CLASSIFICATION DES MODÈLES (G2, J)
(07 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

Dans chaque **domaine de connaissance**, le scientifique cherche à donner une interprétation et une représentation aussi fidèles que possible de tout **phénomène** qui en relève (cf **science**, **représentation statistique**). La conception probabiliste dans lesquels un phénomène est « plongé » conduit à la notion générale de **loi scientifique**. Celle-ci peut se concevoir :

(a) soit comme la **loi de probabilité** du phénomène considéré (ou du **système aléatoire** qui le représente) ;

(b) soit comme une **caractéristique** particulière de cette loi (cf **relation fonctionnelle**).

La formalisation de cette approche probabiliste est généralement appelée **modèle statistique** (théorique) ou **représentation statistique** (théorique).

Il existe différentes façons, complémentaires entre elles, de classer les modèles. Ceci s'effectue à partir de leurs principales particularités.

(i) Selon l'**objectif** ou la **finalité** recherchés par l'**homme de l'art**, on distingue entre :

(a) **modèle descriptif**. Ce type de modèle relie des grandeurs entre elles sans prétendre fournir réellement une explication théorique des manifestations du phénomène décrit. Il peut jouer un rôle de préparation à une analyse causale ;

(b) **modèle explicatif** ou **modèle causal**. Ce second type cherche à illustrer (ou à justifier) une **théorie** particulière (dans laquelle une notion de **causalité** intervient), en utilisant notamment des propriétés de la **statistique inférentielle**. Une **spécification** de base met en relation diverses variables et les raisonnements s'effectuent notamment dans l'**espace des variables**.

(ii) Selon le **caractère aléatoire** ou non aléatoire du phénomène étudié, on distingue entre :

(a) **modèle déterministe**. Ce premier type est souvent directement issu d'une théorie elle-même, du moins lorsque celle-ci est suffisamment formalisée par l'homme de l'art. Ainsi (sociologie : économie), le modèle keynésien élémentaire à deux équations :

$$\begin{aligned} c &= \alpha\, y + \beta, \\ y &= c + d, \end{aligned} \qquad (2)$$

(dans lequel c est la consommation finale, y le revenu des ménages, d une dépense considérée comme exogène et $\alpha = dc / dy$ la propension marginale à consommer le revenu) constitue une représentation théorique dans l'**espace des variables** (c, y), sans substrat probabiliste a priori ;

(b) **modèle stochastique**. Ce second type de modèle résulte souvent d'une modification de la forme déterministe précédente, par adjonction d'**hypothèses stochastiques**, afin d'autoriser la mise en oeuvre des méthodes statistiques. Ainsi, le modèle (1) peut être « plongé » dans le modèle suivant :

$$\begin{aligned} c &= \alpha\, y + \beta + \varepsilon, \\ y &= c + d, \end{aligned} \qquad (2)$$

dans lequel :

(b)$_1$ ε désigne une **perturbation aléatoire** affectant la première équation, et vérifiant notamment $E\,\varepsilon = 0$. Si l'on raisonne en termes d'**espérance** (ou de valeur moyenne) inconditionnelle, la première équation s'écrit aussi $E\,c = \alpha\, y + \beta$. Si, ce qui est la situation habituelle, l'on raisonne en termes conditionnels, elle s'écrit $E\,c / \varepsilon = \alpha\, y + \beta$. On adjoint souvent au modèle (2) des hypothèses probabilistes supplémentaires tq eg $V\,c = V\,\varepsilon = \sigma^2$ (variance) en termes inconditionnels, ou encore $V\,c / \varepsilon = V\,\varepsilon = \sigma^2$ en termes conditionnels. De même, si l'hypothèse de normalité est admissible, on pose $c \sim \mathcal{N}_1\,(\alpha\, y + \beta,\, \sigma^2)$. Enfin, la disponibilité d'**observations** (temporelles en temps discret) (C, Y, D) relatives aux variables (c, y, d), avec $C = (c_1, ..., c_T)$, $Y = (y_1, ..., y_T)$ et $D = (d_1, ..., d_T)$, conduit à écrire (2) dans l'**espace d'observation** (C, Y, D) selon :

