulière, ce qui crée des biais de sélection, mais le statisticien dispose rarement des informations qui lui permettraient d'en apprécier l'ampleur, et de tenter de les corriger (par exemple, la part de marché d'un opérateur téléphonique ou les caractéristiques des utilisateurs de Twitter). Les problèmes de représentativité de ce type de données sont donc complexes à traiter. Les apparier avec d'autres sources peut permettre d'apporter des précisions (par exemple, redresser des données d'utilisateurs à partir de données exhaustives sur la population). Cependant, les redressements peuvent être plus complexes que dans le cadre classique de la statistique d'enquête. Celui-ci s'appuie sur la notion de population de référence, qui représente typiquement des ménages, des individus ou des entreprises. L'unité d'observation dans le cadre des données massives est le plus souvent un événement, une action de ces individus, et il n'est pas immédiat de définir leur lien avec une population de référence [3].

Par ailleurs, il n'est pas assuré que « beaucoup d'informations faibles » suffisent à obtenir une mesure pertinente, dit autrement la quantité ne peut pas remplacer la qualité [26]. Les données massives sont a priori bruitées, il s'agit de données en général non régulières et hétérogènes, et il est important de définir un cadre pour évaluer l'ampleur de l'accumulation éventuelle des erreurs aux différentes étapes de la construction d'un indicateur statistique : génération des données, extractions et transformations, analyse statistique (voir par exemple [8] sur le modèle de l'erreur totale pour les données d'enquêtes). Pour des raisons techniques et statistiques, l'étape d'analyse peut être à l'origine d'importantes erreurs lorsque les données sont très volumineuses. Nous développons ces derniers points dans la partie suivante.

2 Données massives et méthodes statistiques

2.1 Données distribuées, calculs parallèles

Le traitement des données massives a d'abord été un enjeu technique. Avant même d'envisager une exploration de données massives pour faire émerger de l'information, il faut être en mesure de stocker ces données et de communiquer avec elles à partir de requêtes dans un temps de calcul raisonnable. Pour ce faire, des environnements logiciels adaptés ont été développés depuis quelques années. Nous commençons par en rappeler brièvement le principe, pour évaluer dans quelle mesure ces choix techniques peuvent influer sur les performances et la pertinence des modèles statistiques couramment utilisés. Les questions statistiques soulevées par le grand nombre de données ne s'arrêtent cependant pas à la gestion de stockage des données en parallèle. Nous abordons ces questions dans la suite, ainsi que les solutions qui peuvent leur être apportées. Notons que ce document ne fournit qu'un premier aperçu de ces questions : le lecteur intéressé trouvera des synthèses plus complètes en particulier sur les solutions, notamment logicielles, pour l'analyse statistique des données volumineuses dans [7] ; sur les problèmes statistiques posées par la grande dimension dans [21] ou sur les méthodes de machine learning expliquées aux économètres dans [31].

3. On trouvera dans [29] une proposition de modélisation d'un cadre théorique d'une population de référence dans le cadre de données massives.

2.1.1 Le gestionnaire de base de données distribuée

L'utilisation des données massives a été autorisée par l'augmentation impressionnante des capacités de stockage. Une solution devenue classique pour gérer des données massives est le recours à des systèmes de fichiers distribués. Un système distribué est, dans la pratique, une collection d'ordinateurs autonomes (que l'on appelle aussi serveurs ou nœuds), souvent regroupés en *cluster*, et connectés au sein d'un réseau [7]. Le but de ce procédé est de pouvoir utiliser simultanément les capacités de chacun de ces ordinateurs, en évitant les goulots d'étranglements. Les capacités peuvent se comprendre à la fois comme des capacités de stockage mais aussi comme des ressources de calcul, le calcul est alors réparti sur les entités de calcul des différents ordinateurs. E

Depuis quelques années et pour répondre aux besoins d'entreprises gérant des bases de données gigantesques, des environnements logiciels ont été développés sur ce principe. Pour l'utilisateur, la distribution des données et du calcul est invisible puisque l'environnement est conçu de sorte qu'il puisse gérer ses calculs et la mémoire comme s'il n'utilisait qu'un seul ordinateur. Hadoop [4] est l'un de ces environnements logiciels et HDFS [5], pour *Hadoop Distributed File System* fait référence au système de fichiers distribués autour duquel l'environnement est conçu. Outre sa capacité à stocker et gérer un très gros volume de données, ce système a l'avantage de pouvoir gérer des données dites non structurées.

La distribution des calculs dans l'environnement Hadoop peut se faire selon le mode opératoire « MapReduce ». Ce protocole consiste à distribuer une tâche sur différents nœuds, qui correspondent chacun à une partie de la masse de données [6]. De manière très schématique, dans une première étape (« Map »), les calculs sont effectués au niveau de ces nœuds, sur les données stockées à ce niveau. Les résultats partiels sont ensuite redistribués entre les nœuds ("Shuffle"). Ils subissent finalement une deuxième étape de traitement avant d'être agrégés ("Reduce").

On peut illustrer ce mécanisme à partir d'un exemple simple que nous empruntons à [21]. Il s'agit de dénombrer les différents symboles contenus dans une séquence de type "AGTCGGGGCT" (dans cet exemple, il s'agit de dénombrer les différents nucléotides qui composent un gène, mais le principe serait le même pour compter les occurrences de mots par exemple). L'objet en entrée est donc une séquence, et on souhaite obtenir en sortie une table avec les différents symboles et le nombre d'occurrence de chacun d'entre eux dans la séquence initiale. La démarche "naturelle" consiste à prendre le premier symbole de la séquence, de noter dans la table de sortie son nom et d'initialiser un compteur à 1, puis de revenir à la séquence, d'identifier le second et s'il correspond au premier, d'augmenter de un le compteur correspondant au nombre d'occurrences rencontrées jusqu'ici...

Cette démarche séquentielle est assez simple à décrire, mais demande en pratique de

4. Initié par Google, Hadoop est maintenant développé en version libre dans le cadre de la fondation Apache.

5. On peut noter que la distribution des données sur différents nœuds prévoit une réplication des données en divers endroits du réseau afin d'assurer une robustesse aux pannes.

6. Cette approche ne doit pas être confondue avec la parallélisation d'un calcul sur les cœurs d'un micro-processeur par exemple, au sens où dans ce dernier cas chacun des nœuds impliqués dans le calcul a accès à l'intégralité des données et c'est le calcul lui-même qui est distribué entre différentes unités de calcul [7].

parcourir tout ou partie de la séquence à chaque itération. Cela n'est pas problématique si la séquence est aussi élémentaire que celle de l'exemple... mais le devient si elle correspond au génome humain, constitué de quelques milliards de nucléotides (symboles) : même une opération aussi élémentaire que ce comptage s'avère alors coûteuse en temps. De la même manière et pour utiliser un exemple plus classique pour un statisticien, filtrer une table selon la valeur des attributs d'une variable est une opération a priori simple et classique, mais peut s'avérer fastidieuse dès lors que la table est de taille conséquente. Des opérations légèrement plus élaborées telles que le calcul de statistiques sur ces variables : moyenne, médiane, quantiles peuvent devenir très compliquées et nécessiter des approximations [24].

