Choix d'une méthode de classification de l'économie maritime

Valentin Daniel

Chaire maritime

Chaire de recherche sur les dynamiques des activités humaines en mer et la planification de l'espace maritime

Fiche thématique n°2

Choix d'une méthode de classification de l'économie maritime

Valentin Daniel

Nantes Université

Fondation Nantes Université

Préambule

Cette fiche thématique fait suite à la première fiche et a pour but de présenter à démarche et le choix technique de la méthode de classification de l'économie maritime. Elle sera complétée de la fiche 3 qui détaillera la théorie derrière l'application de la méthode retenue.

Ainsi la lecture des fiches 2 et 3 n'est pas obligatoire pour la compréhension de l'outil. Cependant la mise en ligne d'un tel document est importante dans un souci de transparence méthodologique, d'autant plus lorsque l'application directe des méthodes présentées ici fait appel à la big data et aux données socio-économiques.

Enfin, les travaux de la Chaire maritime ont pour objectif d'apporter un éclairage sur différentes méthodes alternatives. L'explication de la démarche scientifique et la description des méthodes proposées est essentielle pour s'assurer de la reproductibilité de ces expérimentations.

La nécessité de la mise en œuvre d'une classification avancée

SEGMENTATION SIMPLE SUR LA BASE DE CRITERES DISCRIMINANTS

La mise en œuvre d'une segmentation économique n'implique pas nécessairement l'utilisation d'outils statistiques très élaborés. Il est en effet possible de segmenter une population sur la base de paramètres simples, on parle alors de « point de coupure » de cette population (J. Creusier et F. Biétry, 2014). Le découpage peut alors s'effectuer sur la base de variables aussi bien quantitatives que qualitatives. On peut par exemple découper la population entre organismes publics et privés, ou bien selon les tailles d'entreprises (plus ou moins de 50 salariés, etc.). Il est bien entendu possible de mêler plusieurs points de coupures, en reprenant l'exemple précédent nous aurions alors 4 segments :

- Les organismes publics de plus 50 salariés
- Les organismes publics de moins de 50 salariés
- Les organismes privés de plus de 50 salariés
- Les organismes privés de moins de 50 salariés

Les points de coupures peuvent être choisis de différentes façon, soit grâce à une bonne connaissance du secteur, soit à l'aide de statistiques descriptives simple, comme par exemple des moyennes ou des quartiles.

Même si cette méthode peut s'avérer utile dans certains cas, elle reste trop imprécise la plupart du temps pour mettre en place de véritables

LIMITES DE LA SEGMENTATION SIMPLE

Plusieurs inconvénients apparaissent rapidement. Par exemple pour une échelle de 1 à 100 on pourrait imaginer une coupure à 50, dans ce cas un profil noté 1 serait dans le même groupe qu'un autre noté 49 et ce dernier serait donc plus éloigné du 51 que du 1. Cela pose donc un problème puisqu'un groupe peut se définir de la manière suivante : ensemble d'individus partageant des caractéristiques communes.

Pour pallier à cet inconvénient il est possible d'utiliser plus de points de coupure, ou d'autres variables afin de croiser les informations disponibles. Dans ce cas, c'est la lisibilité des profils qui se détériore. En effet, on obtient k groupes avec k=n^x où x est le nombre de variables et n le nombre de groupes pour ces variables. Imaginons 2 groupes par variable, pour 8 variables, nous obtenons ainsi 2^8=256, profils différents. (J. Creusier et F. Biétry, 2014) Il n'est pas possible pour les décideurs de prendre en compte autant de profils et de définir autant de stratégies différentes, cela rendrait l'analyse et le suivi trop complexe. Pour pallier à cela, il existe des méthodes statistiques plus élaborées empruntées à l'économétrie et à la science de la donnée.

Classification non-supervisée : revue des principales méthodes

DRESSER UNE TYPOLOGIE D'ENTREPRISE

Les techniques de classification non-supervisées sont généralement présentées comme étant des méthodes de fouille de données visant à faire ressortir des groupes sur la base des différentes variables observées (C. Robardet, 2002). Cela se prête donc assez bien à ce que l'on pourrait attendre de la classification d'une économie. Mais à la différence de la classification supervisée, les méthodes non-supervisées ne nécessitent pas de jeu de données pré-labellisées. C'est-à-dire qu'on ne suppose pas de groupes à priori, mais on explore les données afin de voir quels liens peuvent être trouvés entre les agents. Cela peut s'avérer très utile pour dresser une typologie des acteurs lorsque la connaissance de la structure de l'économie est assez réduite. De nombreuses familles de méthodes permettent de mettre en place une classification non-supervisée, elles seront présentées ici ainsi que leurs caractéristiques. Puis il sera débattu de leurs avantages et inconvénients dans le cadre de la mise en place d'une caractérisation uniforme de l'économie maritime.

