Méthodologie statistique pour la discrimination et le classement. Application au ciblage des interventions nutritionnelles.

Pierre Traissac

In : Padilla M, Delpeuch F, Le Bihan G, Maire B, editors. *Les politiques alimentaires en Afrique du Nord D'une assistance généralisée aux interventions ciblées*. Paris: Karthala; 1995. p. 393-431 (http://horizon.documentation.ird.fr/exl-doc/pleins_textes/pleins_textes_7/b_fdi_03_05/010004037.pdf).

L'article de Pierre Traissac présente la démarche de modélisation prédictive. Il décrit en particulier 3 techniques statistiques : l'analyse discriminante, la régression logistique et les arbres de décision (arbres de segmentation, CART). Il se place dans un cadre tout à fait particulier, le ciblage des programmes nutritionnels. Cet aspect peut occulter l'intérêt du discours pour nous informaticiens / statisticiens, adeptes du téraoctet et de la maximisation de la marge. Il faut pouvoir dépasser cela. On se rend compte alors que l'auteur nous livre là une trame particulièrement claire et accessible de l'analyse prédictive, si populaire en data science (data mining).

L'auteur commence (**section 1**) en se plaçant dans le cadre générique de la discrimination et du classement (ces termes ne sont plus du tout « sexy » de nos jours, nous dirions plutôt « analyse prédictive » pour être en phase avec l'air du temps). Il définit les notions d'unité statistique (individu, observation, enregistrement), de variable cible qualitative (définissant une partition de la population en groupes dont un en particulier que l'on cherchera à cibler), de variables prédictives (descripteurs, variables explicatives) qui sont de type quelconque. L'objectif est de prédire et expliquer (deux objectifs intrinsèques mais qui se rejoignent – il est difficile de valider une prédiction si on n'arrive pas à se l'expliquer) les valeurs prises par la variable cible à partir des variables explicatives.

On distingue deux phases dans la démarche de modélisation : l'ajustement du modèle prédictif (le calcul de ses paramètres) à partir des données disponibles (échantillon de base, échantillon d'apprentissage), puis son évaluation (mesurer ses performances). Dans la **seconde section**, cette procédure d'évaluation est dépeinte de manière détaillée. L'auteur met l'accent sur les outils et indicateurs valables quels que soient les classifieurs utilisés. On distinguera notamment (a) la matrice de confusion (matrice de concordance, sa matrice est transposée par rapport à la présentation que j'utilise habituellement) et les indicateurs

associés (ex. valeurs prédictives positives que l'on appelle précision dans d'autres communautés), et (b) la courbe ROC (Receiving Operating Characteristic) qui présente l'avantage de dépasser la nécessité de définir un seuil pour l'affectation aux classes (la courbe peut servir en revanche à déterminer le dit seuil). Cette évaluation ne doit pas reposer sur l'échantillon d'apprentissage ayant servi à la construction du modèle. L'auteur présente alors différents schémas d'échantillonnage (ou de ré-échantillonnage) pour obtenir une estimation « honnête » de la qualité de la règle d'affectation (utilisation d'un échantillon de test, validation croisée). Il termine cette section en faisant mention des autres critères permettant de situer la qualité des modèles (compromis entre simplicité et performance, nombre de variables, forme du modèle – combinaison linéaire des facteurs ou modèle à base de règles).

La **section 3** est consacrée à la présentation de trois méthodes de data mining. La finalité, les principes sous-jacents, la forme du modèle obtenu sont tour à tour caractérisés pour l'analyse discriminante, la régression logistique et les arbres de segmentation (arbres de décision dirait un informaticien). L'auteur insiste beaucoup sur les arbres, leur consacrant 6 pages (1,5 pour respectivement l'analyse discriminante et la régression logistique). Peut-être parce qu'à l'époque de l'article (1995), les arbres étaient peu diffusés, peu connus et de fait très peu utilisés dans les études. Sans verser dans une description par trop théorique, il expose les points importants de la construction d'un arbre : la segmentation, le rôle du critère de partitionnement, le post-élagage (méthode CART) pour déterminer la taille adéquate de l'arbre, l'utilisation de l'arbre lors du classement d'un nouvel individu non étiqueté. Clairement, l'arbre est présenté comme une alternative aux méthodes « classiques » que constitueraient l'analyse discriminante et la régression logistique. La simplicité du classifieur (règles logiques) et la capacité à prendre en compte les interactions dans la structure décisionnelle sont mises en avant dans cette optique.

Dans la **section 4**, l'auteur décrit la mise en œuvre des différentes techniques sur des données réelles issues d'une enquête nationale sur le budget et la consommation des ménages réalisée par l'Institut National de la statistique de Tunisie en 1990. Il s'agit d'expliquer la maigreur (au sens malnutrition) d'une fraction de la population à partir de différentes caractéristiques associées aux ménages et aux individus. A priori, le texte est spécifique à un domaine. Il est compliqué pour nous, béotiens en matière de santé publique

et de nutrition, de vouloir y comprendre quoi que ce soit. Je l'ai malgré tout lu en détail et, ici aussi, je me suis rendu compte qu'il y a de nombreux enseignements à en tirer : sur la manière de mener une étude, en matière de présentation des résultats, sur la conversion des coefficients d'un classifieur linéaire en grille de score de ciblage facile à déployer (ex. tableau 6, page 419), sur le choix du seuil d'affectation à partir de la courbe ROC (graphique 7, page 422), sur la comparaison des performances prédictives des méthodes (pages 426 et suivantes). La trame peut servir de source d'inspiration pour toute personne désireuse de mener une étude similaire dans tout autre domaine.

J'ai eu beaucoup de plaisir à lire cet article (ça paraît évident). La publication est assez ancienne (1995, il y a 20 ans), à une période où les termes data mining (ne parlons même pas de data science) et analyse prédictive commençaient à peine à poindre auprès du grand public (initié un peu quand même). On se rend compte néanmoins que le processus et les méthodes statistiques sous-jacentes sont génériques et s'appliquent finalement à de nombreux domaines. Le plus important pour nous est de distinguer ce qui est généralisable à d'autres champs d'application. La clarté, la simplicité et la vocation didactique de l'article de Pierre Traissac le rend particulièrement intéressant. Il l'est d'autant plus qu'il se place dans un domaine (la malnutrition) qui est habituellement très peu investi par le data mining (ah bon, on ne parle pas de banque, d'assurance, de grande distribution, de clients à scorer ?...).