Le principe d'une méthode de type Map Reduce est alors de distribuer ces calculs sur des unités élémentaires qui correspondent aux fichiers distribués mentionnés précédemment. Ce principe est illustré dans le schéma suivant, qui provient encore de [21]. Lors du chargement des données, la séquence a été coupée en plusieurs sous-séquences, il s'agit des nœuds sur lesquels seront effectués les calculs, ces sous-séquences sont recoupées en sous-séquences qui constituent les lignes d'un fichier stocké sur un nœud. La première étape ("Map") consiste alors à appliquer sur chacun de ces nœuds une fonction dont le résultat est, pour chaque ligne lue dans le fichier, une liste de couples constitués chacun d'une clé et d'une valeur. Ici, la clé correspond au symbole, et la valeur à l'occurrence 1 : on a donc une liste de couples $('A', 1)$,... Dans une deuxième étape de triage ("shuffling and sorting"), on rassemble sur un même nœud toutes les paires correspondant à une même clé. Finalement, on applique une fonction d'aggrégation ("Reduce") qui combine (ici somme) les valeurs correspondantes à chaque clé. On aboutit dans cet exemple à quatre paires constituées d'un symbole et du nombre d'occurrences de ce symbole sur l'ensemble de la séquence. Le nœud qui collecte les résultats, les enregistre dans la base.

Pour certains traitements, cette décomposition peut permettre de réduire le temps de calcul, d'un ordre de grandeur correspondant au nombre du nœuds qui effectuent cette tâche (on parle de scalabilité). Si les traitements effectués sur les donnés sont des dénombrements ou des filtrages, le recours à une architecture distribuée et à la parallélisation des requêtes s'avèrent donc très efficace à partir d'un certain volume des données.

En pratique, une couche logicielle supplémentaire traduit les requêtes que l'utilisateur souhaite implémenter pour interroger sa base de données. Celles-ci peuvent être, pour les traitements les plus simples, écrites sous la forme d'un langage proche du SQL tel que HiveQL [7]. Pour des traitements plus sophistiqués, il est possible de communiquer avec un environnement comme Hadoop par le truchement de multiples langages de programmation (R, Perl, Python, C, Java) mais les temps de calculs pourront se révéler très différents d'un langage à l'autre. En R, si le volume des données est trop important pour qu'une amélioration virtuelle de la mémoire vive suffise (par le biais de packages tels que ff ou bigmemory), alors il est possible d'avoir recours à Hadoop et de formuler les instructions par le biais de Rhadoop, un ensemble de librairies R sous licence libre servant d'interface avec cet environnement. Mais cette option n'est pas forcément la plus efficace en comparaison à d'autres environnements.

En effet, l'architecture de fichiers distribuée de Hadoop est particulièrement adaptée

7. HiveQL est un langage développé par Facebook pour interroger une base Hadoop avec une syntaxe proche du SQL.

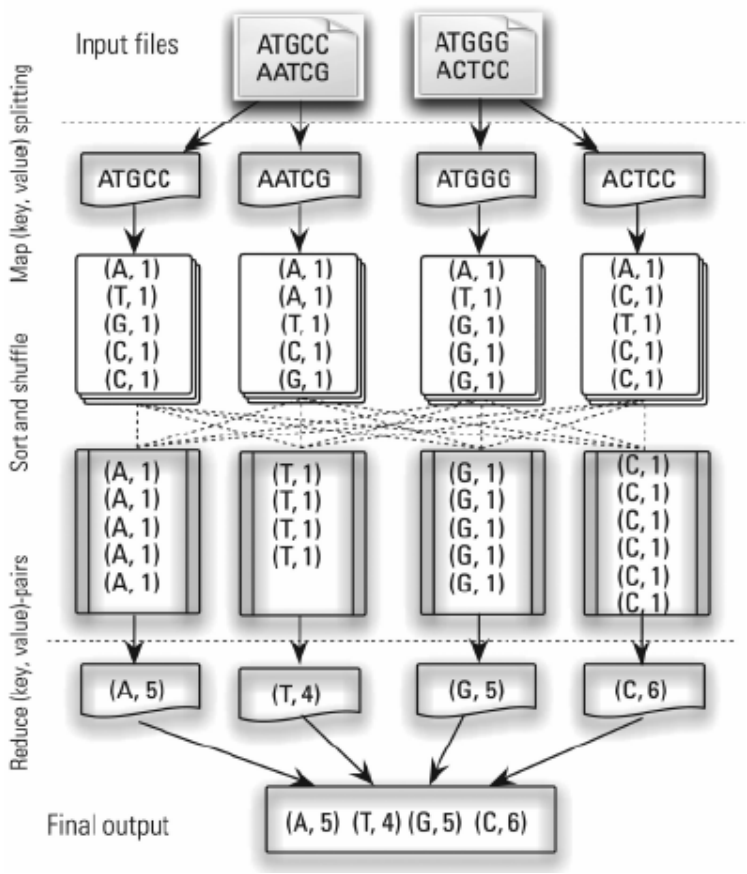

FIGURE 1 – Illustration du fonctionnement de MapReduce [7]

au stockage et à la parallélisation de tâches élémentaires. Cependant, comme chaque tâche réalisée au sein d'un nœud ne peut être gardée en mémoire, et qu'elle requiert la lecture et l'écriture des données dans la base HDFS, un algorithme itératif nécessitera un important nombre d'opérations de lecture-écriture. Celles-ci peuvent devenir très coûteuses en temps d'exécution à mesure que les tâches visées se complexifient. Le projet Apache Spark [8] a été développé dans le but de pallier ces contraintes et permet de garder en mémoire les données entre deux itérations. Mahout propose également une librairie d'algorithmes principalement destinés à l'apprentissage statistique, initialement prévu pour le modèle de programmation MapReduce, mais qui s'adapte désormais à des modèles de programmation plus modernes à l'instar de Spark.

