STA211 : Exploitation des méthodes exploratoires pour le pré-traitement des données

V. Audigier, N. Niang

04 février, 2025

1 Introduction

Nous avons précédemment présenté différentes méthodes de statistique exploratoire de base (univariée, bivariée, multivariée) très utiles en data-mining pour visualiser et comprendre les données. En effet, la modélisation des données nécessite de s'assurer de l'absence d'anomalie, de mesurer les liaisons entre les variables et plus généralement, d'examiner tous les points susceptibles de devoir être pris en compte pour l'analyse. Dans ce document, nous mettons en avant comment ces méthodes peuvent aussi être utilisées pour le pré-traitement des données. En particulier, elles permettront d'effectuer des transformations des variables adaptées vis-à-vis de la structure des données, mais également de gérer un nombre important de lignes ou de colonnes dans un tableau de données.

2 Transformation des variables

La modélisation des variables impose souvent certaines contraintes sur les données (normalité, linéarité, nature,...). En fonction de l'objectif et des méthodes envisagées, il pourra donc être utile de transformer certaines des variables. Nous distinguons par la suite le cas des variables quantitatives de celui des variables qualitatives.

2.1 Variables quantitatives

2.1.1 Discrétisation

Une première transformation qui peut être apportée sur une variable quantitative est sa discrétisation de façon à la rendre qualitative (ordonnée). Un premier intérêt d'apporter cette transformation est de disposer de variables toutes qualitatives alors que les données initiales étaient de nature mixte. Ceci peut être pertinent si on envisage d'utiliser une méthode d'analyse dédiée uniquement aux variables qualitatives (ACM par exemple).

Une deuxième raison est qu'il est généralement plus facile d'interpréter un modèle portant sur les modalités des variables qualitatives, que sur des variables continues. C'est par exemple le cas en analyse de la variance (variables explicatives qualitatives) où le sens d'un coefficient du modèle sera plus simple que celui d'un modèle de régression linéaire (variables explicatives quantitatives). Un autre exemple où il peut être pertinent de discrétiser une variable est lors du calcul d'un score : il est généralement plus simple de présenter le score sur une échelle discrète (e.g. bas, moyen fort), plutôt que d'utiliser une échelle continue.

Une troisième raison est que cela permet de gérer aisément les données manquantes. En effet, une fois la variable rendue qualitative, il est possible de considérer une modalité supplémentaire codant pour les individus n'ayant pas répondu. De cette façon, il devient possible d'appliquer directement les méthodes de data-mining en dépit des données manquantes (pourvu qu'elles puissent s'appliquer sur des données qualitatives).

Enfin, un autre intérêt de ce type de transformation est de gérer la non-linéarité dans les données. Nous revenons sur ce point en Section 2.1.2.

Il existe plusieurs façons de discrétiser une variable continue. Une première façon est de faire une discrétisation ``métier'', i.e. une discrétisation a priori, couramment employée pour l'analyse qu'on en fait ensuite. Celle-ci dépend naturellement des variables qui sont discrétisées et du contexte dans lequel elles seront analysées. Par exemple, la variable âge sera probablement discrétisée différemment si l'on s'intéresse à un diagnostic du cancer du sein chez des patientes ou à l'accord d'un prêt bancaire auprès d'un client. Ce type de discrétisation présente un intérêt pour l'interprétation.

Une autre façon est de faire un découpage selon les quantiles empiriques. C'est par exemple ce qui est fait quand on construit un histogramme à pas variable (les classes considérées définissant la discrétisation de la variable continue). Ce type de découpage a l'avantage d'éviter d'obtenir des classes d'effectif trop faible, problématique pour de nombreuses méthodes comme la régression logistique, les SVM, voire réseaux de neurones (Tufféry (2015)), ou l'ACM.

Une troisième méthode de discrétisation consiste à effectuer un découpage ``naturel'' déformant le moins possible la distribution de la variable (du point de vue uni-dimensionnel). Pour cela, on peut utiliser un algorithme de clustering (k-means ou CAH par exemple) qui assurera des classes homogènes. Par ailleurs, ces stratégies fourniront un choix du nombre de classes (en particulier la CAH).

