WinSitu, un nouveau paradigme pour l'analyse exploratoire de données basée sur des projections

Michaël Aupetit*

*CEA, LIST, Laboratoire Information Modèles et Apprentissage,
F-91191 Gif-sur-Yvette, France.
michael.aupetit@cea.fr

Résumé. Dans cet article, nous discutons des limites pratiques de l'analyse exploratoire de données basée sur les techniques de projection non linéaires continues. Nous montrons que ces méthodes de projection sont inutilisables en l'état pour permettre une inférence quelconque sur les données originelles. Nous présentons une méthode de visualisation *in situ* et montrons au travers de différentes expériences, qu'elle est indispensable à leur interprétation. Ce processus implémente le paradigme *WinSitu* d'analyse exploratoire visuelle basée sur des projections que nous introduisons pour la première fois dans ce travail. Ce changement de paradigme permet de rendre aux méthodes de projection toute leur utilité.

1 Introduction

L'objectif essentiel de l'analyse exploratoire visuelle de données est d'inférer de leur représentation graphique des propriétés sur leur structure originelle. Ces données sont fournies sous forme d'un tableau individus-variables $N \times D$ ou d'une matrice de dissimilarités inter-individus $N \times N$. Deux types d'informations sont recherchées :

- Des classes de variables similaires, ou des variables atypiques, mises en évidence par un ensemble d'individus.
- Des classes d'individus similaires, ou des individus atypiques, mis en évidence par un ensemble de variables.

La similarité entre variables peut être mesurée par exemple par la corrélation de Fisher ou de Pearson. Celle entre individus peut être mesurée par exemple par la distance euclidienne, ou par la distance d'édition ou fournie directement sous forme matricielle. Lorsque les individus sont représentés sous forme d'un nuage de points dans un repère cartésien orthonormé à deux dimensions, la mesure de corrélation entre les variables codées par les axes de ce repère peut être estimée visuellement, de même que la présence d'individus atypiques ou de classes d'individus similaires apparait aussi immédiatement. Les difficultés surviennent lorsque l'on cherche à analyser des données multi-dimensionnelles.

L'analyse exploratoire de données multi-dimensionnelles passe soit par l'estimation automatique de paramètres d'un modèle fournissant un résumé de telle ou telle caractéristique recherchée dans les données (adéquation à une loi de densité de probabilité, degré d'appartenance à des classes de structure prédéfinie, classification des variables ou des individus...)

(Bishop 2008), ou bien par leur projection "explicite" dans un plan (Jollife 1986; Kaski et al. 1998b) afin d'en permettre l'analyse visuelle. Parmi les méthodes de projections, celles continues[1] et non linéaires[2] ont fait l'objet de nombreux travaux (voir par exemple (Demartines et al. 1997) (Kruskal 1964) (Sammon 1969) (Torgerson 1952)). A chaque individu est associé un point image dans l'espace plan de projection. La méthode de projection tente alors en générale par la minimisation d'une fonction d'énergie (ou de stress), de positionner ces points images de telle sorte que leurs distances euclidiennes relatives reproduisent au mieux celles mesurées entre les individus d'origine auxquels ils correspondent. Dans la plupart de ces méthodes, les distances ou dissimilarités d'origine entre les individus sont calculées à partir de leurs coordonnées dans l'espace d'origine à D variables, ou fournies directement sous forme matricielle. Les méthodes diffèrent essentiellement par l'information qu'elles tentent de préserver : les petites distances dans l'espace de projection plutôt que les grandes par exemple pour l'Analyse en Composantes Curvilignes (ACC) (Demartines et al. 1997).

Dans cet article, nous mettons en perspectives nos travaux antérieurs dont nous dégageons les principes d'un nouveau paradigme de visualisation pour l'analyse exploratoire. Nous apportons des éléments démontrant les fondements rationnels de ce paradigme.

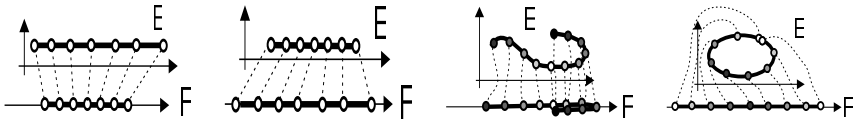
2 Les artefacts

L'objectif principal des projections non-linéaires pour l'analyse visuelle est de fournir un résumé à 2 dimensions d'un ensemble de variables d'origine, préservant au mieux les classes d'individus pouvant exister dans l'espace défini par cet ensemble. Si ce résumé est fiable alors les classes détectées visuellement dans ce résumé devraient aussi exister dans cet espace d'origine.

Malheureusement, toutes les méthodes de projection induisent généralement une perte d'information (ici nous synthétisons l'analyse que nous avions faite dans (Aupetit 2007)). Cette perte d'information se manifeste par des distorsions entre les distances d'origine et les distances après projection. Si l'on considère que les individus sont situés au voisinage d'une certaine structure (un espace topologique) dont elles sont un échantillon bruité, alors on peut interpréter ces distorsions comme le résultat de transformations appliquées à la structure elle-même par la projection. Idéalement, une projection préserve au moins l'information topologique, ce qui n'est le cas que si elle définit un homéomorphisme entre la structure d'origine et son image par la projection. Par exemple, lors de la projection d'une courbe sur une droite, on cherchera à préserver l'ordre des points de la courbe après leur projection sur la droite (topologie) et éventuellement les distances entre les points (géométrie). Les distorsions qui apparaissent lors de la projection de structures plus complexes, sont imagées sur la figure 1 et sont de deux types :

[1] Est continue une projection qui associe à chaque individu un point image distinct dans l'espace de projection. Les cartes auto-organisées de Kohonen (Kohonen 1988) sont des méthodes de projection discrètes car elles incluent une phase de quantification vectorielle.

[2] Est linéaire une méthode de projection qui définit un espace de projection dont les vecteurs unités des axes sont des combinaisons linéaires de ceux des axes de l'espace d'origine. L'Analyse en Composante Principale (Jollife 1986), le Grand Tour (Asimov 1985) ou le "Projection Pursuit" (Friedman et al. 1974) sont des exemples de méthodes continues et linéaires.



(a) Compression (b) Etirement (c) Recollement (d) Déchirure

FIG. 1 – **Distorsions topologiques et géométriques :** *Lors d'une projection d'un espace d'origine E vers un espace de projection F, nous distinguons quatre type de distorsions. Les compressions (a) et étirements (b) correspondent à des distorsions géométriques où les distances relatives entre individus proches dans l'espace d'origine ont changé mais pas suffisamment pour modifier la topologie au voisinage de chaque point. Recollements (faux voisinages) (c) et déchirures (d) correspondent à des distorsions topologiques où les distorsions géométriques au voisinage de certains points d'origine sont telles, que la topologie de ce voisinage est modifiée. En (c), deux parties initialement non connexes, se recouvrent après projection. En (d), une discontinuité apparaît au voisinage des points blancs initialement connectés. Les distorsions topologiques peuvent être rendues visibles : si la distance dans E d'un point quelconque par rapport à un point de référence (blanc) est visualisée dans F par un niveau de gris du clair (proche) au foncé (éloigné), alors deux points voisins dans F avec un fort contraste clair-foncé sont le signe d'un recollement (c) ; et deux points de F avec une couleur claire mais séparés par des points plus foncés, sont le signe d'une déchirure (d). C'est sur ce principe qu'est basée la mesure de "proximité", que nous présentons (section 5).*

- Distorsions géométriques : les distances peuvent être compressées ou étirées par la projection (une demi-sphère en caoutchouc pourra être mise à plat en étirant le voisinage de son bord et en compressant le voisinage de son pôle).
- Distorsions topologiques : les compressions et étirements peuvent être tels qu'un recollement (faux-voisinage) ou une déchirure surviennent modifiant localement la topologie de la structure sous-jacente (les deux hémisphères d'une sphère en caoutchouc se recolleront l'une sur l'autre lors de l'applatissement, alors qu'un ruban de Moëbius devra être nécessairement déchiré si l'on souhaite l'applatir sans recollement).

