Une analyse bayésienne d'hydrologie fréquentielle selon un modèle de copule bivariée

Eric Parent, Anne-Catherine Favre

Une analyse bayésienne d'hydrologie fréquentielle selon un modèle de copule bivariée

Éric Parent & Anne-Catherine Favre

Équipe Modélisation, Risque, Statistique, Environnement (MORSE), ENGREF, AgroParisTech

Chaire en Hydrologie statistique Hydro-Québec/CRSNG, INRS-ETE

Dans plusieurs secteurs de la statistique appliquée, comme l'hydrologie, l'analyse d'événements multivariés est d'un intérêt particulier. Les concepteurs de barrages doivent choisir la taille de leurs structures hydrauliques selon les débits des rivières qui augmentent pendant les inondations de printemps. L'inondation est le plus souvent caractérisée par trois quantités principales : le débit de pointe, le volume et la durée. Comme ces variables sont dépendantes, trois analyses univariées exécutées d'une manière indépendante ne sont pas capables de proposer une évaluation complète de la probabilité d'un événement dommageable rare. De plus une analyse monovariée peut mener ou à une sous-estimation du risque(De Michele et al., 2005). Nous analysons 47 années de couples pointe-volume de crues de la rivière Romaine au Québec à l'aide d'un modèle paramétrique de copule. Les marges sont modélisées soit par une loi gamma ou normale, et la dépendance par une famille paramétrique de copules (Arch12 et Clayton). L'inférence et la sélection de modèles sont alors réalisées sous la perpective bayésienne. Cette approche permet naturellement de s'appuyer sur la théorie de la décision statistique (Berger, 1985) au moyen de l'optimisation d'une fonction de coût, ce qui présente un intérêt particulier en ingéniérie hydraulique si l'on se place dans un contexte multivarié.

Mots-Clès : Analyse bayésienne, Copule, hydrologie fréquentielle, ingéniérie de l'Environnement

In several applied statistical areas, like hydrology, the analysis of multivariate events is of particular interest. The dams designers have to size their hydraulic structures according of the river flows which increase during the spring floods. Flood is more often characterized by three main quantities: peak, volume and duration. As these variables are correlated, three univariate analysis carried out independently are not able to give a complete assessment of the event of interest's probability of occurrence. Moreover a univariate analysis can lead either to an underestimation of risk. We study 47 years of a peak/volume dataset for the Romaine river with a parametric copula model. The margins are modelled with a normal or gamma distribution and the dependance is depicted through a parametric family of copulas (Clayton or Arch 12). Parameter joint inference and model selection is performed under the Bayesian paradigm. This approach allows to rely on the theory of statistical decision theory by means of a utility function optimization , which is of particular interest for hydraulic engineering in a multivariate context.

1 Introduction

En ingéniérie hydraulique, le concept de période de retour est depuis longtemps le concept-clé pour dimensionner un ouvrage de protection : on recommande aux ingénieurs de choisir une crue de projet dite centenale (période de retour = 100 ans) lorsque les enjeux sont modérés , comme pour une digue de protection d'un petit bourg par exemple, ou bien d'aller jusqu'à la crue millénale ou décamillenale lorsque les conséquences de défaillance de l'ouvrage sont catastrophiques (par exemple le canal évacuateur de crue d'un grand barrage hydro-électrique). La période de retour $T(y)$ du quantile y d'un événement aléatoire Y de fonction de répartition F est définie comme

$$T(y) = \frac{1}{1 - F(y)}$$

C'est la durée moyenne d'attente sous l'hypothèse *iid* pour qu'un événement Y ayant dépassé le seuil y à l'instant origine, se reproduise avec la même intensité. Considérons un événement mesurée annuellement (par exemple le debit maximum annuel du débit journalier d'une rivière) : une valeur correspondant à une période de retour 100 ans signifie donc que l'événement a une chance sur 100 d'être dépassé, c'est donc le quantile 0.99.

Cette pratique de l'ingéniérie est contestable à deux égards. D'abord la répartition F est en général inconnue et l'on se contente d'une estimée $\hat{T}(y)$ de la période de retour : il est donc important de quantifier les incertitudes d'estimation et de comprendre leur influence ; elles ont en général de très lourdes conséquences puisque la décision (qui s'appuie sur cette inférence) concerne généralement des événements rares, c'est à dire une extrapolation du modèle hors de sa gamme de fonctionnement ordinaire. Ensuite la situation multivariée fait disparaître le lien biunivoque entre quantile et fonction de répartition (Salvadori and De Michele, 2007). Or, une crue est un événement multivarié : quand on cherche à s'en protéger, il faut à la fois considérer le volume, le débit au pic de l'événement et de la durée de l'éventuelle submersion de la zone inondée, et une infinité de triplets (volume, pic, durée) correspondent à une même probabilité au dépassement.