Il existe d'autres méthodes de discrétisation, plus sophistiquées, qui prendront mieux en compte l'objectif final de l'analyse effectuée (par exemple la prédiction d'une variable). On peut par exemple citer les arbres binaires, consistant à partitionner l'espace afin de constituer des régions homogènes à l'intérieur desquelles les individus sont homogènes vis-à-vis de la variable réponse. L'algorithme MDLPC[1] de Fayyad and Irani (1993) est la méthode de référence utilisant ce type d'approches (disponible dans le package R *discretization* (Kim 2012)). Nous aborderons ces méthodes ultérieurement, sans insister nécessairement sur leur utilisation pour du pré-traitement. Certains logiciels de data-mining proposent aussi des méthodes automatiques de discrétisation. Leur principe est de considérer un très grand nombre de découpages et de choisir parmi ceux-ci, celui qui optimise un critère de performance vis-à-vis de la variable réponse (par exemple minimiser un taux d'erreur de mauvais classement dans le cas d'une variable réponse qualitative).

On pourra compléter le sujet de la discrétisation en consultant Rakotomalala (2010).

2.1.2 Linéarité

Il est aussi fréquent de transformer les variables de façon à se ramener à des liaisons de nature linéaire. En effet, les modèles les plus facilement interprétables supposent généralement une relation linéaire entre les variables. C'est le cas par exemple du modèle de régression linéaire, de la régression logistique, de l'analyse discriminante : le modèle linéaire suppose que la moyenne d'une variable réponse est combinaison linéaire des variables explicatives ; la régression logistique suppose que le logit de la moyenne de la variable réponse (binaire) est combinaison linéaire des variables explicatives ; l'analyse discriminante suppose quant à elle que conditionnellement aux modalités d'une variable binaire, les variables quantitatives sont distribuées selon une loi gaussienne multivariée, dont la particularité est que la moyenne de chaque variable est combinaison linéaire des autres. Dès lors, l'emploi de ces modèles nécessite d'assurer une relation linéaire entre les variables.

La linéarité entre deux variables sera généralement appréciée via la visualisation du nuage de points ou à la suite de la modélisation via des outils de diagnostic spécifiques à la méthode. Quand le nombre de variables est grand, ceci devient fastidieux, on préférera alors comparer des indicateurs de liaisons supposant la linéarité, comme le coefficient de corrélation linéaire, et d'autres ne la supposant pas, comme le coefficient de Spearman. Un écart important pourra suggérer une non-linéarité. Néanmoins, apprécier l'écart n'est pas toujours simple. Une approche statistique comparant le coefficient de corrélation linéaire (au carré) et le rapport de corrélation est proposée dans Aïvazian (1978), on pourra en trouver un résumé en ligne dans Rakotomalala (2009). Cette comparaison nécessite que l'une des deux variables soit discrétisée car le rapport de corrélation évalue la liaison entre une variable qualitative et une variable quantitative. Dans le cas où une des deux variables est quantitative discrète, on pourra considérer les classes définies par chaque valeur de cette variable, dans le cas contraire, on pourra utiliser une discrétisation classique (cf Section 2.1.1).

Quand une non-linéarité est détectée, on pourra linéariser en appliquant une transformation plus où moins complexe sur les colonnes. L'amélioration de la linéarité pourra être vérifiée a posteriori en visualisant le nuage ou en analysant le modèle construit. Pour cela, on pourra procéder en intégrant dans les modèles les variables transformées et non-transformées, et voir si les variables transformées améliorent significativement l'ajustement ou non (via des tests locaux sur ces nouvelles variables). Ces transformations ne sont généralement pas simples à déterminer sans connaissance métier, mais la visualisation des données pourra orienter ces choix (voir Rakotomalala (2009)).

Néanmoins, l'intérêt des modèles linéaires repose essentiellement sur leur interprétation, or ces transformations amèneront naturellement à complexifier l'interprétation de la liaison entre les variables. Une autre alternative consiste à considérer simplement une discrétisation d'une des variables.

Prenons comme exemple les variables *Age* et *Creditability*. On cherche à connaître la nature du lien entre l'âge des clients et la proportion de bons payeurs. En raison des faibles fréquences de chacune des valeurs de la variable *Age*, on effectue une discrétisation en 20 classes selon les quantiles. Cette évolution est alors représentée en Figure 2.1.

