– Texte –
Le niveau du lac Huron

Mots-clefs : Séries temporelles, processus autorégressifs, test de normalité.

1 Le problème et sa modélisation

Le lac Huron est l'un des cinq grands lacs d'Amérique du Nord. Son niveau a été relevé chaque année de 1875 à 1972 (cf. Figure 1). Le but est de modéliser cette série temporelle de façon aussi fine que possible. On fait intervenir ici un processus autorégressif, modèle que l'on retrouve dans de nombreux domaines : en démographie pour la dynamique des populations, en économétrie pour la prédiction d'indices ou encore en automatique via le filtrage de Kalman.

A partir du site `http://w3.bretagne.ens-cachan.fr/math/OrauxBlancs/`, il suffit d'enregistrer le fichier `niveauhuron.txt` dans le répertoire où vous ouvrez Scilab, puis de taper : `>x=fscanfMat('niveauhuron.txt')` pour importer le jeu de données.

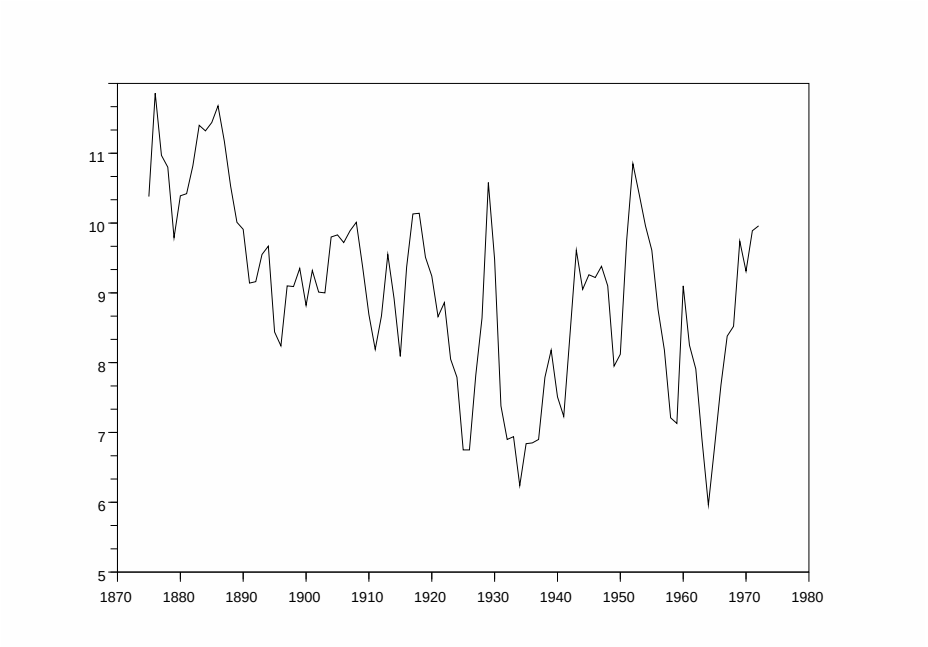

Figure 1 – Niveau du lac Huron entre 1875 et 1972.

2 Moindres carrés et test d'indépendance

Notons $(X_t)_{1\leqslant t\leqslant 98}$ le niveau du lac au cours du temps. La tendance décroissante incite à proposer comme modèle :

$$\forall t \in \{1,\dots,98\} \qquad X_t = a_0 + a_1 t + Y_t.$$

Une idée naturelle est d'estimer a_0 et a_1 par la méthode des moindres carrés, c'est-à-dire de déterminer :

$$(\hat{a}_0, \hat{a}_1) = \arg\min_{(a_0,a_1)} \sum_{t=1}^{98} (X_t - a_0 - a_1 t)^2.$$

On s'intéresse maintenant à la série temporelle des résidus centrés $(Y_t)_{1\leqslant t\leqslant 98}$. En particulier, on aimerait savoir s'ils correspondent à un bruit blanc. On effectue pour cela un test appelé test des extrema et dû à Bienaymé.