Cette communication propose les moyens de résoudre ces deux difficultés en réalisant une analyse hydrologique fréquentielle décisionnelle avec un modèle paramétrique de copule bivariée. Le contexte bayésien permet de mettre facilement en évidence les incertitudes d'estimation des paramètres et les crédibilités relatives des modèles en compétition. Sur l'exempel de la rivière Romaine, la copule semble introduire de fortes corrélations entre les estimées des grandeurs caractéristiques des lois marginales de chacun des modèles. La théorie statistique de la décision se présente comme une extension naturelle de l'analyse bayésienne réalisée: le dimensionnement de la protection ne passe plus par le calcul d'une crue de projet associée à une forte période de retour, mais s'obtient ici comme la solution d'un problème de décision sous incertitude pour lequel nous proposons une fonction de dommage qui tienne compte de la nature multivariée de l'événement contre lequel il faut se prémunir.

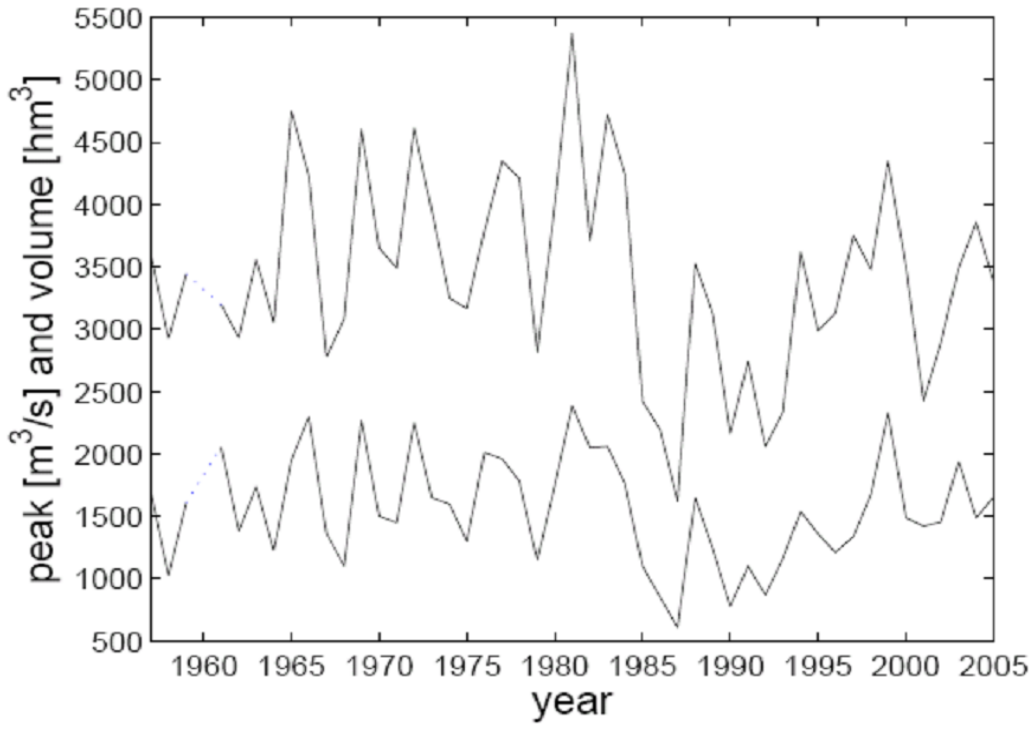

Figure 1: Série chronologique des pointes de débit et de volumes de crue de printemps de la rivière Romaine de 1957 à 2003

2 Un modèle de copule sur les données de la Romaine

On dispose de couples (x_t, y_t) pointe-volume de crues sur les $t = 1$ *à* 47 ans de données de la Romaine et les données de 1960 sont manquantes. La figure 1 montre une co-evolution marquée des deux variables d'intérêt sur cette période (Genest et al., 2007).

Les séries temporelles univariées sont stationnaires et ne montrent pas d'autocorrélation, elles peuvent être assimilées en première à des échantillons *iid* de distributions univariées. Nous effectuons d'abord l'estimation bayésienne conjointe des paramètres $(\theta_x, \theta_y, \lambda)$ d'un modèle *complet* de copule, constitué par les lois de ses marges (paramétrées respectivement par les quantités θ_x et θ_y) et la structure de dépendance de la copule décrite par le paramètre λ. Par la suite, nous porterons une attention particulière à des modèles paramétriques formés par l'association de deux structures marginales (normale ou gamma) et de deux types de copules (Clayton ou Arch12).

