Le secret statistique

La recherche médicale mobilise souvent plusieurs centres de recrutement afin d'accroître la taille des études cliniques et la portée de leurs conclusions. Le partage de données aux fins d'analyse soulève toutefois des préoccupations quant à la protection des renseignements personnels. La statistique offre des outils permettant ce partage dans le respect de l'anonymat.

Dylan Spicker
Université du Nouveau-Brunswick, St-Jean
Erica EM Moodie
Christian Genest
Université McGill

Nos téléphones intelligents et autres appareils électroniques d'usage courant génèrent un flux constant d'informations sur nous-mêmes et sur nos habitudes de vie. Ils fournissent de précieuses indications quant à nos allées et venues, aux lieux ou sites virtuels que nous fréquentons, aux produits que nous consommons, etc. Il en va de même pour toutes les transactions que nous réalisons au moyen d'une carte de débit, de crédit ou de fidélité, mais aussi chaque fois que nous nous rendons chez un professionnel de la santé.

Ces données constituent une mine de renseignements pour tous ceux qui cherchent à prédire le comportement des consommateurs, anticiper leurs besoins et leur proposer des biens et services qui comblent leurs attentes. Elles sont aussi fort utiles dans un contexte médical, notamment pour réguler la prise de médicaments et en mesurer les effets thérapeutiques. Toutefois, leur secret est vital pour la protection de la vie privée.

La nécessité de protéger les données à caractère personnel s'est imposée par suite de scandales retentissants qui ont soulevé l'indignation du public. Dans une affaire largement médiatisée, on est entre autres parvenu à apparier des données dépersonnalisées de la plate-forme de diffusion en continu Netflix et des profils publics de la base de données cinématographiques IMDb pour identifier des abonnés et tenter d'en déduire leurs opinions politiques. Par ailleurs, Latanya Sweeney, professeure d'informatique à Harvard, a démontré que plus de 85 % des Américains peuvent être identifiés de manière unique à partir de leur date de naissance, de leur sexe et du code postal de leur résidence principale.

Dès lors, comment s'étonner que l'on puisse déduire l'identité de quelqu'un à partir d'un fichier de données individuelles, même si ce dernier est dépourvu de renseignements nominatifs ? Par exemple, qui sait si l'exploitation d'une base de données généalogiques en ligne ne saurait permettre d'inférer des faits ou comportements d'ordre privé concernant des personnes ? Divers organismes publics et privés ont donc légiféré et instauré un train de mesures coercitives visant à protéger la vie privée des utilisateurs d'applications mobiles et autres « générateurs de données » en tous genres.

Le secret statistique

De prime abord, le concept de secret statistique paraît intuitif, mais il est difficile d'en donner une définition mathématique rigoureuse. Pour y parvenir, il est utile de se demander ce qui pourrait constituer une violation de la vie privée.

Supposons que l'on s'intéresse au profil socio-économique d'une ville de banlieue et que l'on mène une enquête auprès d'un échantillon de 100 résidents choisis au hasard. Comment faire pour s'assurer que la mise à disposition d'informations est conforme au secret statistique ? Hors de question, bien sûr, de publier les données brutes. Advenant que des noms, adresses ou numéros d'assurance

Latanya Sweeney

Scientifique engagée dans la protection de la vie privée et l'atténuation des risques à l'ère des données massives, passionnée de mathématiques depuis son enfance, Latanya Sweeney a toujours rêvé de concevoir un « ordinateur pensant ». Après avoir entrepris des études en génie électrique et en informatique au Massachusetts Institute of Technology (MIT), elle les a interrompues pour fonder une entreprise de conception de logiciels d'intelligence artificielle. Sa passion pour la recherche et l'innovation l'ont éventuellement poussée à se perfectionner et à entreprendre une carrière universitaire. Elle a obtenu une maîtrise de Harvard en 1997 et a été la première femme afro-américaine à obtenir un doctorat en informatique au MIT en 2001.

