# INTERACTIONS DISTANTES ENTRE CARTE GEOSTRATEGIQUE, CLASSIFICATION, SEGMENTATION ET NORMALISATION

**Bernard DOUSSET, Samuel GOURLAIN**
Institut de Recherche en Informatique de Toulouse, équipe SIG-EVI,
Université Paul Sabatier - 118, route de Narbonne 31062 Toulouse cedex 9
tél. (33) 05 61 55 67 81. ; fax. (33) 05 61 55 62 59 ; email : dousset@irit.fr ; w3 : http://atlas.irit.fr

**Mots Clés**
Découverte de connaissance, Classification ascendante hiérarchique, Clustering, Classification par partition, Cartes géographiques, Arbres planaires, Communication inter processus, Travail distant, Collaboration distante.
**Keywords**
Knowledge discovery, ascending hierarchical classification, Clustering, Classification by partition, Maps, Trees planar septal process, remote work, remote collaboration.

**Résumé**
L'analyse des informations externes utiles à la conduite de stratégies a très souvent recours aux méthodes de classification afin de construire des modèles synthétiques bien adaptés à la psychologie et à la démarche des différents profils de décideurs. Mais quel type de modèle choisir, quel degré de granularité, quelle métrique utiliser, comment dialoguer sur ce sujet avec un expert distant, quelle interprétation peut-on en faire. D'autre part, un ensemble de classes ne suffit pas à faire passer de façon attractive le message de l'analyste vers le destinataire. C'est pour cet ensemble de raisons, que nous proposons un système intégré de communications entre différentes méthodes d'analyse qui permette à la fois de faire collaborer les hommes et les démarches via le réseau grâce aux système d'exploitation Unix ou Linux. En synchronisant les investigations de plusieurs acteurs utilisant diverses démarches, en visualisant simultanément les résultats graphiques de chacun et en faisant collaborer les méthodes, il est possible de confronter les points de vue et donc d'arriver rapidement à un consensus sur une problématique complexe. Techniquement nous montrerons qu'il est possible d'importer le résultat d'une classification dans une carte géographique distante, d'afficher une classification de pays sur une carte du monde, de reprendre cette classification dans un tableur pour en donner une validation et une interprétation, de naviguer dans les arbres de classification et de choisir le bon niveau de coupure. Le clustering et les classifications par partition seront aussi abordés sur le plan de l'interactivité. Enfin, nous illustrerons par des exemples toutes les possibilités d'utilisation de ce nouveau type de découverte basé sur la collaboration à distance entre plusieurs utilisateurs comme par exemple : l'analyste, l'expert et le décideur.

**Abstract**
The analysis of external information relevant to the conduct of strategies often used classification methods to construct synthetic models suited to the psychology and approach of different profiles of decision makers. But what kind of model to choose, what degree of granularity, what metric to use, how to dialogue on this subject with a distant expert, what interpretation can we make of it. On the other hand, a set of classes is not enough to convey the message so attractive to the analyst to the recipient. It is for all these reasons, we propose an integrated system of communications between different methods of analysis that allows both men to work and approaches through the network with the Unix operating system or Linux. By synchronizing the investigations of several players using various approaches, simultaneously visualizing the graphical results of each and making the methods work, it is possible to compare their points of view and thus reach an early consensus on a complex issue. Technically we will show that it is possible to import the result of a classification in a remote geographical map, displaying a

