PRÉVISION (G10, H6, J9, N6)
(10 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

« Une intelligence qui, à un instant donné, connaîtrait toutes les forces dont la nature est animée, la position respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers, et ceux du plus léger atome. Rien ne serait incertain pour elle, et l'avenir comme le passé seraient présents à ses yeux »
P.S. de LAPLACE (cf aussi **Statistique et hasard**)

Le concept de **prévision** est une notion fondamentale en **Statistique** : elle constitue un aboutissement important de la **modélisation** et de la **décision statistique**. L'idée de base (cf P.S. de LAPLACE) est que, au sein de chaque **domaine de connaissance**, la connaissance du **passé** et du **présent** d'un **phénomène** relevant de ce domaine, ainsi que des **lois** qui gouvernent celui-ci, permettent de déterminer son **futur**. Ceci n'est quasiment jamais réalisé, mais diverses procédures ont pour objet d'approcher une situation optimale de ce point de vue.

On distingue généralement entre :

(a) **prévision inconditionnelle** (ou prévision endogène, ou encore prévision autogène) et **prévision conditionnelle** :

$(a)_1$ une prévision inconditionnelle se fonde sur un **modèle** de type « autonome » : ie les variables endogènes considérées ne dépendent pas d'autres variables (« exogènes » ou « prédéterminées »), ni d'autres événements, mais seulement de leurs propres valeurs retardées (cf **processus autorégressif**). Elles peuvent aussi ne dépendre que du **temps** lui-même : dans ce cas, on parle plutôt de **projection** - dans ce sens - ou d'**extrapolation** ;

$(a)_2$ une prévision conditionnelle met en oeuvre un modèle dans lequel on distingue entre **variables endogènes**, qui sont les variables d'intérêt relatives à un phénomène, et **variables exogènes**. Pour prévoir les endogènes, les exogènes doivent elles-mêmes être « prédites », ie évaluées (eg variables de commande d'un **système**) : ainsi, dans un modèle de simulation de ce type, le **statisticien** se donne les valeurs « futures » des exogènes afin d'en déduire celles des endogènes.

Chacun de ces deux types de prévision peut s'effectuer à l'aide d'une **variable aléatoire** (**variable observable** ou non) (étude théorique) ou à l'aide d'un **échantillon aléatoire** qui peut être seulement partiellement observé (étude empirique, ou « opérationnelle ») ;

(b) **prévision paramétrique** (cf **modèle paramétrique**), **prévision non paramétrique** (cf **Statistique non paramétrique**) et **prévision semi-paramétrique** (cf **modèle semi-paramétrique**). La méthode de prévision est généralement liée à une méthode d'**estimation** statistique de l'un de ces types (estimation paramétrique, non paramétrique ou semi-paramétrique).

Par ailleurs, la plupart des méthodes de prévision sont basées sur la notion de **modèle dynamique**.

(i) **Prévision inconditionnelle**

Soit $(\Omega, \mathcal{T}, \mathcal{P})$ un **modèle statistique** fondamental, $(\mathcal{X}_0, \mathcal{B}_0)$ un **espace d'observation** et $\xi : \Omega \mapsto \mathcal{X}_0$ une **variable aléatoire** dont la **loi** P^ξ, qui appartient à l'image $\mathcal{P}^\xi$ de la famille $\mathcal{P}$ par l'**application** ξ, est supposée connue :

(a) la **prévision inconditionnelle**, ou **prévision non conditionnelle**, ou encore **prévision « absolue »**, la plus générale, ou la plus complète, relative à ξ est celle qui fournit le plus d'**information** sur ξ : c'est donc la **loi de probabilité** P^ξ elle-même. Cependant, cette loi est « neutre », et ne fournit donc pas, d'emblée, une indication sur la vraisemblance d'évènements « à venir » ;

(b) si l'on ne connaît qu'une **caractéristique** de **centralité** (ou **paramètre de position**) γ_ξ (eg **espérance**, **médiane**, **mode**) de P^ξ, une **prévision inconditionnelle ponctuelle** de ξ est γ_ξ elle-même ;