Les conséquences de ces distorsions sur les structures de classes qui pourraient exister dans l'espace d'origine sont évidentes. Une compression ou un recollement peuvent aggréger artificiellement des structures originellement disjointes. A l'inverse, étirements et déchirures peuvent morceler des structures originellement connexes.

Ces distorsions sont des artefacts de la projection. Elles ont deux causes possibles :

- une cause intrinsèque (artefact structurel) : la structure sous-jacente aux individus ne peut se projeter dans l'espace de projection sans déformation (e.g. une sphère sur un plan) ;
- une cause extrinsèque (artefact technique) : malgré l'absence d'artefact structurel, le paysage de recherche induit par la fonction d'énergie de la projection et sa méthode d'optimisation, soit possède des optima locaux dans l'un desquels reste piégé le processus d'optimisation créant des distorsions qui n'existeraient pas à l'optimum global, soit est tel que même à l'optimum global, la méthode de projection (inadaptée) ne peut éviter les distorsions. C'est la méthode de projection qui est en cause et non la structure à projeter.

Ces deux causes sont souvent présentes simultanément. Distinguer la cause des artefacts est intéressant dans la mesure où un artefact technique peut théoriquement être supprimé pour améliorer la qualité de la projection, alors qu'il est vain de tenter de supprimer un artefact structurel. En pratique cependant, la distinction est délicate tant que l'on ne connaît pas la nature de la structure sous-jacente aux individus, structure que les méthodes d'analyse exploratoire ont justement pour objectif de permettre à l'analyste de découvrir et de caractériser. On peut aussi s'appuyer sur notre connaissance des propriétés que chaque méthode a tendance à préserver mais cela reste, nous le verrons, insuffisant pour une analyse fiable.

A défaut de pouvoir supprimer ces distorsions, nous devons au moins détecter leur présence et les localiser.

3 Aveugles qui s'ignorent

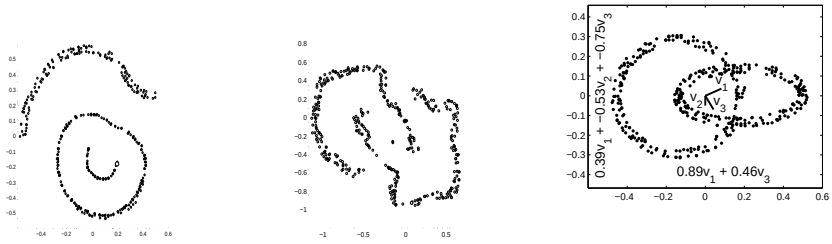
Les méthodes de projection non linéaire ont une particularité supplémentaire : contrairement aux projections linéaires, les axes de l'espace de projection n'ont aucun lien explicite avec les axes ou variables de l'espace d'origine. Ceci a un impact majeur dont l'analyste doit avoir pleinement conscience : aucun lien ne peut plus être établi visuellement à partir de la seule vue du nuage de points projeté, entre les positions des points dans l'espace de projection et les positions ou distances originelles que l'on souhaite préserver, aucune inférence n'est possible de l'espace de projection vers l'espace d'origine, l'interprétation visuelle de la projection est objectivement vouée à l'échec[3]. La projection pouvant superposer des points, l'unique information qui subsiste est triviale : le nombre de points du nuage projeté est inférieur ou égal au nombre d'individus...

Ainsi, a priori, rien ne différencie du point de vue sémantique, le résultat d'une projection non linéaire de celui d'une projection *aléatoire*. Les projections obtenues par l'optimisation d'une fonction d'énergie, produisent bien sûr plus souvent qu'une projection aléatoire, des ensembles de points dont l'oeil s'empresse d'imaginer la forme. Mais rien ne permet d'inférer de cette seule projection que l'une quelconque de ces formes subjectives, existe intègre ou altérée d'une quelconque manière dans l'espace d'origine. En particulier, savoir que la méthode de projection utilisée privilégie la préservation de telle propriété n'est d'aucun secours, tant qu'aucune quantification intelligible de cette préservation n'est fournie à l'analyste en sus de la représentation elle-même. La cause n'en est pas la subjectivité de l'analyste, mais bien l'absence de lien sémantique explicite de la projection avec les variables ou individus d'origine.

Utilisées sans autre précaution, les méthodes de projection non linéaire sont sources de conclusions probablement erronées sur le regroupement des individus en classes ou la présence d'individus atypiques, sans que ce «probablement» soit quantifié. Ces méthodes donnent l'illusion de fournir une information authentique sur les données d'origine, parce qu'elles dessinent des structures qu'il serait improbable d'obtenir par hasard et parce que l'analyste espère voir apparaître des structures indices que les données ont une signification latente, mais elles sont en pratique objectivement inutilisables. L'analyste qui en reste là est aveugle qui croit voir.

La figure 2 fournit des exemples typiques d'illusions. Les données d'origines ne sont pas décrites volontairement afin de placer le lecteur dans la situation de l'analyste explorateur.

[3]Pour autant nous ne nous faisons pas l'avocat des méthodes linéaires qui ont pour défaut majeur de fortement dégrader la topologie des structures originelles lors de leur projection dans le plan. L'engouement pour les méthodes non linéaires nous paraît donc tout à fait légitime.



(a) ACC 1 (b) ACC 2 (c) ACP

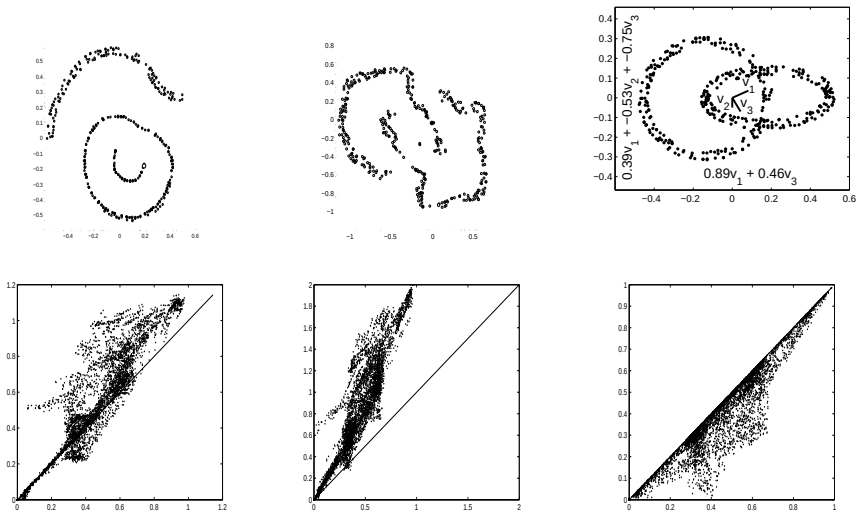
FIG. 2 – **Aveugles qui s'ignorent : les illusions projectives.** *Trois projections différentes d'un même ensemble de données. En (a) et (b) l'Analyse en Composantes Curvilignes (Demartines et al. 1997) piégée dans des optima locaux, peut converger vers des solutions très différentes. En (c) l'Analyse en Composantes Principales (Jollife 1986) (projection sur les deux premiers axes principaux) fournit une solution unique. Quelle inférence peut-on vraiment faire sur les données d'origine à partir de ces trois vues seulement ? Strictement aucune avec l'ACC ! Au contraire, avec l'ACP, les vecteurs unités des axes du plan de projection (les composantes principales) sont des combinaisons linéaires explicites entre les vecteurs unités d'origine v_1, v_2 et v_3, et les vecteurs unités de l'espace d'origine peuvent être projetés dans l'espace de projection, rendant intelligible la vue proposée (même si l'interprétation n'est pas immédiate, elle est possible).*

4 Aveugles se sachant tels

L'analyste soucieux de rigueur utilise donc des outils de mesure de qualité pour diagnostiquer la projection qu'il observe. Ces outils fournissent soit un nombre comme la valeur optimale de la fonction d'énergie utilisée par la projection elle-même, ou comme les mesures de continuité (déchirure) ou de faux voisinage (recollement) (Venna et al. 2001) ; soit un graphique secondaire comme le diagramme de Shepard qui affiche chaque paire d'individus sous forme d'un point dans un plan dont un axe indique les distances relatives dans l'espace de projection, et l'autre les distances d'origine (Les points alignés sur la diagonale principale signifient une projection sans distorsion de ces distances).