Tous les algorithmes et tous les traitements statistiques ne sont pas aujourd'hui disponibles dans Mahout, Rhadoop ou Spark. Ils ne sont en effet pas tous aptes à être parallélisés simplement et lorsqu'ils le sont, ils doivent être manipulés avec précaution. En effet, un très grand nombre d'observations génère des défis en optimisation numérique tant du point de vue de stabilité et de la vitesse de convergence des algorithmes que de leur stabilité dans un contexte de données distribuées. Un important volume de données contraint le nombre d'itérations que l'on peut s'autoriser et un algorithme devra être évité s'il nécessite un grand nombre d'itérations pour converger (algorithmes devant être lancés

8. Cet environnement est accessible en Java, Scala (langage récent adapté à la programmation d'applications parallélisables), Python et R (librairie SparkR encore récente).

plusieurs fois avec des points de départ différents par exemple) ou simplement pour être réglés (validation croisée). Sans quoi la fiabilité des résultats qu'il produit ne pourra être assurée.

Cependant, la littérature relative à la parallélisation, qui n'est pas très récente, est en constante évolution et les environnements tels que Spark ou Mahout se complètent à ce rythme. La gestion des données massives ne motive pas seulement le développement de nouvelles infrastructures de stockage et de modèles de programmation, mais également les recherches dans le domaine de l'optimisation numérique afin de rendre disponibles et efficaces de plus en plus d'algorithmes. S'il est difficile d'avoir une vision complète à chaque instant des algorithmes disponibles, les algorithmes suivants sont implémentés dans au moins une de ces librairies aujourd'hui : système de recommandation, SVD, k-means, classifieur bayésien naïf, forêts aléatoires, SVM...

Les méthodes économétriques sont pour l'instant moins développées, mais les régressions linéaires et logistiques sont disponibles. Ces deux méthodes faisant partie des outils de base des économètres, nous présentons ici en guise d'illustration comment ils peuvent être mis en œuvre sur des données parallélisées. Il est important de souligner que le fait que ces deux méthodes soient "parallélisables" ne signifie pas qu'elles sont les plus adaptées au traitement des données massives. Nous discutons des problèmes liés à la grande dimension plus loin.

2.1.2 Paralléliser les régressions linéaires et logistiques

Une régression linéaire s'intéresse au lien entre une variable principale, y, et des variables explicatives. On s'intéresse au modèle

$$y = X\beta + e$$

où X correspond aux p variables explicatives (y compris la constante). Elle peut donc se représenter sous la forme d'une matrice de taille (n, p) où n est le nombre d'observations. Sous des hypothèses classiques (pas d'endogénéité ni de colinéarité entre les variables explicatives en particulier), l'estimateur des moindres carrés ordinaires est le meilleur estimateur sans biais, et a une forme close : $\hat{\beta} = (X'X)^{-1}X'Y$.

Néanmoins, si les données sont très grandes, les calculs matriciels, et en particulier l'inversion d'une matrice, peuvent être fastidieux. Dans les cas où la grande taille des données provient du nombre d'observations n et non du nombre des variables p, ces calculs peuvent être distribués simplement. En effet, dans la formule précédente l'inversion de matrice concerne la matrice $X'X$ qui n'est que de taille (p, p). Les produits matriciels $X'X$ et $X'y$ sont linéaires et donc facilement distribuables. Concrètement, les lignes de la matrice X, c'est-à-dire les observations, sont distribuées sur les nœuds. On peut calculer pour chaque nœud a correspondant aux observations $i = n_1 \dots n_2$ les produits :

$$X_j^{(a)'} X_k^{(a)} = \sum_{i=n_1}^{n_2} X_{i,k} X_{i,j}$$

pour $\forall j = 1, .., p$ et $\forall k = 1, .., p$ et $X^{(a)}$ les observations stockées sur le nœud a. Le coefficient (j, k) de la matrice produit $X'X$ correspond alors à la somme des produits

précédents :

$$X'_j X_k = \sum_{i=1}^{n} X_{i,k} X_{i,j} = \sum_a X_j^{(a)'} X_k^{(a)}$$

La matrice $X'X$ peut alors raisonnablement être inversée sur un seul nœud.

Pour une régression logistique, il n'existe pas de forme close comme dans le cas de la régression linéaire. L'estimateur classiquement utilisé est obtenu par minimisation de la vraisemblance du modèle. Il s'agit d'un programme de minimisation tout à fait standard, la fonction de coût étant convexe. Elle peut donc reposer sur un algorithme de descente de gradient qui procède par itération en suivant la pente de la fonction objectif. Ce dernier est simplement parallélisable de la même façon. Formellement, si on note la fonction de lien $h_\theta(x) = g(\theta' x) = \frac{1}{1+e^{-\theta' x}}$, on peut écrire la vraisemblance

$$L(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}, \theta)$$

$$L(\theta) = \prod_{i=1}^{n} (h_\theta(x^{(i)}))^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$$

En passant au logarithme, les calculs se simplifient :

$$l(\theta) = logL(\theta)$$

$$l(\theta) = \sum_{i=1}^{n} y^{(i)} log h_\theta(x^{(i)}) + (1 - y^{(i)}) log(1 - h_\theta(x^{(i)}))$$

En utilisant la propriété de g : $g'(z) = g(z)(1 - g(z))$, on peut montrer que la dérivée partielle de la log-vraisemblance en θ_j vaut : $\sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)})) x_j$. Si on note α le pas (ici il est positif car on maximise la fonction donc on suit le gradient), l'algorithme de descente du gradient consiste ainsi à calculer itérativement θ, et la somme qu'il fait intervenir peut être distribuée sur les différents nœuds comme pour la régression linéaire. Ainsi, à la $(k+1)^{ieme}$ étape, on a :

$$\theta_j^{k+1} = \theta_j^k + \alpha \sum_{i=1}^{n} (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)} = \theta_j^k + \alpha \sum_a (y^{(a)} - h_\theta(x^{(a)})) x_j^{(i)}$$

Notons que calculer un gradient peut-être extrêmement coûteux en temps de calcul lorsque le nombre de paramètres est élevé. Une alternative possible consiste à évaluer le gradient suivant la direction d'une coordonnée à la fois [21].

Plus généralement, la question de la dimension (c'est-à-dire des variables pertinentes pour l'analyse) doit être traitée. Comme souligné par [31], s'il est techniquement possible d'utiliser ces méthodes économétriques standards, les outils développés dans les méthodes d'apprentissage automatique peuvent être plus adaptés à la grande dimension, c'est-à-dire en présence de nombreuses variables. Par exemple, l'algorithme de "forêt aléatoire" peut afficher des performances au moins équivalentes à une régression logistique classique, et est intéressant à double titre dans le contexte des données massives [6] : il repose sur un principe d'agrégation d'un ensemble de modèles estimés sur des échantillons indépendants

("bagging") et donc facilement parallélisable. En outre, il est robuste au sur-apprentissage en grande dimension, c'est-à-dire qu'il ne tend pas à aboutir à un ajustement parfait artificiel (voir plus loin), et ne requiert donc pas de réglage fin. Nous revenons sur ces questions dans la partie 2.2.

