Informations limitées, dimensions et contradiction :
application à la gestion des connaissances industrielles par l'emploi de méthodes statistiques.

Michel Lutz[1], Rodolphe Le Riche[2,1], Xavier Boucher[1]
[1] *Ecole Nationale Supérieure des Mines de Saint-Etienne*
[2] *CNRS UMR LIMOS*

Résumé

Le développement des approches statistiques d'aide à la décision à partir de l'instrumentation de systèmes artificiels ou naturels met en évidence un processus de création de connaissances qui, après avoir recueilli des données, procède à une réduction des caractéristiques autonomes de ces données pour les transformer en informations puis en connaissances. Nous montrons dans cet article comment cette réduction de dimensions, qui est nécessaire à l'interprétation humaine, peut induire des connaissances contradictoires. Le lien entre information limitée et contradiction est illustré par un exemple provenant d'une usine de composants micro-électroniques fortement instrumentée. Notre analyse suggère que la contradiction peut être résolue par des cycles de construction de connaissances alternants acceptations des informations contradictoires et réductions dimensionnelles.

1. Création de connaissances par la constitution de structures d'aide à la décision

Dans le cadre de nos recherches appliquées dans une usine de fabrication de puces électroniques (au sein de l'entreprise STMicroelectronics), nous cherchons à aider les gestionnaires d'un système informatique industriel à prendre des décisions. En tant que statisticiens, nous utilisons des outils quantitatifs qui doivent permettre de constituer des « connaissances » utiles à la prise de décision. Pour cela, nous exploitons un important volume de données, stockées au sein de bases de données (plusieurs centaines de séries chronologiques, représentant plusieurs dizaines de téraoctets de données). Ces données permettent de caractériser l'activité du système informatique industriel.

Afin de formaliser la création de connaissances industrielles à partir de données, nous nous appuyons sur un présupposé théorique, qui distingue les notions de données, d'informations et de connaissances (Tuomi, 1999 ; Gordon et Paugam-Moisy, 1997 ; Alavi et Leidner, 2001 ; Tsoukas et Vladimirou, 2001). Chacune de ces notions correspond à une des étapes de la création des connaissances appliquées, qui nous intéressent dans cet article :

- Les « données » sont les chiffres et faits bruts. Elles constituent un codage du réel perçu à travers un espace de mesure construit. Chez STMicroelectronics, les données correspondent aux mesures telles que directement disponibles dans les bases de données. Ce sont des signaux recueillis d'après les choix d'experts et dont l'objet est de mesurer l'activité du système informatique industriel. Le recueil de données est la première étape de tout travail statistique (Desrosières, 2010) ;

- Les « données » deviennent des « informations », lorsqu'elles sont traitées et mises en relation. On considère en effet que les variables brutes ne prennent sens que l'une par rapport aux autres, pour répondre à un problème : les signaux doivent être organisés au sein d'un système (Ermine, 1989). Chez STMicroelectronics, l'étape de constitution d'information correspond à un traitement des variables quantitatives issues de bases de données. On utilise pour cela des méthodes statistiques. Comme l'indique Desrosières (2010), elles permettent « *D'une part,* [de] *fixer des objets, en convenant de leurs équivalences par des définitions standardisées* […] *D'autre part,* [de] *décrire les relations entre les objets ainsi construits, et éprouver la consistance de ces liens.* » ;

- Les « informations » deviennent des « connaissances », lorsqu'elles sont interprétées par des individus, et mises en contexte et/ou en théorie. En effet, comme l'indique Meusnier (*in* Barbin et Lamarche,

2004), l'étape de passage à la connaissance implique d'évaluer si « *le modèle adopté simule convenablement la réalité et si la réalité va se comporter comme le modèle* ». En pratique, ce travail aussi est fait avec les experts de l'entreprise.

Nous considérons les deux phases de recueil de données et de constitution d'information comme une étape plus générale de structuration du réel. Nous appelons structure d'aide à la décision un ensemble {données, informations, connaissances} particulier. Ces structures, permettant de prendre des décisions ultérieures, sont donc constituées par l'utilisation de méthodes statistiques et l'implication d'experts, comme le résume la figure 1.

