Réponses et commentaires aux TP 1 et 2

TP1

1. Premier exercice

   (a) Pour générer la série X_1, une méthode est de travailler d'abord vectoriellement en générant la série des temps, puis celle des X_1, et alors de transformer le tout en série chronologique.

   (b) X_1 est une fonction de t complètement déterministe (pas de bruit). Donc la partie déterministe de $X_1(t)$ est $X_1(t)$.

   (c) Comme X_1 est déterministe, $\mathbb{E}X_1(t) = X_1(t)$ et $\mathrm{var}X_1(t) = 0$ (attention à ne pas confondre espérance et variance avec moyenne et variance empirique, que l'on aurait du calculer à partir des données).

   (d) Les variables $X_1(t)$ sont indépendantes. En effet, quelque soit la fonction borélienne f, on a pour tout $n \in \mathbb{N}^*$, pour tout $(t_1, \cdots, t_n)$,

   $$\mathbb{E}\{f(X_1(t_1))f(X_1(t_2)) \times \cdots \times f(X_1(t_n))\} = \mathbb{E}\{f(X_1(t_1)\}\mathbb{E}\{f(X_1(t_2)\} \times \cdots \times \mathbb{E}\{f(X_1(t_n)\}.$$

   Peutêtre que l'on comprend mieux cette indépendance en prenant l'exemple suivant : est-ce que $X = 3$ et $Y = 9$ sont des variables indépendantes? Oui, et ceci même si $Y = X^2$ ou $Y = 3X$ par exemple.

2. Deuxième exercice

   (a) Pour générer la série X_2, on travaille encore vectoriellement en générant le vecteur des temps, puis le vecteur des $(a(k))$, puis celui des $(S(k))$ (obtenu par exemple avec la fonction *rep*) et enfin celui des (ε_k). X_2 s'obtient alors en sommant tout ceci, puis en transformant le vecteur en série chronologique.

   (b) La partie déterministe de X_2 est donnée par $\mathbb{E}X_2(k)$, soit $a(k) + S(k)$.

   (c) Les variables $X_2(k)$ sont indépendantes car $X_2(k) = f_k(\varepsilon_k)$ et les (ε_k) sont indépendantes.

3. Troisième exercice

   (a) Pour générer X_3, on peut utiliser une boucle. On commence par poser $X_0 = 0$ et $X_1 = 0$, puis on peut lancer la boucle en rajoutant à chaque étape le bruit ε_k.

   (b) Contrairement aux séries X_1 et X_2, X_3 n'a pas de tendance ni de saisonnier, ce qui apparaît clairement en lançant plusieurs trajectoires.

   (c) Les variables $X_3(k)$ ne sont pas indépendantes. En effet, on a $X_3(2) = \varepsilon_2$ et $X_3(3) = 0.6\varepsilon_2 + \varepsilon_3$; donc, $\mathrm{cov}(X_3(2)X_3(3)) \neq 0$: $X_3(2)$ et $X_3(3)$ ne sont pas indépendantes. Plus généalement, on peut écrire $X_3(k)$ comme une combinaison linéaire de $(\varepsilon_2, \varepsilon_3, \cdots, \varepsilon_k)$, et ainsi $X_3(k)$ et $X_3(k')$ ne sont pas indépendantes.

TP2

1. On travaille avec une série (*hstart*) réelle, c'est-à-dire non simulée, issue d'un cas concret. On ne sait donc pas à l'avance tout ce que l'on doit retrouver, même si on suspecte par exemple qu'un phénomène saisonnier de période une année va sans doute avoir lieu et que le nombre de déménagements devrait en gros un peu augmenter avec les années car la population augmente et les gens ont de plus en plus tendance à changer de lieu de travail. En revanche, on n'a a priori aucune idée plus précise de la tendance, d'un bruit possible, du saisonnier.

2. Exercices

   (a) on vérifie que plus le degré du polynôme augmente, plus la tendance est proche des données. On chosira donc un degré pas trop faible (pour suivre suffisamment l'allure des données), mais pas trop élevé non plus (pour éviter une trop grande irrégularité de la tendance). On s'arrête quand apparaissent les petites fluctuations dans la tendance, c'est-à-dire pour un degré du polynôme de l'ordre de 6 (l'appréciation de ce degré est pour l'heure assez subjective). En enlevant cette tendance, on s'aperçoit alors de la périodicité annuelle des données, le saisonnier pouvant alors se retrouver en moyennant par mois.

   (b) Les régressions locales (loess et lowess) s'adaptent mieux aux données sans pour autant être trop irrégulière: elles sont donc plus indiquées pour filtrer les données afin de faire apparaître le saisonnier + le bruit. En revanche, pour la prévision, les régressions polynômiales globales restent intéressantes.

   (c) Pour choisir la taille de fenêtre satisfaisante (c'est-à-dire celle qui mettra le mieux en évidence la saisonnalité), on est encore dans l'ordre du subjectif. On veut que la régression suivent suffisament les principales variations de la série chronologique, mais on ne veut pas pour autant que la série lissée par régression soit trop irrégulière (on évite par exemple les petits pics). Dans le cas de la série *hstart*, on obtient ainsi une taille de fenêtre de l'ordre de $a\times$(longueur de x) avec $0.3 \leq a \leq 0.4$.

   (d) Lorsque l'on reprend les trois séries simulées lors du premier TP, le lissage lowess convient bien à la première série (prend en compte la partie fraction rationnelle au bord et la partie linéaire ailleurs), mais une régression linéaire suffit pour la seconde, quant à la troisième, si on prend suffisament de données (500 ou +), on vérifie que la tendance est nulle.