2.1.3 Echantillonner une base distribuée

Même si l'on dispose d'un grand nombre d'observations, on peut souhaiter effectuer un traitement ou produire un indicateur qui ne justifie pas de garder tous les points ou, au contraire, on peut chercher à appliquer un algorithme sophistiqué qui n'est pas aujourd'hui disponible dans le librairies des environnements RHadoop, Spark ou encore Mahout. Dans ce cas, on va chercher à échantillonner.

Dans le contexte Big Data, la particularité réside dans le fait que l'on ne sait pas forcément de combien de données on dispose. Pour tirer un échantillon de taille k dans la base de taille n, on peut procéder selon l'algorithme de *reservoir sampling* proposé par [32] qui assure un tirage avec équiprobabilité et n'impose qu'une seule lecture de toute la base [7]. Si la base tenait dans un nœud, cela reviendrait à remplir une matrice (ou vecteur) R de k lignes par les k premiers enregistrements, puis pour tout enregistrement i, si $j < k$ pour j un nombre entier aléatoire compris entre 1 et i, alors on remplace la ligne i de R par cet enregistrement. L'idée sous-jacente à cet algorithme est de dire qu'il est équivalent de prendre un échantillon aléatoire de taille k ou de générer une permutation aléatoire des éléments de sorte à ne garder que les k premiers éléments.

L'astuce pour distribuer ce calcul consiste à générer des identifiants pour chaque élément de chaque nœud : l'étape "Map" associe à chaque élément un couple clé-valeur où la clé est l'opposé de cet identifiant et la valeur est l'élément. Comme par construction Hadoop présente automatiquement les couples clé-valeur aux Reducers en les classant par ordre des clés les plus petites aux plus élevées, il suffit de n'envoyer que les k premiers éléments issus des étapes "Map" de chaque nœud dans un seul nœud qui contient donc les mk éléments aux identifiants les plus élevés, pour m le nombre de nœuds. L'étape "Reduce" consiste à la sélection des k éléments de plus grandes clés parmi ceux-ci. L'échantillon est ainsi construit.

2.2 Inférence statistique et données massives

La sous-partie précédente met l'accent sur les solutions qui peuvent être apportées au stockage de données volumineuses, et au calcul sur celles-ci, une fois qu'elles sont distribuées. Au-delà des questions liées au traitement du grand nombre d'observations (n grand), la présence de très nombreux régresseurs potentiels élevés (p grand) pose parfois question dans le contexte des données massives. L'objectif principal du statisticien est toujours d'extraire l'information pertinente des données, même si il peut se décliner de manière différente selon les disciplines : l'accent est souvent mis sur la capacité prédictive des modèles (par exemple en prévision ou en marketing) ou la mise en avant de corrélations ou de relations de causalité entre les variables (pour des études économiques par exemple).

Des outils adaptés ont été développés pour répondre à ces différents objectifs. Les méthodes de *datamining* (en particulier les algorithmes d'apprentissage statistique évoqués plus haut) répondent plutôt au premier objectif : elles permettent d'explorer la base de

données, de manière "agnostique" sur les liens entre les variables, et d'identifier un modèle optimal au sens d'un critère de performance choisi (classification des individus par exemple). L'économétrie en revanche s'attache surtout à la possibilité de comprendre des phénomènes économiques et cherche à identifier des liens causaux entre les variables. Les spécifications retenues sont en général relativement simples, souvent linéaires.

Ces approches se complètent plus qu'elles ne s'opposent. Dans le contexte de données massives, il peut s'avérer intéressant voire indispensable de les combiner. Disposer de nombreuses observations peut permettre d'enrichir les spécifications économétriques classiques, au-delà d'une simple relation linéaire, ou de prendre en compte des interactions entre variables. L'exploration des données peut s'avérer très intéressante pour faire émerger des relations que l'on ne suspectait pas initialement. Même dans une étude économétrique, les méthodes de sélection de variables deviennent donc nécessaires. Cependant, certains problèmes, classiques (sur-apprentissage, causalité) demeurent, et leur traitement peut devenir plus complexe en grande dimension (voir aussi pour une discussion [6]). Nous proposons un rappel des problèmes qui se posent ici, comment ils peuvent être amplifiés par la grande dimension, et évoquons des pistes pour y répondre.

2.2.1 Des méthodes de réduction de la dimension

Accumulation des erreurs En grande dimension, lorsque l'on doit estimer simultanément un grand nombre de paramètres, des erreurs d'estimation peuvent s'accumuler à tel point qu'elles peuvent fausser ou brouiller l'interprétation d'un résultat. Ce phénomène d'accumulation du bruit peut être illustré par un exemple simple que nous empruntons la encore à [21]. Ces derniers simulent deux vecteurs aléatoires dans deux classes distinctes : $X_1, ..., X_n \sim N_p(\mu_1, I_p)$ et $Y_1, ..Y_n \sim N_p(\mu_2, I_p)$. On fixe $n = 100$ et $p = 1000$, $\mu_1 = 0$ et seules les dix premières composantes de μ_2 sont non nulles. Dit autrement, seuls les dix premiers attributs différencient les deux classes. On effectue une analyse en composantes principales à la sous-matrice $X^{(m)}_{(i,j)}$ pour $i = 1 \dots n$ et en retenant plus ou moins de variables $j = 1 \dots m$. Les deux premiers axes de l'analyse en composantes principales pour m =2, 40, 200 et 1000 sont représentés ci-dessous. On peut voir que pour $m = 2$, on est capable de discriminer les deux populations facilement, alors que l'opération devient de plus en plus ardue à mesure que l'on augmente la dimension du problème. Pourtant, dans le processus de génération des données construit ici, seuls les premiers attributs doivent contribuer à la classification. Au-delà de 10, garder des covariables n'apporte donc plus d'information supplémentaire et introduit du bruit.

Plus généralement, on peut montrer que des algorithmes d'apprentissage statistique pour la classification peuvent présenter des performances similaires à celles d'un simple tirage aléatoire en présence d'un trop grand nombre de variables (voir [16, 19] par exemple).