Ces deux types d'outils sont pourtant insuffisants, car ils ne sont pas visuellement corrélés à l'espace de projection. Ils ne renseignent pas l'analyste sur la distribution des erreurs *in situ, i.e.* leur répartition dans l'espace de projection. Cette information supplémentaire très sommaire ne permet que de porter à la conscience de l'analyste que ce qu'il voit n'est pas fiable, qu'il fait face au résultat d'une boîte noire qui détruit la sémantique en lui interdisant l'accès direct aux données sources. Généralement, il cherche à améliorer la qualité de la projection en modifiant les méta-paramètres de sa méthode de projection ou en changeant de méthode, ou bien il utilise d'autres vues qu'il espère complémentaires et ensemble exploitables par la technique du lien interactif[4]. Cependant, tant qu'elles sont basées sur ce même principe de boîte noire, aucune de

[4]Technique d'analyse interactive, où la sélection de points dans une vue, met en avant les images de ces points dans les autres vues, permettant d'établir des corrélations entre variables d'origine ou de déterminer l'existence de groupes de points ou de points atypiques.



(a) ACC 1 (b) ACC 2 (c) ACP

FIG. 3 – **Aveugles se sachant tels : preuves de l'inutilisabilité.** *Les mêmes projections que sur la figure 2 mais accompagnées du diagramme de Shepard (*25% *des* $N(N-1)/2$ *distances entre paires de points distincts tirées aléatoirement sont affichées (avec* $N = 400$ *le nombre de points). L'axe des abscisses représente les distances d'origine, l'axe des ordonnées les distances dans l'espace de projection. Dans les trois cas, le diagramme de Shepard ne nous apporte qu'une information partielle : une partie des petites distances sont préservées par l'ACC, mais on ne sait pas lesquelles. Les distances sont compressées par l'ACP, mais assez mal préservées. On apprend donc que les projections sont difficiles à interpréter. Le seul diagramme de Shepard vraiment utilisable en analyse exploratoire par projection, est celui qui montrerait tous les points alignés sur ou très proche de la diagonale, signe d'une projection sans perte d'information.*

ces vues ne permet de séparer l'information authentique d'origine, des artefacts de projection, la vérité de l'illusion, la partie utile de la partie trompeuse qui induit visuellement de fausses caractéristiques. Ainsi de l'accumulation de vues fausses ne peut émerger une information authentique intelligible.

La figure 3 révèle à l'analyste l'inutilisabilité des projections.

L'idée d'utiliser une approche de rééchantillonnage apparaît alors évidente : puisque les méthodes de projection engendrent des artefacts, générons autant de vues que nécessaire par des méthodes ou des paramètres différents choisis aléatoirement indépendamment et identiquement distribués, et la vérité apparaitra dans ce que ces vues auront en commun. Mais contrairement à l'apprentissage supervisé où un score est calculé automatiquement pour chaque échantillon et leur moyenne et ses propriétés de convergence sont largement maitrisées, demander à un analyste humain de dégager subjectivement de ces multiples vues ce qui leur est commun, apparait impratiquable.

En tant qu'analyste s'ignorant, puis se sachant aveugle, nous avons proposé dans (Aupetit 2007) un outil en mesure de redonner la vue, un outil de visualisation des artefacts de projection *in situ*, qui permet la visualisation simultanée ou "co-visualisation" des proximités entre points projetés fournies par leur positions relatives, et points d'origine mettant ainsi en évidence le degré d'authenticité des proximités affichées.

5 Aveugles recouvrant la vue

5.1 La mesure de proximité

Dans (Aupetit 2007), nous avons proposé une mesure de proximité m_i définie pour chaque individu comme la *distance d'origine* X_{is} entre cet individu x_i et un individu de référence x_s, dont le point image y_s est un point d'intérêt du nuage projeté, sélectionné arbitrairement par l'analyste.

5.2 Les cellules de Voronoï

Nous associons à chaque point y_i du nuage dans l'espace de projection, une région de cet espace qui permet de visualiser facilement par sa coloration la mesure de proximité m_i associée à ce point. Ces régions sont les cellules de Voronoï (Okabe et al. 1992) de chaque point du nuage selon la métrique euclidienne. La cellule de Voronoï d'un point du nuage est le lieu des points du plan dont ce point est le plus proche au sens euclidien, parmi tous les point du nuage. Formellement, si les points du nuage sont notés $\underline{y} = (y_1, ..., y_N) \in (\mathbb{R}^2)^N$, la cellule de Voronoï V_i du point y_i est l'ensemble $\{v \in \mathbb{R}^2 | \forall y_j \in S, d(v, y_i) \leq d(v, y_j)\}$ où d est la distance euclidienne.

Les raisons du choix de cette région sont multiples :

- Les cellules de Voronoï pavent le plan donc aucun arrière plan ne perturbe l'analyse visuelle, il n'y a pas de couleur de fond à définir, toutes les couleurs visualisées sont porteuses d'une information authentique sur les données d'origine. Il n'y a pas de recouvrement donc aucune cellule ne masque l'information portée par une autre.
- Si un point quelconque du plan est coloré, il est naturel d'inférer que le point projeté le plus proche porte probablement cette même couleur. Donc la couleur associée à tout point de la cellule de Voronoï d'un point projeté est naturellement supposée celle de ce point projeté, et donc est attribuée à tous les points de cette cellule, la couleur que l'on veut que l'utilisateur associe visuellement à ce point projeté.
- Les cellules sont convexes, contiennent le point projeté associé, et leur forme s'adapte à la densité des points projetés (petites cellules dans les régions denses) ce qui permet de préserver cette information portée par la projection continue (qui la démarque des projections discrètes). En effet, des glyphes de taille fixe suffisament larges pour que leur couleur soit visible, perturberaient l'évaluation visuelle de la densité dans les régions très denses où ces glyphes se chevaucheraient.

Il faut bien noter que l'objectif de la coloration des cellules de Voronoï n'est pas de fournir par interpolation ou extrapolation, une estimation de la mesure de proximité en tout point de l'espace de projection. D'une part car la projection réduisant la dimension, chaque point de l'espace de projection peut être l'image de l'ensemble des points d'une ou plusieurs régions

non connexes de l'espace d'origine, et donc cette mesure de proximité n'a pas de signification évidente pour d'autres points que les points-individus représentés explicitement. D'autre part parce que ce serait introduire une information artificielle supplémentaire qui surchargerait inutilement la mémoire de travail de l'analyste et engendrerait probablement de fausses interprétations auxquelles il est déjà suffisamment enclin. Il s'agit au contraire simplement de faciliter la perception visuelle de la mesure m_i qui n'a de sens qu'au point-individu y_i en étalant au mieux la couleur l'encodant, au voisinage de ce point, pour la rendre apparente tout en préservant visuellement son lien sémantique avec l'individu qu'elle caractérise. Ce type de représentation est utilisé avec le même objectif dans (Balzer et al. 2005).

5.3 Le cas fil rouge

La figure 4 fournit les mêmes exemples que précédemment avec la mesure de proximité. On constate que la co-visualisation de cette mesure et des points projetés permet de retrouver en grande partie la structure originelle des données : deux anneaux entrelacés visibles sur la figure 7g.