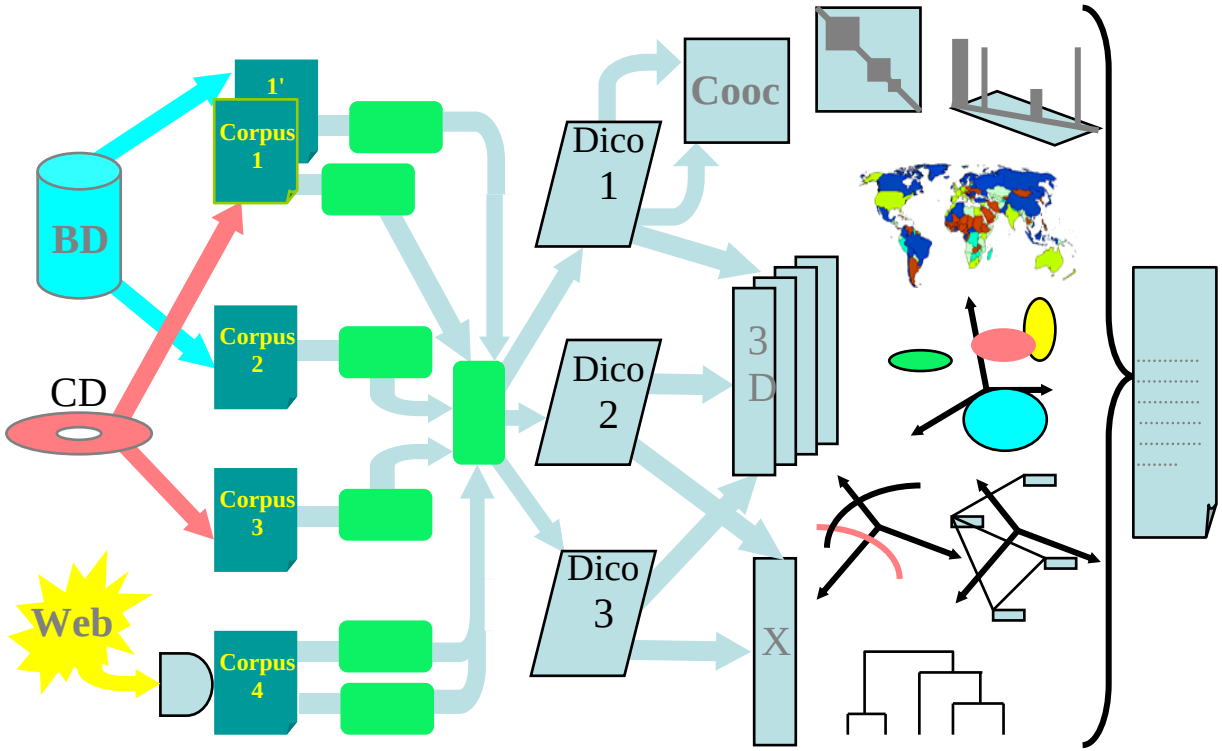
classification of countries on a world map, to take this classification into a spreadsheet to provide validation and interpretation, navigate the tree classification and choose the right level of cuts. Clustering and classification by partition will also be discussed in terms of interactivity. Finally, we illustrate with examples the possibilities of using this new type of discovery-based remote collaboration between multiple users such as: analyst, expert and decision maker.

# 1 Introduction

Le travail qui est présenté ici décrit la philosophie d'utilisation interactive de la plate-forme « Tétralogie » qui a été développée à l'IRIT. En effet, il est possible de faire collaborer plusieurs acteurs (documentaliste, informaticien, statisticien, expert, décideur, …) au cours des différentes phases d'une analyse d'un grand corpus d'information scientifique ou technique. Nous supposons dans notre présentation que les phases de recherche d'information, de préparation des corpus, d'homogénéisation des formes orthographiques, d'extraction des dictionnaires et de croisement de l'information sont déjà réalisées. Nous nous positionnons donc dans la phase d'analyse proprement dite où les différents interlocuteurs sont amenés à étroitement collaborer pour donner un sens à leurs investigations. Nous illustrons notre propos en montrant comment faire collaborer entre elles des cartes géographiques distantes mais aussi comment elles intègrent dynamiquement des résultats de classification, de segmentation ou de normalisation réalisées en s'appuyant sur un ensemble de données externes sur des pays ou des régions : population, pnb, surface, rejets de $CO^2$, regroupements significatifs.

# 2 Principe d'une analyse

Un corpus scientifique ou technique (éventuellement économique ou juridique) est constitué d'un recueil d'information semi structurée issue de sources multiples sous des formats différents. Dans un premier temps, des méta données (ci-contre en vert) permettent de décrire logiquement ces formats et de les mettre en correspondance (forme et contenu) afin d'en réaliser une analyse globale. Après un traitement morphologique (synonymies), des dictionnaires sont constitués pour l'ensemble des acteurs (auteurs, inventeurs, organismes, villes, régions, pays, journaux, sites, sources, …) et des éléments sémantiques (mots clés, index, descripteurs, codes de classification, thesaurus, ontologies [3], segments répétés, texte libre, …). Un système de croisement permet ensuite la génération de matrices de présence absence, de cooccurrences ou de contingences qu'il est possible d'analyser par les méthodes traditionnelles de l'analyse de données ou par des méthodes innovantes: cartes factorielles en 4D, arbres de classification interactifs, réseaux de liens, analyse de trajectoires, rotations procustéennes, dessins de graphes [5], …



La figure ci-contre illustre les différentes étapes d'une telle analyse.

