# Méthode de visualisation de regroupement statistique à relativement haute dimension

Jean-Charles Risch, Jean Brunet, Eddie Soulier, Francis Rousseaux

# Méthode de visualisation de regroupement statistique à relativement haute dimension

**Jean-Charles Risch**
URCA & CAPGEMINI
31000 Toulouse, France
jean-charles.a.risch@capgemini.com

**Jean Brunet**
Capgemini Technology Services
75000 Paris, France
jean.brunet@capgemini.com

**Eddie Soulier**
UTT
10000 Troyes, France
eddie.soulier@utt.fr

**Francis Rousseaux**
URCA
51100 Reims, France
francis.rousseaux@irc am.fr

**RESUME**
La visualisation de regroupement d'individus statistiques se présente souvent comme une série de nuage de points à analyser dimension par dimension. Cependant, leur comparaison devient de plus en plus coûteuse en temps à mesure que le nombre de dimension augmente, jusqu'à devenir hors de portée pour l'être humain. Afin de palier ce problème, nous proposons une méthode de visualisation complète allant du traitement statistique des données à leur affichage graphique. Les traitements statistiques se basent sur des méthodes de réduction de dimension et de regroupement de données. La visualisation des données, elle, est une représentation graphique unique en deux dimensions. Elle se construit autour des groupes et non pas des individus comme un nuage de points classique pourrait le faire. Ainsi, nous obtenons une liste d'objets représentant les groupes disposés dans un espace à deux dimensions connectés par des liens de similarités et dissimilarités. Cette méthode de visualisation a été expérimentée dans le cadre du projet européen COMPOSITE (Comparative Police Studies In The EU) et s'est avérée utile pour comparer sans effort une soixante-dizaine de force de polices selon dix-sept dimensions. Notre méthode propose une représentation en deux dimensions (extensible sur trois) moderne, interactive et intuitive s'éloignant des représentations classiques en nuage de points.

**Mots Clés**
Data visualisation; classification non supervisée; clustering; regroupement statistique.

**ACM Classification Keywords**
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

## INTRODUCTION

Les objets de notre quotidien sont aujourd'hui de plus en plus connectés entre eux et partagent leurs données via le web. Les réfrigérateurs, les voitures, les routes et même des cadenas de vélo sont connectés au réseau Internet. Ces nouveaux périphériques viennent ajouter de la quantité et de la complexité à l'ensemble des données qui constituent le web. A titre d'exemple, le site Facebook, à lui seul, produisait pas moins de 10 téraoctets de données par jour en 2013, ce qui correspond, par analogie, à plus de quatre millions de photographies prises avec un appareil photo standard. Ainsi, les données qui constituent Internet sont de plus en plus nombreuses et diverses. Ces données correspondent à ce que l'on appelle le Big Data (méga données). Elles sont si nombreuses et diverses qu'elles remettent en question les méthodes classiques d'analyse de données et de visualisation.

Sans se limiter au domaine du Big Data, la visualisation de données est l'intermédiaire entre les méthodes de calculs statistiques et les analyses des utilisateurs. Elle a pour but de présenter des informations utiles de manière claire, précise et rapide pour l'utilisateur, ce dernier n'ayant pas nécessairement de connaissance sur les traitements effectués auparavant. Différentes méthodes de représentation existent : nuages de points, histogrammes, camemberts, boites à moustache. En ajoutant de l'interactivité à la représentation, il est possible d'ajouter de l'information, entre autres, en combinant ces méthodes de représentation entre elles (Par exemple, au survol d'une barre d'un histogramme, il est affiché un camembert affichant plus d'information sur les données en question). Parfois, ces méthodes de représentation ne suffisent plus. En effet, certaines données sont si complexes dimensionnellement parlant qu'elles ne peuvent pas être résumées par un graphique classique. Ainsi, le statisticien se retrouve face à un problème : créer une méthode de représentation visuellement compréhensible et intégrante des données complexes.