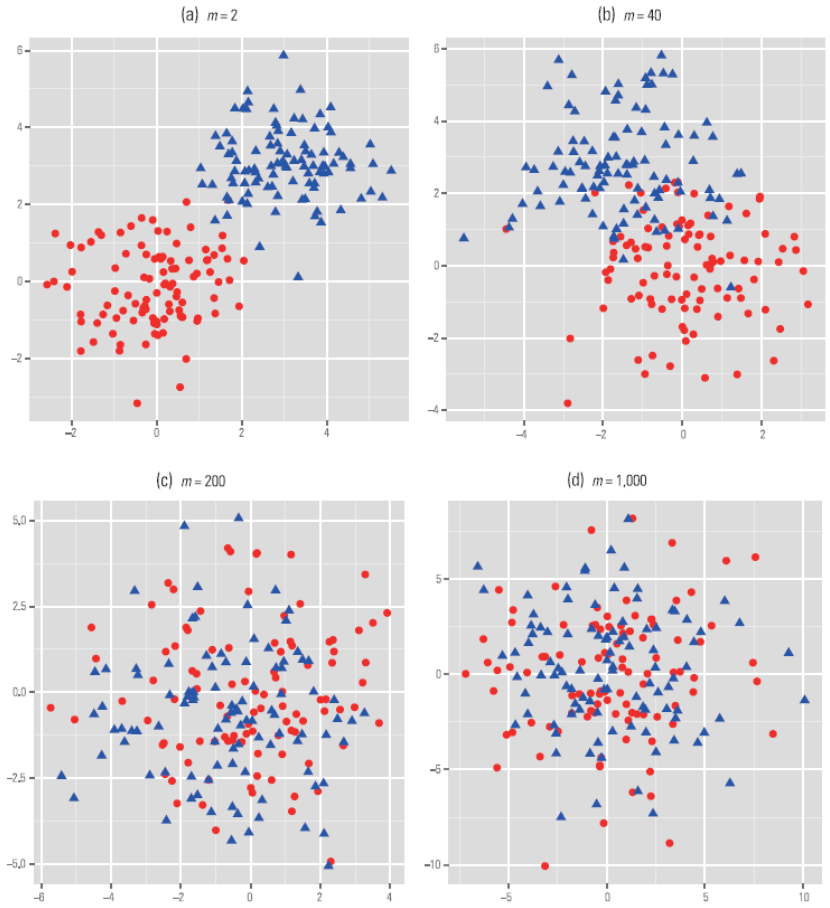

FIGURE 2 – Premiers axes de l'ACP réalisée sur 2, 40, 200 ou 1000 covariables [21]

Sparsité Les méthodes de réduction de la dimension sont une réponse classique à ce problème d'accumulation d'erreur. Ces méthodes sont usuelles en prévision ou pour des méthodes de classifications. Elles reposent sur la notion de "sparsity" : l'idée sous-jacente est que parmi les très nombreuses variables disponibles, seul un petit nombre d'entre elles sont en réalité pertinentes pour expliquer le processus qui nous intéresse. Formellement, si on suppose une relation générale linéaire entre une variable y et des régresseurs x_j, $j \in (1 \dots p)$, avec p grand (soit parce que de nombreuses variables sont disponibles dans la base, soit parce qu'on a adjoint à celles-ci diverses transformations autorisant des non-linéarités ou des interactions dans le but de ne pas trop contraindre la forme du modèle) :

$$y_i = x_i'\beta + \epsilon_i \tag{1}$$

Alors, l'hypothèse de "sparsité" stricte ou de parcimonie implique que parmi ces régresseurs potentiels, seuls certains sont utiles pour l'analyse, c'est à dire qu'un nombre réduit de coefficient β_j sont non nuls. On peut aussi retenir l'hypothèse moins forte que le modèle est seulement "approximativement" parcimonieux [5] : un modèle avec un nombre réduit de régresseurs peut fournir une approximation correcte du vrai modèle. On cherchera donc à déterminer les régresseurs les plus importants parmi l'ensemble de régresseurs potentiels. Il est aussi possible qu'une grande partie des régresseurs soient pertinents mais qu'ils partagent des facteurs communs dont l'identification suffit à estimer correctement le processus y (voir la partie variables synthétiques plus loin). Dans tous les cas, le fait

de devoir déterminer les régresseurs pertinents (ou de résumer l'information contenue par l'ensemble des variables par le biais de facteurs) plutôt que de les identifier a priori distingue ces approches de l'économétrie classique.

LASSO et méthodes de sélection de variables Il existe une multitude d'algorithmes visant la recherche de la parcimonie dans un modèle, mais les deux familles principales sont les suivantes : les algorithmes (backward, forward, stepwise...) de sélection de modèle par sélection de variables et minimisation de critères pénalisés (Cp, AIC, BIC), et les algorithmes de sélection de modèle par pénalisation (Ridge, LASSO, elastic net). Les premiers procèdent de façon itérative en t ou retirant une variable à la fois, le nombre total de variables conservés étant limité grâce au critère d'information pénalisé. Les secondes forcent l'annulation de nombreux coefficients par rapport à la solution optimale des moindres carrés en contraignant la norme du vecteur des paramètres lors de l'estimation. Ces méthodes de pénalisation, dont nous allons décrire un exemple plus loin, sont plus robustes [20]. D'autres approches plus sophistiquées existent, comme les approches bayésiennes d'agrégation de modèles (Bayesian Model Averaging) par exemple, qui permettent également de procéder à une sélection de variables mais de façon indirecte. Les variables les plus pertinentes sont celles qui sont sollicitées dans les modèles sélectionnés par la procédure en fonction de leur probabilité a posteriori d'avoir pu générer les données (pour une application voir [9]). L'objectif commun à l'ensemble de ces approches reste de permettre une meilleure interprétation des résultats ou de meilleures performances en prévision, en jouant sur le nombre de paramètres introduits dans la modélisation.

Un des algorithmes les plus connus dans la famille des méthodes de sélection de modèles par pénalisation est le *Least Absolute Shrinkage and Selection Operator*, soit *LASSO* initialement proposé par Tibshirani [30], dont plusieurs variantes ont été développées depuis.

Formellement, on part du modèle classique décrit en 1. L'hypothèse de sparsité sous-jacente au processus générant les données entraîne que β est un vecteur dont la majorité des composantes sont nulles. Pour l'estimer, la procédure classique consiste à minimiser la quasi-vraisemblance pénalisée pour la norme L_1 (somme des valeurs absolues des composantes) :

$$-QL(\beta) + \lambda\|\beta\|_1$$

λ est un paramètre positif qui contrôle l'arbitrage entre biais et variance. Cette expression peut être estimée par :

$$l_n(\beta) + \sum_{j=1}^{d} P_{\lambda,\gamma}(\beta_j)$$

où $l_n(\beta)$ mesure la qualité de l'ajustement pour le paramètre β et $\sum_{j=1}^{d} P_{\lambda,\gamma}(\beta_j)$ est la pénalité qui entraîne la parcimonie du modèle.

