Chapitre 1

Analyse préliminaire

1.1 Processus stochastiques

L'objectif de ce livre est de fournir les outils pour l'étude des quantités qui évoluent dans le temps : population d'un pays, température à un site donné, activité sismique, cours boursier... On se place dans le cadre où les phénomènes ne sont pas déterministes, c'est-à-dire qu'il existe des facteurs physiques, géologiques et/ou socio-économiques, non maîtrisables, et qui rendent l'évolution de la composante mesurée imprévisible. Les processus stochastiques sont tout à fait adaptés pour modéliser de tels phénomènes.

Définition 1.1.1
On appelle ***processus stochastique*** *une collection de variables aléatoires indexées sur le temps, notées* $(Z_t)_{t\in\mathcal{I}}$. *Ainsi, à chaque instant* $t \in \mathcal{I}$, Z_t *est une variable aléatoire définie sur l'espace de probabilité* $(\Omega, \mathcal{F}, \mathbb{P})$. *Et Chaque observation sur la durée* $\mathcal{I}$, *notée* $z_t = Z_t(\omega)$, *est une réalisation du processus* $(Z_t)_t$.

Généralement, le phénomène sous-jacent évolue en continu sur une durée $\mathcal{I}$ qui est un intervalle ($\mathcal{I} = [a, b]$, avec $0 \leq a < b$). Mais les données ne sont récoltées que ponctuellement, aux instants $t_1, \cdots, t_n \in \mathcal{I}$, avec $0 \leq t_1 < \cdots < t_n$. Les valeurs observées aux instants $t_1, \cdots, t_n$ forment la série chronologique (appelée encore série temporelle) et sont notées $z_1, \cdots, z_n$. Chaque observation z_k est une réalisation de la variable aléatoire Z_{t_k}. On notera $\mathbf{Z}_n = (Z_1, Z_2, ..., Z_n)$ le vecteur aléatoire $(Z_{t_1}, \cdots, Z_{t_n})$ dont est issue la série des observations $z_1, \cdots, z_n$.
On se restreindra au cas où le pas de temps $t_{k+1} - t_k$ est constant, et on supposera que la série ne présente pas de données manquantes, ou du moins que les données manquantes ont pu être comblées, par interpolation linéaire par exemple.

Il faut noter que les observations $z_1, \cdots, z_n$ ne constituent pas un échantillon (observations issues de variables indépendantes de même loi), car les variables $Z_{t_1}, \cdots, Z_{t_n}$ ne sont *a priori* pas indépendantes. Les outils classiques tels que les tests statistiques ou les modèles de régression ne peuvent donc pas être utilisés, en tout cas pas directement.

1.2 Visualisation des données

1.2.1 Chronogramme

Lorsqu'on dispose d'une série chronologique, la première chose à faire, avant de tenter tout calcul, consiste à tracer son évolution, le chronogramme. On parle aussi de trajectoire du processus, au sens où on visualise une de ses réalisations.

Exemple :

On considère la série temporelle ***nidd.annual*** *qui représente le niveau maximal de la rivière Nidd dans le Yorkshire (série constituée de 35 observations) et qui se trouve dans le package* ***evir*** *de* ***R****.*

```
> library(evir)
> data(nidd.annual)
> Nidd <- nidd.annual
> plot(Nidd, type = "l", main = "Trajectoire de la série Nidd")
```

On obtient la Figure 1.1

Dans **R**, la classe `ts` permet de gérer facilement les séries temporelles, grâce à l'attribut `time` qui stocke les dates successives des observations.

Exemple :

Etudions la population (en milliers) d'Australiens résidant dans leur pays. Le recensement est effectué tous les trimestres de juin 1971 à juin 1993.

```
> data(austres)
> class(austres)

[1] "ts"
```

Ces données sont de type ***ts****, avec un attribut* ***frequency*** *qui indique la fréquence annuelle des relevés. La commande* ***time(austres)*** *fournit les temps d'observation. Il s'agit ici de données trimestrielles.*

```
> frequency(austres)
[1] 4
> time(austres)[1:14]
 [1] 1971.25 1971.50 1971.75 1972.00 1972.25 1972.50 1972.75 1973.00
 [9] 1973.25 1973.50 1973.75 1974.00 1974.25 1974.50
```