2.1 Des marges Normales ou gamma

On considère des lois Normales ou gamma pour les distributions des marges x (pointes) et y (volumes). Pour la loi gamma, on peut trouver un paramétrage en moyenne/écart type, ce qui sera utile pour s'appuyer sur les mêmes informations a priori quand, dans l'approche bayésienne, on spécifiera les distributions a priori pour les paramètres marginaux gamma ou normal. Pour spécifier ces priors, on considère ici une approche très empirique: comme l'année 1960 est manquante, on exclut les trois premières années $(1957 - 1959)$ de l'échantillon des pointes et des volumes et, on ajuste un prior conjugué (gamma Normal) peu informatif ayant les même caractéristiques statistiques marginales que ces trois années, et compte tenu de ce contexte peu informatif, on fait l'hypothèse d'indépendence a priori entre les jugements probabilistes portant sur les pointes de crues et ceux à propos de leurs volumes, soit en utilisant la notation [] de Gelfand (Gelfand and Smith, 1990) pour les distributions de probabilités: $[\theta_x, \theta_y] = [\theta_x] \times [\theta_y]$.

2.2 Une copule paramétrique Clayton ou Arch 12

On pose $x = F_{\theta_x}(u)$ et $y = G_{\theta_y}(v)$ les fonctions de répartition des marges. Si on prend des copules (Sklar, 1959) $C(u, v, \lambda)$ archimédiennes , les travaux préliminaires (Genest et al., 2007) sur la Romaine suggèrent de choisir

1. soit une forme de type *Clayton* $C(u, v; \lambda) = \varphi^{-1}(\varphi(u) + \varphi(v))$, $\lambda \geq 1$, avec $\varphi(u) = \frac{(u)^{-\lambda}-1}{\lambda}$
2. soit une forme dite *Arch12* avec $\varphi(u) = (u^{-1} - 1)^{\lambda}$.

Les mesures de concordance, notamment le τ de Kendall donnent une reparamétrisation plus explicite des structures monoparamétriques courantes de copules . En effet, ce coefficient τ, contrairement à d'autres comme le coefficient de corrélation linéaire de Pearson, ne dépend pas des lois marginales. Il s'exprime uniquement en fonction de la copule. Il est donc révélateur de la structure de dépendance. De plus, pour nombre de copules paramétriques, ce τ de Kendall peut être exprimé facilement en fonction du paramètre de la copule. Pour la copule de Clayton, τ de Kendall est tel que $\lambda(\tau) = \frac{2*\tau}{1-\tau}$ tandis que pour Arch12, on a $\lambda(\tau) = \frac{2}{3(1-\tau)}$. On sait que $-1 \leq \tau \leq 1$ et l'on connait la loi normale asymptotique de τ sous l'hypothèse nulle que l'on pourrait prendre comme prior informatif pessimiste dans un contexte d'inférence bayésienne. Néanmoins, si on suppose une concordance *a priori* positive , il est commode mais réaliste de prendre un prior uniforme $[\tau] = 1$ sur $(0, 1)$.

3 Analyse bayésienne

En s'appuyant sur la décomposition du modèle de copule, la loi *a posteriori* conjointe des inconnues $[\mathbf{x}, \mathbf{y}|\theta_x, \theta_y, \lambda]$ fait apparaître la vraisemblance propre à la copule $[\mathbf{u}, \mathbf{v}|\theta_x, \theta_y, \lambda] = \prod_{i=1}^{n} \frac{\partial^2 C(F(x_i), G(y_i)}{\partial u \partial v}$ et les vraisemblances marginales $[\mathbf{x}|\theta_x][\mathbf{y}|\theta_x] = \prod_{i=1}^{n} \frac{\partial F(x_i)}{\partial x} \frac{\partial G(y_i)}{\partial y}$. Il vient donc

$$[\theta_x, \theta_y, \lambda|\mathbf{x}, \mathbf{y}] = \frac{[\mathbf{u}, \mathbf{v}|\theta_x, \theta_y, \lambda] \times [\mathbf{x}|\theta_x][\mathbf{y}|\theta_y] \times [\theta_x, \theta_y, \lambda]}{[\mathbf{x}, \mathbf{y}]}$$

