avancées 28/07/08 10:37 Page 47

VISUALISATION POUR L'ANALYSE DE DONNÉES GÉOPHYSIQUES

M. AUPETIT
CEA - DAM - Île-de-France

Dans le cadre de la surveillance des traités internationaux (Traité d'interdiction complète des essais), *les équipes du CEA - DAM enregistrent en continu les signaux sismiques mesurés sur les stations du système de surveillance internationale. Elles analysent ces signaux, détectent les événements, et déterminent leur origine naturelle ou artificielle. Nous avons développé deux méthodes d'analyse de données, qui forment les bases d'un outil d'aide à la décision destiné aux analystes. La première est une méthode de visualisation basée sur la projection des données sur un plan. Elle fournit une visualisation immédiate des zones d'intérêt, et des zones présentant un risque d'erreur d'interprétation. La seconde est une méthode d'analyse exploratoire des données combinant statistique et topologie, évitant les distorsions associées aux méthodes de projection usuelles.*

Les événements sismiques étant caractérisés par leur localisation, leur magnitude, leur date, et d'autres paramètres issus des signaux mesurés, ils sont considérés comme des points dans un espace de dimension égale au nombre de ces caractéristiques. L'ensemble forme un nuage de points dans cet espace initial. Nous disposons aussi pour chaque évènement, de la classe d'origine fournie par l'analyste. Intuitivement, la probabilité que deux événements appartiennent à la même classe est d'autant plus forte qu'ils sont proches dans l'espace initial, et les méthodes de discrimination automatiques attribuent à un nouvel événement une probabilité d'appartenance à chaque classe. Mais, bien que les taux d'erreurs soient faibles [1], les analystes ne peuvent se satisfaire d'un score dont ils ne maîtrisent pas l'élaboration.

Il est primordial que l'analyste ait confiance en l'outil d'aide à la décision. Pour l'aider dans sa prise de décision, il est souhaitable qu'il appréhende lui-même la configuration géométrique des classes d'événements. Comment montrer à l'analyste à quoi ressemblent les classes dans l'espace, et comment s'y positionne un nouvel événement à classer ?

Deux méthodes sont proposées :

- par projection de ce nuage de points dans le plan, fournissant à l'analyste une vue directe mais déformée des données. Dans ce cas, l'analyste est mis en confiance parce qu'il utilise sa propre vision pour l'analyse. Ce qu'il voit pouvant l'induire en erreur, nous proposons une méthode de coloration pour juger de la fiabilité locale de la projection.
- par modélisation directement dans l'espace initial des relations de voisinage des classes les unes par rapport aux autres, et présentation d'une synthèse à l'analyste. Dans ce cas, l'analyste n'utilise plus sa vision directe, et il doit interpréter une représentation plus sommaire des données, mais la synthèse présentée est plus fiable car c'est le reflet direct de la structure du nuage initial.

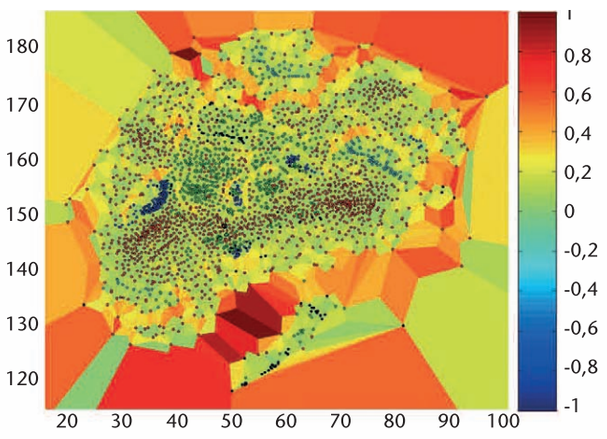

Figure 1
Analyse par projection.
Les données sismiques sont projetées dans le plan. La couleur des points représente la classe de l'événement *(bleu : effondrement minier ; rouge : séismes ; vert : tirs de carrière)*. La couleur de fond indique si la distance observée entre deux points projetés est similaire *(vert)* ou très différente *(rouge)* de la distance séparant ces deux points dans l'espace initial. Les zones vertes sont fiables, les groupes de points qui y sont observés existent aussi dans l'espace initial.

avancées 28/07/08 10:37 Page 48

Analyse par projection

Dans ce cas, le nuage de points initial est projeté dans le plan à deux dimensions, en tentant de préserver au mieux les distances entre toutes les paires de points. En privilégiant la préservation des petites distances aux dépens des grandes, nous obtenons une représentation qui tend à positionner au voisinage l'un de l'autre deux points effectivement voisins dans l'espace initial. De nombreuses méthodes de projection existent. L'analyse en composantes principales est la plus connue. Cependant, les techniques de projection présentent nécessairement des distorsions, car le nuage de point de grande dimension doit être "compressé" pour être représenté dans le plan. En conséquence, la fidélité de la projection aux données doit être diagnostiquée afin que cet outil soit crédible, et que son résultat soit interprétable visuellement par l'analyste.