Attention, on ne peut visualiser correctement les temps d'observation que si le logiciel ***R*** *fournit un affichage avec suffisamment de digits (au moins 6), car sinon il n'affiche que des arrondis. On vérifie la précision d'affichage avec la commande* `getOption("digits")`*, et on peut la modifier avec la commande* `options(digits=6)`*.*

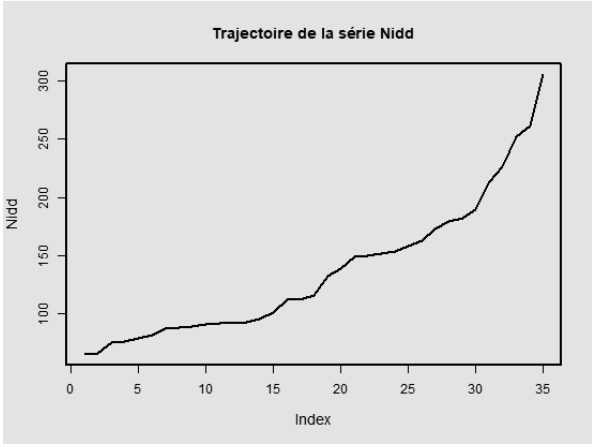

FIGURE 1.1 – Trajectoire de la série `Nidd`

Les données `austres` *étant de type* `ts`*, le graphique s'ajuste automatiquement sur les années.*

```
> plot(austres, ylab = " ", main = "Nb habitants Australie")
```

On obtient la Figure 1.2

La série `nidd.annual` n'étant pas de classe `ts`, on a tracé le chronogramme de la Figure 1.1 en utilisant l'argument `type= "l"` de la fonction `plot`. On peut s'en affranchir si on utilise à la place la fonction `ts.plot`. La syntaxe suivante produit le même graphique que la Figure 1.1.

```
> ts.plot(Nidd, main = "Trajectoire de la série Nidd")
```

Mais généralement, on ne se contente pas de tracer les chronogrammes des séries. On les étudie en détail afin de les modéliser. Il est alors plus commode de transformer en classe `ts` les séries qui ne le sont pas.

```
> Nidd <- as.ts(Nidd, frequency = 1)
```

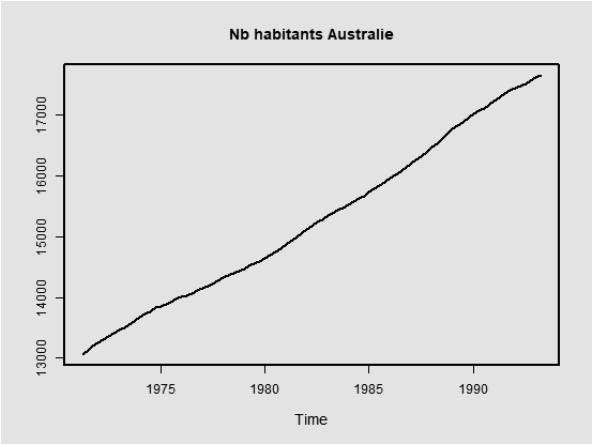

FIGURE 1.2 – Trajectoire de la série `austres`

1.2.2 Décomposition d'une série temporelle

Quelques exemples

Les Figures 1.1 et 1.2 montrent des quantités qui ont une forte tendance à croître avec le temps, modulo quelques variations pour la série `nidd.annual`. Il en est de même pour la série `wagesuk` du package `fma`, avec des variations encore plus marquées.

```
> library(fma)
> data(wagesuk)
> plot(wagesuk, main = "Salaire journalier moyen en Angleterre",
+     ylab = " ")
```

On obtient la Figure 1.3

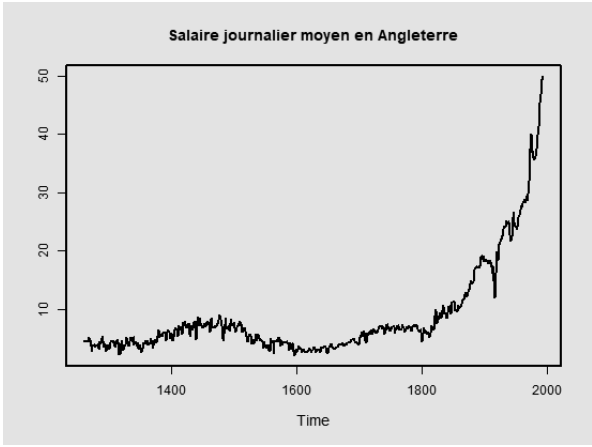

FIGURE 1.3 – Trajectoire de la série `wagesuk`

La tendance d'une série représente son orientation à long terme. D'autres composantes entrent aussi parfois en jeu, comme dans l'exemple suivant :

```
> data(nottem)
> plot(nottem, ylab = " ", main = "Température à Nottingham")
```