Ce papier, présente une méthode de représentation s'appuyant sur un ensemble de données préparées au préalable par des méthodes statistiques. Dans une première partie, nous présentons une synthèse des travaux réalisés sur le sujet puis nous discutons dans une seconde partie des données utilisées, des différents algorithmes de réduction de dimensions et de regroupement. Dans une troisième partie, nous présentons en détail notre méthode de visualisation et enfin, nous conclurons sur les perspectives de cette recherche.

## TRAVAUX CONNEXES

Nos travaux se basent sur des méthodes statistiques de regroupement. Le regroupement statistique est un domaine largement étudié. L'algorithme des K-Moyennes, également appelé algorithme de Lloyd-Max est le plus connu dans le domaine. Pavel Berkhin le qualifie d' « algorithme de regroupement de loin le plus populaire et le plus utilisé dans les applications scientifiques et industriels » dans [1]. Le but est de diviser une population en k groupes dans lesquels chaque individu est associé au groupe dont la moyenne est la plus proche. Il existe d'autres techniques de regroupement. Nous pouvons citer par exemple les méthodes de regroupements hiérarchiques. Ces techniques sont elles aussi très connues et très utilisées, notamment l'algorithme de classification ascendante hiérarchique. Ces méthodes de regroupement sont en règle générale couplées avec des méthodes de visualisation classique : nuage de points, histogramme, etc. Les données s'étant complexifiées ces dernières années (décennies) avec une augmentation de la dimensionnalité des données, nous devons remettre en question la qualité de ces graphiques dits classiques. C'est ainsi qu'intervient le domaine de la visualisation de données.

La visualisation de données est un sujet très important dans le domaine des statistiques. La visualisation correspond au bout de la chaine d'analyse de données. Elle est l'interface avec l'utilisateur et permet à celui-ci de comprendre les données préparée par le statisticien en amont. C'est ce dont il est question dans [2] où le but était de présenter aux internautes un moyen simple, rapide et intuitif permettant d'évaluer la qualité d'un article sur Wikipédia via différents graphiques.

Représenter des données à forte dimensionnalité est un problème connu dans la littérature [5], [10]. En effet, plus la dimensionnalité des données est grande, plus le nombre de graphique pour analyser ces dernières va être important. Effectivement, un graphique en trois dimensions ne permet d'afficher que trois dimensions à la fois. Ainsi, pour combiner toutes les variables entre elles, il est nécessaire d'avoir un nombre de graphique conséquent. [12] présente une méthode de visualisation interactive en 3 dimensions et se basant sur des sous espaces à dimensions réduits. Cette méthode est une amélioration de la représentation sous forme de nuage de points d'individus statistique avec des outils modernes permettant l'interaction avec l'utilisateur. Cependant, cette méthode ne résout pas le problème de la dimensionnalité des données, elle contourne le problème à l'aide d'outils récents. [5] présente une série de graphiques permettant la visualisation de données à plus de trois dimensions. Un chapitre entier est réservé à la visualisation de regroupements statistiques. On y retrouve des représentations par nuage de points, radar etc. Ces méthodes de représentation ont toutes un défaut en commun : elle nécessite un apprentissage utilisateur important à cause du nombre important d'informations par graphique.

En résumé, les méthodes statistiques permettant la réduction de la dimension ont déjà fait leurs preuves. De plus, des outils performants permettent de les exploiter facilement. D'un autre coté, il y a eu des efforts concernant la visualisation de regroupement de données complexes. Cependant, ces méthodes ont toutes une même limite concernant la dimensionnalité. C'est ce constat qui nous a mené vers la mise en place notre méthode de représentation.

## TRAITEMENT DES DONNEES

Dans cette partie, nous allons discuter de la partie technique permettant de produire les données utiles à la représentation. Pour commencer, nous décrirons les données d'entrée "brutes" en introduisant leur contexte, puis nous expliquerons les méthodes de regroupements et de réduction de la dimension utilisées. Enfin, nous présenterons la méthode utilisée pour différencier deux groupes de données.