5.4 La complexité

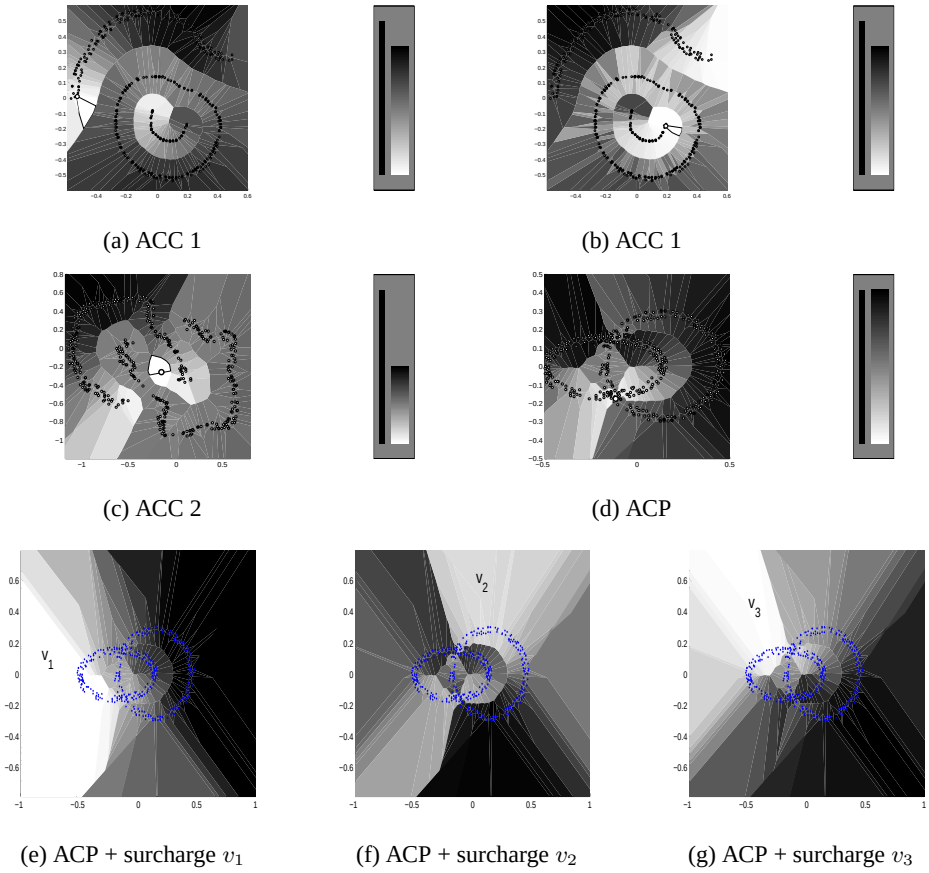
En termes de complexité, ces cellules sont calculables en temps $O(N \log(N))$ dans le plan (Barber et al. 1996) et leur coloration est obtenue par simple lecture et standardisation des distances déjà calculées pour réaliser la projection. Aussi est-il possible de traiter par cette méthode des données de très grande dimension (Jusqu'à 617 dimensions dans nos expériences) car le coût en temps de calcul est négligeable devant celui de la méthode de projection elle-même qui est au mieux en $\Omega(N^2)$ lié au nombre d'éléments de la matrice de distance à estimer par la projection.

6 Les raisons du miracle

La mesure de proximité est en fait la mesure de distance ou de similarité d'origine entre deux points, c'est donc la valeur authentique fournie en entrée, avant la projection et les artefacts qu'elle engendre. Ainsi, on réinjecte dans l'espace de projection une information intègre sur les données d'origine que la projection boîte noire a altérée, ce qui rend interprétable la vue obtenue.

En effet, si les projections linéaires sont interprétables en dépit de la perte d'information qu'elles aussi induisent, c'est uniquement parce que les axes de l'espace de projection sont des combinaisons linéaires entre variables d'origine. Ces combinaisons sont parfaitement intelligibles à l'analyste initié qui peut inférer visuellement l'importance relative de la contribution de ces variables dans la constitution des formes qu'il observe soit pour évaluer les corrélations entre variables, soit pour déterminer les sous-espaces principaux dans lesquels existent ces formes. En résumé, l'information qui subsiste, le lien entre l'espace d'origine et l'espace de projection, est maitrisée et suffisament simple pour être interprétable visuellement, donc exploitable pour l'inférence par l'analyste.

Dans le cas des projections non linéaires, les axes de projection n'ont plus de *sémantique* par rapport aux variables d'origine, variables d'origine et de projection ne sont plus reliées de



(a) ACC 1 (b) ACC 1

(c) ACC 2 (d) ACP

(e) ACP + surcharge v_1 (f) ACP + surcharge v_2 (g) ACP + surcharge v_3

FIG. 4 – **Aveugles recouvrant la vue.** *Les mêmes projections que sur la figure 2 mais en co-visualisant la mesure de proximité. Le point blanc est le point projeté d'intérêt, sélectionné par l'analyste. Les cellules de Voronoï portent un niveau de gris indiquant la distance d'origine entre les point projetés et le point d'intérêt. Plus la cellule est foncée plus la distance est grande. La hauteur de l'échelle de couleur fournie à droite de chaque vue et la hauteur de la barre noire parallèle, sont proportionnelles avec un même facteur, la première à la distance maximale au point d'intérêt dans l'espace d'origine, la seconde à la hauteur de la vue projeté. Sur les vues (a) et (b), la mesure de proximité co-visualisée permet de détecter une déchirure, en montrant que les trois groupes de points en forment en fait seulement deux dans l'espace d'origine (les cellules claires sont en fait connectées). La vue (c) indique par exemple que le point d'intérêt qui semblait isolé, n'est en fait pas un individu atypique, car il est connecté à l'une des structures en U qu'il referme en une structure en O. La vue (d) permet de détecter un recollement de deux structures originellement disjointes ici au niveau du croisement du bas. Les vues (e), (f) et (g) sont les projections ACP surchargées de la coordonnée v_1, v_2 ou v_3 de l'espace d'origine (valeur plus forte, couleur plus claire), ce qui montre que les deux structures en O sont entrelacées (passage "devant" puis "derrière" de l'une par rapport à l'autre très visible pour v_2 et v_3). Les individus d'origine sont issus de deux sructures en anneaux entrelacées et bruitées dans un espace à trois dimensions montré sur la figure 7g.*

manière intelligible. Les distances elles-mêmes représentées par l'intermédiaire du nuage de points, ne sont que des estimations très fortement biaisées des similarités X d'origine comme le montre l'éloignement des points à la diagonale sur les diagrammes de Shepard de la figure 3. En visualisant les similarités d'origine de tous les individus à l'un d'entre eux x_s dont l'image y_s est sélectionnée par l'analyste dans l'espace de projection, on régénère cette connexion qui fait sens, on donne à l'analyste une vue directe des similarités authentiques entre individus.

Ces N similarités $X_{.s}$ correspondent à la ligne s de la matrice X. Si les données sont fournies sous forme d'un tableau individus-variables à N lignes et D colonnes (dont X se déduit par l'application d'une fonction distance), alors il est possible de visualiser de la même manière les N lignes d'une colonne s de ce tableau, donc la coordonnée de chaque individu selon la variable s. La figure 4efg montre le cas de l'ACP surchargée respectivement des coordonnées v_1, v_2 et v_3 de l'espace d'origine utilisé pour l'exemple. Dans le cas présent, cela permet d'inférer que les deux structures en anneaux découvertes en explorant les données par la mesure de proximité, sont en fait entrelacées.

7 Proposition d'un nouveau paradigme

7.1 La fin justifie les moyens

La méthode d'analyse exploratoire par projection décrite ci-dessus s'inscrit dans un nouveau paradigme : la projection n'est plus une fin en soi, car elle n'est pas interprétable ainsi, mais elle est un moyen. En effet, elle permet de construire une mosaïque de pixels colorés par une information authentique, une image, une fenêtre ouverte sur les données d'origine telles qu'elles sont, qui apporte le complément d'information nécessaire à l'interprétabilité de cette projection, et la rend finalement utilisable pour l'analyse exploratoire.

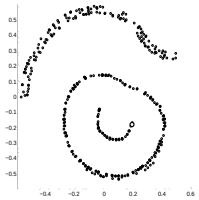
Dans ce paradigme il y a donc trois ingrédients essentiels :

1. (i) des données d'origine représentées par un tableau $N \times D$ individus-variables et une métrique qui permettent de calculer une matrice des distances relatives $N \times N$ entre individus, ou bien représentées seulement par cette matrice de distances ou similarités $N \times N$;
2. (ii) une méthode de projection linéaire ou non, de l'ensemble des données, tendant à minimiser les distorsions ;
3. (iii) une mesure sans perte d'information sur les données d'origine, co-visualisée avec les points projetés, dans l'espace de projection.

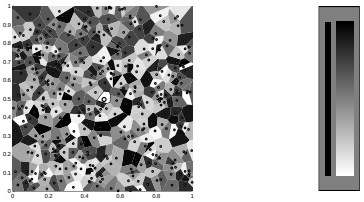
La figure 5 illustre la complémentarité des deux derniers ingrédients.