Nous avons donc développé une méthode de coloration de l'arrière-plan *(figure 1)* [2] qui fait ressortir, en vert, les régions dans lesquelles les distances sont très bien préservées *(les points visiblement proches le sont effectivement dans l'espace initial)*, et en rouge celles où les distances ne sont pas du tout préservées. La couleur des points indique l'origine des événements *(bleu : effondrement minier ; rouge : séismes ; vert : tirs de carrière)*. La coloration de l'arrière plan est indispensable pour déterminer visuellement si une structure géométrique *(un groupe de points)* existe effectivement dans l'espace initial, ou n'est qu'un artefact de la projection, entraînant un risque d'erreur d'interprétation.

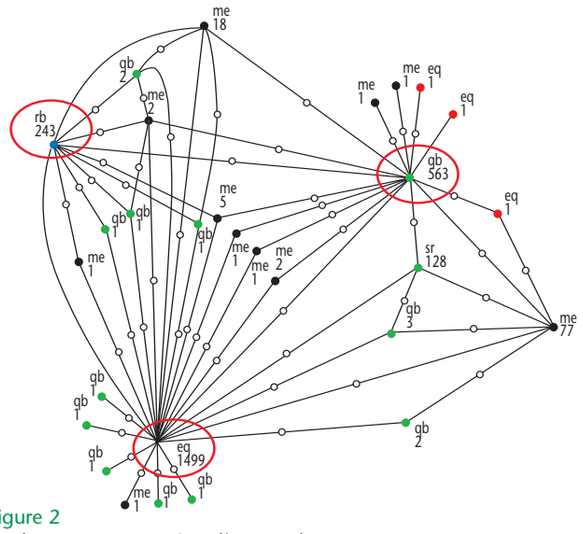

Figure 2
Analyse par construction d'un graphe.
Un graphe relie les composantes des différentes classes d'événements sismiques. Ce graphe est représenté dans le plan, et montre la topologie des classes telles qu'elles sont dans l'espace initial. Par exemple, la classe des "effondrements miniers" *(rb)* contient 243 événements, elle est d'un seul tenant dans l'espace initial ; la classe des "tirs de carrière" *(qb)* est morcelée en un groupe principal de 563 événements, et une quinzaine d'autres événements isolés ; enfin, 3 événements naturels *(eq)* sur 1502 sont isolés du groupe principal.

Analyse par construction d'un graphe

Les projections sont "parlantes" pour l'analyste mais potentiellement trompeuses, car elles déforment nécessairement le nuage de points initial. Aussi, puisque l'information recherchée concerne ce nuage, nous avons proposé une méthode d'analyse directe. Il s'agit de retrouver la topologie des classes, et de fournir à l'analyste une représentation synthétique et interprétable de cette information.

Nous avons proposé de construire un graphe basé sur des critères géométriques et statistiques [3], [4], dont la topologie reproduit celle des classes dans l'espace initial *(figure 2)*. Ce graphe nous indique comment les groupes de données de chaque classe sont connectés les uns aux autres dans l'espace initial. Ce graphe montre que :

- la classe des "effondrements miniers" *(rb)* est d'un seul tenant. Elle contient 243 événements tous au contact les uns des autres dans l'espace initial ;
- la classe des "tirs de carrière" *(qb)* est morcelée en un groupe principal de 563 événements, et en une quinzaine d'événements isolés de ce groupe principal par la présence de données des autres classes ;
- 3 événements naturels *(eq)* sur 1502 sont isolés du groupe principal.

Les caractéristiques utilisées sont pertinentes puisque chaque classe est bien regroupée. La classe et la taille des groupes, auxquels un nouvel événement de classe inconnue est connecté, permettent de conforter ou non l'analyste dans son choix de la classe à attribuer à cet événement.

Suite à ces recherches, nous poursuivons nos études dans le domaine de l'extraction et de l'exploitation des caractéristiques topologiques d'un nuage de points [5].

Références

[1] D. MERCIER, P. GAILLARD, M. AUPETIT, C. MAILLARD, R. QUACH, J.-D. MULLER, "How to help seismic analysts to verify the French seismic bulletin?", *Engineering Appl. of Art. Intel.*, **19**, p. 797-806, (2006).

[2] M. AUPETIT, "Visualizing distortions and recovering topology in continuous projection techniques", *Neurocomputing*, **70**, p. 1304-1330 (2007).

[3] M. AUPETIT, T. CATZ, "High-dimensional labeled data analysis with topology representing graphs", Neurocomputing, **63**, p. 139-169, (2005).

[4] P. GAILLARD, M. AUPETIT, G. GOVAERT, "Learning topology of a labeled data set with the Supervised Generative Graph", *Neurocomputing*, **71**, p. 1283-1299 (2008).

[5] http://www.dase.cea.fr/topology learning/index.html