### Description des Données d'Entrée

COMPOSITE est un projet Européen ayant pour objectif de comparer les forces de police européennes entre elles. Il s'est déroulé de Août 2010 à Juillet 2014 et à regrouper différentes équipes de recherche et industriels d'une dizaine de pays différents. Les données d'entrées correspondent aux données produites par l'entreprise Capgemini France, membre du consortium.

Les données représentent un comptage simple de mots à partir des différents sites web des forces de polices analysées. Ces mots sont regroupés dans une ontologie à différents niveaux hiérarchiques. L'un de ces niveaux est composé de 17 concepts permettant de décrire une force de police (Justice, Mission, …). C'est ce niveau de hiérarchie que nous avons analysé du fait du nombre de dimensions suffisamment élevé sans être dans l'excès.

La méthode de représentation a donc initialement été mise en place pour ce projet. Elle a notamment été utilisée par l'analyste de l'équipe et a permis de produire des résultats intéressants.

Les données d'entrée sont composées d'environ 70 individus statistiques. Ce nombre est faible, cependant la complexité de nos travaux n'est pas le nombre d'individu mais celle de ses caractéristiques. En effet, chaque individu est décrit par 17 variables numériques. Ces variables sont exprimées en pourcentage, ainsi, la somme de chacune des variables d'un individu est égale à 100.

Pour information, ces variables résultent d'une extraction d'entités nommées par le biais d'une ontologie sur des sites web. Les variables correspondent à une certaine hiérarchie de l'ontologie et les individus correspondent aux sites web analysés.

**Réduction de la Dimension et Regroupement des Individus**

Le nombre de variable de nos données est suffisamment élevé pour rendre impossible une visualisation correct des individus les uns par rapport aux autres. Egalement, visualiser les individus variable par variable serait une tâche longue à exécuter. En effet nous nous retrouvons avec une combinaison de 136 variables deux à deux et donc 136 graphiques différents à analyser.

La première étape de notre analyse a donc été de réduire le nombre de dimension en cherchant des corrélations entre les variables. Les données étant numériques, nous avons procédé à une analyse en composante principale (ACP).

Ensuite, afin de grouper les individus ayant le maximum de caractéristiques en commun, nous avons appliqué l'algorithme des K-Moyennes sur les données résultantes de l'ACP. Cet algorithme prend en entrée un nombre k de groupes souhaité par l'utilisateur est un ensemble d'individus à regrouper. Il émet en sortie une liste de k partitions contenant les individus. Cependant, le nombre de groupes n'est pas forcément connu par l'utilisateur. Ainsi, nous avons utilisé l'algorithme de Classification Ascendante Hiérarchique (CAH). Cet algorithme de regroupement permet de donner des estimations sur le nombre de groupes optimal permettant de séparer au mieux les données.

Ainsi, une fois les données préparées, nous obtenons une liste de groupes contenant les individus. Chaque groupe est décrit par des variables significatives. Une variable significative, indique que la moyenne de cette variable pour les individus du groupe est significativement plus forte ou plus faible que la moyenne de cette variable sur l'ensemble des individus.

Ces travaux de préparation des données ont pu être possibles grâce au package *FactoMineR* développé pour R.

**Distance entre Deux Groupes**

Nous disposons désormais d'un ensemble de groupes contenant chacun un certain nombre d'individus et d'une liste de variables descriptives pour chacun des groupes. Nous souhaitons maintenant trouver une méthode pour connaitre les distances entre les groupes et ainsi positionner les groupes les uns par rapport aux autres dans un espace à deux dimensions.

Chercher une distance entre deux groupes peut se ramener à calculer la distance entre les centres des groupes. Différentes distances sont usuellement utilisées. Nous pouvons citer la distance euclidienne, de Manhattan ou encore de Minkowski. Ces distances prennent en considération chacune des dimensions des données et tente de chiffrer la similarité entre deux données différentes.

Cependant, notre jeu de données contient 17 variables comme nous l'avons dit. Calculer une distance entre deux points se résume à faire une moyenne des distances sur chacune des dimensions de ces points. Ainsi, lorsque le nombre de dimension est élevé, l'effet de moyenne aplatit les différences significatives qu'il peut y avoir entre ces deux points. Ainsi, calculer des distances entre deux points sur 17 dimensions risque de ne pas représenter correctement l'écart réel entre deux données.