Dans une entrevue accordée à la Fondation Ford[1], Latanya Sweeney explique que son intérêt pour la protection de la vie privée a été éveillé par un commentaire fortuit sur les méfaits possibles de l'informatique. En couplant un fichier de données médicales anonyme à une liste électorale municipale qu'elle avait acquise pour la modique somme de 20 $, elle a pu identifier le dossier médical du gouverneur de l'État. « Cette petite expérience m'a permis de montrer que l'on pouvait reconstituer un jeu de données. J'ai aussitôt été invitée à témoigner devant le Congrès. Ce fut un tournant dans ma vie : nous voulions que le monde entier bénéficie des bienfaits de l'informatique, mais personne ne s'était méfié de ses répercussions possibles. »

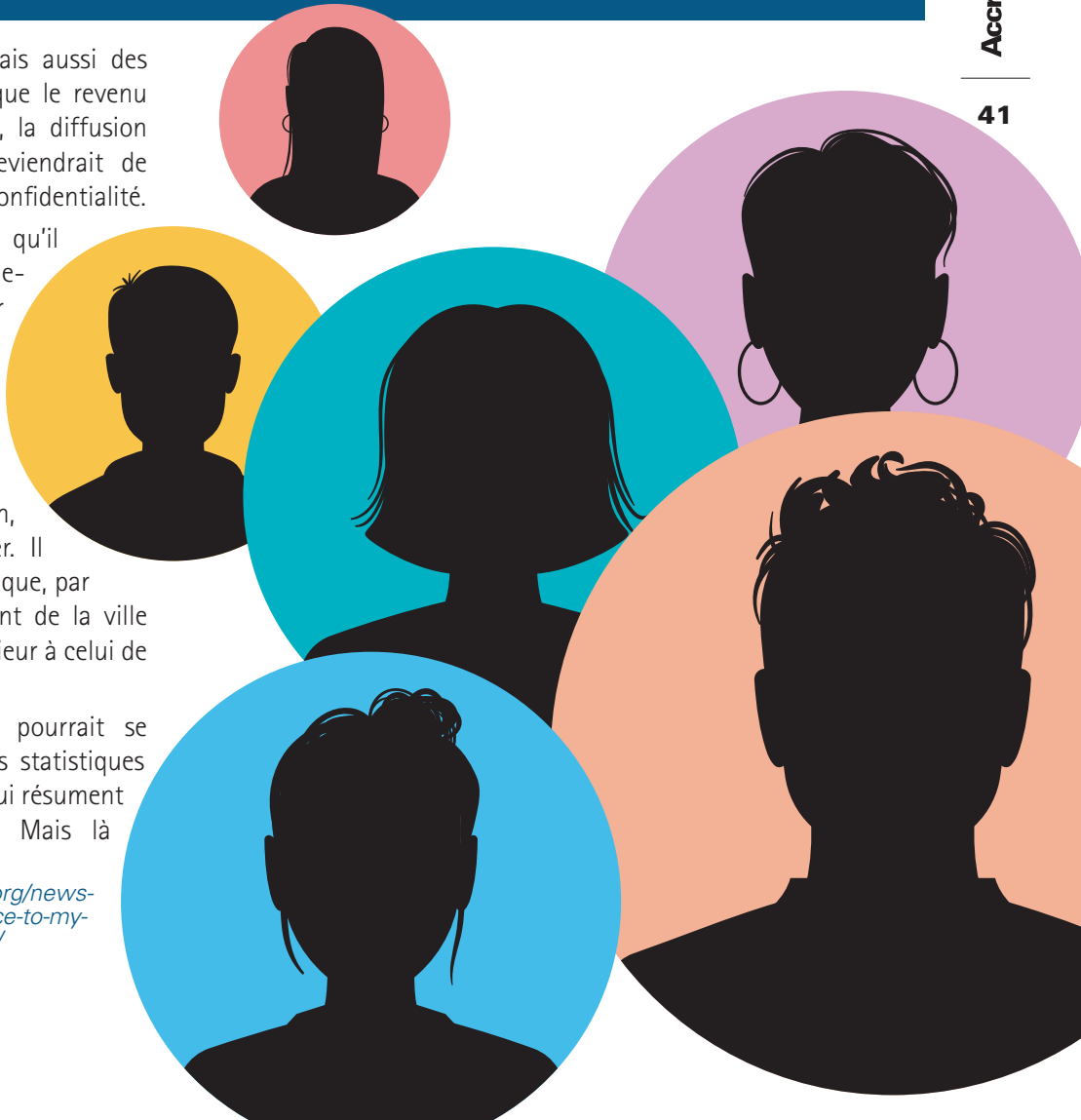
Mme Sweeney a occupé divers postes administratifs et de recherche dans les secteurs public et privé. Elle est titulaire de la Chaire Daniel-Paul sur les pratiques gouvernementales en matière de technologie à la Harvard Kennedy School. Elle a fondé et dirige le « Public Interest Tech Lab », le « Data Privacy Lab » et le « Tech Science Program » à Harvard. Elle est aussi rédactrice en chef de la revue *Technology Science*.