Figure 1 : Processus de découverte de connaissances

Mais à chaque étape de l'analyse, des différents acteurs interviennent. Ils sont le plus souvent distants les uns des autres, ils doivent se concerter, leur dialogue est souvent assez bref mais doit être étayé par des visualisations graphiques [1] sur les quelles ils doivent pouvoir successivement s'exprimer et prendre la main. Le mode opératoire pour une analyse coproduite est alors le suivant:

- Le documentaliste met au point une équation de recherche avec l'aide de l'expert et interroge les bases d'informations les plus pertinentes.
- Un ou plusieurs corpus sont alors déchargés et mémorisés sur un serveur distant accessible par tous.
- Une première étude assez rapide est effectuée par l'analyste.
- Une concertation entre le documentaliste et l'analyste permet de corriger les biais de la première interrogation.
- Le corpus est alors corrigé, complété et validé.
- Une seconde étude plus poussée est ensuite réalisée. Elle débouche sur une première évaluation avec l'expert.
- L'expert peut demander de pousser l'étude plus loin, de faire des zooms sur des points techniques ou stratégiques.
- Des sorties graphiques des résultats les plus marquants sont alors sélectionnées pour pouvoir les communiquer au décideur.
- La dernière étape met en jeu l'expert et le décideur qui font une synthèse de l'étude et déterminent la conduite à tenir.

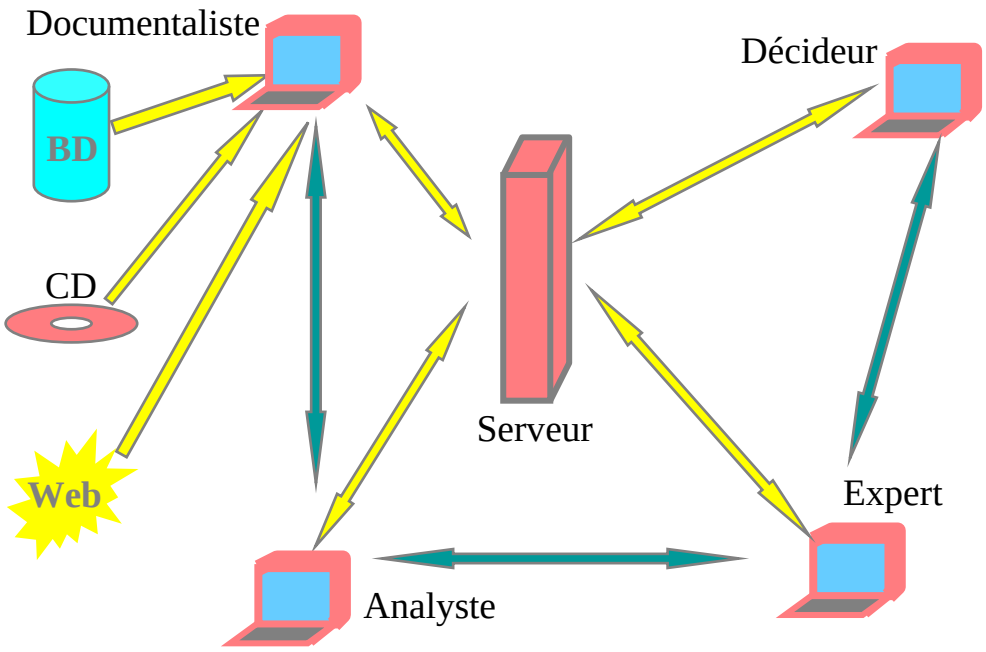



Figure 2 : Interactions distantes entre les différents acteurs d'une analyse

# 3 Classification ascendante hiérarchique

## 3.1 Principe

La CAH consiste à regrouper des points (individus) dans des classes imbriquées. Au départ chaque point représente une classe, puis on regroupe les deux classes les plus proches selon une mesure de distance. Les classes obtenues à l'étape précédente seront à leur tour agrégées en fonction de leurs distances deux à deux. Le processus est réitéré jusqu'à l'obtention d'une seule classe regroupant tous les points.
Cette classification génère un arbre planaire que l'on peut couper à différents niveaux pour obtenir des classes plus ou moins grandes ou plus ou moins nombreuses. L'examen de l'arbre hiérarchique est un moyen efficace pour déterminer un nombre de classes raisonnable, ce qui est toujours difficile en classification. On peut, en

effet, juger immédiatement de la qualité de la classification obtenue, car il faut pouvoir couper l'arbre le plus bas possible (distance inter classes grande, distance intra classes petite) et ce sans générer trop de classes.
Pour faciliter cette coupe, il faut donc évaluer correctement les paliers de la hiérarchie ; c'est à dire, leur associer une hauteur significative dans l'arbre correspondant, d'où la notion de hiérarchie indicée [1].

## 3.2 Algorithme général de la CAH

Cet algorithme consiste à construire, à l'aide de l'indice d'agrégation $\delta$ choisi, une suite de partitions de moins en moins fines dont les classes forment la hiérarchique H cherchée. L'algorithme est le suivant:

1) Partir de la partition initiale dont les classes sont réduites à un seul élément ou à des groupes d'éléments confondus.
2) Construire une nouvelle partition en réunissant les deux classes de la partition précédente qui minimisent $\delta$.
3) Recommencer 2) jusqu'à ce que toutes les classes soient réunies en une seule.