(c) de même, une **prévision inconditionnelle ensembliste** de ξ au seuil préalablement fixé $\alpha \in]0, 1[$ (ie au seuil de confiance 1 - α) est une partie $C \in \mathcal{B}_\Gamma$ (**tribu de parties** de l'ensemble Γ des caractéristiques de centralité ou de position associé à ξ) tq (cf **région de confiance**) :

(1) $\quad \Gamma \subset \mathcal{X}_0$, $\quad \gamma_\xi \in C$, $\quad P^\xi(C) \geq 1 - \alpha$.

Si l'on ne connaît pas P^ξ, on peut généralement disposer d'un N-**échantillon aléatoire** $X = (X_1, ..., X_N) : \Omega \mapsto \mathcal{X}_0^N$. Celui-ci est souvent un **échantillon iid** selon P^ξ. Etant donné une **fonction de perte** L et une **fonction de risque** associée R, on peut estimer P^ξ à l'aide d'un **estimateur** $P_N^\#$ optimal au sens de R, qui dépend en principe de la **loi empirique** P_N associée à X (cf **statistique naturelle**). Par suite, la meilleure prévision inconditionnelle de ξ peut se définir à l'aide de $P_N^\#$ lui-même :

(a) si l'on considère une caractéristique de centralité ou de position $c_N(X)$ de $P_N^\#$ et si $c_N(X)$ est un estimateur ponctuel optimal de γ_ξ , on dit que $c_N(X)$ est une **prévision inconditionnelle ponctuelle** de ξ ;

(b) une **prévision inconditionnelle ensembliste** de ξ est définie à l'aide d'une **partie aléatoire** $S(X) \in \mathcal{B}_\Gamma$ tq $P([S(X) \supset \gamma_\xi]) \geq 1 - \alpha$, où $\mathcal{B}_\Gamma$ est encore une tribu de parties de l'ensemble Γ des valeurs possibles de γ_ξ , et où l'on note, comme il est d'usage, $[S(X) \supset \gamma_\xi] = \{\omega \in \Omega : \gamma_\xi \in S(X(\omega))\}$.

Une prévision tq P^ξ (resp γ_ξ, resp B) est parfois dite **prévision théorique** (elle peut résulter d'une **simulation**). Par distinction, une prévision tq $P_N^\#$ (resp $c_N(X)$, resp $S(X)$) est dite **prévision empirique**.

Les développements précédents sont notamment à la base des méthodes de prévision (endogènes ou autogènes) d'un **processus stochastique** ou d'une **série temporelle**, scalaires ou vectoriels, considérés isolément (cf **analyse harmonique**, **analyse spectrale**) : on suppose souvent, dans ce cadre, que l'indice $n \in N_N^*$ représente le **temps** (et on le note plutôt $t \in N_T^*$).

On peut rattacher cette approche aux approches, dites « empiriques », relatives aux **séries temporelles** : cf **méthode des chaînes de rapports**, **méthode des moyennes cycliques**, **méthode des moyennes groupées**, **méthode des moyennes mobiles**, **méthode des moyennes périodiques**, **méthode des rapports à la tendance**.

(ii) **Prévision conditionnelle**

Soit $(\Omega, \mathcal{T}, \mathcal{P})$ un **modèle statistique** fondamental, $(\mathcal{X}_0 \times \mathcal{Y}_0, \mathcal{B} \otimes \mathcal{C})$ un **espace d'observation** et $(\xi, \eta) : \Omega \mapsto \mathcal{X}_0 \times \mathcal{Y}_0$ une va de loi $P^{(\xi, \eta)}$. On est souvent conduit à étudier la **loi conditionnelle** de η pr à ξ, notée eg $P(. / \xi)$. Une démarche analogue à la précédente peut être développée :