Source de la photo : www.iq.harvard.edu/people/latanya-sweeney

sociale aient été recueillis, mais aussi des informations sensibles telles que le revenu annuel ou l'identité de genre, la diffusion intégrale des données contreviendrait de façon flagrante aux règles de confidentialité.

On pourrait naïvement croire qu'il suffit de masquer les renseignements nominatifs pour régler le problème. Certaines des variables restantes pourraient néanmoins être liées d'assez près à un seul individu pour que, mises en rapport avec d'autres bribes d'information, elles permettent de l'identifier. Il pourrait être de notoriété publique, par exemple, qu'un certain habitant de la ville dispose d'un revenu bien supérieur à celui de tous les autres.

Conscient de ce risque, on pourrait se résoudre à ne publier que des statistiques (moyennes, proportions, etc.) qui résument les informations pertinentes. Mais là



1. *https://www.fordfoundation.org/news-and-stories/stories/posts/advice-to-my-younger-self-latanya-sweeney/*

encore, plus ces valeurs agrégées sont nombreuses et variées, plus on ouvre la porte à des recoupements qui permettent d'identifier des répondants, voire de reconstituer une partie des données… Pour cette raison, on procède généralement au « contrôle de la divulgation statistique » (CDS) en supprimant volontairement certaines observations ou valeurs particulièrement instructives.

En 2019, le Bureau du recensement des États-Unis a révélé qu'en dépit de ses règles de CDS, le recours à des informations publiques et à un système d'équations géant lui avait permis de reconstituer un jeu de données nominatif ayant un haut degré de précision pour plus de 70 % de la population ! Il a donc révisé ses normes avant le recensement de 2020. Depuis lors, Statistique Canada a tenté de reproduire cette étude dans le cadre du recensement canadien, mais les résultats obtenus à ce jour suggèrent que les techniques de CDS employées au Canada offrent une meilleure protection que celles en vigueur aux États-Unis.

Outre la reconstruction de données, il importe de se prémunir contre les tentatives de repérage visant à déterminer si des individus spécifiques font partie d'un jeu de données. Le problème ne date pas d'hier, mais il a pris une toute nouvelle dimension lorsqu'en 2008, une équipe de recherche dirigée par Nils Homer a pu identifier le matériel génétique d'un individu dans le cadre d'études génomiques à grande échelle. Dans un tout autre contexte, Joseph Calandrino et son équipe ont pu déduire les habitudes de consommation de clients d'Amazon en associant les avis marchands de la compagnie à ceux de son système de recommandation. Des exploits similaires ont aussi été réalisés sur les sites web *Last.fm et LibraryThing.*

La confidentialité différentielle

Pour mesurer l'efficacité d'une modalité de partage de l'information au regard du secret statistique, on peut recourir au concept mathématique de *confidentialité différentielle*, que nous allons maintenant présenter.

Comme mise en contexte, considérons un jeu de données x et un objet $\mathcal{M}(x)$ que l'on souhaite rendre public. En général, x est une matrice dont les lignes représentent des individus et dont les colonnes représentent des variables. Quant à $\mathcal{M}(x)$, ce pourrait être une moyenne ou un intervalle de confiance autour d'une estimation, mais il pourrait aussi s'agir d'un tableau d'estimations ou d'un jeu de données dans son ensemble.

Dans le cadre de l'enquête portant sur la ville de banlieue, par exemple, x pourrait être le vecteur $(x_1, \ldots x_{100})$ des revenus annuels des 100 répondants choisis au hasard et

$$\mathcal{M}(x) = \frac{1}{100}(x_1 + \cdots + x_{100}) \qquad (1)$$

l'estimation du revenu annuel moyen. Supposons maintenant que dans cette ville, 80 % des habitants gagnent 50 000 $ par an, 20 % des gens n'ont aucun revenu et qu'un haut salarié gagne 1 000 000 $ par an. Comme $\mathcal{M}(x)$ dépend de chacune des composantes de x, il est difficile d'en énumérer toutes les valeurs possibles, mais on peut avoir une idée de leur répartition en générant un grand nombre d'échantillons sur ordinateur et en traçant un histogramme des valeurs obtenues. Le résultat se trouve à la figure 1 ci-dessous.