Si à l'étape 2 il y a plus d'un couple de classes qui minimise $\delta$, on en choisi un au hasard ; il n'y a donc pas toujours unicité pour la hiérarchie obtenue. On peut remarquer aussi que la hiérarchie donnée par cet algorithme est nécessairement binaire. Pour $n$ individus, cet algorithme fournit une hiérarchie de $2n - 1$ paliers.

## 3.3 Visualisation interactive de la classification

Après l'étape d'analyse intégrée au processus de découverte de connaissances, il est nécessaire de fournir des moyens efficaces pour représenter ses résultats. Dans le principe de l'approche interactive, il s'agit de visualiser l'arbre planaire ou dendrogramme correspondant à la hiérarchie indicée telle qu'elle est définie précédemment, la hauteur d'un palier de la hiérarchie est équivalente à la valeur de son indice d'agrégation. Une telle méthode est bien adaptée à des hiérarchies de taille moyenne. Cependant, elle ne s'adapte pas directement à de grosses structures. Le système Tétralogie intègre cette méthode. Dans le cas de données de taille moyenne le résultat est particulièrement satisfaisant. En effet, un seul coup d'œil suffit à juger de la qualité de la classification, il est possible de couper l'arbre de façon interactive et d'exporter les classes ainsi obtenues vers les vues multidimensionnelles en 4D (colorations), vers le tableur (numérotation des classes), vers des dessins de graphes ou vers une carte du monde si on a classé des pays ou des cartes régionales (Etats Unis, Brésil, Russie, Inde, Chine, Union Européenne) pour des granularités plus fines.

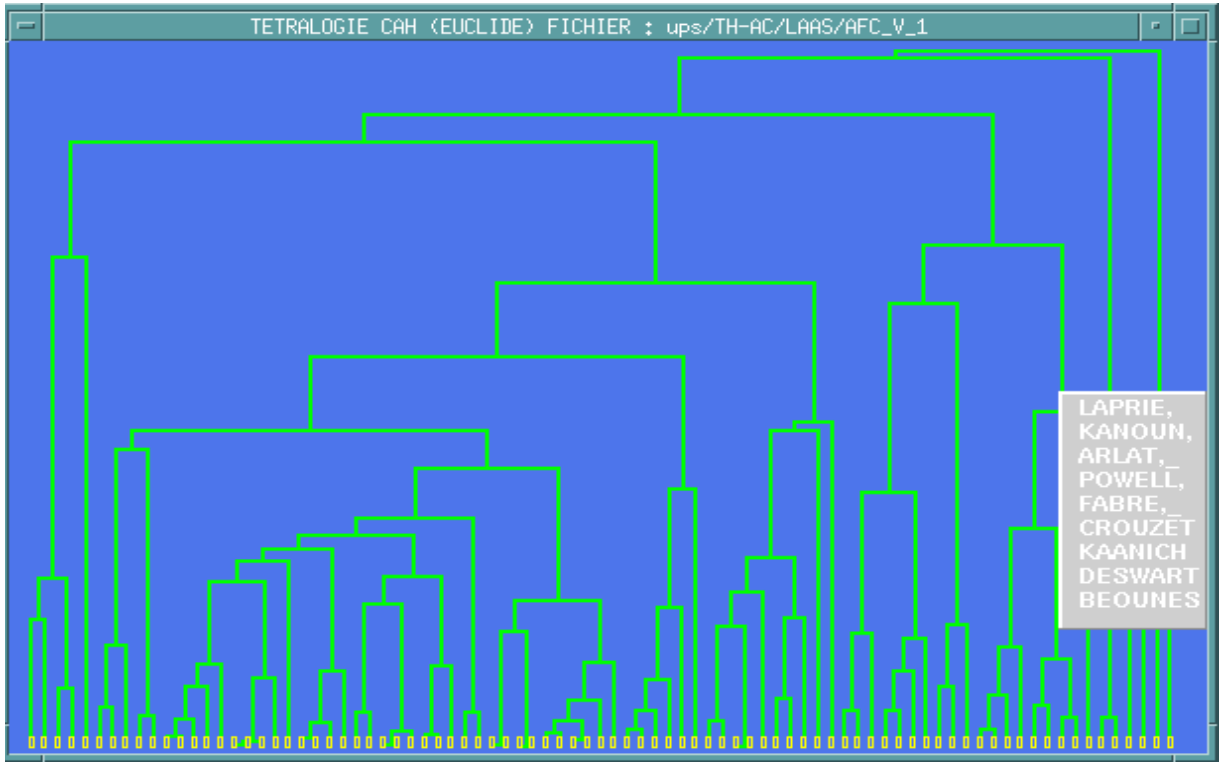



Figure 3 : Dendrogramme d'une classification

ascendante hiérarchique

# 4 Interactivité avec les classifications

## 4.1 Tableur

En important les classes dans le tableur, il est possible de conserver ces classes sous forme de numéros et de proposer ensuite des graphiques dont les couleurs sont en correspondance. Un tri par classe peut aussi être opéré afin d'avoir en contigu les éléments d'une même classe et ainsi pouvoir les isoler, les comparer et en trouver les caractéristiques communes. Les autres tris du tableur (connexité, blocs relatifs ou absolus) permettent le même type d'analyse.