7.2 Des fondements rationnels

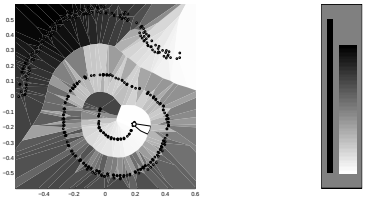
Dans le paradigme usuel, la projection est une fin en soi, la course à la projection idéale est sans fin, il faudra toujours faire un choix sur l'information qui doit être préservée et celle qui sera perdue, chaque méthode pouvant être meilleure que les autres suivant le critère choisi pour les comparer (Lespinats et al. 2009). Dans le nouveau paradigme, il n'est plus primordial d'obtenir une projection parfaite, il suffit qu'elle tende à limiter les déchirures et les faux voisinages et qu'elle soit raisonnablement efficace dans cette tâche, afin de former un écran



(a) ACC 1



(b) aléatoire + proximité



(c) ACC 1 + proximité

FIG. 5 – **La fin justifie les moyens : un nouveau paradigme.** *(a) Une projection non linéaire seule (ici l'ACC des anneaux entrelacés) ne permet objectivement aucune inférence dont la fiabilité soit quantifiable. (b) La mesure de proximité (point d'intérêt sélectionné au centre) co-visualisée avec une projection aléatoire des mêmes données, ne permet pas davantage l'inférence visuelle. (c) La mesure de proximité co-visualisée avec la projection non linéaire de la vue (a) où le point d'intérêt sélectionné est le même qu'en (b). Seule la corrélation visuelle d'une projection non linéaire tendant à minimiser les distorsions (ici les faux voisinages ou recollements) avec une mesure d'information authentique* in situ*, apparaît intelligible à l'analyste, justifiant la définition du nouveau paradigme de visualisation. La projection ne peut être une fin en soit, elle n'est qu'un moyen de rendre intelligible une information authentique non déformée par elle.*

constitué de groupes de pixels adjacents qui soient aussi localement ordonnés (*i.e.* voisins) dans l'espace d'origine. La figure 5b montre qu'un cas extrême, où déchirures et faux voisinages sont à leur paroxysme du fait d'une projection aléatoire, ne permet aucune inférence sur les structures d'origine dont la fiabilité soit quantifiable.

Nous émettons l'hypothèse qu'il vaut mieux en pratique choisir des méthodes de projection privilégiant plutôt les déchirures que les recollements, donc plutôt de larges zones fiables où le voisinage perçu reproduit le voisinage originel, éventuellement séparées par quelques failles où les zones de part et d'autre ne sont pas originellement voisines. En effet, une dissymétrie existe dans la perception des similarités entre individus projetés. Les similarités sont transitives : si A est proche de B et B proche de C alors A est proche de C. Tandis que les dissimilarités ne le sont pas : Si A et B sont éloignés et B et C sont éloignés, on ne peut rien inférer quant à la similarité de A et C. Il est donc très facile grâce à la transitivité de construire visuellement des groupes l'un après l'autre par aggrégation d'individus proches entre eux, alors qu'il est pratiquement impossible de construire visuellement des groupes par exclusion mutuelle d'individus éloignés du fait de l'effort de mémoire et de raisonnement considérable que cela nécessiterait. Il est donc primordial que les zones voisines de l'espace de représentation représentent le plus souvent possible des régions voisines de l'espace d'origine, donc des colorations claires plutôt que sombres par la mesure de proximité, puisque c'est cette proximité visuelle qui sera le support de la construction des groupes. Au contraire, l'alternance trop fréquente de motifs colorés décorrélés de leur proximité spatiale dans l'espace de représentation comme le montre la figure 5b, ne permet pas d'appréhender les structures génératrice de cette coloration alors même qu'elles existent. Les travaux de Le Roux *et al.* (Le Roux et al. 2007) sont éclairant à ce sujet puisqu'ils montrent que les pixels initialement désorganisés d'une rétine artificielle (physiquement agencés en grille, mais sans lien de voisinage explicitement pré-établi au niveau logique) s'organisent automatiquement sous forme d'une carte à deux dimension par l'observation répétées d'images naturelles en appliquant un simple critère de corrélation jugé biologiquement plausible (deux pixels tendent à devenir voisins au niveau logique, si les signaux qu'ils perçoivent sont corrélés, et ces derniers le sont naturellement plus souvent lorsqu'ils sont issus de deux éléments voisins d'un même objet physique). En d'autres termes, nous ne pourrions appréhender la topologie des structures du monde physique par leur projection sur notre rétine que parce que celle-ci s'organise topologiquement pour reproduire au mieux les voisinages physiques perçus par cette corrélation.

Nous donnons le nom de *WinSitu* à ce paradigme pour signifier *Window in situ*[5], ou fenêtre *in situ* car le troisième ingrédient permet l'analyse des données authentiques, telles qu'elles sont *in situ* dans l'espace d'origine, et sa visualisation par surcharge sous forme de coloration des cellules de Voronoï dessine une mosaïque de pixels à l'instar d'une *fenêtre* donnant à voir cet espace.

8 Exemple d'application à des données de grande dimension

Le paradigme *WinSitu* permet l'analyse des connexités intra-classes et inter-classes des données «Isolet» (Merz et al. 1998) de dimension 617 (figure 6). L'expérience est décrite plus

[5]Le terme de visualisation *in situ* qui paraissait le plus approprié a été récemment utilisé (Ma 2009) pour dénommer les méthodes de visualisation opérant *au coeur* des calculs de simulations numériques plutôt qu'en post-traitement pour visualiser les résultats de simulations générant des Péta-Octets de données.

en détail dans (Aupetit 2007). Le graphe des classes construit sur la figure 6be l'est de manière empirique et subjective, par l'accumulation de points de vue différents fournis par les sélections successives d'individus de référence y_s et la visualisation de la mesure de proximité associée. Les régions simultanément claires sont plutôt voisines donc les sommets représentatifs de ces régions dans le graphe des classes tendent à être connectés.

Les données «Oil flow» (Bishop et al. 1998) de dimension 12, sont aussi analysées par cette approche dans (Aupetit et al. 2008) et en complément d'autres méthodes dans (Gaillard et al. 2008).

Nous avons récemment proposé dans le cadre du projet ERITR@C (Euritrack 2006) de détection de traffic de marchandises illicites ou dangereuses, d'utiliser le paradigme *WinSitu* dans un système d'aide à la décision (Allano et al. 2010) pour assister les douaniers dans l'interprétation des mesures effectuées sur un conteneur portuaire. Les contenus d'une centaine de conteneurs de référence, sont caractérisés par les proportions de 15 éléments chimiques mesurés par un système d'inspection neutronique. Chaque conteneur est donc représenté par un point dans un espace d'origine à 15 dimensions. Ces points sont projetés non linéairement sur une carte de telle sorte que les conteneurs au contenu similaire y sont voisins. Le conteneur requête en cours d'analyse n'est pas projeté sur la carte mais sa similarité aux conteneurs de référence est visualisée par la coloration de leurs cellules de Voronoï. Le douanier accède ainsi visuellement à l'ensemble des conteneurs les plus ressemblant au conteneur requête et pour lesquels il connait le contenu en termes de matériaux (et non d'éléments chimiques). Il peut ainsi en complément d'autres éléments dont il dispose, appuyer sa décision d'ouvrir le conteneur pour un contrôle manuel ou bien de le libérer. La carte de référence étant fixe (seules ses couleurs changent en fonction du conteneur analysé), elle peut servir de support au douanier pour se forger au cours du temps une image mentale de son espace de décision, qui l'assiste de mieux en mieux pour accomplir sa tâche.