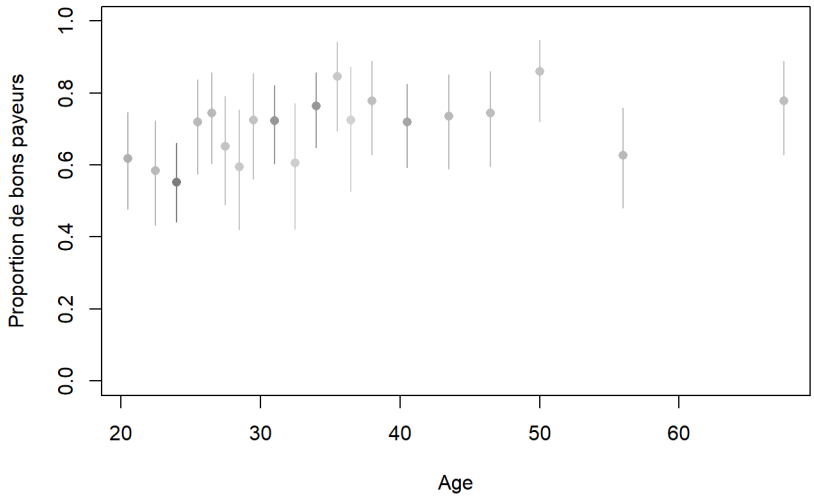

Figure 2.1: Proportion de bons payeurs en fonction de leur classe d'âge. Chaque proportion au sein d'une classe est représentée par un point selon une échelle de gris indiquant le nombre d'individus sur lequel cette fréquence est calculée. L'intervalle de confiance associé est représenté par un segment.

Il apparaît que le lien entre la proportion de bons payeurs et l'âge n'est pas linéaire. On observe en effet un plateau entre 20 et 24 ans, puis une croissance de cette proportion jusqu'à 35 ans, puis une stabilisation au-delà. Une discrétisation de la variable *Age* en 3 classes ($]0; 24],]24; 35],]35; 100]$) permettrait d'améliorer la linéarité (cf Figure 2.2).

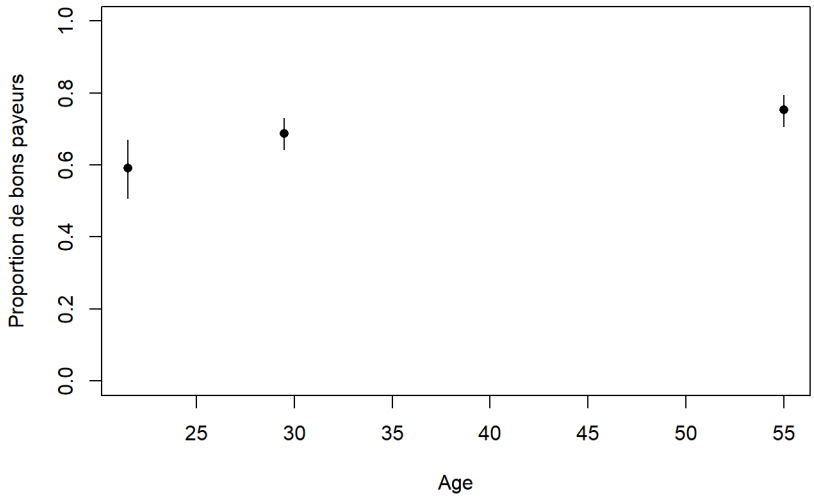

Figure 2.2: Proportion de bons payeurs en fonction de leur classe d'âge. Chaque proportion au sein d'une classe est représentée par un point selon une échelle de gris indiquant le nombre d'individus sur lequel cette fréquence est calculée. L'intervalle de confiance associé est représenté par un segment.

2.1.3 Normalité

La normalité d'une distribution est une hypothèse très classique quand on souhaite modéliser des variables quantitatives. C'est par exemple le cas de l'analyse discriminante, mais aussi en classification non-supervisée (en particulier parce qu'une distribution gaussienne amène à des classes sphériques).

La détection de la non-normalité d'une distribution peut être effectuée visuellement, à l'aide de diagramme qq-plot, consistant à comparer les quantiles empiriques de la variable à ceux d'une loi normale, ou à l'aide d'un simple histogramme, ou d'une courbe de densité, mais à nouveau, ce type d'investigation devient vite

fastidieux quand le nombre de variables est grand. Dans ce cas, on pourra analyser des indicateurs de formes (coefficients d'aplatissement et d'asymétrie pour lesquels des tests statistiques de nullité existent (Rakotomalala (2011)).