Dans l'illustration ci-dessous que nous empruntons à [12], on voit une vue en coupe d'une surface Q convexe à minimiser. On peut voir que sans contrainte, le minimum serait obtenu pour une grande valeur de β. On rajoute un terme de pénalité L_1, pour un λ fixé. Si λ est suffisamment élevé pour que $Q + \lambda\|\beta\|_1$ soit croissante pour $\beta > 0$, on obtient la courbe rouge, convexe mais non différentiable en zéro puisque la norme L_1 ne l'est pas. Cette géométrie ramène brutalement le minimum à zéro ("shrinkage").

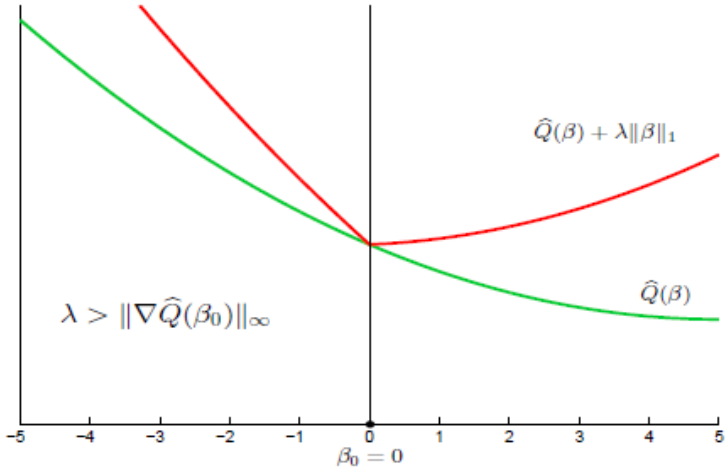

FIGURE 3 – Géométrie de la fonction de coût à minimiser avec la pénalité Lasso [12]

La norme L_1 peut être remplacée par une autre norme ou une combinaison de normes (Elastic net), entraînant plus ou moins facilement l'annulation d'un ou de nombreux paramètres. Dans le cas du Lasso, la géométrie de la norme L_1 entraîne une sélection particulièrement agressive des variables, ce qui peut être très utile en très grande dimension.

Variables synthétiques Les méthodes citées précédemment utilisent les variables "telles qu'elles". Une autre approche classique pour réduire la dimension est de résumer l'information contenue dans de nombreuses variables dans un nombre réduit de variables synthétiques. L'analyse en composantes principales est la méthode la plus connue en statistique descriptive, mais c'est également le principe des modèles à facteurs en séries temporelles (cf [2] pour une revue de la littérature). L'idée consiste à projeter les données sur l'espace orthogonal de plus faible dimension capturant le plus de variance possible. Pour cela il faut calculer les valeurs propres de la matrice de covariance empirique. Les algorithmes de décomposition en valeurs singulières d'une grande matrice creuse adaptés au paradigme MapReduce et nécessaires pour une analyse en composantes principales ne sont proposés que par Mahout à ce jour [7]. La complexité du calcul de l'ACP est exponentielle en p : $O(p^2n + p^3)$ où n est le nombre d'observations et p le nombre de variables. La détermination des valeurs propres peut donc rapidement poser problème.

Cependant, lorsque p est vraiment grand, il est possible de montrer qu'une projection aléatoire sur un sous-espace de dimension $k < n$ permet de conserver la structure des données et de réduire la dimension [21] (le problème devient en $O(pnk)$). Les auteurs qui ont développé cette méthode justifient la démarche en montrant que pour k bien choisi la projection aléatoire conserve assez bien les distances entre les points. En principe, la projection doit être orthogonale, or orthogonaliser une matrice est coûteux en temps de calcul. Cependant, en très grande dimension, un nombre fini de vecteurs (constitués de 0 et de 1) pris aléatoirement dans cet espace sont presque orthogonaux entre eux.

Matrices creuses Une autre conséquence de la sparsité est que les données peuvent parfois être représentées par de grandes matrices creuses. On peut alors appliquer la méthode de réduction de la dimension par factorisation sous contrainte de non négativité des

facteurs. Contrairement à l'ACP, les facteurs ne sont pas orthogonaux et ne permettent pas de représenter les informations dans un espace de taille plus petite, mais ils permettent l'application d'algorithmes d'apprentissage ou de modèles de prévision. Cette technique a depuis été depuis largement utilisée dans de très nombreux domaines avec pour objectif d'étudier la structure des très grandes matrices creuses [9] : imagerie, reconnaissance de formes, fouille de textes, systèmes de recommandations, génomique. Il existe plusieurs algorithmes itératifs principalement conçus pour fournir une solution efficace mais pas forcément unique à ce problème de décomposition.

Formellement, pour X une matrice de taille (n, p), il s'agit donc, pour $r < p$, de rechercher deux matrices ne contenant que des valeurs positives ou nulles et dont le produit s'approche de X :

$$x \approx W_{(n,r)} H_{(r,p)}$$

La factorisation est résolue par la recherche d'un optimum local du problème d'optimisation suivant :

$$min_{W,H \geqslant 0} \mid L(X, WH) + P(W, H) \mid$$

Avec L une fonction de perte mesurant la qualité de l'approximation et P une fonction de pénalisation optionnelle.

2.2.2 Grande dimension, corrélation fortuite et sur apprentissage

Corrélation fallacieuse Disposer de nombreux régresseurs pose des questions d'inférence statistique que les méthodes précédentes ne permettent pas de régler complètement. La première est celle de la corrélation fortuite. En grande dimension, la probabilité d'observer des corrélations importantes entre variables n'ayant en réalité aucun lien augmente. On peut être tenté d'interpréter à tort une corrélation fortuite (*spurious correlation* ou corrélation fallacieuse) ou d'intégrer à une modélisation des variables qui n'apportent en réalité pas d'information mais réduit artificiellement la variance des résidus. On parle alors de sur-apprentissage (voir plus bas) pour la prédiction.

Ce phénomène a été illustré par [21] à travers une simple simulation. On simule un grand nombre d'échantillons de taille n, à partir d'un tirage d'un processus gaussien x de dimension p : $X = (X_1, ..., X_p)^T \sim N_p(0, I_p)$. Il s'agit donc de p variables indépendantes entre elles. Pour chacun des échantillons simulés, on peut estimer la corrélation empirique obtenue entre une variable X_1 et les autres variables de l'échantillon prise une à une. On peut alors calculer la corrélation maximale observée pour un tirage donné lorsque le nombre de variables indépendantes p vaut respectivement 800 et 6400 (dans les deux cas, on suppose un échantillon réduit $n = 60$). Les auteurs constatent que la corrélation absolue maximale moyenne pour mille tirages augmente nettement pour $p = 6400$ par rapport au cas où $p = 800$. Pour chaque couple de variables la "vraie" corrélation est en réalité nulle par construction. Statistiquement, il est cependant vraisemblable qu'on puisse par pur hasard obtenir une valeur élevée de la corrélation empirique entre une variable et une autre. Plus on introduit de variables dans le modèle, et plus le risque est grand d'exhiber une telle corrélation.