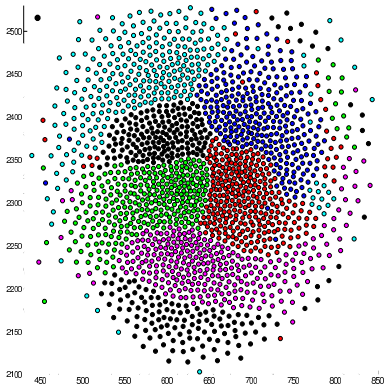
9 Le paradigme WinSitu face à l'état de l'art

9.1 Les méthodes classique de représentation graphique

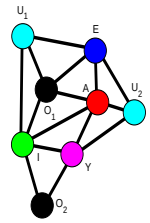
Les méthodes de projection linéaire ne sont pas exclues du paradigme WinSitu, elles en font naturellement partie par construction, car elles présentent déjà une information authentique qui permet l'inférence visuelle. Le dernier ingrédient correspond pour elles à l'explicitation des axes de l'espace de projection en termes de combinaison linéaire des axes d'origine. Il est possible de compléter cette information, en co-visualisant la mesure de proximité ou toute autre mesure authentique par exemple par la coloration des cellules de Voronoï du nuage de points projetés comme nous l'avons fait sur la figure 4defg.

Notons aussi que les méthodes de représentation graphiques largement utilisées comme la représentation par coordonnées parallèles (Inselberg 1985) et la projection linéaire consistant à définir le plan de projection par deux des axes d'origine, sont encore deux autres formes de visualisation implémentant le paradigme *WinSitu* par construction. La figure 7 présente ces méthodes usuelles ainsi que la méthode de surcharge déjà utilisée sur la figure 4.

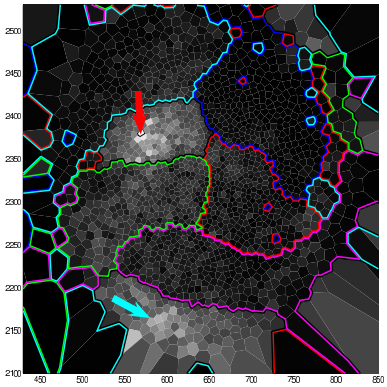
Aussi, la nouveauté de ce paradigme ne tient pas tant à la nouveauté des méthodes qui l'implémentent, qu'à la nouveauté de son expression explicite en trois points. Le dernier point est ainsi posé comme nécessaire à l'intelligibilité de l'information visualisée par la projection, et



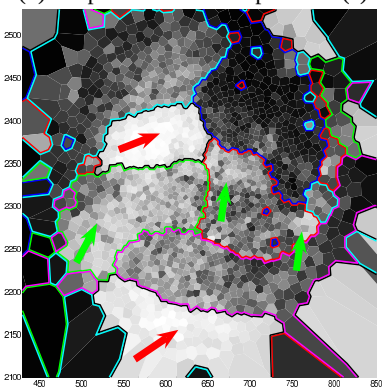
(a) ACC seule



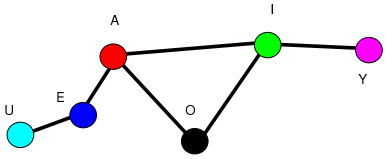
(b) Graphe des classes à partir de (a)



(c) Connexité intra-classe de la classe 'O'



(d) Connexité inter-classe de la classe 'O'



(e) Graphe des classes à partir de l'analyse *in situ*

FIG. 6 – **Exemple d'application aux données Isolet.** *(a) Les voyelles 'A', 'E', 'I', 'O', 'U', 'Y' de la base Isolet à l'origine en* 617 *dimensions, sont projetées par ACC dans le plan. Sans autre information, l'analyste est tenté de conclure en termes de connexité intra-classe, que la classe 'O' (en noir) est en deux composantes connexes séparées l'une de l'autre par la classe 'I' (en vert) et la classe 'Y' (magenta). Il est aussi tenté de conclure en termes de connexité inter-classe, que les différentes classes sont adjacentes dans l'espace d'origine de la manière montrée par la projection. (b) On peut synthétiser par un «graphe des classes» les connexités intra-classe (nombre de sommets de chaque classe, seules les composantes connexes contenant beaucoup de points sont représentées) et inter-classes (liens entre ces sommets) découvertes visuellement par l'analyste. Ici le graphe obtenu par une analyse naïve de la vue (a). Malheureusement ces conclusions sont fausses ! Ce n'est pas un problème de subjectivité de l'analyste mais bien de non authenticité de l'information qui lui est présentée. (c) La mesure de proximité appliquée à l'un des points de classe 'O' (flèche rouge) montre que cette composante est connectée à l'autre (flèche bleu clair). (d) la mesure de proximité moyennée sur tous les points de la classe 'O' (flèches rouges), montre un faux voisinage, la classe 'E' (bleu foncé) n'est pas voisine de la classe 'O'. (e) En poursuivant ce type d'analyse pour les autres classes, on obtient le graphe des classes in situ différent de celui obtenu en (b) à partir de la seule vue (a). Cette analyse visuelle est en soi subjective, mais elle se nourrit d'une information objectivement authentique transmise par la co-visualisation de la mesure de proximité dans l'espace de projection. D'un point de vue sémantique, le graphe des classes (e) indique que le son 'E' (orateurs anglophones) est plus proche du son 'U' que du son 'O', ce qui signifierait par exemple que pour passer du son 'U' au son 'I', un orateur produirait des sons phonétiquement plus proches d'un 'E' puis d'un 'A', que d'un 'O'.*

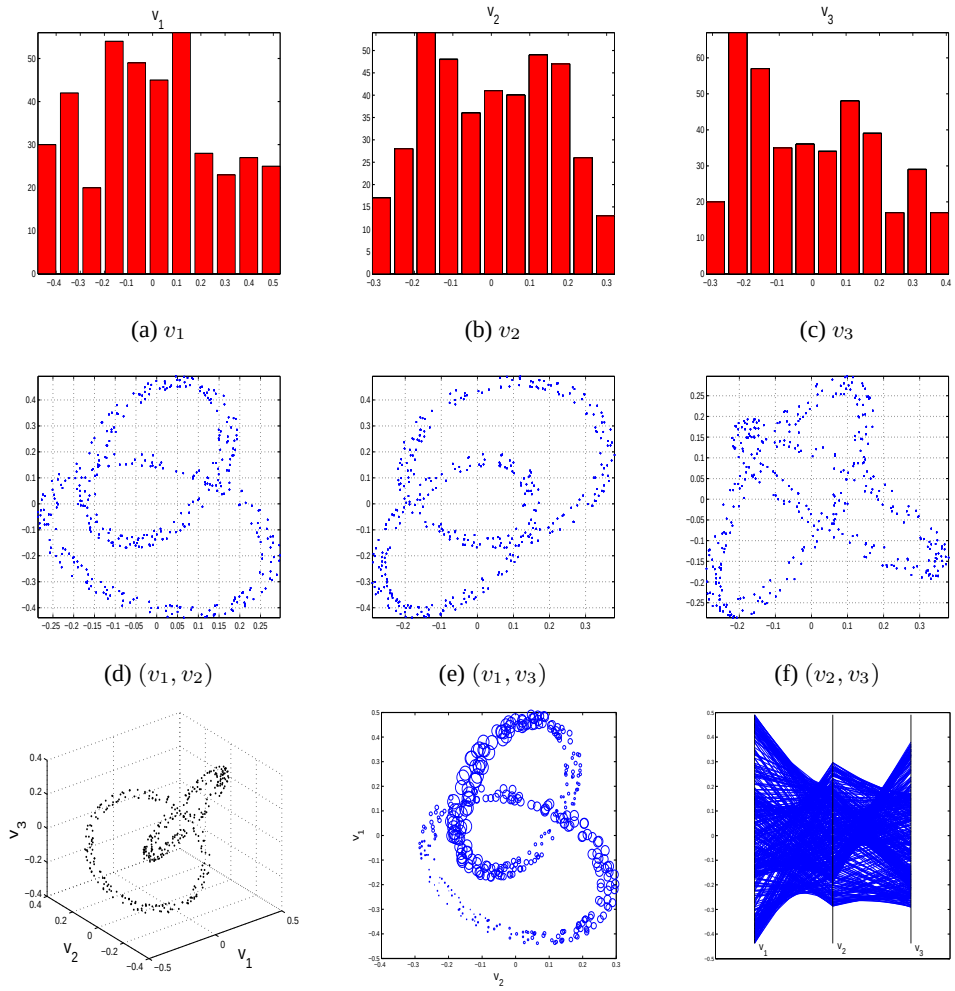
permet d'identifier comme incomplètes, pour ne pas dire ineptes, les projections non linéaires telles qu'elles sont encore trop souvent utilisées.

