IMPORTANCE DES VARIABLES DANS LES METHODES CART

Badih GHATTAS

GREQAM - Université de la Méditerranée

Ghattas@lumimath.univ-mrs.fr

1. Introduction

Un arbre de régression ou de classification A obtenu par les méthodes CART (Breiman *et al*, 1984) permet de visualiser des variables dites *actives* qui participent directement à sa construction, et donc à la procédure de discrimination et de prévision correspondante.

Cela dit certaines variables explicatives qui ne jouent plus aucun rôle lorsque l'arbre est construit ont pu être pour plusieurs nœuds *concurrentes* des variables actives. La connaissance de ces variables concurrentes est utile à différents niveaux et peut permettre d'établir une *hiérarchie* de l'ensemble des variables explicatives. Cette hiérarchie peut servir pour mettre en œuvre d'autres méthodes statistiques avec un nombre de variables réduit.

L'exemple présenté sur la figure 1 est un arbre de régression pour la prévision quotidienne du maximum de l'ozone dans la station d'Istres des Bouches du Rhône. Le tableau 1 donne les 10 variables les plus importantes obtenus par la hiérarchie établie sur cet arbre, ainsi qu'un *indice d'importance* sur une échelle de 0 à 100. La valeur 100 est attribuée à la variable la plus importante ayant servi à la construction de ce modèle.

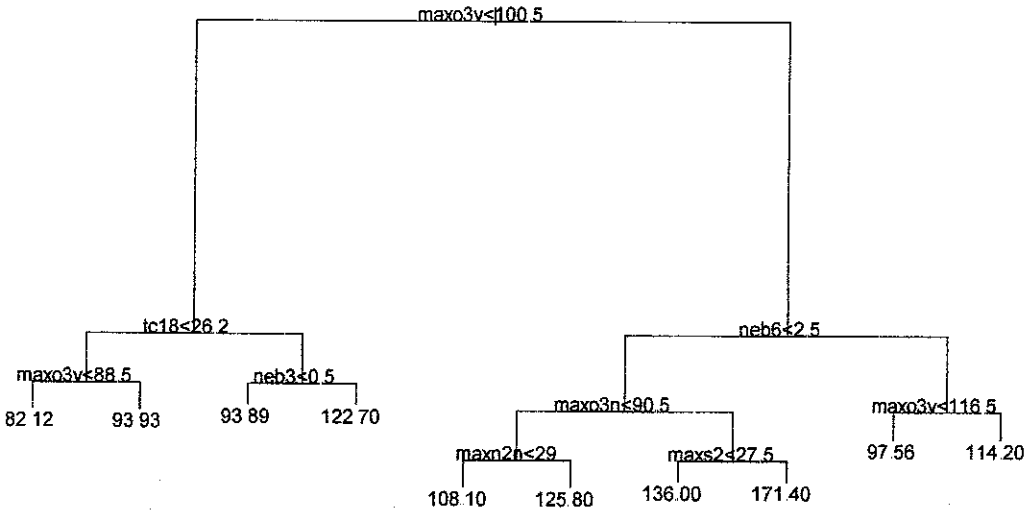

Figure 1 : Les variables actives sont : le maximum de l'ozone la veille "maxo3v" et la nuit "maxo3n" (exprimés en $\mu g/m^3$), la température à 6h et à 18h "tc6" et "tc18" (exprimés en degrés Celsius), la nébulosité à 3h et à 6h "neb3"et "neb6" (ayant neuf niveaux ordonnés de 1 à 9), le maximum et le minimum de NO2 la nuit "maxn2"-"minn2n" et le maximum de SO2 la nuit "maxs2" (exprimés tous deux en $\mu g/m^3$).