On obtient la Figure 1.4

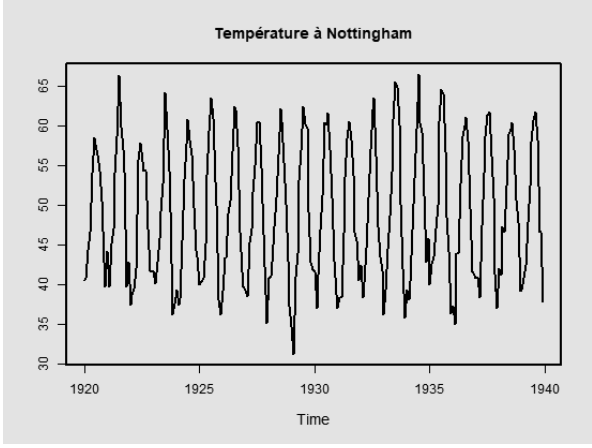

FIGURE 1.4 – Trajectoire de la série `nottem`

Cette fois-ci il ne semble pas y avoir de tendance, ni à croître, ni à décroître. Les données sont périodiques, avec une faible variabilité. Les séries issues de relevés mensuels (c'est le cas de la série `nottem`) ou trimestriels présentent généralement une telle structure, et il peut s'y ajouter simultanément une tendance.

```
> data(uselec)
> plot(uselec, main = "Production d'électricité aux USA",
+     ylab = " ")
```

On obtient la Figure 1.5

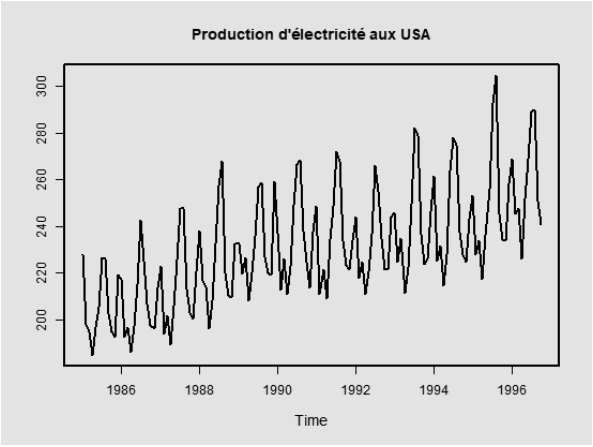

FIGURE 1.5 – Trajectoire de la série `uselec`

Décomposition type

La plupart des séries présentent une structure commune constituée d'une composante déterministe (la tendance, la saisonnalité) et d'une composante aléatoire (les variations). En fait on décompose une série $(Z_t)_t$ de la façon suivante :

$$Z_t \;=\; m_t \,+\, s_t \,+\, B_t\,,$$

avec

- m_t la tendance généralement définie par une fonction polynomiale du temps,
- s_t la composante périodique (de période r),
- et B_t le bruit (variations de faible intensité, de courte durée et de nature aléatoire).

Pour une bonne identifiabilité des termes, on suppose que $(B_t)_{t\in\mathcal{I}}$ est centré et que

$$s_{t+1} + \cdots + s_{t+r} \;=\; 0\,.$$

La fonction `decompose()` de **R** permet d'obtenir cette décomposition de façon automatique.

Exemple :

```
> plot(decompose(austres))
```

On obtient la Figure 1.6

A cause du zoom sur chaque sous-graphique, toutes les composantes semblent jouer un rôle équivalent. Seule la lecture des échelles permet d'identifier la ou les composantes prépondérantes (la tendance ici).

La méthode de décomposition implémentée dans la fonction `decompose()` a été proposée par Kendall et Stuart (voir [58], pages 410 à 414) et repose sur des calculs de moyennes mobiles.