Notre idée est différente. Nous allons chercher à mettre en valeur les similarités et les différences que les groupes peuvent avoir entre eux. Pour calculer ces similarités/différences, nous allons nous appuyer sur les variables significatives qui décrivent chacun des groupes. Comme nous l'avons dit plus tôt, une variables peut être significative négativement (moyenne plus faible dans le groupe que pour l'ensemble des individus) ou significative positivement (moyenne plus forte dans le groupe que pour l'ensemble des individus).

De ce fait, nous pouvons lister l'ensemble des similarités/ différences que deux groupes peuvent avoir entre eux :

- Les groupes A et B ont une variable significativement positive en commun;
- Les groupes A et B ont une variable significativement négative en commun;
- Les groupes A et B ont une variable significative en commun mais elle l'est positivement pour le groupe A et négativement pour le groupe B;
- Les groupes A et B ont une variable significative en commun mais elle l'est négativement pour le groupe A et positivement pour le groupe B.

Les deux premiers points correspondent donc à des similarités entre deux groupes et les deux suivant à des différences.

Nous pouvons donc calculer la matrice des similarités entre les groupes.

$$Soit\ A_{mn}\ une\ matrice\ carrée$$

$$de\ dimension\ (m,n)$$

En ligne (m) et en colonne (n) se trouve les différents groupes. Une case de la matrice correspond donc à la distance entre le groupe i et le groupe j.

Ainsi, nous posons le premier calcul :

$$A_{ij} = NB(Similarité)_{ij} - NB(différence)_{ij}$$

A noter que calculer la similarité/différence d'un groupe avec lui même n'ajoute pas d'. Ainsi, nous notons 0 dans la diagonale de la matrice. D'où :