L'hypothèse d'indépendances des priors des marges et de la copule s'écrit $[\theta_x, \theta_y, \lambda] = [\theta_x] \times [\theta_y] \times [\lambda]$. Si on écrit $[\mathbf{x}|\theta_x] \times [\theta_x]$ comme $[\mathbf{x}] \times [\theta_x|\mathbf{x}]$, le posterior devient:

$$[\theta_x, \theta_y, \lambda|\mathbf{x}, \mathbf{y}] = \frac{[\mathbf{x}][\mathbf{y}]}{[\mathbf{x}, \mathbf{y}]} \times [\mathbf{u}, \mathbf{v}|\theta_x, \theta_y, \lambda] \times ([\lambda][\theta_x|\mathbf{x}][\theta_y|\mathbf{y}]) \tag{1}$$

La forme de l'équation (1) suggère naturellement un algorithme d'inférence fondé sur un échantilonnage d'importance. La loi d'importance sera construite en assemblant les tirages dans le prior de λ en conjonction avec les lois *a posteriori* (en considérant les marginales indépendantes $[\theta_x|\mathbf{x}]$ and $[\theta_x|\mathbf{y}]$. Comme le rapport $\frac{[\mathbf{x}][\mathbf{y}]}{[\mathbf{x},\mathbf{y}]}$ ne dépend pas des paramètres, les poids d'importance seront simplement fournis par la vraisemblance de la copule:

Ainsi, la figure 2 donne la loi jointe *a posteriori* des 5 paramètres $\alpha_x, \beta_x, \lambda, \sigma_y, \mu_y$ correspondant à une marge gamma en x , une Normale pour y et une copule de Clayton pour relier les deux marges. L'inférence bayésienne montre ici que tous les paramètres sont *a posteriori* fortement corrélés spécialement $\theta_x|x, y$ et $\theta_y|x, y$, conséquence évidente de la liaison introduite par la copule. Ce résultat nous met en garde contre une pratique courante d'inférence classique, qui consisterait à estimer séparément les marges, par un estimateur du maximum de vraisemblance par exemple, puis à ajuster ensuite la copule, par une technique non paramétrique commode, par exemple en s'appuyant sur une estimation via le tau de Kendall. On peut comparer les intervalles de crédibilités obtenus quand l'inférence est conduite de façon séparée et quand elle est menée conformément à la formule 1 : quoique le lien dû à la copule introduise de forte corrélations, il ne semble pas changer de façon drastique la moyenne *a posteriori* de chacun des paramètres du modèle. Les variances *a posteriori* ont tendance à étre plus petites quand on considère ensemble les deux séries des volumes et pointes, puisque cela conditionne les inférences par une information plus riche. Enfin, la moyenne *a posteriori* du paramètre de la copule de Clayton dépend des formes de distribution choisies pour chacune des marges. Ce comportement *a posteriori* avec fortes liaisons peut avoir d'importantes répercussions en termes de sélection de modèles ou d'aide au dimensionnement d'ouvrages de protection, tels les évacuateurs de crues.

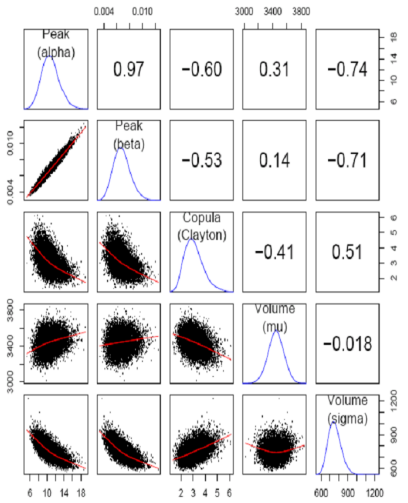

Figure 2: Loi conjointe *a posteriori* avec marge gamma sur les volumes, marge Normale sur les pointes de crues et copule de Clayton. L'inférence a été réalisée par échantillonnage d'importance.

References

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition.

De Michele, C., Salvadori, G., Canossi, M., Petaccia, A., and Rosso, R. (2005). Bivariate statistical approach to check adequacy of dam spillway. *J. Hydrologic Eng.*, 10(1):50–57.

Gelfand, A. and Smith, A. (1990). Sampling based approach to calculating marginal densities. *J. Am. Stat. Ass.*, 85:398–409.

Genest, C., Favre, A.-C., Béliveau, J., and Jacques, C. (2007). Meta-elliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resources Research*, 43(W09401):doi:10.1029/2006WR005275.

Salvadori, G. and De Michele, C. (2007). On the use of copulas in hydrology: theory and practice. *J. Hydrologic Eng.*, 12(4):369–380.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231.