(a) une **prévision conditionnelle**, ou **prévision relative**, de la va η sachant (une information contenue dans) la va ξ est définie par la loi conditionnelle $P(. / \xi)$ elle-même ;

(b) si l'on connaît une caractéristique de centralité ou de position $\gamma_{\eta / \xi}$ (espérance, médiane, mode) conditionnelle de $P(. / \xi)$, une **prévision conditionnelle ponctuelle** de η sachant ξ est définie par $\gamma_{\eta / \xi}$;

(c) une **prévision conditionnelle ensembliste** de η sachant ξ pour un seuil de confiance donné $1 - \alpha$ est une partie $C \in \mathcal{B}_\Gamma$ (tribu des parties de l'ensemble Γ des caractéristiques de η) tq :

(2) $\quad \Gamma \subset \mathcal{Y}_0, \qquad \gamma_{\eta/\xi} \in C, \qquad P(C / \xi) \geq 1 - \alpha,$

où Γ désigne l'ensemble des valeurs possibles de $\gamma_{\eta / \xi}$ et où :

(5) $\quad [S(Y / \xi) \ni \gamma_{\eta / \xi}]) = \{\omega \in \Omega : \gamma_{\eta / \xi} \in S(Y(\omega) / \xi)\},$

où le symbole ∋ signifie « contient ».

L'approche conditionnelle est à la base des méthodes de prévision associées à des modèles usuels : modèle de **relation fonctionnelle**, **modèle de régression** ou **modèle d'interdépendance**. Ces modèles peuvent être linéaires ou non, simples ou multiples. ils peuvent d'ailleurs relier entre eux plusieurs **processus** ou séries.

D'autre part, elle concerne aussi l'étude prévisionnelle des processus vectoriels ou des séries vectorielles (cf **analyse cospectrale**, pour des séries multiples, réalisations d'un **processus vectoriel**).

(iii) **Approche paramétrée**

Les deux approches précédentes ont été présentées dans un cadre non explicitement paramétré. Si le modèle initial (Ω, $\mathcal{T}$, $\mathcal{P}$) est explicitement paramétré, ie si (Ω, $\mathcal{T}$, $(P_\theta)_{\theta \in \Theta}$), elles s'appliquent encore (transposition directe).

(iv) **Théorie de la prévision**

Les principales notions qui interviennent dans cette théorie sont donc celles de **loi conditionnelle** et de **loi d'échantillonnage**, de **caractéristique de centralité** (ou de **caractéristique de position**), ou encore de **partie centrale** (conditionnelles) :

(a) la **théorie de l'estimation** intervient donc directement dans la résolution d'un **problème de prévision**.

On distingue cependant entre **prévision** et **estimation**. Prévoir une va (conditionnellement ou non) revient surtout à étudier son comportement **en dehors** de l'ensemble des valeurs (théoriques ou observées) (eg en fonction de l'échantillon disponible) : ceci revient à lier des variables inobservables (les « **prédictions** ») à des variables observables (données ou observations proprement dites) (les **« prédicteurs »**) (cf **prévision conditionnelle**).

(b) la nature d'une prévision dépend de la **nature des données** et, plus précisément, du type d'observation de ces données par un **système statistique**. Ces données sont le plus souvent des séries temporelles implicitement générées par des processus.

La notion de prévision ne se restreint cependant pas à ce cas : il est ainsi possible d'effectuer une prévision à partir d'une **coupe instantanée** (prévision transversale), ie d'une **série spatiale**. Sur des données en coupes instantanées, une **méthode de prévision** peut s'apparenter à une méthode d'**interpolation** (ou parfois d'**extrapolation**), ou encore à une méthode de **classification** (ou parfois de **discrimination**) : l'observation des variables exogènes relatives à une **unité statistique** supplémentaire donnée permet de « prévoir » une certaine valeur endogène « typique », laquelle peut donc différer de la valeur endogène effectivement observée sur cette unité (**erreur de prévision**).