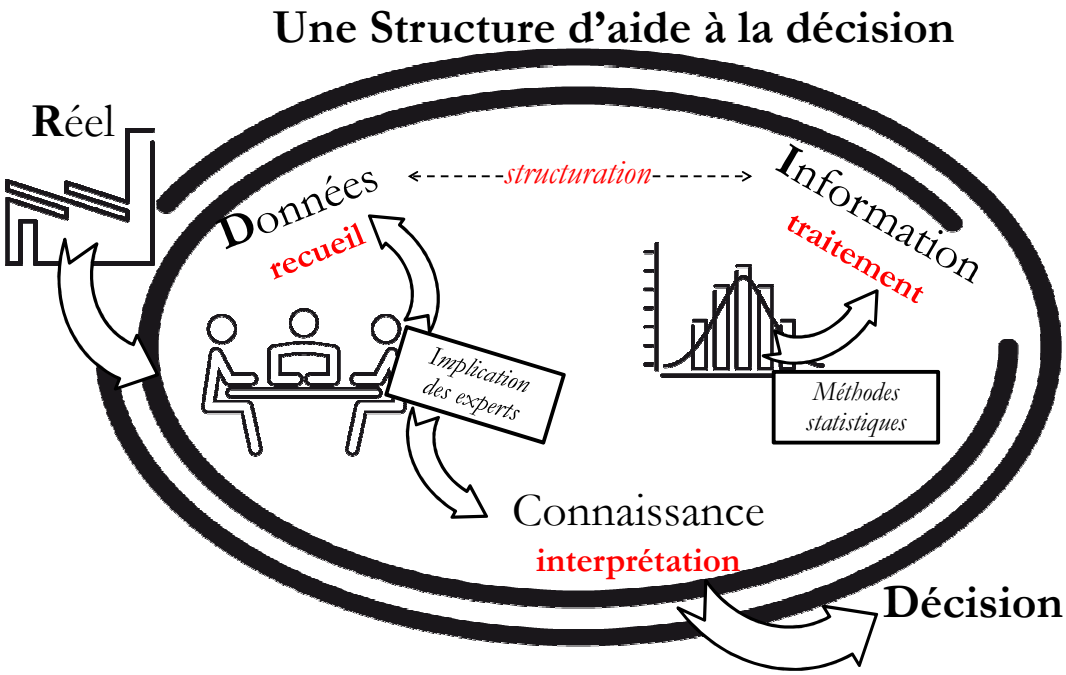


Figure 1. La constitution d'une structure d'aide à la décision chez STMicroelectronics.

Nous suggérons que la création de telles structures engendre une perte de dimensions de la réalité observée.

2. Structuration et perte de dimensions

Tout objet et *a fortiori* tout système réel peut potentiellement être observé selon un nombre infini de caractéristiques. Par exemple, une chaise peut être décrites à travers ses dimensions, positions spatiales, matériaux (des échelles microscopiques aux échelles macroscopiques), âge, style, histoire, conditions de fabrication, valeurs économiques, valeurs sentimentales, …, auxquelles il faut ajouter les caractéristiques pas encore connues de la science contemporaine. Nous appelons ce nombre de caractéristiques la « dimension » de l'objet, notée R. Selon nous, les phases de recueil de données et de constitution d'informations impliquent nécessairement une réduction dimensionnelle :

- Lors du recueil des données, l'accès à la réalité du système se fait par le choix d'un nombre limité de données d'observation D ($D < R$) ;

- La constitution d'informations vise à réduire davantage la dimension des données, pour pouvoir décrire le système suivant I dimensions ($I < D$). Dans le cadre d'applications statistiques, de nombreuses méthodes existent pour cela : voir par exemple Fodor (2002), pour un panorama des techniques disponibles.

Cette réduction dimensionnelle est nécessaire car :

- Les capacités d'observation (détection et mesure d'un phénomène réel), de stockage (taille des bases de données) et de traitements statistiques (complexité algorithmique et puissance de calcul disponible) sont limitées ;

- Le volume de données nécessaires à l'apprentissage artificiel et à la décision statistique croit très vite avec le nombre de dimensions traitées. On parle parfois de la « malédiction de la dimension » : voir par exemple Donoho (2000), qui explique notamment la lenteur de convergence d'une estimation statistique en haute dimension ;

- Il est cognitivement difficile pour un être humain d'appréhender des espaces en dimensions supérieures à trois. Une telle capacité nécessite des apprentissages spécifiques et ne relève pas de l'intuition commune.