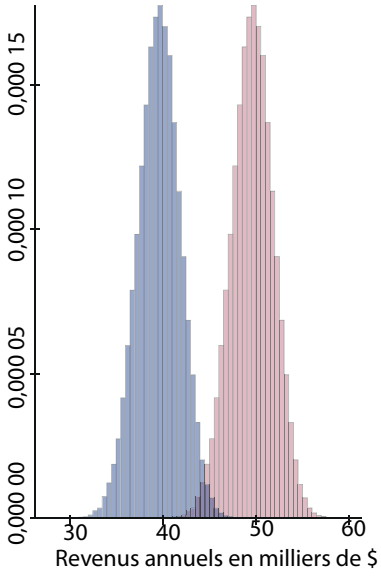


Figure 1

Histogramme des valeurs de la variable $\mathcal{M}(x)$ calculée par la formule (1), selon que le haut salarié en fait partie (histogramme rouge) ou non (bleu).

Comme on peut le constater, la valeur de $\mathcal{M}(x)$ varie autour de 50 000 $ quand le haut salarié est inclus dans l'échantillon (histogramme en rouge), alors que $\mathcal{M}(x)$ tourne autour de 40 000 $ si ce haut salarié est remplacé par un non-salarié (histogramme en bleu). Comme les deux histogrammes se chevauchent à peine, le fait de publier la valeur de $\mathcal{M}(x)$ pourrait donc permettre à une personne bien informée de savoir si le haut salarié faisait partie de l'échantillon et, éventuellement, de déterminer son revenu annuel.

Soit x' le jeu de données dans lequel le haut salarié est remplacé par une personne sans revenu. On dit que les jeux de données x et x' sont voisins car ils ne diffèrent que par un seul individu. Pour que l'objet $\mathcal{M}(x)$ réponde à l'exigence de confidentialité différentielle, on exige alors que les valeurs de $\mathcal{M}(x)$ et $\mathcal{M}(x')$ soient « relativement proches » l'une de l'autre lorsque x et x' sont voisins. À défaut, on court le risque de violer le secret statistique.

Par « relativement proche », on entend que pour tous jeux de données voisins x et x', et pour toute valeur y que puissent prendre $\mathcal{M}(x)$ et $\mathcal{M}(x')$, on ait

$$\frac{P\{\mathcal{M}(x) = y\}}{P\{\mathcal{M}(x') = y\}} \leq e^{\varepsilon}$$

pour un niveau de confidentialité prédéterminé e^{ε}. Plus la valeur de ε est proche de zéro, meilleure est la protection. Typiquement, on prend ε plus petit ou égal à 1, de sorte que les probabilités d'avoir $\mathcal{M}(x)$ et $\mathcal{M}(x')$ soient proches l'une de l'autre.

Pour vérifier si la formule (1) fournit un bon degré de confidentialité différentielle, on doit calculer le ratio $P\{\mathcal{M}(x) = y\}/P\{\mathcal{M}(x') = y\}$ pour toutes les valeurs possibles de y dans le cas où le haut salarié fait partie de l'échantillon x mais est remplacé par un non-salarié dans l'échantillon x'. Dans la figure 1, il est clair que ce ratio peut être très grand ; il suffit de le calculer quand y = 55 000 $, par exemple. Par conséquent, le fait de publier la valeur de $\mathcal{M}(x)$ calculée en (1) ne répond pas au critère de confidentialité différentielle pour $\varepsilon \leq 1$.

Une manière simple d'induire la confidentialité différentielle consisterait à prendre

$$\mathcal{M}(x) = \frac{1}{100}(x_1 + \cdots + x_{100}) + \tau, \qquad (2)$$

où τ est une variable aléatoire d'espérance nulle dont la variance est fonction du niveau de confidentialité souhaité. Ceci induit un bruitage qui atténue l'impact de chacun des répondants sur le résultat, offrant ainsi un démenti plausible à leur inclusion dans les données publiées. La figure 2 ci-dessous illustre l'effet de ce bruitage : on y voit clairement que l'ajout d'une variable τ (ici, une variable de Laplace centrée) fait en sorte que les deux histogrammes se chevauchent alors considérablement, semant le doute quant à l'inclusion du haut salarié dans l'échantillon. De fait, le critère de confidentialité différentielle est vérifié pour ε = 1 quand l'écart-type de la loi de τ est égal au revenu maximal divisé par $\sqrt{2}$ fois n, où n = 100 est la taille de l'échantillon.