9. `http://wikistat.fr/pdf/st-m-explo-nmf.pdf`

Sur-apprentissage En outre, introduire de nombreux régresseurs "fallacieux" va conduire à sous-estimer la précision du modèle, et donc aboutir à des choix de variables et des tests statistiques biaisés (voir [17] pour une discussion et une méthode reposant sur de la validation croisée). Cela a un impact important dans les modèles en prévision par exemple : les régressions factices permettent de minimiser l'erreur du modèle sur l'échantillon, mais risquent de donner des résultats incorrects en dehors de l'échantillon. C'est le risque de sur-apprentissage.

Ce phénomène peut être illustré simplement : supposons par exemple, qu'on s'intéresse aux données générées à partir d'une fonction h et d'un aléa e telles que $y(x) = h(x) + e$, avec une dépendance de y en x quadratique. Le praticien, qui ignore en principe cette « vraie » relation, peut chercher un modèle le plus simple $g(x)$ qui consiste en une dépendance simplement linéaire. Ce modèle fait intervenir peu de paramètres mais génère un biais important : l'ajustement est mauvais (graphique de gauche). Inversement, le prévisionniste peut, en ajoutant un grand nombre de paramètres dans la modélisation, effectuer un ajustement parfait (cf. graphique de droite). Mais cet ajustement parfait capte, à tort l'aléa intervenant ici dans le processus de génération de ces données, par définition imprévisible. Pour éviter le « sur-apprentissage » illustré dans le graphique de droite, mais réduire le biais que l'on constate sur le graphique de gauche, un arbitrage doit être fait entre biais et variance. En général, on utilise des critères pour juger des bonnes performances sur des observations n'ayant pas servi à l'estimation. Dans cet exemple, il s'agirait par exemple de regarder de quelle façon évolue le RMSE hors échantillon (root-mean-square error, c'est-à-dire la moyenne des écarts au carré entre prévision et réalisé) avec le nombre de paramètres retenus dans le modèle, et de conserver le paramétrage qui minimise ce critère (on parle de validation croisée). Ces méthodes sont cependant coûteuses et peuvent être difficiles à mener en grandes dimensions.

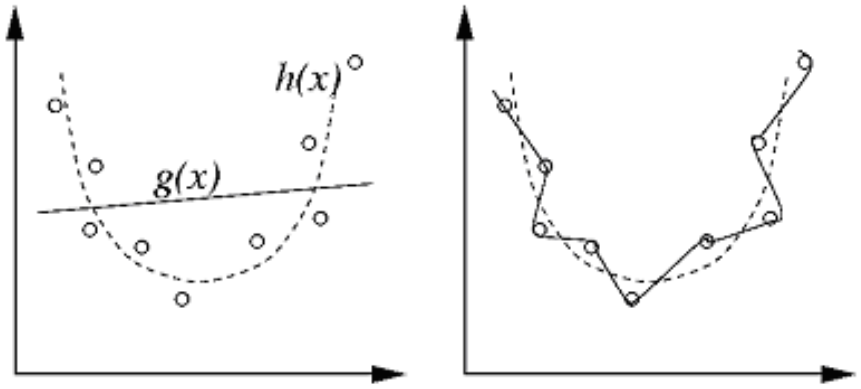

FIGURE 4 – Arbitrage biais variance

Comme analysé par [22], la défaillance de l'application très médiatisée de Google connue sous le nom de Google Flu, est sans doute liée à ce problème de sur apprentissage, du fait de corrélations fortuites. Cet outil, lancé en 2008 aux Etats-Unis puis étendu à une douzaine de pays européens par la suite, avait pour objectif de prévoir la progression de l'épidémie de grippe. A ces débuts, l'outil parvenait à fournir un indicateur avancé d'une à deux semaines par rapport aux publications officielles aux Etats-Unis, tirant parti de la disponibilité quasi immédiate des requêtes Google. Il a ensuite été mis en défaut une

première fois en 2009 lorsqu'il n'a pas été capable de prévoir l'épidémie non saisonnière de grippe porcine (H1N1) puis lors des hivers 2011-2012 et 2012-2013, en surestimant largement les épidémies de grippe aux États-Unis. Pour produire son indicateur, Google a cherché à identifier parmi des milliers de mots-clé les plus corrélés aux évolutions des indicateurs officiels fournis par des organismes de veille sanitaire. Les termes de recherche qui connaissaient des pics de fréquence d'utilisation identiques à ceux de la progression de la grippe saisonnière ont été sélectionnés. Or, certains termes utilisés pour prévoir l'épidémie étaient bien corrélés aux données sanitaires en raison de comportements saisonniers similaires mais ne présentaient pas de rapport direct avec la maladie [10].

Endogénéité Une autre question plus familière aux économistes est celle de l'endogénéité des variables explicatives : dans le cadre simple d'un modèle linéaire, les estimations peuvent être biaisées lorsque les variables explicatives sont corrélées avec le terme résiduel du modèle. La corrélation n'implique pas la causalité, et si l'objet de l'étude est d'évaluer un lien de cause à effet (par exemple, l'efficacité d'une campagne de publicité sur les achats), ne pas tenir compte de ces dimensions peut conduire à des conclusions erronées. En particulier, une sélection aveugle de variables peut entraîner l'ajout de variables qui n'ont aucun lien avec la variable à expliquer, mais sont "fortuitement" corrélées avec le résidu, et le risque est d'autant plus grand que le nombre de régresseurs potentiels est élevé. Les résultats de l'estimation seront alors biaisés et la modélisation invalide [21].

Cette dimension est traditionnellement peu prise en compte dans les méthodes usuelles de data mining, dont les performances sont souvent jugées en fonction des qualités prédictives. Cependant, l'endogénéité peut aussi avoir des conséquences sur ces performances si elles conduisent à retenir à tort certainement variables. Des travaux récents s'intéressent à la modélisation en grande dimension prenant en compte l'endogénéité des variables. Par exemple, [4] proposent plusieurs approches permettant de faire de l'inférence causale lorsque de nombreux régresseurs (ou instruments) potentiels sont présents. Citons également [10] qui développent une méthode d'estimation bayésienne des évolutions contrefactuelles d'une variable d'intérêt, nécessaire à estimer un effet causal, dans le cas de séries temporelles à haute fréquence.