DATA CLUSTERING

Le partitionnement de données, plus généralement appelé «data clustering» est l'une des diverses techniques d'analyse de données, et fait l'objet de nombreux travaux en machine learning (M. Dziechciarz-Duda, 2007). Le clustering est utilisé afin de séparer un jeu de données en plusieurs groupes homogènes, ces groupes sont appelés « clusters ». Dans le cadre d'une segmentation de l'économie maritime, ce sont ces clusters qui représenteront nos segments d'entreprises. En effet les individus (au sens statistique) d'un même groupe partageront théoriquement des caractéristiques identiques ou proches, et il sera alors possible de définir un ou plusieurs profils types pour chaque segment. Les clusters sont généralement définis à partir de critères de distances entre les individus ou les classes. La distance est ici calculée sur la base de l'ensemble des variables relatives au observations (ici des entreprises maritimes). Au sein d'un cluster on cherchera à minimiser la distance entre les individus. A l'inverse, entre les groupes on cherchera à maximiser cette même distance.

Il existe de nombreux algorithmes de partitionnement de données, nous présenterons ici les pistes ayant été explorées par la Chaire.

Classification non-supervisée : revue des principales méthodes

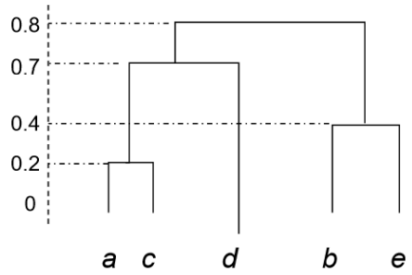
K-MEANS

Cet algorithmes appartient à la famille des méthodes de partitionnement, il fournit une partition unique des éléments à classer. Il nécessite cependant de spécifier au préalable le nombre de classes que l'on veut obtenir. Ce dernier point nécessite donc de poser des hypothèses sur la structure de l'économie à segmenter. L'objectif des k-means va être de minimiser l'inertie à l'intérieur des classes. On peut en schématiser le fonctionnement de la manière suivante :

- On choisit au hasard plusieurs individus, autant qu'il y a de classe. Ils seront les représentants initiaux de nos classes, et donc dans un premier temps leurs centres respectifs.
- On introduit ensuite l'ensemble des individus restant en les associants au centre de classe le plus proche
- Les centres des classes sont ensuite recalculés, on les appelle les centres de gravité des classes. Suite à se recalcul on regarde si des individus doivent changer de groupe en étant rattaché à un centre devenu plus proche.
- On répète la dernière étape jusqu'à obtenir nos classes finales. Le partitionnement par les k-means ne donnera donc pas toujours les mêmes résultats pour un même jeu de données.

CLASSIFICATION HIÉRARCHIQUE

La classification hiérarchique se base sur les mesures de ressemblance entre les individus. A partir de cela, elle procède à une partition emboitée de manière hiérarchique, sous forme de dendrogramme :



Pour se représenter de manière plus concrète ce que peut donner ce genre de classification, on peut regarder du côté de la biologie avec le système d'espèces et de sous-espèces. La classification y est bien hiérarchiquement emboîtée, d'un côté on aurait par exemple les mammifères, qui se rediviseraient en deux avec les mammifères marins et les terrestres et ainsi de suite. Il existe deux grandes familles de classification hiérarchique :

- La classification ascendante hiérarchique : On construit les classes en partant des individus, en regroupant entre eux les plus ressemblants.
- La classification descendante hiérarchique : On part de la population globale pour ensuite scinder les individus les plus opposés.

Les méthodes vues précédemment peuvent être utilisées de manière complémentaire grâce à la Classification Hiérarchique sur Composantes Principales.