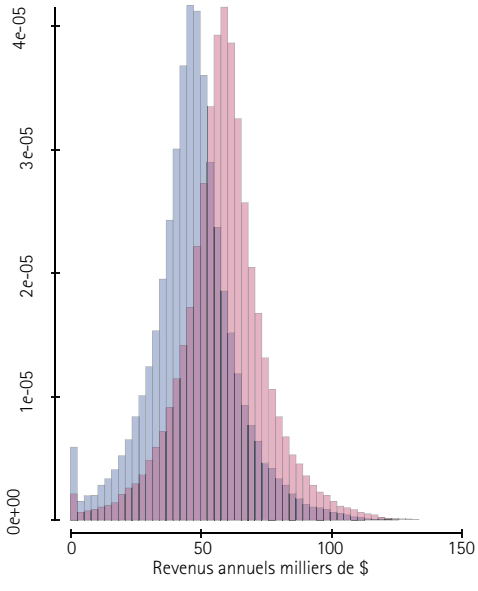


Figure 2

Histogramme des valeurs de la variable $\mathcal{M}(x)$ calculée par la formule (2), selon que le haut salarié en fait partie (histogramme rouge) ou non (bleu).

Le secret statistique en recherche

Malgré son importance, le secret statistique s'avère parfois trop contraignant. C'est notamment le cas en recherche médicale où, dans le respect de l'anonymat des participants aux essais cliniques, les chercheurs ont besoin d'informations bien plus fines et détaillées que celles qui figurent dans des fichiers « grand public ». Cependant, l'accès aux données brutes dénominalisées comporte des risques de divulgation d'informations sensibles.

Au Canada, par exemple, les soins de santé et les données qui en sont issues sont de juridiction provinciale. Dans le cadre d'études cliniques pancanadiennes, les administrateurs locaux sont souvent les seuls à pouvoir y accéder. Les exigences en matière de secret statistique étant généralement axées sur la protection des informations relatives aux patients en amont de l'analyse, le partage de données s'en voit limité.

Quelle que soit la norme de protection de la vie privée que l'on souhaite adopter, un défi majeur consiste à trouver un équilibre entre l'accès aux données et leur confidentialité.

Statistique Canada et la loi

En vertu de la Loi sur la statistique, la participation à certaines enquêtes de Statistique Canada, dont celle portant sur le recensement de la population, est obligatoire. En cas de refus ou de fausse déclaration, le contrevenant était jadis passible d'une peine d'emprisonnement ; il s'expose maintenant à une amende pouvant aller jusqu'à 500 dollars.

La loi sur la statistique impose également des responsabilités à Statistique Canada : toutes les informations qui lui sont divulguées doivent rester confidentielles et ne peuvent être utilisées qu'à des fins statistiques. Tous les employés de Statistique Canada sont tenus de prêter un serment d'office en vertu duquel ils s'engagent à protéger les renseignements reçus. Ils s'exposent à des sanctions sévères en cas de violation de la confidentialité des données.

Des spécialistes en mathématiques, en statistique et en informatique continuent de chercher des solutions. Voici un petit tour d'horizon des techniques dont ils disposent actuellement.

Contamination : Comme on l'a déjà vu, une approche courante et simple de protection de la vie privée consiste à contaminer les données. Plutôt que de fournir les variables telles que mesurées, on les bruite en en perturbant les valeurs par des variables aléatoires soigneusement sélectionnées. Pour des variables telles que le sexe ou l'origine ethnique, il s'agit de remplacer la bonne catégorie par une autre pour une petite proportion des répondants ; l'idée est semblable à l'ajout d'un terme d'erreur τ, mais elle est appliquée au niveau individuel, de façon à modifier les renseignements nominatifs en tout ou en partie.