$$Pour\ tout\ i = j, on\ a$$

$$A_{ij} = 0$$

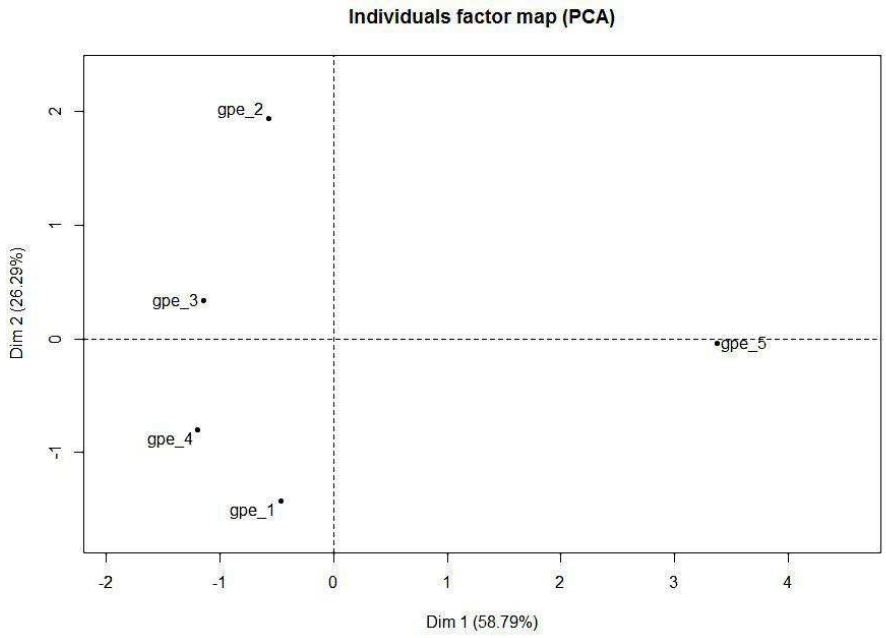


**Figure 1. Position de 5 groupes après une ACP**

Ainsi, la matrice peut contenir des valeurs négatives (plus de différences que de similitudes), des valeurs positives (plus de similitudes que de différences) et de valeurs nulles (ni similitude, ni différences ou autant de similitudes que de différences). Cependant, nous souhaitons trouver des distances entre les groupes, or une distance ne peut pas être négative. Egalement, annoncer une distance entre deux groupes comme étant égale à 0 signifierait que les deux groupes sont confondus (ce qui n'est pas le cas excepté pour la distance d'un groupe à lui-même). Ainsi, nous allons remonter l'ensemble des valeurs de la matrice jusqu'à obtenir une valeur minimale à 1. Pour cela, nous allons chercher la plus petite valeur inférieure à 1 (sauf les valeurs contenues dans la diagonale) et nous allons additionner sa valeur absolue agrémentée de 1 à l'ensemble de la matrice. Nous obtenons donc :

$$Si\ MIN(A_{ij}) < 1$$
$$Quelque\ soit\ i <> j$$
$$A_{ij} = [1 - MIN(A_{ij})] + A_{ij}$$
$$Sinon$$
$$A_{ij} = A_{ij}$$

Enfin, nous devons calculer l'inverse de chacune des distances de la matrice car plus il y a de différences entre deux groupes, plus nous souhaitons que ceux ci soient distants. Bien sûr, nous n'appliquons pas ces calculs à la diagonale. Une distance entre un groupe et lui même est 0, nous assignons donc 0 à la diagonale. Nous obtenons donc la matrice finale :

$$Quelque\ soit\ i <> j$$
$$A_{ij} = 1/A_{ij}$$

Nous obtenons donc une matrice des distances entre chacun des groupes. Afin de pouvoir donner des coordonnées sur deux dimensions à ces groupes, nous appliquons de nouveau une analyse en composante principale et sélectionnons les deux premiers axes obtenu.

La figure 1 présente le résultat d'une ACP sur une matrice des distances de 5 groupes. Les groupes sont disposés sur les deux axes principaux résultants. Comme nous pouvons le constater en additionnant les pourcentages de représentation de chacun des deux axes, ce graphique expose 85.08% de la réalité.

Désormais, nous avons les informations nécessaires pour créer une représentation simple en deux dimensions. Dans la partie suivante, nous allons expliquer nos choix concernant la mise en forme de notre représentation.

## REPRESENTATION GRAPHIQUE EN 2D

Dans cette partie, nous allons présenter notre représentation réalisée sur les données précédemment décrites. Dans un premier temps, nous allons exposer l'aspect global de la représentation, puis nous nous intéresserons à la représentation d'un groupe simple, ensuite nous décrirons les liens entre les groupes et enfin nous exposerons l'interactivité que présente notre représentation.

### Aspect Général

Comme nous l'avons présenté dans l'introduction, nous souhaitons créer une représentation liant la complexité des travaux statistiques et la modernité des infographies actuelles.