La première proposition que nous développons dans cet article est la suivante : les opérations de réductions dimensionnelles d'un système donné peuvent engendrer des connaissances contradictoires du réel.

3. Perte de dimensions et contradiction

Par connaissance contradictoire, nous entendons une connaissance du réel étant à la fois *K* et *non-K*. Pour un système réel et des mesures données, des structurations différentes peuvent conduire à des connaissances différentes et contradictoires. Nous nous intéressons en particulier à la phase de constitution d'informations (passage de *D* à *I*). Remarquons que le passage des informations aux connaissances pourrait également engendrer ses propres contradictions, mais cette thématique ne sera pas abordée ici.

Géométriquement, on montre aisément comment la perte de dimension peut créer toutes sortes de contradictions. Considérons la figure 2, qui représente deux espaces (x_1,x_2) et (*P*,*s*) d'observation originels en deux dimensions, dans lesquels sont observés trois individus A, B et C.

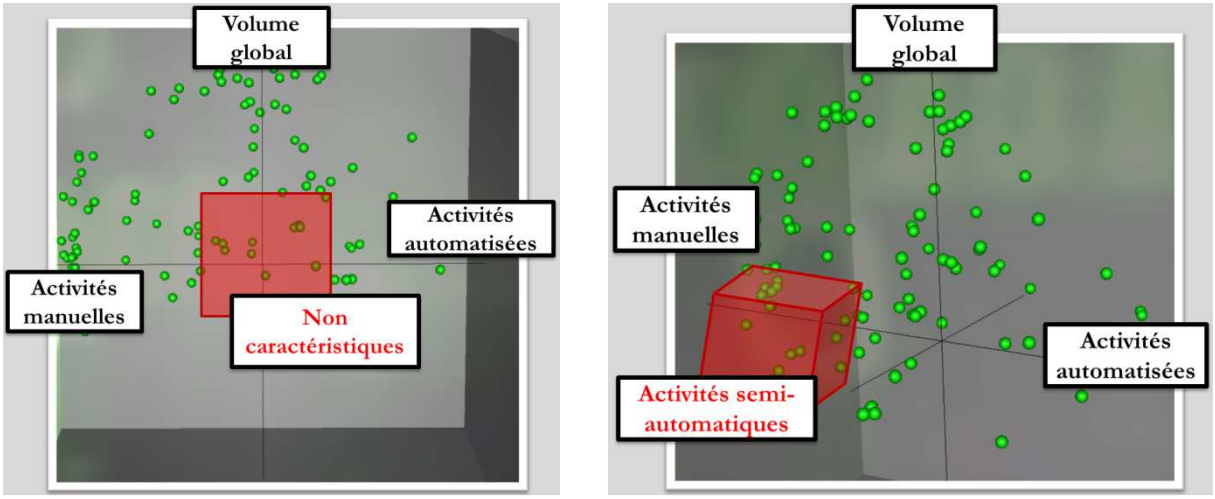
Figure 2. (x_1,x_2) et (P,s) : deux espaces d'observation de A, B et C.

Observons les distances euclidiennes entre les trois individus. Dans (x_1,x_2), A est plus loin de B que de C. Pourtant, si l'on observe le même système selon la seule dimension x_1, A sera plus proche de B que de C, la réduction de dimension créant dans ce cas un faux rapprochement ; au contraire, A paraîtra plus loin de B que de C si l'on observe les individus selon la seule dimension x_2 : des connaissances contradictoires peuvent ainsi être produites du fait du passage de deux à une dimension. *Idem* pour le système non-linéaire *(P,s)*, où apparaît cette fois un faux éloignement : A est plus proche de B que de C, mais A est plus loin de B que de C suivant *s* alors que A est plus proche de B que de C suivant *P*.

4. Illustration par un cas industriel

Nous proposons ici un cas réel issu des travaux conduits chez STMicroelectronics. On cherche à comprendre l'activité transactionnelle d'une application du système d'information industriel d'une usine (*R*). Les données récoltées correspondent aux volumes de transactions appelées quotidiennement (*D*). On observe simultanément 450 variables environ : d'un point de vue cognitif, ces données ne peuvent pas être directement exploitées. Il faut être capable de réduire la dimension de cet espace d'observation. Une méthode statistique exploratoire multidimensionnelle est utilisée pour cela : l'Analyse en Composantes Principales (ACP). Cette méthode vise à construire des axes factoriels (ou composantes principales), permettant de réduire la complexité des données originales, en les projetant dans un espace de dimension réduite (Fodor, 2002).