9.2 Liens avec les cartes auto-organisées

Le paradigme proposé ici n'apparaîtra pas comme original du point de vue des utilisateurs de cartes auto-organisées. En effet, dans ces cartes, les neurones forment une grille régulière qui ne porte donc pas d'information sur leurs positions relatives dans l'espace d'origine. De nombreux outils de visualisation complémentaires ont donc été développés (Kaski et al. 1998a; Rousset et al. 2001; Ultsch 1993; Tasdemir et al. 2009; Vesanto 1999), modifiant la taille, la couleur ou la forme des neurones et des liens pour transmettre une information sur les données d'origine. La mesure de proximité décrite ici n'a pas été proposée sous cette forme pour les cartes de Kohonen, mais des méthodes similaires ont été développées (Kaski et al. 1998a; Rousset et al. 2001; Tasdemir et al. 2009) dans le but identique de rendre objectivement interprétable la projection. Pourtant, nous n'avons pas connaissance de travaux les utilisant alors qu'elles nous paraissent indispensables à l'interprétabilité objective en particulier lorsque l'on utilise ces carte en classification non supervisée. En effet, des groupes de neurones identifiés par leur taille ou leur couleur homogène mais séparés sur la carte comme avec la méthode U-matrix (Ultsch 1993), ne le sont pas nécessairement dans l'espace d'origine s'il y a déchirure, donc une telle carte n'exprime pas visuellement la possibilité que ces groupes n'en forme en fait qu'un unique. Un diagnostic *in situ* devrait donc impérativement être effectué afin de rendre légitime une quelconque inférence sur la topologie des données à partir d'une telle représentation.

Il existe pour les cartes auto-organisées, une méthode de visualisation appelée "component plane" qui consiste à visualiser la distribution des neurones variable par variable (Vesanto 1999), en codant par une couleur la position des neurones selon l'un des axes de l'espace d'origine. En juxtaposant une telle vue pour chaque variable, donc D vues au total, on peut détecter des groupes de points similaires et des corrélations linéaires ou non entre variables. Cette méthode est en quelque sorte l'analogue des coordonnées parallèles, ou chaque axe vertical est remplacé par une carte, elle permet de visualiser une information *in situ* sur les neurones. Elle est utilisée sur la figure 4efg dans le cas d'une projection continue.

Concernant l'analogie des cellules de Voronoï du nuage de points projetés, avec les pixels d'un écran, les cartes auto-organisées forment naturellement un écran dont la structure est régulière au lieu d'être amorphe. Cette régularité ne nous semble pas avantageuse pour deux raisons. D'une part, les deux degrés de liberté que procurent la position des neurones dans le plan, ne sont pas exploités autrement que pour déterminer un voisinage topologique (aspect discret de la carte). De ce point de vue, les projections continues sont aptes à fournir plus d'information, car en plus du voisinage topologique (visuellement inductible par l'adjacence des cellules de Voronoï des points projetés) elles permettent de transmettre une information sur les distances d'origine entre individus. D'autre part, la régularité de la position des neurones est la contre-partie de la quantification vectorielle qu'ils imposent. Cette quantification implique une perte d'information sur les données d'origine : dans une carte auto-organisée, on n'analyse pas les individus, on analyse les neurones représentant ces individus. C'est un ingrédient supplémentaire qui apporte avec lui son lot d'incertitudes quant à l'interprétation.



(a) v_1 (b) v_2 (c) v_3

(d) (v_1, v_2) (e) (v_1, v_3) (f) (v_2, v_3)

(g) (v_1, v_2, v_3) (h) (v_2, v_2) + surcharge v_3 (i) Coordonnées parallèles

FIG. 7 – **Les méthodes usuelles de représentation graphique.** *(abc) Les histogrammes univariés ne permettent pas d'appréhender des structures de données complexes. (def) Les projections sur les paires de variables sont pour le cas présent (g) beaucoup plus informatives. (h) un autre moyen de visualiser des variables supplémentaire est de surcharger la représentation en codant ces variables au niveau de chaque glyphe (ici la taille code v_3) plutôt qu'en termes de position spatiale. Dans ce cas, on détecte aisément l'entrelacement des anneaux (i) Les coordonnées parallèles sont difficiles à interpréter dans le cas présent. Elles permettent cependant de juxtaposer et d'appréhender de l'ordre d'une dizaine de variables simultanément. Un individu y est représenté par une ligne brisée coupant chaque axe de coordonnée à la valeur prise par l'individu pour cette variable. Idéalement les groupes d'individus similaires apparaissent comme des faisceaux de lignes regroupées suivant approximativement un même trajet.*

Globalement, les cartes auto-organisées munies de la visualisation par proximité ou par variable, implémentent le paradigme de la visualisation *in situ*, à la nuance près que les objets de la visualisation ne sont pas les données à analyser mais les neurones qui les quantifient.

10 Conclusion

Les méthodes de visualisation par projection sont attractives, elles semblent faciliter l'analyse exploratoire, mais elles sont trompeuses, elles peuvent générer des artefacts, de fausses informations indiscernables des vraies. Elles sont donc en pratique inexploitables en l'état.

Nous avons présenté un nouveau paradigme de l'analyse exploratoire par projection dans lequel la projection n'est pas une *fin* en soi, mais un *moyen*, un moyen de construire un écran permettant d'afficher une information authentique (*i.e.* non détériorée par la projection) sur les données multi-dimensionnelles à analyser. Nous appelons *WinSitu* ce paradigme de l'analyse exploratoire. Les méthodes de projection linéaire en font partie par construction, l'information authentique visualisée étant dans ce cas l'exact positionnement des points projetés par rapport aux axes principaux combinaisons explicite de variables d'origine. Nous avons montré ici comment les méthodes de projection non linéaire, sous réserve de leur complétion par une mesure authentique co-visualisée, peuvent implémenter ce paradigme et ainsi redevenir exploitables.

Dans les outils d'aide à la décision, l'utilisateur final non expert ne peut faire confiance à cette machine qui l'assiste (et donc légitimer son existence) que dans la mesure où le résultat qu'elle fournit lui est intelligible. Implémenter les méthodes de projection non linéaire dans le cadre *WinSitu*, rend intelligible les vues que ces méthodes fournissent. On peut dès lors légitimement les intégrer aux outils d'aide à la décision, ce qui ouvre de nouvelles perspectives dans le développement de ces outils qui font l'interface avec l'homme dans de nombreux projets industriels.

Nous avons proposé dans (Aupetit et al. 2008) d'autres mesures que la mesure de proximité décrite ici, mais la porte est maintenant grande ouverte pour la proposition de mesures alternatives à co-visualiser dans l'espace de projection. Dans nos travaux futurs, nous envisageons d'étudier la visualisation des données incertaines (incertitude sur la position des individus d'origine, ou sur leurs distances relatives) dans ce nouveau paradigme (Pang et al. 1997). Il serait aussi important de mesurer objectivement le gain en information que ce paradigme apporte, trouver une telle mesure reste encore une question ouverte.

De plus, l'extension de cette méthode à des projections dans un espace à 3 dimensions paraît naturelle et pose des problèmes spécifiques de représentation graphique, mais nous pensons qu'elle vaut la peine d'être expérimentée du fait de la plus grande richesse des informations *in situ* que de telles projections permettraient de visualiser, d'une part parce que le système visuel humain est naturellement adapté à la perception de la profondeur en 3 dimensions (malgrès les problèmes qui se posent d'occlusion entre objets d'avants et d'arrières plans), d'autre part parce que cette dimension supplémentaire permettrait de révéler des configurations spatiales plus complexes avec moins de distorsions (emboîtements, structures surfaciques en plus des structures linéiques et ponctuelles visualisables en 2 dimensions...).

Lorsque le nombre N d'individus devient grand, il n'est plus possible de visualiser l'ensemble des N vues distinctes fournies par la mesure de proximité associée à l'un des N points d'intérêts. Il faudra donc s'attacher à définir un critère d'ordonnancement de ces vues afin de proposer à l'analyste les plus pertinentes, *i.e.* les plus révélatrices de structures particulières.