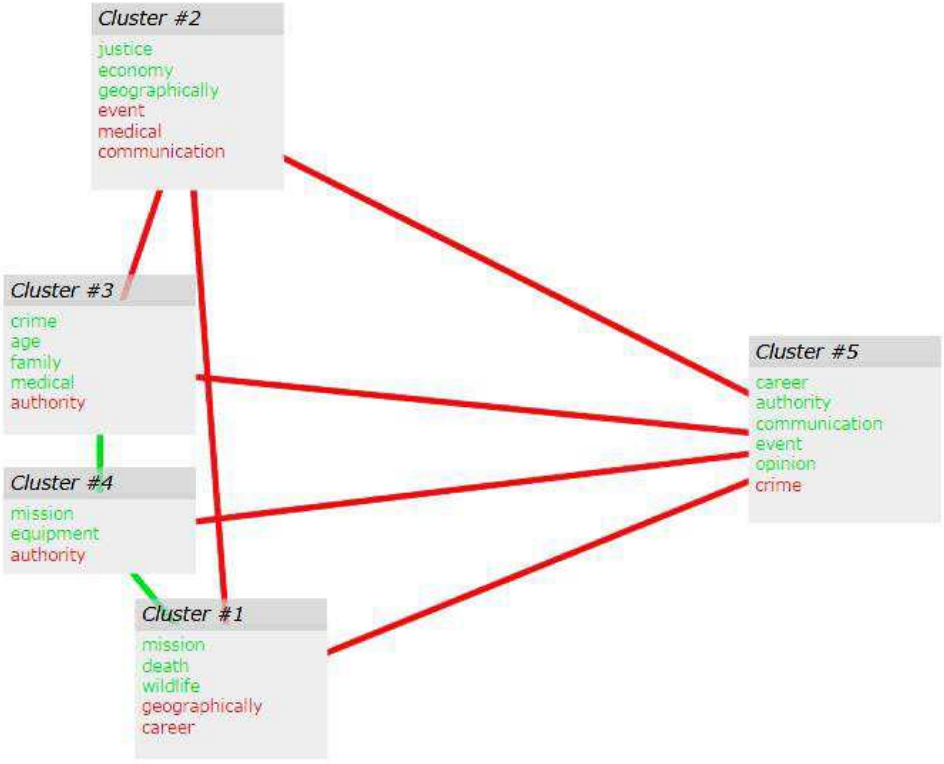


**Figure 2. Exemple de représentation finale (figée) d'un regroupement**

C'est en s'inspirant des applications web modernes et des récentes *infoviz* que nous avons décidé d'introduire de l'interactivité dans l'application. Effectivement, les sites web tendent de plus en plus à devenir des applications dynamiques et interactives avec l'essor des technologies telles que CSS3, HTML5 ou encore SVG, AJAX et Javascript.

C'est donc tout naturellement, que nous avons créé une application web générant une représentation en SVG (Scalable Vector Graphics) via la bibliothèque Snap SVG[1]. Les données permettant de construire le graphique sont stockées dans une base de données relationnelle classique. Dans un souci de portabilité, les données relatives au graphique sont également stockées dans des fichiers XML. C'est d'ailleurs en traitant ces fichiers que la représentation est construite.

La figure 2 est un exemple de représentation d'un regroupement statistique d'individus. Ci-après, nous allons décrire en détails chaque élément de cette représentation.

### Un Groupe

Un groupe est représenté par un rectangle gris. Il a un nom (affiché en haut de son rectangle) et contient la liste de ses variables significatives. Comme on peut l'observer sur la figure 2, une variable significative peut avoir deux couleurs :

- Vert : la variable est significativement positive pour le groupe;
- Rouge : la variable est significativement négative pour le groupe.

### Liens Entre les Groupes

Les liens entre les groupes représentent les similitudes et différences qu'il peut y avoir entre les groupes. Il existe 3 types de lien possible, comme on peut en observer deux sur la Figure 2 :

- Lien vert : Les deux groupes liés ont au moins une similitude et aucune dissimilitude;
- Lien rouge : Les deux groupes liés ont au moins une dissimilitude et aucune similitude;
- Lien bleu : Les deux groupes liés ont au moins une similitude et une dissimilitude.

Le code couleur n'a pas été choisi au hasard. Il correspond aux codes généralement utilisés dans la vie de tous les jours. Par exemple, la couleur bleue est utilisée en électricité pour désigner un conducteur neutre.

### Interactivité

Différents éléments sont affichés à la demande de l'utilisateur. Mais pas seulement, l'utilisateur peut manipuler le graphique comme il le souhaite. Ci après, nous exposons la liste des différentes interactivités possibles.

1. [1] Adobe. http://snapsvg.io/

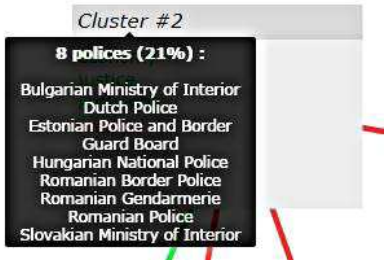


**Figure 3. Liste des individus lors du passage de la souris sur le nom du groupe**

[1] *Liste des individus* : La liste des individus par groupe est disponible en passant la souris au dessus du nom du groupe. Elle est accompagnée du pourcentage d'individus contenu dans le groupe par rapport au nombre total d'individu. Figure 3 est un exemple.