Définition 1.2.1
Soit $A \;:\; [\,a_{-q_1},\cdots,a_{-1},\underline{a_0},a_1,\cdots,a_{q_2}\,]$ *un ensemble de réels, appelé un* ***filtre*** *ou une* ***moyenne mobile****. Soit* $S \;:\; z_1,z_2,\cdots,z_n$ *une série observée. Appliquer le filtre* A *à la série* S*, c'est calculer les valeurs*

$$A(z_k) \;=\; a_{-q_1}\,z_{k-q_1} \,+\, \cdots \,+\, a_{-1}\,z_{k-1} \,+\, a_0\,z_k \,+\, a_1\,z_{k+1} \,+\, \cdots \,+\, a_{q_2}\,z_{k+q_2}\ ,$$

pour tout $k \;=\; q_1+1,\, q_1+2,\, \cdots n-q_2$.

q_1+q_2+1 *est appelé l'ordre du filtre* A *et on dit que le filtre est*

- *un* ***lissage*** *si* $\displaystyle\sum_{j=-q_1}^{q_2} a_j \;=\; 1$ *;*
- *une* ***moyenne arithmétique*** *si* $\forall j,\ a_j \;=\; \frac{1}{q_1+q_2+1}$ *;*
- *un filtre* ***centré*** *si* $q_1 = q_2$ *;*
- *un filtre* ***symétrique, centré***
  si $q_1 = q_2 = q$ *et si* $\forall\, j = 1,\cdots,q,$ *on a* $a_{-j} \;=\; a_j$.

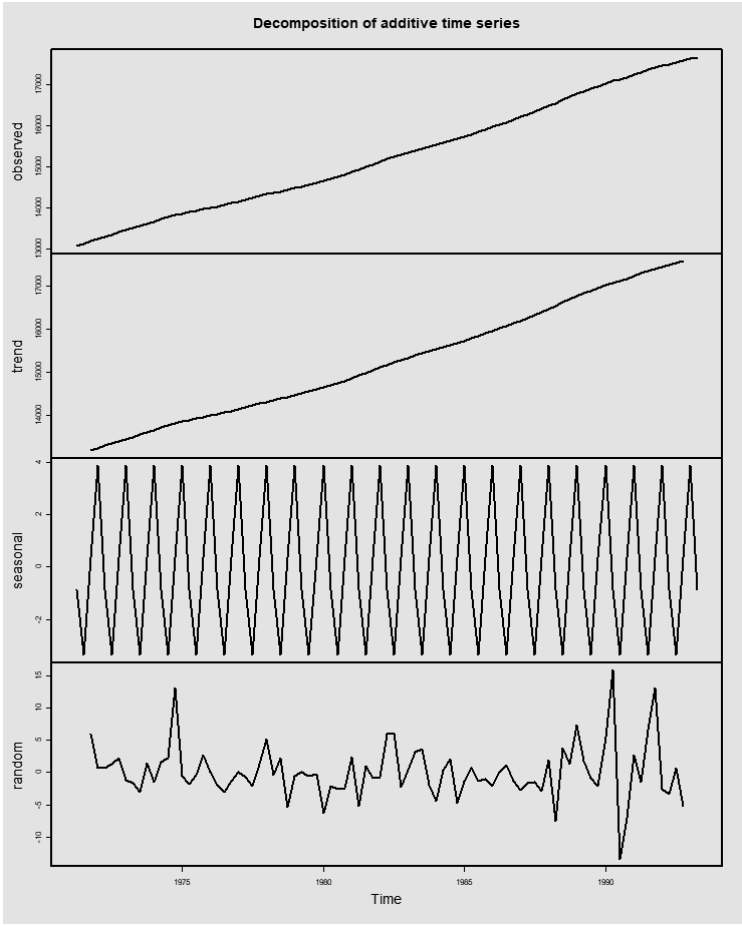

FIGURE 1.6 – Décomposition de la série `austres`

Théorème 1.2.2
Soit $A : [a_{-q_1}, \cdots, a_{-1}, \underline{a_0}, a_1, \cdots, a_{q_2}]$ *un filtre et* $S : z_1, z_2, \cdots, z_n$ *une série observée.*

Le filtre A conserve les séries constantes $\iff \sum_{j=-q_1}^{q_2} a_j = 1.$

Si de plus A est symétrique et centré, alors il conserve aussi les tendances linéaires :

$$A(ak+b) = ak+b\,.$$