水　利　学　报
2011年4月　SHUILI　XUEBAO　第42卷　第4期

文章编号:0559-9350(2011)04-0483-07

基于Copula函数的水量水质联合分布函数

张　翔[1]，冉啟香[1]，夏　军[2]，宋星原[1]

（1. 武汉大学 水资源与水电工程科学国家重点实验室，湖北 武汉　430072；
2. 中国科学院 地理科学与资源研究所 陆地水循环及地表过程重点实验室，北京　100101）

摘要：由于河流的水量水质是相互联系、相互影响的，如果只进行单一变量的分析，难以全面反映事件的真实特征。本文以淮河蚌埠闸为例，利用Copula函数构造了水量水质的二维和三维(分为对称型和非对称型)联合分布函数，对蚌埠闸的水量水质联合分布频率进行了分析。结果表明，蚌埠闸的Copula函数水量水质联合分布的经验累积频率和理论累积频率的一致性很高，拟合精度很好。说明应用Copula联结函数来构建水量水质联合分布是可行的，可为水量水质综合管理的风险分析提供有效的途径。

关键字：Copula函数；水环境；水量水质联合分布；多变量

中图分类号：X143　　**文献标识码**：A

1　研究背景

在水文频率分析事件中，往往不是单一变量，而是具有两个甚至更多的变量，而且这些变量之间很少是完全独立的，它们是有关系的。如果只进行单一变量的分析，则难以全面反映事件的真实特征。近年来，水文多变量联合分析已经成为相关领域的一个研究热点。

自然界中，环境水文学和水力学中的大部分现象都是随机现象。很多环境和水资源工程需要构建出随机变量的联合频率分布。20世纪80年代以来，淮河流域的水质呈逐年恶化趋势。水污染事故频繁发生，严重影响了工农业生产，破坏了水生生态系统。淮河的水量水质情况是联合在一起的，水量水质相互联系、相互影响，所以在进行该流域的水污染治理中，需要把淮河流域的水量和水质结合起来考虑。目前国内外研究中，把水量和水质情况联合起来考虑的研究还比较少。

本文利用Copula函数构造淮河流域蚌埠闸水量水质的二维及三维联合分布，该函数不需要假设变量是独立的或者正态分布的或者它们具有相同的边缘分布，可以用来描述水文变量之间的相关性结构，能够灵活地构造边缘分布为任意分布的水文变量联合分布。熊立华等[1]应用Copula联结函数对同一河流上下游的两个水文站建立了最大洪水联合分布函数；许月萍等[2]描述多元联合分布和各边缘分布之间的耦合关系；莫淑红等[3]进行了基于Copula函数的河川径流丰枯遭遇分析；闫宝伟等[4]建立了基于Copula函数的一阶非平稳时间序列模型，并与季节性模型进行比较。Copula函数理论上适用面很广，在水文事件多变量分析计算中具有良好的应用前景。

2　Copula函数简述

2.1　定义及性质　Copula函数是定义域为[0,1]均匀分布的多维联合分布函数，它可以将多个随机变量的边际分布连接起来构造联合分布，它的表述如下：

收稿日期：2010-07-15

基金项目：国家水体污染控制与治理科技重大专项(2009ZX07210-006)；高等学校博士点基金(20100141111003)

作者简介：张翔(1969-)，男，北京人，教授，主要从事基于生态水文的可持续水资源管理研究。E-mail：zhangxiang@whu.edu.cn

$$F(x_1, x_2, \cdots, x_n) = C_\theta(F_1(x), F_2(x), \cdots, F_n(x)) \quad (1)$$

式中：$C()$为Copula函数；θ为Copula参数；F_1，F_2，……，F_n为各随机变量的边际分布。

基于Sklar定理[5]，令$H(\bullet, \bullet)$为具有边缘分布$F(\bullet)$和$G(\bullet)$的联合分布函数，那么存在一个Copula函数$C(\bullet, \bullet)$，满足：

$$H(x, y) = C(F(x), G(y)) \quad (2)$$

若$F(\bullet)$、$G(\bullet)$连续，则$C(\bullet, \bullet)$唯一确定；反之，若$F(\bullet)$、$G(\bullet)$为一元分布函数，$C(\bullet, \bullet)$为相应的Copula函数，那么由上式定义的$H(\bullet, \bullet)$是具有边缘分布$F(\bullet)$、$G(\bullet)$的联合分布函数。