## 4.2 Cartes factorielles

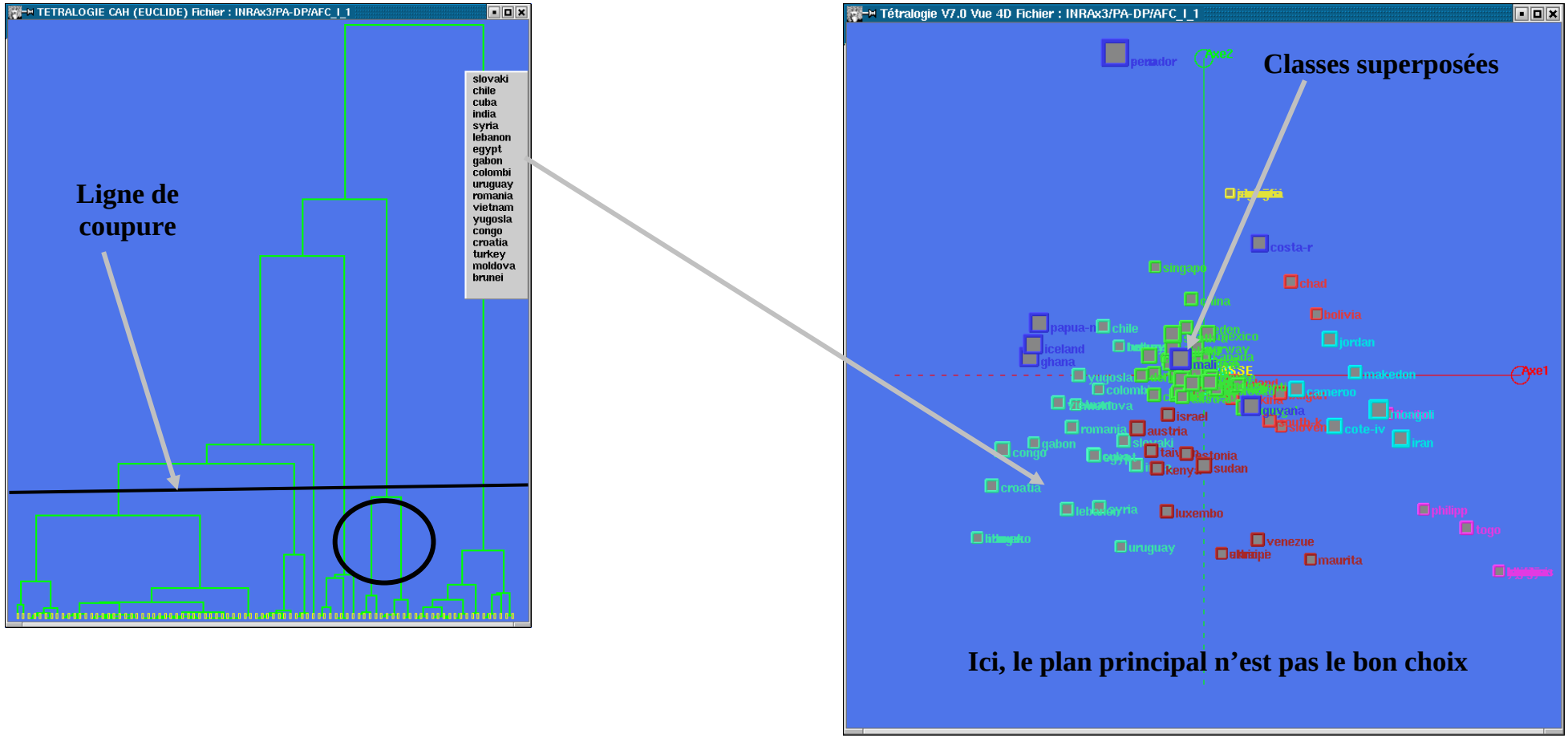



Figure 4 : Interaction distante entre classification et carte factorielle

Dans le cas d'une carte factorielle, il est possible de visualiser simultanément une vue 4D et la classification de ses items: coordonnées des individus (ACP), corrélations des variables (ACPr), contributions (AFC) des individus, des variables ou des deux. Ensuite, pour un niveau de coupure donné, il est possible d'exporter les classes obtenues dans la carte et de les y visualiser par différentes couleurs. La vue peut alors apparaître comme mal choisie pour la classification proposée (classes confondues ou non représentées), des rotations interactives et des glissements d'axes peuvent en améliorer notablement la qualité perçue