Classification non-supervisée : revue des principales méthodes

CLASSIFICATION HIERARCHIQUE SUR COMPOSANTES PRINCIPALES (HCPC)

Cette technique permet d'allier les principales méthodes d'analyses de données (F. Husson, J. Josse et J. Pagès, 2010). Trois éléments vont être utilisés dans ce cas, la famille des analyses en composantes principales et multiples (ACP, ACM, etc.), la classification hiérarchique et les k-means. La mise en œuvre s'effectue de la manière suivante :

- Création d'une analyse en composantes principales ou multiples en fonction du jeu de données.
  Ces analyses sont très utiles pour synthétiser de l'information, notamment en présence d'un grand nombre de variables
- Sélection du nombre de dimensions à retenir de la même manière que si l'on faisait une analyse en composante seule.
- Effectuer une classification hiérarchique ascendante sur les résultats précédents.
- Améliorer le résultat de la classification en appliquant un partitionnement en k-means.

La mise en place d'une telle méthode est utile dans plusieurs cas de figures, qui correspondent aux cas d'usages des analyses factorielles (A. Kassambara, 2017) :

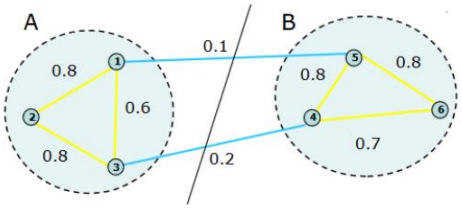
- Lorsque l'on dispose d'un grand nombre de variables quantitatives continues, l'ACP peut permettre d'effectuer une réduction de dimension des données en quelques variables latentes. La classification peut ensuite être réalisée pour mieux comprendre les liens entre les individus et analyser les différents groupes. Dans ce cas, l'analyse en composantes principales permet de réduire le bruit dans les données et donc de mener à une classification plus fiable.
- Pour classifier des données qualitatives, passer par une ACM au préalable s'avère être une solution efficace. En effet, l'ACM va permettre de projeter l'ensemble des variables qualitatives sur les axes principaux. Puis comme c'était le cas pour l'ACP, une classification peut être appliquée sur les coordonnées des points dans chaque dimension.
- Enfin, quand le jeu de données présente des données à la fois quantitatives et qualitatives, le mieux est de passer par les analyses factorielles multiples ou bien par les analyses factorielles de données mixtes, puis d'effectuer la classification dans un second temps.

Classification non-supervisée : revue des principales méthodes

SPECTRAL CLUSTERING

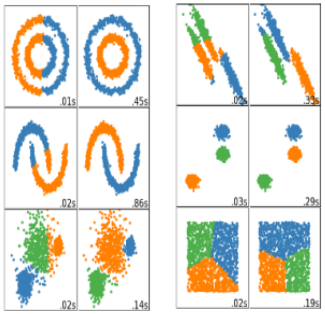
Cette technique repose sur la structure d'une matrice de similarité pour partitionner des points en grappes avec des individus d'un même groupe fortement similaires et ceux de différents groupes faiblement similaires (F. Bach et M. Jordan).

Lors d'une analyse par clustering spectral, les données sont traitées à l'aide de graphes, puis le graphe est sectionné pour former les groupes.



Source : Introduction to spectral clustering

Le clustering spectral permet, grâce à cette technique, d'être très efficace sur les jeux de données dits non conventionnels. La figure suivante compare à droite les résultats obtenus par les k-means et à gauche ceux obtenus grâce au clustering spectral :

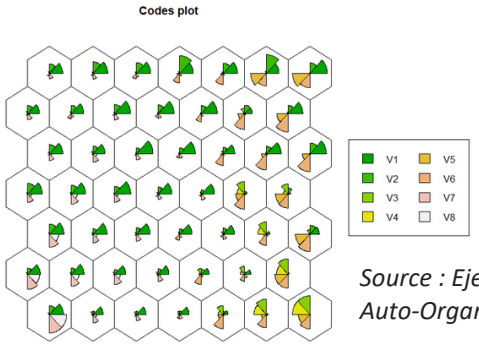


Source : Python – Scikit learn

On remarque ici que pour un jeu de données « standard » (4ème jeu de données), les performances des deux algorithmes sont similaires. La différence est cependant très marquée pour les 1er, 2eme et 4eme cas. Cependant, de telles formes dans les données sont assez rares dans le domaine de la segmentation d'une économie. De plus, certains agents d'un même cluster peuvent paraitre assez éloignés et il n'est donc pas forcément optimal de leur appliquer les mêmes décisions politiques.

