Les erreurs associées aux données : dysfonctionnement, révélateur de sens et matière première

Pascal Rivière, Insee, chef de l'inspection générale

Nous sommes désormais dans un monde de *data*. Elles s'immiscent partout dans nos vies quotidiennes, constituent un carburant pour les processus opérationnels et de décision, et apparaissent fréquemment comme un véritable actif des entreprises. Le fait qu'elles puissent être erronées engendre de nombreuses difficultés : impacts en termes d'image, décisions inappropriées, procès intentés, etc. Analyser la qualité des données, c'est notamment évaluer le coût de la non-qualité, donc des erreurs. Elles sont donc perçues en premier lieu comme un phénomène indésirable, un désagrément, un dysfonctionnement.

Mais qu'est-ce qu'une donnée en erreur ? La comparer à une supposée « donnée vraie » poserait un problème à la fois philosophique et opérationnel. Il s'agit plutôt de l'écart à une norme, à « ce qui devrait être », et donc de manière générale du non-respect d'un certain nombre de règles, explicites ou non, et non exhaustives. Utiliser la notion d'erreur présuppose donc l'existence de conventions. Elles sont de toutes natures : certaines sont simples et déterministes (règles de forme sur les dates), d'autres bien plus complexes (vraisemblance d'évolutions brutales dans le temps).

Chercher à réduire les erreurs, à améliorer la qualité des données, oblige à s'interroger sur le processus qui les a fait naître : les données ne sont pas données, elles sont construites. Elles proviennent d'un enchaînement complexe et largement méconnu d'opérations de capture, saisie, vérification, modification, normalisation, de traitements automatiques divers et de boucles de rétroaction. On réalise alors que, loin de l'image d'Epinal de l'erreur de saisie, les erreurs ne sont pas nécessairement dues au fait que «quelqu'un s'est trompé » . En comprendre l'origine permet de faire émerger les normes et conventions employées, de découvrir que celles-ci se transportent mal d'un univers à un autre, et de constater que le non-respect d'une norme signifie parfois que c'est la norme, et non la donnée, qui est à revoir. On passe ainsi de l'erreur comme désagrément à l'erreur en tant que révélateur de sens.

Enfin, l'erreur peut être utilisée de manière positive dans la démarche scientifique. Il s'agit notamment de l'écart à un *modèle* mathématique (décrivant un fonctionnement idéalisé et simplifié) : le modèle joue alors le rôle de convention, de norme. L'erreur est alors formalisée, on lui associe un comportement, on la dote d'une représentation probabiliste. Elle devient une matière première pour une démarche essais-erreurs, pour déterminer la pertinence d'un modèle économétrique, pour calculer la précision d'une statistique fondée sur un sondage. L'erreur est ici domestiquée : prévue, normale, elle n'est plus indésirable, et se mue au contraire en allié indispensable.