## 4.3 Carte du monde

Si une des informations croisées est le pays (en ligne ou en colonne), il est possible d'exporter le résultat d'une classification vers une carte du monde et ainsi situer les signaux forts, l'évolution et les émergences du second critère. Des pondérations sont possibles en faisant intervenir des données géographiques externes (population, pnb, surface, $CO^2$, …) afin de relativiser certaines phénomènes perçus (prédominance américaine, pays riches, très peuplés, très grands, …). Ces cartes sont elles mêmes interactives et chaque opinion peut être partagée et discutée en temps réel, images à l'appui. Des éléments géostratégiques sont ainsi facilement décelables et leur découverte peut être facilitée par l'utilisation de filtres spécifiques (G7, G8, tiers ou quart monde, Europe, Francophonie, zone Euro, Commonwealth).

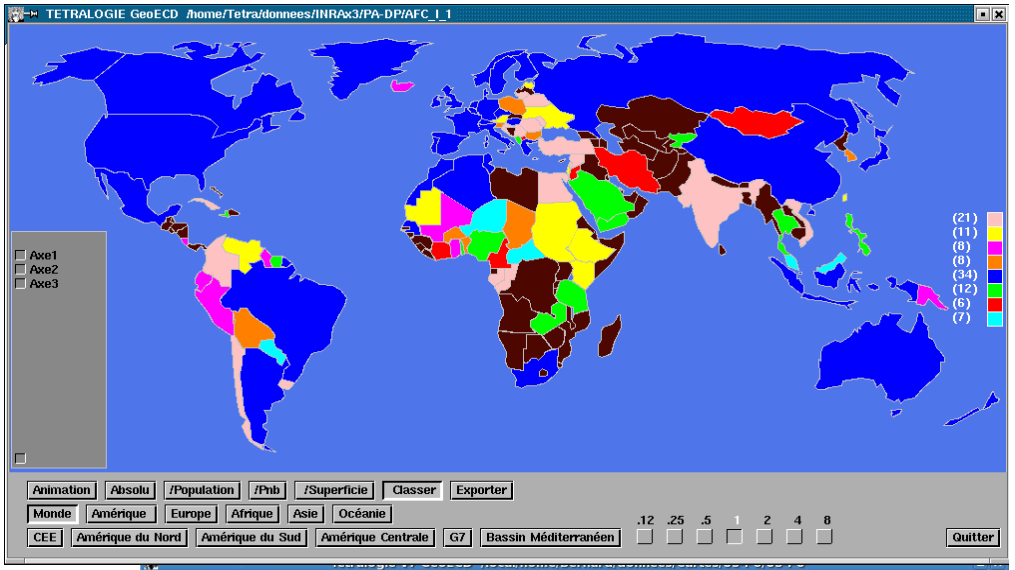



Figure 5 : Importation distante d'une classification dans une carte.

## 4.4 Autres cartes géographiques

Etant donné que certains pays sont très vastes et souvent divisés en plusieurs régions ou états d'une taille comparable aux pays européens, il est pertinent de diminuer, à leur niveau, la granularité afin d'affiner les résultats géostratégiques obtenus. Une difficulté apparaît immédiatement, la notion de région ou d'état (Californie, Nord d'Este, Sichuan) est le plus souvent absente des données sources que nous exploitons : adresses des auteurs de publications scientifiques, de déposants de brevets, d'entreprises, de laboratoires. Il peut même y avoir des confusions gênantes comme par exemple dans la base de brevets USPTO entre le sigle CA désignant tour à tour le Canada (Wancouver, CA) ou la Californie (Los Angeles, CA). Cette absence d'information ou ces ambiguïtés peuvent être levées par l'utilisation de tables de correspondances, qui viennent transformer les données avant l'analyse. Les corrections ou transformations apportées obéissent au même principe que les synonymes des orthographes des auteurs ou des mots clés. Voici des exemples de ressources lexicographiques fort utiles en data mining qui permettent d'adapter la granularité géographique en fonction des besoins.