L'idée de ce test est la suivante : on considère une suite $(X_i)_{1\leqslant i\leqslant n}$ de variables aléatoires et on compte le nombre d'extrema, c'est-à-dire

$$T_n = \#\left\{i,\ 2 \leqslant i \leqslant n-1\ :\ (X_i - X_{i-1})(X_i - X_{i+1}) > 0\right\}.$$

On a alors le résultat suivant.

Théorème : *Si $n \geqslant 3$ et si $X_1,\dots,X_n$ sont des variables aléatoires i.i.d. de fonction de répartition continue, alors*

$$\sqrt{n}\left(\frac{T_n}{n} - \frac{2}{3}\right) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \frac{8}{45}\right).$$

Preuve. Pour tout j entre 2 et $(n-1)$, on note I_j la variable aléatoire :

$$I_j = \mathbb{1}_{\{X_j < \min(X_{j-1}, X_{j+1})\}} + \mathbb{1}_{\{X_j > \max(X_{j-1}, X_{j+1})\}}.$$

On a donc $T_n = \sum_{j=2}^{n-1} I_j$. Les I_j sont identiquement distribuées et forment une suite 2-dépendante, puisque I_j et I_k sont indépendantes dès que $|k-j| > 2$. Notons $\gamma(0) = \mathrm{V}(I_2)$, $\gamma(1) = \mathrm{Cov}(I_2, I_3)$ et $\gamma(2) = \mathrm{Cov}(I_2, I_4)$ les coefficients d'autocovariance. Pour le premier, on a :

$$\gamma(0) = \mathbb{E}[I_2^2] - \mathbb{E}^2[I_2] = \mathbb{E}[I_2] - \mathbb{E}^2[I_2],$$

Pour voir que $\mathbb{E}[I_2] = 2/3$ il suffit de remarquer que le triplet (X_1, X_2, X_3) peut être ordonné de $3! = 6$ façons différentes, chaque configuration étant équiprobable et 4 d'entre elles correspondant à un point pivot en X_2. Les calculs de $\gamma(1)$ et $\gamma(2)$ font intervenir le même raisonnement. On peut alors appliquer le théorème central limite suivant (admis)

pour les suites de variables (X_n) centrées, identiquement distribuées et m-dépendantes : en notant $v_m = \gamma(0) + 2\sum_{k=1}^{m}\gamma(k)$ et $\bar{X}_n = (\sum_{i=1}^{n} X_i)/n$, on a

$$\sqrt{n}\bar{X}_n \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, v_m).$$

■

En particulier, une faible valeur de T_n signifie que la série (X_n) fluctue moins vite que prévu et indique une corrélation entre valeurs voisines. Le résultat ci-dessus permet de construire un test d'hypothèse très simple basé sur les quantiles de la loi normale.

Sur notre exemple, le test amène à rejeter l'hypothèse d'indépendance des résidus $(Y_t)_{1\leqslant t\leqslant 98}$. Le tracé du nuage (Y_t, Y_{t+1}) corrobore ce résultat, puisqu'il suggère une relation linéaire entre Y_t et Y_{t+1}.

3 Processus autorégressif

Schématiquement, on adopte donc le modèle suivant pour les résidus :

$$\forall i \in \{0, \dots, n-1\} \qquad Y_{i+1} = rY_i + \sigma Z_{i+1} \qquad (A.R.)$$

où (Z_n) est une suite de variables aléatoires i.i.d. centrées réduites indépendantes de Y_0. On dit dans ce cas que (Y_n) est un processus autorégressif d'ordre 1. L'estimation de r peut à nouveau se faire par la méthode des moindres carrés :

$$\hat{r}_n = \frac{\sum_{i=0}^{n-1} Y_i Y_{i+1}}{\sum_{i=0}^{n-1} Y_i^2}.$$

La variance σ^2 est alors estimée par :

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=0}^{n-1}(Y_{i+1} - \hat{r}_n Y_i)^2.$$

Dans notre exemple, l'hypothèse d'indépendance est acceptée pour les nouveaux résidus (Z_t) par le test des extrema. Mieux, on aimerait savoir s'ils suivent une loi normale centrée réduite. On effectue pour cela un test de Kolmogorov-Smirnov.