Cependant, c'est surtout sur des données temporelles (processus ou séries) que des méthodes spécifiques ont été développées : cf **analyse spectrale** ou **analyse cospectrale**, **autocorrélation**, **méthodes de BOX-JENKINS**, **prévision d'un processus**, etc.

(v) La distinction entre prévision inconditionnelle et prévision conditionnelle peut s'illustrer à l'aide de deux processus classiques :

(a) le **processus autorégressif** ar (p) :

(6) $X_t = \Sigma_{i=1}^{p} \alpha_i . X_{t-i} + u_t , \quad \forall t \in T,$

où $X = (X_t)_{t \in T}$ et $u_t = (u_t)_{t \in T}$ sont deux processus donnés (à 1 ou plusieurs dimensions).

La forme autorégressive (6) est de type « autogène », car l'information utilisée dans cette équation de définition n'est apportée que par X. C'est donc sa structure statistique « interne » qui permet de le « projeter » dans le futur : les seules valeurs passées de X déterminent son avenir (prévision inconditionnelle) ;

(b) le **processus de moyenne mobile** mm (q) :

(7) $X_t = \Sigma_{j=0}^{q} \beta_j . V_{t-j} , \quad \forall t \in T,$

où $X = (X_t)_{t \in T}$ et $V = (V_t)_{t \in T}$ sont deux processus donnés.

La forme mobile (7) est de type « mixte », ie l'information utilisée dans cette équation de définition concerne à la fois X et V. C'est donc leur structure statistique « croisée » qui permet de « projeter » X dans le futur : les valeurs passées de V déterminent l'avenir de X (prévision conditionnelle).

(vi) Dans la littérature, on rencontre de nombreux termes plus ou moins équivalents à celui de prévision :

(a) **prédicteur** ou **prédiction**. Ce terme est repris de l'anglais « *prediction* » (mais n'a aucun rapport avec NOSTRA DAMUS ni avec l'astrologie !). Le **statisticien** ne prétend pas « connaître » l'avenir de cette manière (d'ailleurs vague et généralement purement subjective), mais seulement préciser (dans une certaine mesure) les évolutions futures d'un phénomène donné ;

(b) **projecteur** ou **projection**, **extrapolateur** ou **extrapolation** : ces termes se réfèrent plutôt à des méthodes de prévision relativement simples, ie dont les modèles statistiques sous-jacents sont peu élaborés (ce qui peut souvent suffire) : modèles « rapides », modèles « omnibus ». Ainsi, dans les cas les plus simples, on peut vouloir « projeter » une population humaine dans les 50 années à venir en « extrapolant » sa **tendance** passée simplement en fonction du temps. Une méthode de projection plus élaborée (conditionnelle) ajoutera des hypothèses de natalité et de mortalité (mouvement naturel), voire des hypothèses de flux migratoires (immigration, émigration), dans la mesure où ces hypothèses peuvent permettre d'améliorer les **schémas** élémentaires précédents ;

(c) **rétropolation** ou « **prévision arrière** ». Dans la pratique, les séries temporelles disponibles sont plus ou moins « longues », mais généralement de « longueur » finie : elles possèdent donc un instant d'observation initial. Le « temps » étant un **ensemble** ordonné (cf **relation d'ordre**), on peut vouloir « remonter » les données avant cet instant initial. Ainsi, si $T \in \mathbf{Z}$, et si $x = (x_t)_{t=1,...,T}$ est la série temporelle observée (disponible), on peut vouloir « évaluer » eg x_0 , x_{-1} ,…, x_{-H} ($H \in \mathbf{N}^*$ étant un entier donné), à l'aide d'une méthode donnée, qu'elle soit inconditionnelle ou non. C'est le but d'une rétropolation. Ainsi, la rétropolation des populations humaines des diverses régions du monde a aidé à l'évaluation du nombre d'êtres humains ayant vécu sur terre depuis l' « origine » : selon l'INED, quelques 80 Gh = $80 . 10^9$ humains vivent ou ont vécu sur Terre avant l'an 2000 ;

(d) **interpolation** : si les observations sont connues selon un degré de « finesse » donné (eg données trimestrielles), on peut vouloir les « connaître » avec un degré de finesse supérieur (eg données hebdomadaires). Il existe ainsi des méthodes de « **lissage** » (numérique) des données dans le but de « prévoir » les valeurs intermédiaires ;

(e) **anticipation** : ce terme (de connotation psychologique) se réfère à ce que l'on peut s'attendre à observer (ou à décider) pendant une durée de temps à venir (cf eg **anticipation rationnelle**, **modèle à correction d'erreur**).