Pour le même jeu de données, nous réalisons deux ACP différentes, interprétées avec deux groupes d'experts différents. Nous montrons comment ces deux ACP, qui correspondent donc à deux structurations différentes, peuvent aboutir à formuler deux connaissances contradictoires concernant l'activité industrielle sous-jacente à l'activité applicative observée (on considère que les composantes principales caractérisent différents aspects de l'activité de production). La figure 3 fournit une visualisation de l'espace des variables, selon les deux ACP.



ACP Groupe 1
(espace à 2 dimensions)

ACP Groupe 2
(espace à 3 dimensions)

Figure 3. Applications de deux ACP différentes au même jeu de données (espaces des variables).

Le groupe 1, qui perçoit l'activité applicative selon une ACP à deux dimensions, conclut à trois grands groupes de transactions représentant les principaux aspects de l'activité industrielle. D'une part, un groupe qui dépend du volume d'activité globale (*V*) de l'usine. D'autre part, deux groupes en opposition : les activités manuelles (*M*) *versus* les activités automatiques (*A*), qui correspondent à deux modes de production différents (dans le cas du mode manuel, les opérateurs réalisent eux-mêmes un certain nombre d'opérations, tandis que ces opérations sont effectuées directement par le système informatique dans le cas du mode automatique). Les autres transactions, non caractérisées par les axes factoriels considérés, ne peuvent être interprétées et ne correspondent donc pas à une réalité exprimable. D'où la connaissance suivante : activité industrielle = {*M, V, A*}.

En observant l'activité selon un espace des variables en trois dimensions, le groupe 2 aboutit à d'autres conclusions. Il constate en effet un troisième mode de production, correspondant à des activités *SA* de production semi-automatique (l'opérateur réalise lui-même des opérations, tout en étant assisté par le système informatique). On formalise alors la connaissance suivante : activité industrielle = {*M, V, A, SA*}.

En conséquence, à partir de données similaires, deux connaissances contradictoires concernant la réalité de l'activité du système sont constituées. Par ailleurs, d'autres analyses, en dimensions supérieures à trois, aboutissent encore à d'autres conceptions de la réalité du système observé. Par exemple, l'ajout d'une quatrième dimension permet de discerner les activités liées aux tests qualités, une cinquième caractérise le traitement des activités de R&D, etc.

5. Dimensions et évolution de la connaissance

Il nous semble qu'il existe une voie à suivre quand de telles contradictions sont constatées. Il s'agit de procéder à une concaténation des structures divergentes et de recommencer l'analyse dans le cadre d'une structure plus générale. Dans l'exemple de l'usine de composants électroniques, on constate aisément que l'ACP à deux dimensions n'est qu'un sous-ensemble de l'ACP à trois dimensions. Les causes de la contradiction sont donc faciles à comprendre : le groupe 1 observe le système avec moins de finesse que le groupe 2. Dans l'exemple des projections de la figure 1, la concaténation des dimensions reconstitue l'espace de départ et résout la contradiction.

Constituer une telle structure concaténée correspond cette fois à une montée en dimension. Ce genre de procédé est bien connu des mathématiciens, qui passent parfois volontairement dans des espaces à plus hautes dimensions pour y structurer des données de manière plus simple. Un exemple est donné par les machines à support vectoriel (Vapnik, 1995) qui sont utilisées pour automatiquement séparer deux familles de points. Les machines à support vectoriel utilisent l'astuce du noyau (« *kernel trick* », Aizerman et *al*., 1964) pour passer dans un espace à haute dimension et y réaliser la classification. En effet, comme illustré sur la figure 4, si une frontière non linéaire est nécessaire pour séparer deux ensembles de points, passer dans un espace à plus haute dimension permet de les séparer par une frontière linéaire, ce qui simplifie les calculs. Mais la montée en dimension a également l'inconvénient de rendre les points séparables par un très grand nombre de frontière, si bien que l'on ne peut pas vraiment en tirer une connaissance : pour inférer statistiquement de la connaissance en haute dimension, il faut soit beaucoup plus de points qu'en basse dimension (la malédiction de la dimension), soit compenser le manque de points par l'ajout d'une connaissance supplémentaire. Dans les machines à support vectoriel, la connaissance injectée pour rendre unique la frontière linéaire est de choisir la frontière qui maximise la marge entre les points.