Généralement, si l'asymétrie positive, alors on se ramène à la normalité en utilisant une transformation par le log népérien (on pourra gérer les valeurs nulles en ajoutant 1 à la série) ou la racine carrée. Dans le cas inverse on pourra utiliser les puissances 2 ou 3. Dans le cas d'une proportion, on envisage généralement une transformation par la fonction $x \mapsto arcsin(\sqrt{x})$ (Tufféry (2007)). Néanmoins, ces transformations ne donnent pas toujours satisfaction, et il est difficile d'identifier la transformation optimale sans une méthodologie rigoureuse. Box et Cox ont ainsi proposé des transformations plus génériques qui sont paramétrables. La plus simple d'entre elles est

$$x \mapsto f^{\lambda}(x) = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & \text{si } \lambda \neq 0 \\ ln(x) & \text{sinon} \end{cases}$$

En définissant correctement le paramètre λ, il est alors possible de rapprocher la distribution de la normalité. Pour le choisir, une possibilité est d'évaluer, en fonction du paramétrage, l'évolution du coefficient de corrélation linéaire entre les quantiles empiriques de la distribution transformée et ceux d'une loi normale centrée-réduite[2] (voir Rakotomalala (2011)). On retiendra alors le paramètre λ rendant le coefficient de corrélation le plus grand possible.

Naturellement, il conviendra de vérifier l'apport de la transformation effectuée à l'aide de graphique, ou d'indicateurs de forme.

2.2 Variables qualitatives

2.2.1 Regroupement de modalités

Une première transformation qui peut être apportée sur une variable qualitative est le regroupement de modalités. Il s'agit d'affecter une même modalité à des individus prenant des modalités différentes sur une même variable. Ceci peut s'avérer très utile en présence de petits effectifs pour différentes méthodes.

Dans le cas particulier d'une variable qualitative ordonnée, on regroupera des modalités proches de façon à préserver l'ordre. Par exemple, la variable *Job* possède 4 modalités, dont la première *UnemployedUnskilled* dont la fréquence relative est de 2.2% (cf Table 2.1). On pourra alors la regrouper avec la seconde *UnskilledResident*, en affectant la modalité "unskilled" aux individus prenant les modalités *UnemployedUnskilled* ou *UnskilledResident* (cf Table 2.2).

Table 2.1: Effectifs des modalités de la variable Job.

UnemployedUnskilled	UnskilledResident	SkilledEmployee	Management.SelfEmp.HighlyQualified
22	200	630	148

Table 2.2: Effectifs des modalités de la variable Job après regroupement des deux premières modalités.

Unskilled	SkilledEmployee	Management.SelfEmp.HighlyQualified
222	630	148

Toutefois, quand la méthode d'analyse que l'on veut mettre en oeuvre vise à expliquer ou prédire une variable réponse, il est important de préserver au mieux la liaison entre la variable qualitative (explicative) et la variable réponse. En effet, il serait préjudiciable de fusionner des modalités si la proportion de bons payeurs est très différente selon ces deux dernières. La table 2.3 montre que cette proportion est similaire selon les modalités de la variable *Job*. De ce point de vue, il n'y a donc pas d'objection à fusionner les deux premières modalités de cette variable.

Table 2.3: Proportion de bon et mauvais payeurs en fonction des modalités de la variable Job

	good	bad
UnemployedUnskilled	0.68	0.32
UnskilledResident	0.72	0.28
SkilledEmployee	0.70	0.30
Management.SelfEmp.HighlyQualified	0.66	0.34

2.2.2 Analyse factorielle

Comme la discrétisation des variables quantitatives permet de mettre en oeuvre des méthodes réservées à des données qualitatives, il est parfois nécessaire d'effectuer des transformations des variables qualitatives pour y appliquer des méthodes dédiées aux variables quantitatives. Ceci n'est cependant pas trivial (il n'y aurait pas de sens à simplement renommer les modalités par des nombres !). La façon la plus simple de procéder consiste à appliquer une ACM sur ces variables, puis d'effectuer l'analyse à partir des composantes principales plutôt que des variables d'origine. Chaque composante étant de nature quantitative (coordonnées des individus sur un axe), on obtient bien un recodage des variables qualitatives en variables quantitatives. Retenir l'intégralité des composantes permettra de travailler sur la même information que celle contenue dans les données initiales (les intérêts à utiliser un nombre réduit de composantes est développé en Section 3.2.1). Notons qu'en général, le nombre de composantes excédera le nombre de variables qualitatives ce qui conduira à avoir un jeu de données comportant davantage de colonnes que le tableau initial.