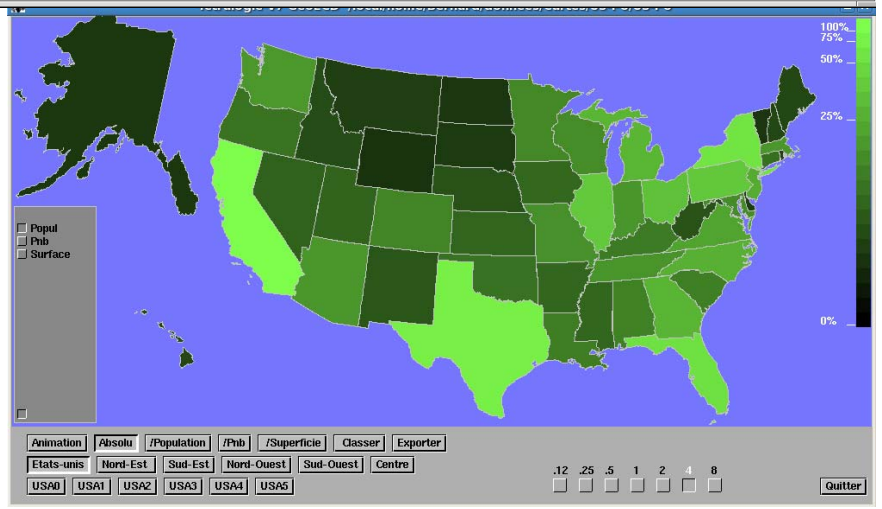



Figure 6 : Carte des USA avec une échelle de couleurs non linéaire.

**Pour retrouver le pays**:

| | |
|---|---|
| ALAMEDA, CA | USA |
| BALTIMORE, MD | USA |
| LONDON, GB | UK |
| MAYNARD, MA | USA |
| MINNEAPOLIS, MN | USA |
| ONTARIO, CA | CANADA |
| PASADENA, CA | USA |
| PORTLAND, OR | USA |
| QUEBEC, CA | CANADA |
| RICHMOND, CA | USA |
| ROCHESTER, MN | USA |

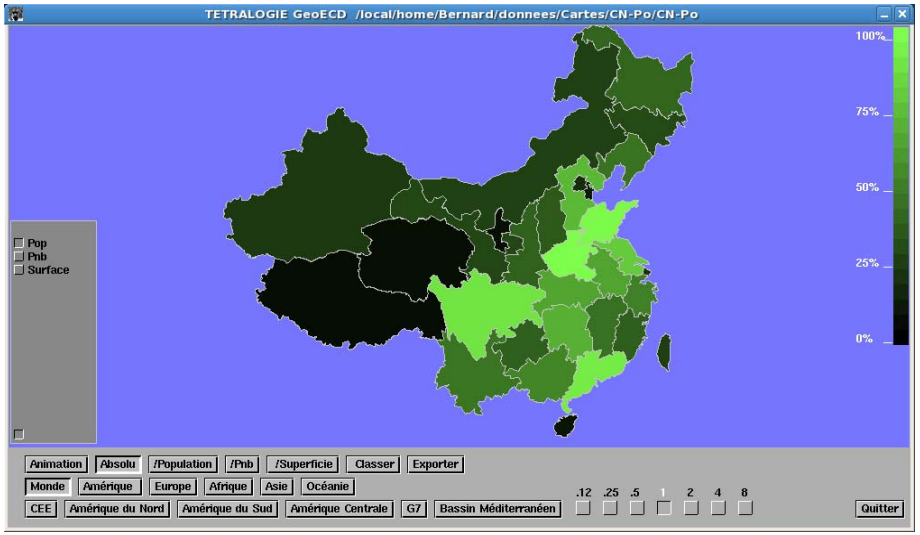
| | |
|---|---|
| SALT LAKE CITY, UT | USA |
| SHATIN, HK | HONG KONG |
| ST. LUCIA, AU | AUSTRALIA |
| STANFORD, CA | USA |
| TORONTO, CA | CANADA |
| UNIVERSITY PARK, PA | USA |

**Pour retrouver l'état ou la région** :

| | |
|---|---|
| ALAMEDA | CALIFORNIA |
| BALTIMORE | MARYLAND |
| BATON ROUGE | LOUISIANA |
| BOTHELL | WASHINGTON |
| COLUMBIA | MARYLAND |
| COLUMBUS | OHIO |
| EXTON | PENNSYLVANIA |
| FOSTER | CALIFORNIA |
| HERCULES | CALIFORNIA |
| KING OF PRUSSIA, PA | PENNSYLVANIA |
| LONDON | UK |
| MAYNARD | MARYLAND |
| MINNEAPOLIS | MINNESOTA |
| ONTARIO | CANADA |
| PASADENA | CALIFORNIA |
| PORTLAND | OREGON |
| QUEBEC | CANADA |

**Pour rétablir un format unique**:

| | |
|---|---|
| TENNESSEE | USA |
| TEXAS. | USA |
| TEXAS | USA |
| TN USA | USA |
| TX USA | USA |
| UNITED STATES OF AMERICA | USA |
| UNITED STATES | USA |
| UNITED STATES. | USA |
| UNITED-STATES | USA |
| USAA | USA |
| U.S.A. | USA |
| USA. | USA |
| UTAH. | USA |
| UTAH | USA |
| UT USA. | USA |
| UT USA | USA |
| VA USA. | USA |