CARTES AUTO-ORGANISATRICES DE KOHONEN

Les cartes auto-organisatrices de Kohonen (Self-Organizing Map, en anglais, communément appelée « SOM ») sont un type d'algorithme utilisé pour l'apprentissage non-supervisé appartenant à la famille des réseaux de neurones. Introduites par Teuvo Kohonen en 1981-1982 (T. Kohonen, 1982), ces cartes s'inspirent du fonctionnement d'un cerveau humain. En effet, celui-ci est découpé en plusieurs zones qui reprennent les mêmes caractéristiques que les capteurs sensoriels qui leurs sont associés. Dans le cortex visuel par exemple, deux zones proches vont correspondre à deux zones également proches sur la rétine. Cette topologie des zones dans le cerveau n'est pas innée mais est due à l'apprentissage. C'est ce processus d'apprentissage que vont essayer de reproduire les cartes de Kohonen. Deux groupes proches sur une carte vont donc être relativement similaires ; la notion de voisinage est ici très importante. Deux agents projetés dans les mêmes groupes vont donc être jugés comme ayant des caractéristiques communes. L'aspect très visuel de la méthode la rend assez simple à interpréter grâce aux vecteurs de poids (*c.f fiche #3*). Le graphique ci-dessous dit de « codebooks », présente par exemple le poids de chaque variable dans la caractérisation des groupes. On y observe que les groupes du coin inférieur droit sont très fortement caractérisés par la variable 4 (en jaune) mais pas du tout concernés par les variables 1, 2, 7 et 8. Ainsi, en plus d'obtenir une classification, l'algorithme SOM permet d'avoir une compréhension globale de la structure du phénomène observé.



Source : Ejemplo de uso de un Mapa Auto-Organizado (SOM) de Kohonen en R

Il est également intéressant de signaler que l'application des cartes de Kohonen ne se limite plus à l'apprentissage non-supervisé, en effet les algorithmes XYF et BDK ont été développés pour permettre d'effectuer une classification supervisée sur la base des réseaux de Kohonen (W. Melssen, R. Wehrens et L. Buydens, 2006). En d'autres termes une fois qu'une typologie est mise en place suite à l'analyse de la structure d'une économie à partir d'un algorithme « exploratoire », on peut ensuite apprendre aux algorithmes XYF et BDK à reproduire cette typologie à travers le temps ou bien sur d'autres données.

Résultat : méthode retenue

L'algorithme SOM : la réponse aux différents besoins

Cette méthode a été retenue pour constituer la base de l'algorithme de caractérisation de l'économie maritime, et sera donc plus détaillée dans la fiche n°3.

Afin de pouvoir mener à bien cette segmentation dans le cadre des travaux de la Chaire, plusieurs impératifs ont été définis, la solution à ces impératifs est détaillée pour chaque point :

- Assurer la reproductibilité du modèle : *Une classification devra être effectuée à intervalle régulier afin de suivre l'évolution de la structure de l'économie. De plus le modèle doit être déclinable sur un ensemble de zones géographiques et de filières économiques.*

Une fois que le modèle a appris, il est possible de l'enregistrer. Chaque nouvelle observation sera alors classifiée sur la base de ce que le modèle a appris lors du processus d'apprentissage. On pourra donc à intervalle régulier entrer un nouveau jeu de données et le modèle sortira toutes les classifications correspondantes. De plus, la représentation donnée par cette méthode nous permet d'appliquer des filtres sur les résultats et donc de décliner ces derniers en fonctions des zones géographiques, des filières ou encore d'appliquer des croisements de manière très rapide sans avoir à relancer un nouveau processus d'apprentissage.

- Garantir la stabilité : *Deux entreprises ayant les mêmes caractéristiques socio-économiques devront être classées de manière similaire. Deux entreprises similaires à deux instants différents dans le temps, toutes choses égales par ailleurs, devront aussi mener à une classification identique. Pour cette raison les k-means ont été exclus.*

Comme évoqué précédemment, le modèle peut être enregistré. Lorsqu'on rejoue l'analyse, le modèle se base sur ce qu'il a appris. Ainsi pour un même jeu de données il ressortira le même résultat.

- Le modèle doit être interprétable et analysable : *Cela permettra de comprendre au mieux la structure des choses et d'appliquer les règles de décisions adéquates. De plus, les résultats du modèle doivent être compréhensibles par tous pour garantir la possibilité d'appropriation de la méthode.*

Les sorties données par les cartes de Kohonen sont très visuelles et simples à expliquer. De plus, les « codebooks » permettent d'analyser la composition des nœuds, ce qui permettra de détailler la structure du secteur étudié.