[2] *Explication chiffrée des variables significatives* : En passant la souris sur chacune des variables significatives, l'utilisateur pourra y visualiser la moyenne de la variable pour le groupe concernée ainsi que la variable pour l'ensemble des individus. Figure 4 est un exemple.

[3] *Liste des variables en commun* : En passant la souris au dessus d'un lien, l'utilisateur pourra y voir la liste des variables qui forment ce lien. Egalement, la taille du lien survolé augmente pour éviter toute ambigüité avec un lien adjacent. .

[4] *Drag and drop des groupes* : En cliquant sur le rectangle gris d'un groupe, l'utilisateur peut déplacer les groupes où il le souhaite et ainsi repositionner les groupes à sa guise.

[5] *Effet de zoom simple* : Deux boutons [+] et [-] permettent à l'utilisateur de zoomer et dé-zoomer sur le graphique.

Avec ces interactivités, l'utilisateur augmente les informations disponibles sur la représentation de base sans se perdre dans la quantité d'information.

## CONCLUSION

Nous avons montré qu'il est possible de représenter des regroupements de données à relativement haute dimension à l'aide d'un unique graphique à deux dimensions.


**Figure 4. Explication chiffrée d'une variable significative**

Ainsi, notre méthode de représentation permet de simplifier les nombreux graphiques nécessaires à l'analyse d'un regroupement de données complexes. De ce fait, nos travaux permettront aux experts et autres analystes de réduire leur temps passé à analyser des données de ce type. Aussi, notre graphique étant simplifié et interactif, il nécessite moins de temps d'apprentissage de la part des utilisateurs, ainsi, il démocratise ce genre d'analyse.

La dimension est l'un des points centraux de notre étude. Les données sont constituées de 17 dimensions numériques. Une limite de notre étude est ce nombre, qui comme le titre l'indique est seulement "relativement" élevé.

Maintenant que nous avons reçu de bons retours de la part de plusieurs consultants analystes de COMPOSITE, nous avons de nouveaux projets pour cette visualisation en particulier le passage de 2 à 3 dimensions (pour gagner en qualité de représentation), la possibilité d'ajouter des notes à n'importe quel objet du graphique (afin de faciliter l'analyse à plusieurs) ou encore l'exportation au format GeoPDF (dans le but de conserver l'interactivité du graphique dans un PDF).

## BIBLIOGRAPHIE

1. Berkhin P. Survey of clustering data mining techniques. Dans Grouping Multidimensional Data pp 25-71; 2006.
2. Chevalier Fanny, Huot Stéphane et Fekete Jean-Daniel. Visualisation de mesures agrégées pour l'estimation de la qualité des articles Wikipedia. 2010.
3. Edward R. Tufte. The Visual Display of Quantitative Information. 2001.
4. Fekete Jean-Daniel. Dataviz & BidData : Mythes et réalité. Microsoft Tech Days 2014.
5. Hoffman Patrick, Grindstein Georges. A survey of visualizations for high-dimensional data mining. Dans Information visualization in data mining and knowledge discovery. 2002.
6. Huron Samuel, Vuillemot Romain, Fekete Jean-Daniel. Visual Sedimentation. 2013.
7. Husson F., Lê S. et Pagès J. Analyse de données avec R. 2009.
8. Keim Daniel A., Panse Christian, Schneidewind Jörn, Sips Mike, Hao Ming C., Dayal Umeshwar. Pushing the limit in Visual Data Exploration: Techniques and Applications. Dans KI 2003 : Advances in Artificial Intelligence.2003.
9. Maccandless David. Datavision. 2011.
10. Saby Claude-Alain. Méthodes de visualisation de données à fortes dimensions dans un espace réduit à 2 D. 2011.
11. McCandless David. La beauté de la visualisation des données. Dans TEDGlobal 2010. Juillet 2010.
12. Waddel A, Oldford. Interactive Visual Clustering of High Dimensional Data by Exploring Low-Dimensional Subspaces. Vis 2012.
13. Yau Nathan. Data visualisation : De l'extraction des données à leur représentation graphique. 2013 (Livre)