Agrégation de données, mise en commun et méta-analyse : Dans un contexte multisite, comme celui des soins de santé au Canada, on peut choisir d'agréger les données en classes plutôt que de fournir des données individuelles. Par exemple, plutôt que de communiquer l'âge de chaque répondant, on peut former des groupes (de 5 à 30 personnes, par exemple) et ne communiquer que les moyennes (l'âge moyen) par groupe.[2]

La forme la plus extrême d'agrégation de données est sans doute la méta-analyse. Dans le contexte d'une étude clinique pancanadienne, par exemple, il est possible

2. *Pour un exemple de mise en commun d'échantillons biologiques plutôt que statistiques, voir l'article de Genest et Rousseau dans* Accromath, vol. 15, no 2.

que chaque centre de recrutement ne consente à fournir qu'une estimation de l'effet considéré, telle qu'une moyenne ou une proportion, couplée à une mesure d'incertitude, tel qu'un intervalle de confiance. Une méta-analyse agrège ces données selon un principe qui tient compte de la taille de chaque échantillon, de sorte que ceux qui sont plus grands – et dont les estimations sont donc plus précises – jouent un rôle prépondérant dans l'estimation combinée.

Ces analyses sont souvent résumées par un « graphique en forêt », dont un exemple est fourni ci-dessous à la figure 3. Dans ce graphique, chaque estimation spécifique à un site est représentée par un rectangle dont la superficie est proportionnelle à la taille d'échantillon du site. La barre horizontale traversant chaque rectangle représente l'intervalle de confiance correspondant (à 95 %). L'estimation agrégée est représentée par le losange au bas du graphique. Les estimations ponctuelles et les bornes des intervalles de confiance sont également fournies.

Effets pervers

Malgré l'intérêt des techniques statistiques permettant d'éviter les fuites de données et de protéger le secret statistique, il est important de souligner qu'elles ont leur prix. En contaminant les variables pour garantir l'anonymat, on court le risque d'induire des distorsions dans les relations estimées. Dans certains cas, l'ajout d'une erreur, même de moyenne nulle, peut affaiblir la relation existant entre deux variables, ce qui peut mener à une sous-estimation de l'impact potentiel qu'elles peuvent avoir l'une sur l'autre.

Dans le cas plus complexe où l'analyse porte sur des variables dichotomiques ou sur plusieurs variables à la fois, l'ajout d'un terme d'erreur peut parfois transformer un effet bénéfique en risque apparent. Il pourrait mener, par exemple, à la conclusion (erronée) que le port du casque accroît les risques de blessure lors d'une chute en vélo !

L'implantation du secret statistique comporte donc de grands défis, qui restent à explorer : trouver un équilibre entre la protection de la vie privée et la nécessité d'éviter de graves distorsions dans l'analyse ou l'interprétation des données.

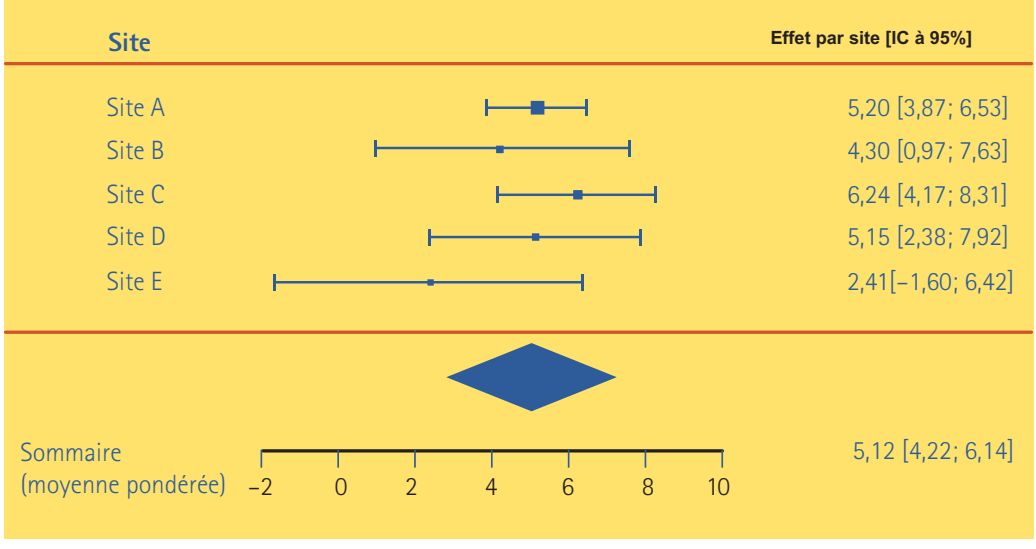


Figure 3. *Un exemple de graphique en forêt.*