Soit donc de façon générale $(X_i)_{1\leqslant i\leqslant n}$ une suite de variables i.i.d., dont on note F la fonction de répartition, supposée continue et connue. La fonction de répartition empirique F_n de l'échantillon est définie pour tout réel x par :

$$F_n(x) = \frac{1}{n}\#\{i, 1 \leqslant i \leqslant n \ : \ X_i \leqslant x\}.$$

On note alors :

$$K_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|,$$

et on a :

$$K_n \xrightarrow[n \to +\infty]{\mathcal{L}} K,$$

où K suit la loi de Kolmogorov-Smirnov, dont les quantiles sont connus. En particulier $\mathbb{P}(K \geqslant 1.22) \approx 0.1$.

L'application de ce test aux résidus (Z_t) amène à accepter l'hypothèse de gaussianité, que nous ferons donc dans la suite de cette section. On considère ainsi le modèle autorégressif suivant :

$$\forall i \in \{0, \ldots, n-1\} \qquad Y_{i+1} = rY_i + \sigma Z_{i+1} \qquad (A.R.G.)$$

où $|r| < 1$ et $Y_0 \sim \mathcal{N}(0, s^2)$ est indépendante de la suite (Z_n) de variables aléatoires i.i.d. gaussiennes centrées réduites. Avec les mêmes estimateurs $\hat{r}_n$ et $\hat{\sigma}_n^2$ que précédemment, on peut alors montrer les résultats suivants.

Propriétés : *Pour le modèle autorégressif (A.R.G.) ci-dessus :*

1. *Les estimateurs $\hat{r}_n$ et $\hat{\sigma}_n^2$ sont les estimateurs au maximum de vraisemblance.*
2. *(Y_n) converge en loi vers une variable gaussienne centrée de variance $\sigma^2/(1-r^2)$.*
3. *Les estimateurs $\hat{r}_n$ et $\hat{\sigma}_n$ convergent en probabilité vers r et σ.*
4. *$\sqrt{n}(\hat{r}_n - r)$ et $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2)$ convergent en loi vers des variables gaussiennes de lois respectives $\mathcal{N}(0, 1-r^2)$ et $\mathcal{N}(0, 2\sigma^4)$.*

Preuve. Les deux premiers points ne posent pas problème. Pour la suite, on réécrit $\hat{r}_n$ comme suit :

$$\hat{r}_n = r + \sigma \frac{\sum_{i=0}^{n-1} Y_i Z_{i+1}}{\sum_{i=0}^{n-1} Y_i^2},$$

et on remarque que la suite $(M_n)_{n \geqslant 0}$ définie par $M_0 = 0$ et

$$\forall n > 0 \qquad M_n = \sum_{i=0}^{n-1} Y_i Z_{i+1}$$

est une martingale par rapport à la filtration $(\mathcal{F}_n)$ engendrée par les variables aléatoires $(Z_i)_{1 \leqslant i \leqslant n}$. Sa variation quadratique $(\langle M \rangle_n)_{n \geqslant 0}$ est donnée par $\langle M \rangle_0 = 0$ et

$$\forall n > 0 \qquad \langle M \rangle_n = \sum_{i=0}^{n-1} Y_i^2.$$

Supposons pour simplifier que $Y_0 = 0$, alors la formule de Cauchy-Schwarz permet d'établir la majoration :

$$Y_n^2 \leqslant \frac{\sigma^2}{1-|r|} \sum_{i=1}^{n} |r|^{n-i} Z_i^2.$$

Si on pose $L_n = \sum_{i=1}^{n-1} Z_i^2$, on a $L_n = \mathcal{O}(n)$ p.s. par la loi des grands nombres, d'où par l'inégalité ci-dessus $\langle M \rangle_n = \mathcal{O}(L_n) = \mathcal{O}(n)$ p.s. Pour la suite, la décomposition suivante sera utile

$$(1-r^2)\langle M \rangle_n = \sigma^2 L_n + 2r\sigma M_{n-1} - r^2 Y_{n-1}^2.$$