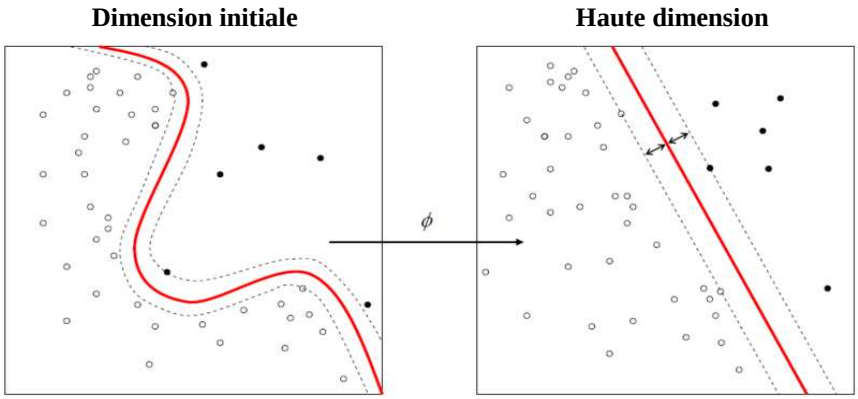


Figure 4. Illustration mathématique du passage en haute dimension. Les machines à support vectoriel se servent de l'astuce du noyau pour passer en haute dimension et mieux séparer deux familles de points. Une frontière simple (linéaire) en haute dimension est équivalente à une frontière complexe (non linéaire) en basse dimension.

Joindre les structures de connaissances contradictoires, *i.e.* accepter les contradictions, est un pas dans la direction de la pensée complexe (Morin, 1990). Cette acceptation dépendra des qualités cognitives des personnes concernées pour deux raisons. D'une part, des résistances culturelles sont susceptibles d'apparaître : le groupe 1 admettra-t-il d'accepter la réalité proposée par le groupe 2 ? D'autre part, il est plus difficile de décider en présence de plus d'informations : comment le groupe 1 va-t-il intégrer l'activité semi-automatique dans son processus décisionnel ? Comment le groupe 2 va-t-il décider alors qu'il sait que le groupe 1 ne tient pas compte de l'activité semi-automatique ? Dans ce dernier cas, on peut alors se demander si cette augmentation dimensionnelle n'est pas un retour au point de départ : ne recréons nous pas une complexité qui obscurcit à nouveau la compréhension du réel ?

Nous proposons alors d'inscrire ces interactions entre données, dimensions et connaissances, dans le cadre d'un cycle de la connaissance. Par structuration, des connaissances nouvelles sont constituées. Ces connaissances peuvent faire émerger des contradictions. Ces contradictions peuvent ensuite être levées, par des concaténations

et des analyses quantitatives supplémentaires, etc. Comme l'indique Tuomi (1999), on peut aller des données vers la connaissance, mais également de la connaissance vers les données. Cette idée est illustrée par la figure 5.

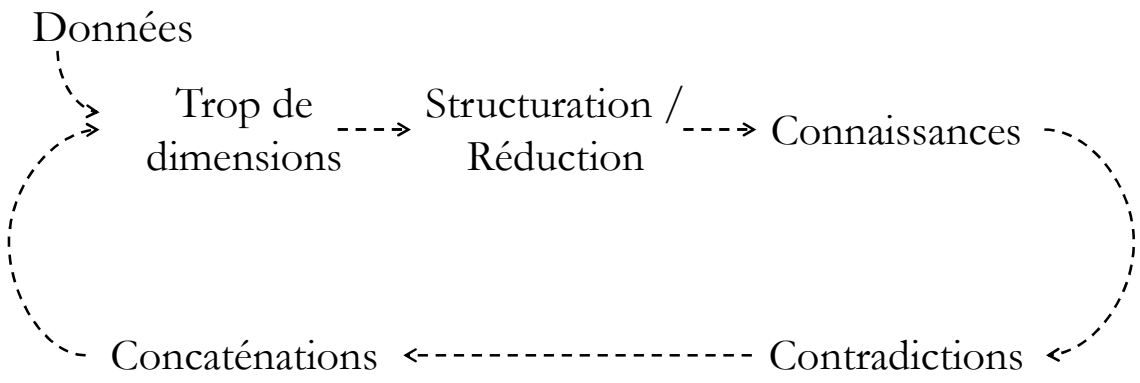


Figure 5. Proposition : un cycle de la connaissance ?