3 Réduction des données

La quantité d'information n'en faisant pas la qualité, un autre type de pré-traitement souvent nécessaire consiste à simplifier le jeu de données en diminuant son nombre de lignes ou son nombre de colonnes. Les intérêts en sont multiples :

- Débruiter les données, en éliminant une partie de l'information non pertinente, i.e. sans lien avec la structure des données
- Faciliter l'interprétation, en réduisant le nombre de variables explicatives dans un modèle, ou en allégeant des représentations graphiques
- Réduire les temps de calcul, ou faciliter le stockage en mémoire du jeu de données

3.1 En lignes

La réduction en ligne d'un jeu de données permettra essentiellement de gérer des problèmes computationnels relatifs au stockage en mémoire ou au temps de calcul nécessaire pour mettre en oeuvre des méthodes, mais également d'améliorer l'ajustement des modèles en considérant des sous-populations

homogènes. A nouveau, cette réduction du nombre de lignes pourra s'effectuer de façon supervisée ou non-supervisée, en fonction de la méthode envisagée pour analyser les données.

3.1.1 Echantillonnage

Une première façon de réduire le nombre de lignes est d'échantillonner les individus. On distingue en particulier :

- l'échantillonnage simple : en effectuant un tirage aléatoire sans remise des individus du jeu de données initial.
- l'échantillonnage systématique : en tirant des individus de façon régulière (en prenant par exemple le 1er, le 101ème, le 20ème, etc.)
- l'échantillonnage stratifié : en définissant une partition de la population selon une variable qualitative, puis en tirant au sein de chaque ensemble, appelé *strate*, une certaine proportion d'individus. Cette proportion pourra correspondre au rapport du nombre d'individus dans la strate sur le nombre total d'individus (on parle d'échantillonnage *proportionnel*) ou non (on parle d'échantillonnage *non-proportionnel*). L'échantillonnage non-proportionnel peut permettre de gérer l'hétérogénéité des classes en tirant davantage d'individus dans des classes hétérogènes (nécessitant un plus grand nombre d'individus pour stabiliser les résultats) que dans des classes homogènes.

Dans une approche supervisée, l'échantillonnage stratifié selon la variable cible permettra de contrôler la distribution de celle-ci. En particulier, dans le cas où la variable à prédire est qualitative avec une modalité rare, l'échantillonnage non-proportionnel sera précieux pour augmenter la fréquence de cette modalité, ce sans quoi il pourrait être bien difficile d'ajuster un modèle.

NB : Au-delà de la réduction du nombre de lignes, l'échantillonnage sera également une étape primordiale pour éviter les problèmes de sur-apprentissage, développés dans un prochain cours.

3.1.2 Classification non-supervisée

Il est également possible de réduire le nombre de lignes à partir des méthodes de classification non-supervisée. Plutôt que d'effectuer un tirage aléatoire des individus, il s'agit d'analyser séparément les sous-populations homogènes d'individus de façon exhaustive. Pour mettre en oeuvre une méthode supervisée, cette approche diminuera la variabilité sur les variables explicatives et pourra améliorer ainsi la stabilité des résultats. Néanmoins, la partition ne sera généralement pas optimale pour l'analyse dans le sens où les classes sont constituées indépendamment de l'objectif final de prédiction ou d'explication. Il est possible de palier à ce problème en utilisant d'autres méthodes de partitionnement avancées (méthodes *clusterwise* notamment, cf Saporta (2017)).

3.2 En colonnes

Quand on souhaite prédire une variable, et que le nombre de variables explicatives est grand, les modèles sont souvent peu performants car il y a trop d'instabilité dans l'estimation de leurs coefficients (c'est par exemple le cas dans les modèles de régression logistique). Ceci se produit également quand les variables explicatives sont très fortement corrélées. Différentes stratégies peuvent être employées pour limiter ce phénomène.

3.2.1 Analyse factorielle

Une façon efficace de réduire le nombre de variables consiste à appliquer une analyse factorielle sur les données, puis à effectuer l'analyse sur les premières composantes seulement. De cette façon, on réduit le nombre de variables tout en perdant le moins possible d'information. En particulier, cette technique pourra même être bénéfique dans le sens où elle permettra de débruiter les données en isolant la variabilité portée par les dernières dimensions et qui n'est généralement pas reliée à la structure des données.