3 Conclusion

Le développement exponentiel des sources de données sur la période récente, et de leur appropriation, ne peut laisser la statistique publique indifférente. L'apparition de nouvelles sources de données peut offrir des opportunités de compléter, affiner ou améliorer la production des indicateurs classiquement produits par la statistique publique ; elle doit aussi l'amener à se positionner face à de nouveaux producteurs d'informations [11], tout en maintenant une forte exigence de rigueur et d'indépendance qui font l'identité des statisticiens publics. Des réflexions ont été déjà engagées dans plusieurs instituts statistiques pour définir une stratégie d'utilisation des données massives (voir [15], [3]). Ce document présente un premier aperçu des questions que soulève l'utilisation des données

10. [22] explique également cette perte de performance par l'évolution du comportement de recherche des internautes en réponse aux évolutions du moteur de recherche lui-même.

11. par exemple, le Billion Prices Project du MIT télécharge, en coordination avec des sites de vente en ligne, les prix journaliers de centaines de milliers de produits et les caractéristiques de ces derniers dans le but de produire des indicateurs de prix quotidiens [11].

massives. Il ne vise évidemment pas à l'exhaustivité, et ne propose qu'un survol des méthodes techniques et statistiques, qui se développent à un rythme rapide. Utiliser ce type de données non structurées peut également nécessiter d'autres compétences non développées ici : analyse textuelle pour gérer des contenus, visualisation, traitement de la confidentialité, appariement de données différentes... Sur la dernière période, les progrès des techniques informatiques avaient pu offrir à la plupart des statisticiens le confort d'oublier les aspects plus techniques de la gestion et du traitement des données. Au moins sur le court terme, les statisticiens gagneront à acquérir une meilleure connaissance de ces questions pour gérer des données volumineuses, les performances des outils pouvant en être affectées. Une intégration plus importante des équipes d'ingénierie informatique et d'analyse statistique sera sans doute nécessaire. Des compétences, notamment juridiques et en communication sont aussi indispensables. Les enjeux dépassent en effet ces seuls aspects techniques, puisqu'il s'agit aussi de garantir un accès à des données, en toute indépendance, dans un respect strict des libertés fondamentales et notamment de la vie privée.

Références

[1] Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.-L., Nurmi, O., Potier, F., Schmucker, D., Sonntag, U., and Tiru, M. Feasibility study on the use of mobile positioning data for tourism statistics -consolidated report. Tech. rep., Consortium, June 2014.

[2] Bai, J., and Ng, S. Large dimensional factor analysis. *Foundations and Trends(R) in Econometrics 3*, 2 (2008), 89–163.

[3] Baldacci, E., Barcaroli, G., and Scannapieco, M. Using big data for statistical production : The Istat strategy. In *DGINS Conference* (2013), Eurostat.

[4] Belloni, A., Chernozhukov, V., and Hansen, C. High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*, 2 (2014), 29–50.

[5] Belloni, A., Chernozhukov, V., and Hansen, C. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies 81*, 2 (2014), 608–650.

[6] Besse, P., Garivier, A., and Loubes, J.-M. Big Data Analytics - Retour vers le Futur 3 ; De Statisticien à Data Scientist. Mar. 2014.

[7] Besse, P., and Villa-Vialaneix, N. Statistique et big data analytics, volumétrie - l'attaque des clones, 2014.

[8] Biemer, P. Dropping the s to tse : applying the paradigm to big data. In *The 2014 International Total Survey Error Workshop* (2014), National Institute of Statistical Science.

[9] Bortoli, C., and Combes, S. Apports de google trends pour prevoir la conjoncture francaise : des pistes limitees. In *Note de Conjoncture*. Insee, Mars 2015, pp. 37–51.

[10] Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., and Scott, S. L. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics* (2014).

[11] Cavallo, A. Scraped data and sticky prices. *Massachusetts institute of technology working paper*, 4976-12 (2012).

[12] Chernozhukov, V., and Hansen, C. Econometrics of high-dimensional sparse models (p much larger than n). National Bureau of Economic Research.

[13] Choi, H., and Varian, H. Predicting the present with google trends. *The Economic Record 88*, s1 (2012), 2–9.

[14] Daas, P., Puts, M., Buelens, B., and van den Hurk, P. Big data and official statistics. In *New Techniques and Technologies for Statistics* (2013), Eurostat.

[15] Eurostat. Big data : an opportunity or a threat to official statistics ? In *Conference of European Statisticians* (2014), United Nations Economic Commission for Europe.

[16] Fan, J., and Fan, Y. High Dimensional Classification Using Features Annealed Independence Rules. *Annals of statistics 36*, 6 (2008), 2605–2637.

[17] Fan, J., Guo, S., and Hao, N. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society : Series B (Statistical Methodology) 74*, 1 (2012), 37–65.

[18] Florescu, D., Karlberg, M., Reis, F., del Castillo, P. R., Skaliotis, M., and Wirthmann, A. Will big data transform official statistics ? Tech. rep., Eurostat, 2014.

[19] Hall, P., Pittelkow, Y., and Ghosh, M. Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society Series B 70*, 1 (2008), 159–173.

[20] Hastie, T., Tibshirani, R., and Friedman, J. *The elements of statistical learning : data mining, inference and prediction*, 2 ed. Springer, 2008.

[21] Jianqing Fan, F. H., and Liu, H. Challenges of big data analysis. *National Science Review 1*, 2 (2014), 293–314.

[22] Lazer, D., Kennedy, R., King, G., and Vespignani, A. The parable of google flu : Traps in big data analysis. *Science 343*, 6176 (2014), 1203–1205.

[23] Leonard, I., Varlet, G., and Sillard, P. Data editing and scanner data. In *Conference of European Statisticians* (2014), United Nations Economic Commission for Europe.

[24] Li, R., Lin, D. K., and Li, B. Statistical inference in massive data sets. *Applied Stochastic Models in Business and Industry 29*, 5 (2013), 399–409.

[25] McLaren, N., and Shanbhogue, R. Using internet search data as economic indicators. *Bank of England Quarterly Bulletin 51*, 2 (2011), 134–140.

[26] Mick, C. Is the sky falling ? new technology, changing media, and the future of surveys. *Survey Research Methods 7*, 3 (2013), 145–156.

[27] Sillard, P. Donnees de caisse : vers des indices de prix a la consommation a utilite constante. Tech. rep., Insee, 2013.

[28] Statistical Commission of the Unece. Big data and modernization of statistical systems. In *Conference of European Statisticians* (2014), United Nations Economic Commission for Europe.

[29] Tam, S.-M., and Clarke, F. Big data, official statistics and some initiatives by the australian bureau and statistics. Tech. rep., Australian Bureau of Statistics, 2014.

[30] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58* (1994), 267–288.

[31] Varian, H. R. Big data : New tricks for econometrics. *Journal of Economic Perspectives 28*, 2 (2014), 3–28.

[32] Vitter, J. S. Random sampling with a reservoir. *ACM Trans. Math. Softw. 11*, 1 (Mar. 1985), 37–57.