$$\begin{aligned} C &= \alpha\, Y + \beta\, e_N + u, \\ Y &= C + D, \end{aligned} \qquad (3)$$

où e_N désigne le premier vecteur bissecteur de $\mathbf{R}^N$ et $u = (u_1, ..., u_N)$ le vecteur perturbant la première équation de (3). Les hypothèses probabilistes correspondant aux précédentes sont alors $E\,C / U = \alpha\, Y + \beta\, e_N$ et $C \sim \mathcal{N}_N\,(\alpha\, Y + \beta\, e_N\,,\, \sigma^2\, I_N)$;

(b)$_2$ la seconde (équation « comptable ») demeure déterministe (cf **identité**).

Un modèle déterministe est aussi parfois appelé **modèle théorique**, un modèle stochastique étant alors dit **modèle statistique**, ou **représentation statistique**.

(iii) Selon la **forme analytique** (ie mathématique) du modèle, on distingue entre :

(a) **modèle linéaire**. La linéarité s'entend, en général, pr aux **paramètres**, car il existe souvent des non linéarités pr aux **variables**, qu'il s'agisse de **variables endogènes** ou de **variables exogènes**, du modèle. Un **modèle linéaire** (au sens précédent) a l'avantage de la simplicité (relative) des calculs, qui peuvent généralement être menés analytiquement, ainsi que celui de fournir des formules exactes sur un certain nombre de questions. Il peut aussi être considéré comme une **approximation locale** (ou **approximation tangentielle**) d'un modèle non linéaire (pr aux paramètres) au voisinage des **« vraies » valeurs** de ces paramètres (cf **linéarisation**, **modèle linéarisé**) ;

(b) **modèle non linéaire**. Cette catégorie prétend mieux approcher la réalité du phénomène étudié mais a l'inconvénient de nécessiter des calculs d'**approximation** : d'où le recours, soit au **calcul numérique**, soit à l'étude des **propriétés asymptotiques**. D'autre part, la théorie sous la main n'explicite pas toujours un tel modèle (ie sa forme analytique n'est pas toujours connue ou donnée) : on doit donc souvent (b_1) soit la déduire des observations ou de raisonnements a priori, (b_2) soit l'interpréter comme modèle non paramétrique (cf **modèle paramétrique**).

(iv) Selon la **nature des observations** disponibles, on distingue entre :

(a) **modèle sur séries temporelles**, ou **modèle sur processus** (cf **modèle de processus**). Cette catégorie de modèles utilise des observations indicées par un ensemble totalement ordonné (le « **temps** ») (cf aussi **indice**) ;

(b) **modèle sur coupes instantanées**, ou **modèle sur données individuelles**. Ce second type utilise des observations indiciées par un ensemble souvent « amorphe », génériquement appelé **espace**. L'instant ou la durée d'observation, fixe, peut alors concerner des **unités statistiques** (souvent appelées individus) ou diverses **zones de l'espace ambiant** $\mathbf{R}^3$ (régions, départements, terrestres vs aquatiques, urbaines vs rurales, micro vs macro ou méso, etc) ;

(c) **modèles mixtes**. Cette situation combine les deux précédentes : ces modèles sont construits sur des séries temporelles de coupes instantanées, donc dans un espace-temps (eg $\mathbf{R}^3$ x $\mathbf{R}$). Les observations s'effectuent au cours du temps sur des **unités statistiques** (ou sur des **variables** observées sur celles-ci) : on parle de modèles à double (ou même à multi-) **indice**, tel le **modèle à erreurs composées** ou le modèle à partie certaine composée.

On parle alors de modèle temporel, de modèle instantané et de modèle mixte (au sens actuel).

