Notes méthodologiques

Dans cette partie, nous allons présenter les méthodes utilisées dans ce papier.
Rappelons que la variable chiffre d'affaires est notée par le processus stochastique Y_t.
Y_t étant supposé comme une fonction du temps t et d'une variable ϵ_t faisant office d'erreur au modèle, représentant la différence entre la réalité et le modèle proposé $Y_t = f(t, \epsilon_t)$

1. Modèle A : lissage exponentiel

La première tentative de prévision des séries temporelles a utilisé des méthodes de lissage exponentiel [Hyndman et al., 2002]. Le premier modèle testé est plus précisément un modèle de prévision par lissage exponentiel double. Ces prévisions seront des moyennes pondérées d'observations antérieures : modèle de Holt-Winters, dans lequel le CA sera prédit grâce à ses valeurs antérieures (un modèle de lissage exponentiel) [Holt, 2004,Winters, 1960].

Ces modèles sont beaucoup utilisés dans la littérature, voir par exemple [Lim and McAleer, 2001]. Dans cet article, ces modèles de lissage exponentiel avaient été utilisés pour prévoir les arrivées de touristes par trimestre en Australie en provenance de Hong Kong, de Malaysia et de Singapour. Nous pouvons aussi citer les travaux de [Holt, 1957] et [Winters, 1960].

Nous posons Y_t une série chronologique observée jusqu'au temps T :

$$\begin{aligned} Y_t &= C_t \times S_t + \epsilon_t && (1) \\ &= (a(t-T) + b)\, S_t + \epsilon_t, && (2) \end{aligned}$$

en supposant que la tendance C_t est linéaire.
Notre objectif est de prédire la série temporelle à un horizon h : $\hat{Y}_t(h)$. Pour se faire, il faudra estimer a, b et S_t :

$$\begin{cases} \hat{a}_t & = (1-\beta)\hat{a}_{T-1} + (1-\beta)(\hat{b}_T - \hat{b}_{T-1}) \\ \hat{b}_t & = \alpha \frac{Y_T}{\hat{S}_{T-P}} + (1-\alpha)(\hat{b}_{T-1} + \hat{a}_{T-1}) \\ \hat{S}_t & = \gamma \frac{Y_T}{\hat{b}_T} + (1-\gamma)\hat{S}_{T-P} \end{cases} \quad (3)$$

Une fois ces valeurs estimées, nous obtenons les prévisions à l'horizon h, de la série Y_t de période P :

$$\hat{Y}_T(h) = \begin{cases} (\hat{a}_T h + \hat{b}_T)\hat{S}_{T+h-P}, & \text{si } 1 \leq h \leq P \\ (\hat{a}_T h + \hat{b}_T)\hat{S}_{T+h-2P}, & \text{si } P+1 \leq h \leq 2P \\ ... \end{cases}$$

2. Modèle B

Nous avons commencé par faire des tests de causalité pour identifier les covariables qui expliquent au mieux la variations du chiffre d'affaires.

Test de causalité

On considère deux séries temporelles $(b_t)_{t\in\mathbb{Z}}$ et $(d_t)_{t\in\mathbb{Z}}$ à valeurs dans $\mathbb{R}$. Nous cherchons à voir si (b_t) cause (d_t) ou inversement, ou ni l'un ni l'autre.
Deux modèles seront alors mis en place :
Modèle M1 :

$$b_t = f_b^1(b_{t-1}) + \epsilon_{b,t}^1 \tag{4}$$
$$d_t = f_d^1(d_{t-1}) + \epsilon_{d,t}^1 \tag{5}$$

Modèle M2 :

$$b_t = f_b^2(d_{t-1}, b_{t-1}) + \epsilon_{b,t}^2 \tag{6}$$
$$d_t = f_d^2(b_{t-1}, d_{t-1}) + \epsilon_{d,t}^2 \tag{7}$$

où $(\epsilon_{b,t}^1)_{t\in\mathbb{Z}}$, $(\epsilon_{d,t}^1)_{t\in\mathbb{Z}}$, $(\epsilon_{b,t}^2)_{t\in\mathbb{Z}}$ et $(\epsilon_{d,t}^2)_{t\in\mathbb{Z}}$ sont des bruits blancs i.i.d., indépendants entre eux et suivent la loi normale.

$\epsilon_{b,t}^1$ (resp. $\epsilon_{d,t}^1$) représente l'erreur de prédiction de b_t (resp. d_t) à partir de son passé.
$\epsilon_{b,t}^2$ (resp. $\epsilon_{d,t}^2$) représente l'erreur de prédiction de b_t (resp. d_t) à partir du passé de b_t et de d_t.
Selon [Granger, 1969], b_t cause d_t si $\mathbb{V}(\epsilon_{d,t}^2) < \mathbb{V}(\epsilon_{d,t}^1)$.

Modèle ARDL (Auto-Regressive Distributed Lag)

Le modèle ARDL [Pesaran et al., 1995] (en français, autorégressif à retard échelonné) est un modèle de régression de séries temporelles. Il sert à étudier l'influence de séries temporelles (variables exogènes) sur une série temporelle (variable endogène). Il est largement utilisé en économétrie. Dans notre cas, le CA est prédit à partir de variables explicatives, soit un indicateur interne à l'EPS soit un indicateur de l'INSEE [Pesaran et al., 1995].

$$Y_t = c + \sum_{i=1}^{r} a_i Y_{t-i} + \sum_{j=0}^{l} b_j X_{t-j} + \epsilon_t \tag{8}$$

avec r le retard sur l'endogène Y et l le retard sur l'exogène X. Ce modèle à retard échelonné, comme son nom l'indique, étudie l'influence des valeurs antérieures de la variable endogène et ceux des variables exogènes sur la variable à expliquer. Dans notre cas, l'ensemble de ces variables sont soit les indices internes soit les indices externes (macroéconomiques) identifiés.

Plusieurs critères peuvent-être utilisés pour trouver les valeurs optimales de retards (r_{opt} et l_{opt}). Dans la littérature économétrique, il existe plusieurs approches de sélection du nombre de pas (retards) optimal, aussi bien pour la variable endogène que pour la variable exogène. D'après les modèles simulés de [Schwert, 1989], vu que nous utilisons des données mensuelles, nous maximisons les retards comme suit :

$$\begin{cases} 1 \preceq r_{opt} \preceq 12(n/100)^{1/4} \\ 0 \preceq l_{opt} \preceq 12(n/100)^{1/4} \end{cases} \tag{9}$$