Nos récents travaux dans cette voie (Lespinats et al. 2010) permettent déjà de distinguer les régions fiables des régions sujettent aux distorsions ainsi que l'intensité et la nature de celles-ci. De même, un trop grand nombre d'individus ou une densité trop forte de points projetés, peuvent masquer la mesure portée par les cellules de Voronoï devenues trop petites dans la représentation graphique : un pixel de l'écran d'affichage pouvant alors contenir plusieurs cellules. Il faudrait proposer un moyen de synthétiser ces multiples mesures afin qu'aucune information importante (VanWijk 2006) ne soit masquée.

Enfin, il n'existe pas à notre connaissance de protocole standard d'évaluation de ces méthodes d'analyse exploratoire par visualisation *in situ*. La mise en place de tels protocoles en partenariat avec des experts en sciences cognitives, et d'un moyen d'enquête (*e.g.* site internet) à partir duquel un nombre suffisant d'utilisateurs experts et non experts pourrait tester et évaluer les différentes méthodes de visualisation, nous semblerait un moyen de pallier au manque d'évaluation scientifique de la qualité de ces représentations visuelles.

Remerciements

Nous remercions les relecteurs pour leurs commentaires constructifs qui ont permis de compléter et d'éclaircir ce travail.

Références

[Allano et al. 2010] L. Allano, M. Aupetit, G.Sannié. Eritr@c : du neutron à la décision pour le contrôle de conteneurs. *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'10)*, Université de Technologie de Troye, 26-27 janvier 2010.

[Asimov 1985] D. Asimov, The grand tour : a tool for viewing multidimensional data. *SIAM journal on scientific and statistical*, 6(1) : 128-143, 1985.

[Aupetit 2007] Aupetit, M. Visualizing distortions and recovering topology in continuous projection techniques. *Neurocomputing* **70(7-9)** :1304-1330, Elsevier 2007.

[Aupetit et al. 2008] M. Aupetit, P. Gaillard. Mesurer et visualiser les distorsions dans les techniques de projection continues. *Revue Sciences et Techniques de l Ingénieur, Revue Intelligence Artificielle, Visualisation et extraction des connaissances*, **22** :443-472, 2008.

[Balzer et al. 2005] M. Balzer and O. Deussen, Voronoi treemaps. *Proc. IEEE Symp. on Information Visualization (INFOVIS'05)*, 2005.

[Barber et al. 1996] Barber, C.B., Dobkin, D.P., and Huhdanpaa, H.T. The Quickhull algorithm for convex hulls *ACM Trans. on Math. Software* vol. 22 :469-483, 1996.

[Bishop et al. 1998] C. Bishop, M. Svensen and C. Williams Developments of the generative topographic mapping. *Neurocomputing* **21** :203-224, Elsevier 1998.

[Bishop 2008] C. Bishop, *Pattern Recognition and Machine Learning.* Springer-Verlag New York 2008.

[Demartines et al. 1997] Demartines, P. & Hérault, J. Curvilinear Component Analysis : a Self-Organising Neural Network for Non-Linear Mapping of Data Sets. *IEEE Trans. on Neural Networks,*8(1) :148-154, 1997

[Euritrack 2006] Euritrack and Eritr@c projects web site : http ://www.euritrack.org

[Friedman et al. 1974] J. H. Friedman, J. W. Tukey, A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers C* 23(9) :881-890, 1974.

[Gaillard et al. 2008] Gaillard P., M. Aupetit, G. Govaert. Learning topology of a labeled data set with the supervised generative Gaussian graph. *Neurocomputing* 71(7-9) :1283-1299, Elsevier 2008.

[Inselberg 1985] A. Inselberg. The Plane with Parallel Coordinates. *Visual Computer*, 1(4) :69-91 ; 1985.

[Jollife 1986] Jollife, I.T. Principal Component Analysis. *Springer Verlag, NY, 1986*.

[Kaski et al. 1998a] S. Kaski, T. Kohonen, J. Venna, *Visual Explorations in Finance using Self-Organizing Maps*, Chapter Tips for SOM Processing and Colorcoding of Maps. Springer-Verlag, London, 1998.

[Kaski et al. 1998b] Kaski, S. , Kangas, J. & Kohonen, T. Bibliography of self-organizing map (SOM) papers : 1981-1997. *Neural Computing Surveys*, 1(3&4) :1-176, 1998.

[Kohonen 1988] T. Kohonen, *Self-Organization and Associative Memory Formation*, Springer-Verlag, 1988.

[Kruskal 1964] J.B. Kruskal, Multidimensional Scaling : A Numerical Method. Psychometrika, 29 :115-129, 1964.

[Le Roux et al. 2007] N. Le Roux, Y. Bengio , P. Lamblin, M. Joliveau et B. Kegl, Learning the 2-D Topology of Images. *Advances in Neural Information Processing Systems*, 20 :841-848, J.C. Platt, D. Koller, Y. Singer et S. Roweis Editors, MIT Press, Cambridge 2008.

[Lespinats et al. 2009] S. Lespinats, M. Aupetit False neighbourhoods and tears are the main mapping defaults. How to avoid it ? How to exhibit remaining ones ? *Quality Issues, Measures of Interestingness and Evaluation of data mining models*, QIMIE09, pp. 55-65. Bangkok, Thaïlande, Avril 2009.

[Lespinats et al. 2010] S. Lespinats, M. Aupetit Mapping without visualizing local defaults is nonsense *European Symposium on Artificial Neural Networks (ESANN'10)*, Bruges, Avril 2010.

[Ma 2009] K.-L. Ma, In Situ Visualization at Extreme Scale : Challenges and Opportunities. *IEEE Computer Graphics and Applications*, 29 :14-19, IEEE 2009.

[Merz et al. 1998] D.J. Newman, S. Hettich, C.L. Blake & C.J. Merz UCI repository of machine learning databases. *Irvine, CA : Dept. of Information and Computer Science, University of California at Irvine*. 1998. Url :http ://www.ics.uci.edu/ mlearn/MLRepository.html.

[Okabe et al. 1992] A. Okabe, B. Boots, K. Sugihara, *Spatial tessellations : concepts and applications of Voronoï diagrams*, John Wiley, Chichester 1992.

[Pang et al. 1997] A. Pang, C. Wittenbrink and S. Lodha Approaches to Uncertainty Visualization, *The Visual Computer*, 13(8) :370-390, 1997.

[Rousset et al. 2001] P. Rousset, C. Guinot Distance between Kohonen classes, visualization tool to use SOM in data set analysis and representation. *IWANN 2001, LNCS 2085*, pp. 119-126, Springer-Verlag Berlin Heidelberg, 2001.

[Sammon 1969] J.W. Sammon, Jr, A nonlinear mapping for data structure analysis, *IEEE Transactions on Computers C* 18(5) :401-409, May 1969.

[Ultsch 1993] A. Ultsch, Self-organizing neural networks for visualization and classification. *Information and Classification - Concepts, Methods and Applications*, pp. 307-313. Springer Verlag, Berlin, 1993.

[Tasdemir et al. 2009] K. Tasdemir, E. Merényi, *Exploiting Data Topology in Visualization and Clustering of Self-Organizing Maps*, IEEE Transactions on Neural Networks, 20(4) :549-562, April 2009.

[Torgerson 1952] W.S. Torgerson, *Multidimensional scaling I - Theory and methods*, Psychometrica, 17 :401-419, 1952.

[VanWijk 2006] J.J. VanWijk, *Views on Visualization*, IEEE Trans. on Computer Graphics, 12(4) :421-432, July-August 2006.

[Venna et al. 2001] J. Venna, S. Kaski, *Neighborhood preservation in nonlinear projection methods : an experimental study*, Artificial Neural Networks - ICANN 2001, LNCS 2130, 485-491, 2001.

[Vesanto 1999] J. Vesanto, *SOM-based data visualization methods*, In Intelligent Data Analysis, 3(2), Elsevier Science, pp. 111-126. IOS Press 1999.

Summary

In this work, we discuss about the practical limits of exploratory data analysis based on continuous projection techniques. We show that it is impossible to infer anything usefull about the original data, from their nonlinear continuous projection. We present an *i*n situ visualization method and we show how it is necessary to make inference possible. This process is part of a new projection-based exploratory data analysis paradigm called *WinSitu*.