(vii) La notion de prévision est inséparable de celle d'**horizon d'une prévision**. En effet, elle implique une notion de cible ou de but, ou encore de délai. Ainsi, dans le cas d'un processus temporel en temps discret (ou observé en temps discret) $X = (X_t)_{t=1,...,T}$ ou d'une série $x = (x_t)_{t=1,...,T}$ associée, cet horizon est ainsi la « date ultérieure » T+H, avec $H > 0$.

De plus, l'intérêt peut porter sur un « cheminement » à venir possible résultant d'une telle prévision. Ainsi la prévision des suites $(X_t)_{t=T+1,...,T+H}$ ou $(x_t)_{t=T+1,...,T+H}$ a aussi une utilité : eg action sur des variables de commande composant X ou x, afin de satisfaire à des contraintes exogènes durant la période T+1 ,..., T+H (« tunnel » de passage obligé, « butoirs », etc).

Une particularité assez générale d'une prévision réside dans la propriété suivante : l'incertitude sur (ie la **variabilité** de) la « valeur-horizon » X_{T+H} augmente avec H (cf **écart de prévision**, **prévision d'un processus**, **prévision des moindres carrés d'un processus**, **prévision des moindres carrés généralisés**,).

(viii) Enfin, le terme de « prévision » désigne aussi bien :

(a) l'**action** (ie la **procédure statistique** consistant à « prévoir ») ;

(b) que le **résultat** (ie la valeur « prévue » elle-même).

(ix) Des prévisions notoires peuvent concerner les domaines de connaissance suivants.

(a) physique (météorologie) : un modèle météorologique peut intégrer des propriétés de la physique (gradients de pression, de température, etc) pour aboutir à diverses

prévisions (spatio-temporelles). Toutes les variables en jeu sont, a priori, endogènes ;

(b) biologie (carcinogénèse) : l'influence de diverses thérapies (doses, combinaisons, fréquences) sur le développement d'un cancer donné a pour but l'établissement d'un pronostic. Il existe ici des variables de commande (ou de contrôle) dont le pronostic va dépendre ;

(c) écologie (biodiversité animale) : les conditions climatiques ou l'exploitation des ressources naturelles animales modifient leur démographie. Les premières variables (climat) ne sont pas modulables, tandis que les secondes (vitesse et ampleur de l'exploitation) peuvent être, dans une certaine mesure, régulées (épargne des individus jeunes, quotas de chasse ou de pêche, etc) ;

(d) psychologie : les modes de gestion d'une ressource humaine, au sein d'un groupe d'individus, peuvent permettre d'aboutir à un accord lors d'une négociation. L'anticipation du résultat (eg accord, désaccord, compromis) s'apparente à une prévision à trois modalités, mais dont l'horizon n'est pas toujours lui-même prévisible ;

(e) sociologie (économie) : l'évolution macroéconomique d'une zone économique dépend, en partie, des variables de politique économique (fiscalité, subventionnement, politique monétaire, encadrement de prix, politique budgétaire et dépense publique, droits de douane, etc). Ces variables jouent des rôles de commandes qui doivent modifier les évolutions futures dans un sens recherché. L'articulation de ces variables au cours du temps, par l'entité gouvernante, détermine généralement des évolutions à venir que celle-ci cherche à contrôler (objectifs d'un plan : réduction des inégalités, rééquilibrage de certains budgets, etc).