Or $\mathbb{E}[L_n] = \mathcal{O}(n)$, $\mathbb{E}[M_{n-1}] = 0$ et $\mathbb{E}[Y_{n-1}^2] = o(n)$, donc $\langle M \rangle_n / n^2$ tend vers 0 dans L^1. Puisque $\mathbb{E}\langle M \rangle_n = \mathbb{E}[M_n^2]$, ceci entraîne que M_n/n tend vers zéro en probabilité. Par suite, du fait que Y_{n-1}^2/n tend vers zéro en probabilité et de l'égalité ci-dessus, on déduit que $\langle M \rangle_n / n$ tend en probabilité vers $\sigma^2/(1-r^2)$. Le quotient $M_n/\langle M \rangle_n$ tend donc en probabilité vers zéro et $\hat{r}_n$ vers r. Le dernier point vient du théorème central limite pour les martingales de carré intégrable.

■

4 Fonction de corrélation

Dans l'étude des séries temporelles, on considère souvent que certaines quantités ne varient pas au cours du temps, typiquement la variance et les corrélations. On parle alors de processus stationnaires (au second ordre).

Définition : *On dit qu'un processus (X_n) centré est stationnaire si :*
(i) X_n est de carré intégrable pour tout n ;
(ii) Pour tout $h \in \mathbb{N}$, $\mathrm{Cov}(X_n, X_{n+h})$ ne dépend que de h.
Pour tout $h \in \mathbb{N}$, on note alors $\gamma(h) = \mathrm{Cov}(X_n, X_{n+h})$ la fonction de covariance et $\rho(h) = \gamma(h)/\gamma(0)$ la fonction de corrélation.

Ainsi le processus autorégressif gaussien (Y_n) de la section précédente est stationnaire si et seulement si $s^2 = \sigma^2/(1-r^2)$. C'est l'hypothèse que nous ferons pour le processus (Y_t) associé à l'exemple du lac Huron. Lorsque l'on dispose simplement d'un échantillon, les fonctions de covariance et de corrélation sont estimées directement à partir de celui-ci :

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=0}^{n-h} X_i X_{i+h} \qquad \text{et} \qquad \hat{\rho}(h) = \frac{\hat{\gamma}(h)}{\hat{\gamma}(0)}.$$

Ces estimateurs sont biaisés, mais pertinents pour n assez grand et h assez petit par rapport à n. Dans notre exemple, on ne s'intéressera ainsi qu'aux 20 premiers coefficients $\rho(h) = r^h$ et $\hat{\rho}(h)$.

Supposons donc $h \ll n$ et notons w_h le coefficient donné par la formule de Bartlett :

$$w_h = \frac{(1+r^2)(1-r^{2h})}{1-r^2} - 2hr^{2h}.$$

On peut alors démontrer un résultat de convergence en loi.

Proposition : *Sous les hypothèses précédentes, on a*

$$\sqrt{n}(\hat{\rho}(h) - \rho(h)) \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0, w_h).$$

Sur notre exemple, ce résultat permet de voir si le modèle de processus autorégressif d'ordre 1 est bien adapté aux données. Pour tout h entre 1 et 20, on regarde si le coefficient de corrélation théorique $\rho(h) = r^h$ est dans l'intervalle de confiance à 95% autour de $\hat{\rho}(h)$:

$$\hat{\rho}(h) - \frac{1.96\sqrt{w_h}}{\sqrt{n}} \leqslant \rho(h) \leqslant \hat{\rho}(h) + \frac{1.96\sqrt{w_h}}{\sqrt{n}}.$$

5 Suggestions

On pourra :
- démontrer le théorème de la section 2 ;
- construire le test des extrema ;
- construire le test de Kolmogorov-Smirnov ;
- détailler la preuve des premières propriétés pour le modèle de la section 3 ;
- représenter les fonction de corrélation empirique, théorique et les intervalles de confiance.