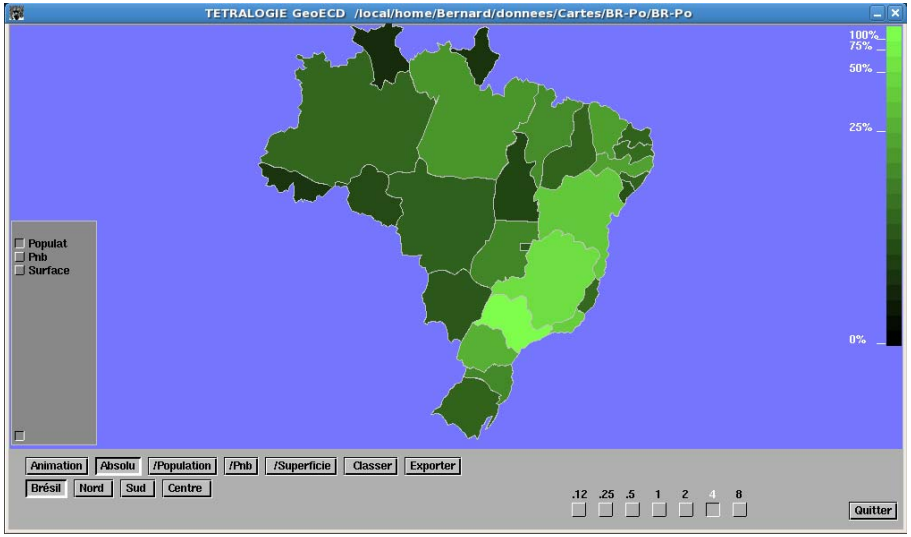




Figure 7 : Répartition de la population par régions pour la Chine (échelle linéaire) et le Brésil (échelle non linéaire)

# 5 Conclusion

L'ensemble d'outils que nous venons de présenter et leur philosophie d'emploi ont été testés dans plusieurs contextes et ce sont révélés particulièrement attractifs pour des non spécialistes de l'analyse de données. Citons:

- L'enseignement des méthodes d'analyse et de l'art de les utiliser,
- La veille scientifique et technique (publications scientifiques, brevets, sites web),
- La prise en compte des facteurs humains (analyse de dialogues, de simulations, de situations, de stress)
- Le dépouillement d'enquêtes et la recherche de profils,
- La bibliométrie et la scientométrie,
- L'intelligence économique en général.

Ils sont disponibles sur le serveur tetralogie.irit.fr afin d'être testés et ils sont régulièrement utilisés à distance pour les TP de nos IUP ou pour des analyses réalisées via le réseau. Nous envisageons d'étendre ces techniques de communication à d'autres types d'analyse et les rendre accessibles par la téléphonie mobile.

# 6 Références

[1] Bonnel N., Chevalier M., Dousset B., Hubert G., - Visualisation en recherche d'information -. Information & visualisation - enjeux, recherches et applications. Sophie Chauvin (Eds.), Cépaduès Editions, 7, p. 1-20, octobre 2008.

[2] Diday E., Celeux G., Govaret G., Lechevalier Y. et Ralambondrainy H. - Classification des données : Environnement statistique et informatique. -Edition Dunod 1989.

[3] El-Haddadi A., Dousset B., Berrada I. et Kassou I., - Construction d'une ontologie de domaine fondée sur le Text Mining -, 2ème séminaire de Veille Stratégique, Scientifique et Technologique VSST'2009, Nancy, 30 -31 mars 09.

[4] Gutwin C., Greenberg S. et Cockburn A. - Awareness through fisheye views in relaxed-wysiwis groupware. In : Proceeding of Graphics Interface, pp 28-38. - Toronto, 1996.

[5] Loubier E., Bahsoun W., Dousset B. - La prise en compte de la dimension temporelle dans la visualisation de données par morphing de graphe - Dans : Colloque Veille Stratégique Scientifique et Technologique (VSST 2007), Marrakech, 21-25/10/2007, IRIT, (support électronique), october 2007

[6] Remede J.R., Gomez L.M. et Landauer T.K. - SuperBook : An automatic tool for information exploration - hypertext. In Proceeding of the ACM Hypertext'87 Conference, pp 189-194. - Chapel Hill, 1987.

[7] Robertson J.R., Makinlay J.D et Card S.K. - Cone Trees : animated 3D visualisations of hierarchical information. In Proceeding of the ACM CHI'91 Conference on human Factors in computing Systems. pp 81-91. - New orleans, Louisiana, 1991.