

Modèles statistiques

Les modèles statistiques sont en général utilisés dans le cadre d'une démarche déductive. La question porte sur la variabilité d'un phénomène particulier, et on cherche à comprendre les composantes de cette variabilité. Si les individus statistiques sont des entités spatiales, cela revient à s'interroger sur une différenciation spatiale. Si les individus statistiques sont des entreprises, on s'interroge sur les différenciations inter-entreprises etc. La démarche consiste à élaborer un certain nombre d'hypothèses sur les causes de cette différenciation et ces hypothèses sont ensuite testées par le modèle statistique. Le modèle le plus classique est la régression multiple qui s'applique lorsque l'ensemble des variables en jeu sont quantitatives. Le modèle s'écrit:

$$Y = a_1 X_1 + a_2 X_2 + \dots + a_p X_p + e$$

où Y désigne la variable " à expliquer "

X_1, X_2, \dots, X_p désignent les variables " explicatives "

a_1, a_2, \dots, a_p sont les coefficients

e est le résidu de la régression.

L'application du modèle permet, à l'aide des statistiques inférentielles, d'évaluer la significativité et les rôles respectifs de X_1, X_2, \dots, X_p pour " expliquer " la variabilité du phénomène décrit par Y. Par exemple, si Y représente le taux de variation moyen annuel de la population des villes d'un pays, les variables " explicatives " pourraient être: la taille initiale de la ville, la qualification de la main d'oeuvre (la part des cadres dans la population active par exemple), le profil économique (la part des actifs travaillant dans différents secteurs d'activité), la situation relative (distance à la ville voisine la plus proche, nombre de villes dans le voisinage).

Des variables explicatives qualitatives peuvent être introduites par l'intermédiaire d'un codage approprié. A l'exemple précédent on peut donc ajouter des variables d'accessibilité comme la présence d'une gare TGV ou d'un aéroport etc.

En revanche si la variable " à expliquer " est qualitative, le cadre de la régression multiple n'est plus approprié. Le modèle le plus couramment utilisé est alors le modèle logit qui exprime la probabilité d'observer telle modalité de la variable à expliquer en fonction des variables explicatives, qualitatives et éventuellement quantitatives.

Bibliographie