Il est très fréquent d'appliquer ce type de transformation avant d'effectuer une classification non-supervisée : éliminer de la classification les dernières dimensions dont on est sûr qu'elles ne représentent que du ``bruit'' est censé permettre d'obtenir une classification plus stable et plus claire (Husson, Lê, and Pagès (2009)). On choisira généralement un nombre de dimensions suffisamment grand pour garder une part importante de l'inertie du nuage initial (disons environ 80% pour fixer les idées). Néanmoins cette approche est imparfaite car elle dissocie la construction des composantes et la classification. Il n'est donc pas certain que les premières composantes retenues soient les plus pertinentes pour la classification. D'autres approches ont ainsi été proposées pour effectuer les deux approches simultanément, citons par exemple les méthodes *Factorial k-means* (Vichi and Kiers (2001)) et *Reduced k-means* (De Soete and Carroll (1994)).

Dans une optique supervisée, cette approche est très classique pour mettre ajuster un modèle de régression. On parlera alors de *régression sur composantes*. Naturellement, il faudra construire les composantes à partir des variables explicatives uniquement, sans utiliser la variable réponse (à moins qu'elle soit considérée comme variable supplémentaire). L'intérêt de cette approche est double : d'une part, elle permet de diminuer le nombre de variables explicatives, mais en plus, les variables explicatives deviennent orthogonales ce qui résout les problèmes de colinéarité. Notons qu'il sera possible de réexprimer le modèle à partir des variables initiales (pourvu que le modèle soit linéaire) et, dans ce cas, toutes les variables seront considérées, ce qui n'en facilitera pas l'interprétation. Le choix du nombre de dimensions pourra être effectué par validation croisée. La limite de cette approche est que les premières composantes ne sont pas forcément celles qui sont les plus liées à la réponse. Une première possibilité est alors d'identifier les composantes dont le rapport de corrélation avec la variable réponse est le plus élevé. Une autre possibilité est d'employer une méthode plus avancée comme la régression PLS (*partial least square*) qui permet de trouver les composantes orthogonales qui maximisent la covariance avec la variable réponse (Tenenhaus (1998),Wikistat (2017)).

NB : Même si elles ne portent pas de nom spécifique, l'application de méthodes supervisées autre que les méthodes de régression sont tout à fait envisageables sur les composantes principales (pourvu qu'elles puissent être appliquées sur des données quantitatives). On parle par exemple d'analyse *Disqual* pour l'analyse discriminante sur les composantes de l'ACM.

3.2.2 Classification de variables

Les méthodes de classification non-supervisée peuvent être assez facilement adaptées à la classification des variables. Comme pour la classification des individus, on distingue 3 stratégies de partitionnement : les méthodes hiérarchiques ascendantes, descendantes, et les méthodes de partitionnement direct. Classifier les variables peut servir à analyser la multicolinéarité entre les variables ou plus généralement à identifier des groupes de variables corrélées entre elles et plutôt indépendantes des autres. Dès lors, si des variables sont extrêmement liées entre elles (voire de façon déterministe), il pourra être pertinent de n'en conserver qu'une. Dans le cas où ces liaisons sont moins fortes, on pourra substituer un petit nombre de variables latentes synthétiques, représentatives de chacune des classes. Ceci pourra être effectué en utilisant une méthode d'analyse factorielle adaptée à la nature des variables du groupe. Les méthodes de classification de variables seront développées plus en détail dans un prochain cours.

3.2.3 Sélection de variables

Les méthodes précédentes visent essentiellement à transformer les variables initiales par d'autres variables synthétiques. Ceci à l'inconvénient de complexifier l'interprétation des résultats in fine (cf transformation par analyse factorielle Section 3.2.1). Dans une optique d'analyse supervisée, une autre stratégie consiste à choisir, selon un certain critère, le meilleur sous-ensemble des variables initiales pour prédire ou expliquer une variable réponse. De cette façon, l'analyse portera sur des variables non-transformées qu'il sera plus facile d'interpréter.

Nous reviendrons ultérieurement sur ces méthodes. Sans trop anticiper sur la suite, on peut déjà remarquer que le modèle qui ajustera le mieux les données sera nécessairement celui avec le plus de variables explicatives. La qualité de l'ajustement ne peut donc pas être le seul critère considéré. Par ailleurs, on peut remarquer que le nombre de possibilités est très grand (2^p pour un jeu de données avec p variables candidates), rendant impossible une recherche exhaustive.