(d) **modèle sur données individuelles (ou locales)**, ou **modèle sur données agrégées** : Selon le cas, les données sont constituées d'observations individuelles (ou locales) ou d'observations agrégées (eg par **agrégation** des premières) (cf aussi **transformation des données**).

(v) Selon que les variables prises en compte sont des **variables contemporaines** ou des **variables décalées** (par application de l'**opérateur avance** ou de l'**opérateur retard**), on parle :

(a) de **modèle statique**. Dans ce type de modèles, les variables sont toutes « contemporaines » ;

(b) de **modèle dynamique** ou **modèle de cheminement**. Ce type de modèles comporte (au moins) une variable (temporellement) décalée pr aux autres.

Dans un modèle dynamique, un **indice** (au moins) t associé aux observations des variables est à valeurs dans un ensemble totalement ordonné : $t \in T$, où $(T, \leq)$ représente le **temps**. Ainsi en est-il du **modèle récursif**, du **modèle autorégressif**,

du **modèle à retards échelonnés**, du **modèle de moyenne mobile** (cf aussi **processus de moyenne mobile**), du **modèle d'interdépendance dynamique**, etc.

(vi) Le **mode d'observation temporelle** des variables conduit à distinguer entre :

(a) **modèle en temps discret** (ie $T \subset \mathbf{Z}$ ou $T \subset \mathbf{N}$) ;

(b) **modèle en temps continu** (ie $T \subset \mathbf{R}$ ou $T \subset \mathbf{R}_+$). Ce modèle utilise, par définition, des **séries temporelles** ou des **processus** qui peuvent être, par nature, en temps discret ou en temps continu, ou encore observés selon un **dispositif d'observation permanent** ou selon un **dispositif d'observation intermittent**. Ainsi, on peut observer en temps discret une **trajectoire** relative à un processus en temps continu ; en temps continu celles d'un processus en temps continu ; en temps continu celles d'un processus en temps discret ; en temps discret celles d'un processus en temps discret (ce dernier cas exige, pour que l'observation soit effective, la réunion de certaines conditions : coïncidences ou synchronisme, etc).

(c) **modèle en temps « mesuré »** $(T, \mathcal{B}_T, \nu)$. Ce modèle utilise des observations qui sont temporellement « cadencées » selon une mesure ν donnée.

(vii) Selon l'**utilisation** d'un modèle, on distingue entre :

(a) **modèle descriptif** (ou **modèle pédagogique**). Un tel modèle constitue souvent une première formalisation résultant de la simple description du fonctionnement (supposé ou apparent) du phénomène en examen (cf équation (1) infra). Dans d'autres situations, il peut servir à l'enseignement d'une science ou à une « explication » destinée à des « décideurs ». Un tel modèle est donc souvent exprimé dans un **espace de variables** ;

(b) **modèle de simulation**. Un modèle de **simulation** (eg tq $\eta = f(\xi)$) est généralement conçu en sorte que, après estimation (eg selon $y^{\#} = f^{\#}(\xi)$), il puisse fournir des images endogènes successives $F_i^{\#} = f^{\#}(E_i)$ obtenues en faisant parcourir aux variables exogènes ξ des « plages » de valeurs (ie des ensembles) donné(e)s E_1 ,..., E_i ,... Une autre acception du terme « simulation » est celle d'un modèle décrivant divers « comportements » individuels en sorte de représenter fidèlement, au niveau macroscopique, l'évolution d'une population donnée;

(c) **modèle décisionnel**. Ce type de modèles distingue :

$(c)_1$ dans la liste des **variables exogènes**, des **variables instruments**, ou **variables de contrôle**. Cette notion diffère de celle de **variable instrumentale**, utilisée dans d'autre contextes ;

$(c)_2$ dans la liste des **variables endogènes**, des **variables objectifs** ou **variables cibles** : le modèle doit permettre de quantifier les niveaux possibles que doivent atteindre les instruments pour que les objectifs puissent atteindre des valeurs fixées a priori.