2.2 函数主要类型 目前来说，Copula函数有很多种类型，总体上可以划分为3种：椭圆型（包括正态Copula函数和学生t-Copula函数）、二次型、Archimedean型[6]。水文领域中常见的是Archimedean型，Archimedean型又分为对称Archimedean型和非对称Archimedean型两种形式。非对称Archimedean型又包含Clayton Copula 、Frank Copula和Gumbel-Hougaard Copula等函数形式。

2.3 函数参数估计 Copula函数的参数估计方法有很多种，主要有矩估计、核估计法、Genest和Rivest的非参数估计法、两阶段估计方法[7]、边际推断法、极大似然估计法等几种，Copula模型的边缘分布参数估计一般采用极大似然估计和矩法估计，其中极大似然估计是最常用的Copula模型边缘分布的参数估计方法。Copula联结函数参数多采用两阶段极大似然估计方法来估计[8]。

2.4 函数的检验和评价 指定的分布模型能否很好的拟合变量的实际分布，对Copula函数能否正确的描述变量间的相关性结构至关重要，因此要建立分布的检验和拟合度评价。目前国内外利用的Copula函数检验方法主要是以下几种：(1)K-S检验[9]和Q-Q图，K-S检验是一类常用的非参数检验方法，可用于检验单一样本是否服从某一特定分布，或者检验两个独立的样本是否服从同一分布。Q-Q图可以直观的表达出经验频率和理论频率的拟合情况，比较简单方便。(2)χ^2检验和“Hit”检验，χ^2检验方法以及“Hit”检验方法在水文中应用的不是很多。Copula函数评价方法[10]主要有离差平方和准则法、AIC信息准则法、Genest-Rivest方法3种。

3 案例分析

以淮河流域蚌埠闸为例，对其进行水量水质联合分布概率计算。蚌埠闸位于淮河中下游、洪泽湖上游，是淮河干流中游重要的控制站。该闸主要承担蓄水灌溉任务，兼有航运发电和供水等作用，是一座综合利用的水利工程，为蚌埠市社会经济发展用水提供了可靠保障。蚌埠闸的水量水质的联合概率分布函数的计算在淮河的水量水质调度风险分析中起着十分重要的作用。

3.1 二维Copula联合分布

3.1.1 *二维边缘分布* 受资料条件限制，本文选择蚌埠闸1986—2005年的实测月水量和水质(采用高锰酸盐指数和氨氮两个指标)资料。在我国水文分析计算中，对于单变量水文数据系列的分析常假定水文变量服从皮尔逊Ⅲ型分布。

采用以下公式计算边缘分布的经验频率。

$$H(x) = P(X \leqslant x_m) = (m - 0.44)/(N + 0.12) \quad (3)$$

式中：P为$X \leqslant x_m$的经验概率；m为x_m的序号；N为样本容量。

利用线性矩法对皮尔逊Ⅲ型分布进行参数估计。蚌埠闸的水量和水质(高锰酸盐指数和氨氮)月总量分布概率拟合情况及参数估计值如图1—3和表1所示。从它们的频率曲线拟合结果可以看出，各边缘分布的拟合曲线能够很好的反映出边缘分布的频率直方图。可以用这些分布及参数来表示边缘分布。

计算得知3个变量边缘分布的经验累积频率与理论累积频率是基本相等的，相关性系数较高，可以进一步的反映出可以用皮尔逊Ⅲ型分布来表示边缘分布，其参数的选择是合理的。

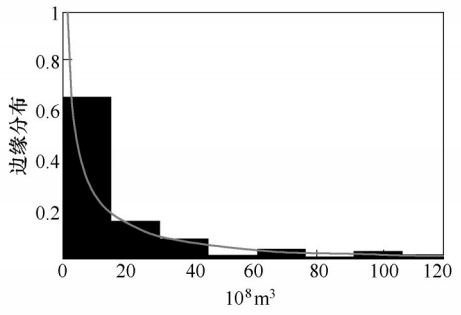

图1　月径流量的概率密度

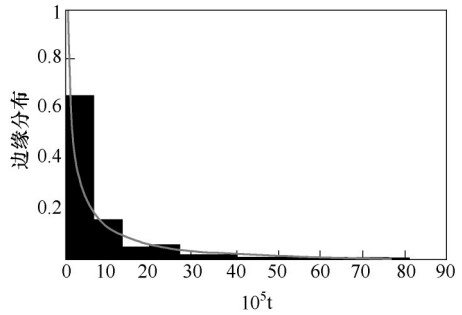

图2　月高锰酸盐指数总量的概率密度

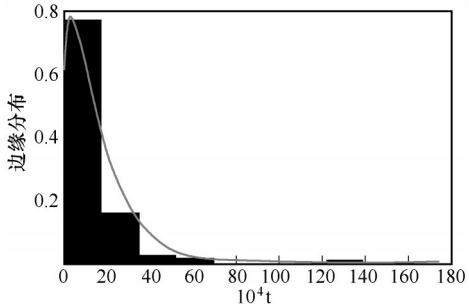

图3　月氨氮总量的概率密度

表1　二维边缘分布函数的参数估计值

边缘分布参数估计值	α	β	a_0
月径流量	0.336	0.016	2.271
月高锰酸盐指数总量	0.338	0.034	0.903
月氨氮总量	1.273	0.076	−0.547