4 Conclusion

L'utilisation des méthodes exploratoires dans une optique de pré-traitement est très différente de leur utilisation classique : alors qu'elles sont généralement utilisées pour visualiser, explorer et comprendre les données, ici l'interprétation importe peu, et ces méthodes sont uniquement utilisées pour hiérarchiser l'information, transformer les variables, isoler la structure des données. Il n'en demeure pas moins que visualiser, explorer et comprendre les données est très important dans un premier temps, mais une fois ce travail effectué, les méthodes exploratoires pourront être réemployées en vue du pré-traitement et l'application futures de méthodes d'analyse.

Il est important de noter que le pré-traitement effectué est toujours relié aux méthodes d'analyse employées. Il serait par exemple inutile de regrouper des modalités d'une variable qualitative ordonnée pour y appliquer un arbre binaire par la suite. En effet, ce type de méthodes gère très bien ce type de variables qu'elles aient ou non des modalités rares. De même, certaines méthodes statistiques peuvent être appliquées en présence de données manquantes, sans qu'il ne soit nécessaire d'imputer préalablement les données (en utilisant par exemple des algorithmes EM). Dans les deux cas, ces pré-traitements pourraient détériorer les performances des méthodes.

Références

Aïvazian, S. 1978. *Étude Statistique Des Dépendances*. Editions de Moscou.

De Soete, Geert, and J. Douglas Carroll. 1994. "K-Means Clustering in a Low-Dimensional Euclidean Space." In *New Approaches in Classification and Data Analysis*, edited by Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand, and Bernard Burtschy, 212–19. Berlin, Heidelberg: Springer Berlin Heidelberg.

Fayyad, Usama M., and Keki B. Irani. 1993. "Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning." In *13th International Joint Conference on Artificial Intelligence*, 1022–29.

Husson, F., S. Lê, and J. Pagès. 2009. *Analyse de Données Avec r*. Presses universitaires de Rennes.

Kim, HyunJi. 2012. *Discretization: Data Preprocessing, Discretization for Classification.* https://CRAN.R-project.org/package=discretization (https://CRAN.R-project.org/package=discretization).

Rakotomalala, R. 2009. "La Regression Dans La Pratique." https://eric.univ-lyon2.fr/ricco/cours/cours/La_regression_dans_la_pratique.pdf (https://eric.univ-lyon2.fr/ricco/cours/cours/La_regression_dans_la_pratique.pdf).

———. 2010. "Discrétisation Des Variables Quantitatives." https://eric.univ-lyon2.fr/ricco/cours/slides/discretisation.pdf (https://eric.univ-lyon2.fr/ricco/cours/slides/discretisation.pdf).

———. 2011. "Tests de Normalité : Techniques Empiriques Et Tests Statistiques." https://eric.univ-lyon2.fr/ricco/cours/cours/Test_Normalite.pdf (https://eric.univ-lyon2.fr/ricco/cours/cours/Test_Normalite.pdf).

Saporta, G. 2006. *Probabilités, Analyse Des Données Et Statistique*. Editions Technip.

———. 2017. "Clusterwise Methods, Past and Present." https://cedric.cnam.fr/fichiers/art_4070.pdf (https://cedric.cnam.fr/fichiers/art_4070.pdf).

Tenenhaus, M. 1998. *La régression PLS: Théorie Et Pratique*. Editions Technip.

Tufféry, S. 2007. *Data Mining Et Statistique décisionnelle: L'intelligence Des Données*. Editions Technip.

———. 2015. *Modélisation Prédictive Et Apprentissage Statistique Avec r:* Éditions Technip.

Vichi, Maurizio, and Henk A.L. Kiers. 2001. "Factorial k-Means Analysis for Two-Way Data." *Computational Statistics & Data Analysis* 37 (1): 49–64.

Wikistat. 2017. “Composantes Principales Et Régressions PLS Parcimonieuses — Wikistat.” https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-sparse-pls.pdf (https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-sparse-pls.pdf).

---

1. minimum description length principle criterion↩
2. le coefficient de corrélation est ici utilisé comme un critère caractérisant l’alignement des points, son interprétation n’a pas d’importance.↩