Ainsi, la connaissance est perçue dans le cadre d'un processus dynamique, récurent et en dépassement permanent.

6. Liens avec l'épistémologie constructiviste

Cette dernière section propose une esquisse de réflexion, concernant des liens entre analyse de données et épistémologie constructiviste. Au cœur de cette épistémologie, qui propose une alternative à l'épistémologie positiviste « classique » basée sur Descartes, Compte, etc. (Le Moigne, 2003, 2007), se trouve la notion centrale de l'observateur et de la non-séparabilité du monde réel et de l'expérimentateur. Dans ce cadre, ce n'est pas le monde qui implique l'expérience, mais l'expérience qui implique le monde (Watzlawick, 1988 ; Segal, 1990). En d'autres termes, notre connaissance du réel est nécessairement conditionnée par notre capacité à percevoir le monde et à traiter ces perceptions, afin d'être capable de les appréhender.

Avec la méthode de construction de connaissances, la perception du monde correspond à la définition de l'espace de mesure, son traitement aux méthodes statistiques employées. Elle peut complètement s'inscrire dans le cadre d'une épistémologie constructiviste. Effectivement, Desrosières (2010) propose un positionnement épistémologique de la statistique qui semble lui faire écho. Selon lui, lors de la construction d'un modèle statistique, l'objectif est d'apporter des éléments de réponse à un besoin de décision. Peu importe que le modèle ne vise pas à fournir une description complète et objective de la réalité. L'essentiel est de constituer un espace cognitif partageable, pour décrire et décider, correspondant à une prise de position sur la contingence. Cette conception téléologique et constructive est propre à l'appréhension des systèmes complexes (Le Moigne, 1999) et, en ce sens, nous pouvons penser que la statistique est un outil crédible pour la modélisation de tels systèmes. D'autres auteurs confirme cette proposition : Shalizi *in* Deisboeck et Kresh, (2006), Laflamme (2008).

Concluons cette section avec un exemple de construction de connaissances scientifiques issu du CERN[1]. Si l'on se réfère à la thèse de Nicquevert (2012), nous apprenons que le protocole ayant permis au CERN de découvrir expérimentalement le boson de Higgs dépendait de trois étapes : disposer d'une énergie de collision suffisante pour créer le boson (générer le phénomène réel), disposer de détecteurs suffisamment performants pour capter l'activité produite par cette collision (percevoir le phénomène réel), récolter et traiter les données issues des détecteurs (traiter les perceptions). Comme l'indique le communiqué du CERN du 4 juillet 2012 (*in* Nicquevert, 2012) : « *Nous observons dans nos données des indices clairs d'une nouvelle particule, au niveau de 5 sigmas, dans la gamme de masses autour de 126GeV* ». La découverte expérimentale du boson de Higgs a donc nécessité de recueillir des données, qu'il a fallu être capable de traiter. La connaissance scientifique expérimentale produite résulte bien d'un processus de type données/information/connaissance, tel que présenté dans la figure 1, et a permis de valider une théorie scientifique préliminaire.

[1] Remarque : nous n'affirmons pas que les scientifiques du CERN sont constructivistes.

Dans le cas du CERN, l'interprétation du traitement quantitatif résulte d'un présupposé théorique précédant le protocole expérimental. Nous observons actuellement le développement d'une tendance visant à renverser cette approche, en inférant directement des connaissances à partir de données collectées à des échelles de plus en plus massives, sans recours à des théories préalables : on se référera par exemple au phénomène « *Big Data* », actuellement très en vogue (Anderson, 2008). Le cadre constructiviste nous invite à être conscients de la subjectivité de la connaissance scientifique et à la considérer avec prudence et modestie : on redoublera donc de précautions lorsqu'on entendra parler de production scientifique sans fondements théoriques *a priori*, surtout si l'interprétation des analyses de données n'est pas réalisée par des experts du domaine considéré.