(viii) Selon la **structure probabiliste**, ou non probabiliste, sous-jacente au modèle, on distingue entre :

(a) **modèle à un régime**. Un modèle à régime unique est considéré comme valide sur tout le champ des valeurs des variables (observations disponibles) et, souvent aussi, en dehors de ce champ (notamment dans le cadre d'une **projection** ou d'une **prévision**) ;

(b) **modèle à plusieurs régimes** (cf **modèle à structure variable**). Un modèle à régime multiple est un modèle eg de la forme $\eta = f(\xi)$, où le couple (ξ, η) est à valeurs dans un **espace d'observation** produit $\mathcal{X} \times \mathcal{Y}$: mais ce modèle est supposé tq, sur chaque classe d'une **partition** (souvent inconnue) $\Pi_{\mathcal{X}}$ de $\mathcal{X}$, la restriction de f soit égale à une fonction de forme donnée (cf **restriction d'une application**). Autrement dit, la liaison f entre ξ et η dépend de certaines plages de valeurs prises par ξ.

(ix) Selon la **structure théorique** sous-jacente au modèle, on distingue entre :

(a) **modèle à paramètres fixes**. Un modèle à paramètres fixes est un cas particulier (très usuel) du suivant, dans lequel les paramètres ne subissent aucune influence (cf **vraie valeur d'un paramètre**) ;

(b) **modèle à paramètres variables**. Un modèle à paramètres variables est tq ses paramètres peuvent dépendre eux-mêmes de variables (internes ou non au modèle considéré).

(x) Selon la **place de l'aléas**, ou la **nature de l'aléas**, on distingue entre :

(a) **modèle à paramètres certains** ;

(b) **modèle à paramètres aléatoires** (parfois aussi appelé **modèle mixte**). Ce dernier correspond eg à la **théorie bayésienne** dans laquelle le **statisticien** est supposé disposer d'**informations** a priori sur la **loi** (ou sur certaines valeurs) du **paramètre** avant de procéder à l'**inférence statistique** (**classification**, **estimation**, **tests**, **prévision**, etc). En particulier, les informations a priori peuvent provenir de l'estimation de modèles semblables (ou de paramètres ayant la même signification) effectuée sur d'autres jeux d'observations.

(xi) Toujours selon la **place (ou nature) de l'aléas**, on distingue parfois entre :

(a) **modèle « intérieurement » aléatoire**. Ce premier type correspond à un modèle de **production statistique**, dans lequel les **observations** sont volontairement, ou par nécessité, engendrées par construction d'un mécanisme aléatoire interne (**sondage**, **plan d'expérience** randomisé, etc) ;

(b) **modèle « extérieurement » aléatoire**. Ce second type correspond plutôt à un modèle d'étude statistique dans lequel, à la différence du précédent, les observations aléatoires résultent de causes externes (nombreuses, ou non identifiées, etc) assimilables, souvent par commodité, à un mécanisme aléatoire : **relation fonctionnelle**, **modèle de régression**, **modèle d'interdépendance**, etc.

Les deux situations précédentes peuvent se combiner (eg **modèle d'analyse de la variance** avec données d'**unités expérimentales**, modèle de **régression avec données de sondage**).

(xi) Encore selon la **place (ou nature) de l'aléas**, on distingue parfois entre :

(a) **modèle avec famille de lois pures** ;

(b) **modèle avec famille de lois mélangées**.

En effet, la famille des lois susceptibles de gouverner un phénomène peut être « contaminée » par des « lois parasites » entrainant la présence, parmi les données, d'observations aberrantes (cf **aberration**).

Par ailleurs, certaines modèles sont spécifiés avec des familles de lois tronquées (cf **troncature**) ou encore des observations censurées (cf **censure**).

En pratique, une **situation statistique** donnée combine souvent, au sein d'un même modèle, certaines des particularités précédentes.