Conclusion

Dans cet article, nous nous sommes interrogés sur les conséquences de la perte de dimensions dans le cadre de la construction de connaissances basée sur l'analyse de données. Ce questionnement est apparu et a été illustré par une étude menée dans une usine instrumentée. Notre analyse dépasse certainement ce contexte industriel, pour devenir plus fondamentalement un modèle de création de connaissance scientifique, ouvrant un certain nombre de questionnements épistémologiques (modélisation des systèmes complexes, paradigme constructiviste, etc.).

Ainsi, nous pensons que la proposition de cycle de la connaissance (figure 5) trouve toute sa place dans l'évolution de la connaissance scientifique expérimentale. Les analyses quantitatives deviennent des éléments de construction d'une connaissance raisonnée, dialectique et dynamique. Chaque progrès engendrera des incompréhensions et des contradictions nouvelles. De manière optimiste, on peut penser que ces problèmes devaient alors être temporaires : en affinant leurs capacités à mesurer le réel, puis à traiter et analyser ces mesures, les scientifiques pourront construire un système explicatif plus général, qui soulèvera de nouvelles difficultés, selon un cycle sans fin.

Références

M. Alavi, D.E. Leidner, Knowledge management and knowledge management systems: conceptual foundations and research issues, MIS quarterly, 25 (1) (2001) 107-136.

C. Anderson, The end of theory: the data deluge makes the scientific method obsolete, WIRED magazine (2008). www.wired.com/science/discoveries/magazine/16-07/pb_theory (dernière consultation le 18 mars 2013).

M. Aizerman, E. Braverman, and L. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning, Automation and Remote Control, 25 (1964) 821-837.

E. Barbin, J.P. Lamarche (Coord.), Histoire de probabilités et de statistiques, Ellipses, Paris, 2004.

T.S. Deisboeck, J.Y. Kresh, Complex systems science in biomedicine, Springer, New-York, 2006.

A. Desrosières, La politique des grands nombres – histoire de la raison statistique, La Découverte, Paris, 2010.

D.L. Donoho, Aide-Mémoire. High-dimensional data analysis: the curses and blessings of dimensionnality, Departement of Statistics, Stanford University, 2000.

J.L. Ermine, Systèmes experts – théorie et pratiques, Lavoisier, Cachan, 1989.

I.K. Fodor, A survey of dimension reduction techniques, U.S. Department of Energy – Lawrence Livermore National Laboratory, 2002.

M.B. Gordon, H. Paugam-Moisy (Dir.), Sciences cognitives – diversités des approches, Hermès, Paris, 1997.

S. Laflamme, Analyse statistique linéaire et interprétation systémique, Nouvelles perspectives en sciences sociales : revue internationale de systémique complexe et d'études relationnelles 4 (1) (2008) 141-159.

J.L. Le Moigne, Les épistémologies constructivistes, PUF, Paris, 2007.

J.L. Le Moigne, Le constructivisme – modéliser pour comprendre, L'Harmattan, Paris, 2003.

J.L. Le Moigne, La modélisation des systèmes complexes, Dunod, Paris, 1999.

E. Morin, Introduction à la pensée complexe, Le Seuil (1990).

B. Nicquevert, Manager l'interface – approche par la complexité du processus collaboratif de conception, d'intégration et de réalisation : modèle transactionnel de l'acteur d'interface et dynamique des espaces d'échanges, Thèse pour obtenir le grade de Docteur de l'Université de Grenoble dans la spécialité Génie Industriel, 2012.

L. Segal, Le rêve de la réalité, Seuil, Paris, 1990.

H. Tsoukas, E. Vladimirou, What is organizational knowledge?, Journal of Management Studies, 38 (7) (2001) 973-993.

I. Tuomi, Data is more than knowledge – implications of the reversed knowledge hierarchy for knowledge management and organizational memory, Journal of Management Information Systems, 16 (3) (1999) 107-121.

V. Vapnik, The nature of statistical learning theory, Springer-Verlag, New-York, 1995.

P. Watzlawick (Dir.), L'invention de la réalité – contributions au constructivisme, Seuil, Paris, 1988.