3.1.2　二维联合分布

(1)二维联合经验分布。把水量和水质资料分别用X和Y来描述，在实测数据中，把$(X_1,\ X_2,\ \cdots,\ X_n)$和$(Y_1,\ Y_2,\ \cdots,\ Y_n)$按照系列$X$的升序排列，挑选排好次序数据中的$X_i \leqslant X_j$，$Y_i \leqslant Y_j (i<j=1,\ \cdots,\ n)$的数据对，通过这些数据对来计算联合分布函数的经验频率值。计算公式如下[11]：

$$H(x_i,\ y_i)=P(X \leqslant x_i,\ Y \leqslant y_i)=\frac{\sum_{m=1}^{i}\sum_{n=1}^{i}N_{mn}-0.44}{N+0.12} \tag{4}$$

式中：P为$X \leqslant x_i$，$Y \leqslant y_i$的二维联合概率值；N_{mn}为$X \leqslant x_i$，$Y \leqslant y_i$的序号；N为总的数据对数目。

(2)二维联合理论分布函数。在水文计算中，一般采用最多的Copula函数是阿基米德Copula函数。本文计算得到的水量和高锰酸盐指数以及水量和氨氮指数之间的Kendall秩相关系数分别为0.816和0.221。根据不同的Copula函数对相关性的适应范围，本文选择Clayton和Gumbel−Hougaard Copula函数来构造联合分布函数，从中选择拟合效果较好的Copula函数。计算参数为见表2。

表2　Copula函数参数的估计值

Copula函数	二维边缘分布	$C(u_1,\ u_2)$	τ与θ	θ
Clayton	水量和高锰酸盐指数	$C(u,\ v)=\left(u^{-\theta}+v^{-\theta}-1\right)^{-1/\theta}$	$\tau=\frac{\theta}{1+\theta}$	8.90
	水量和氨氮			0.57
Gumbel−Hougaard	水量和高锰酸盐指数	$C(u,\ v)=\exp\left\{-\left[(-\ln u)^{\theta}+(-\ln v)^{\theta}\right]^{1/\theta}\right\}$	$\tau=1-\frac{1}{\theta}$	8.78
	水量和氨氮			1.28

θ为两个变量之间关系的值，需要对θ进行估计。对于两变量单参数阿基米德Copula函数的参数估计，常采用Genest等提出的非参数估计方法[13]，通过Kendall秩相关系数τ与Copula函数参数θ的关

系进行估计。

3.1.3　二维联合分布结果分析　将蚌埠闸的水量和高锰酸盐指数、水量和氨氮指数的经验频率和理论联合分布频率值绘于图4—7中。图中$F(x，y)$为理论联合分布，$F_{emp}(x，y)$为经验联合分布。

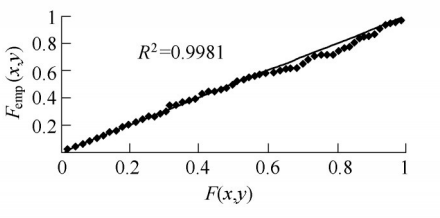

图4　Clayton函数水量和高锰酸盐指数联合观测点的经验分布和理论分布的比较

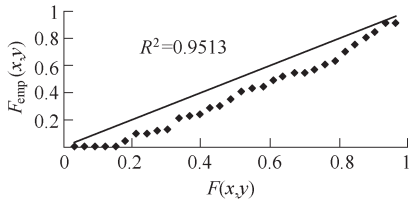

图5　Clayton函数水量和氨氮联合观测点的经验分布和理论分布的比较

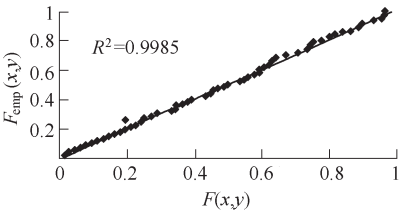

图6　Gumbel-Hougaard函数水量和高锰酸盐指数联合观测点的经验分布和理论分布的比较

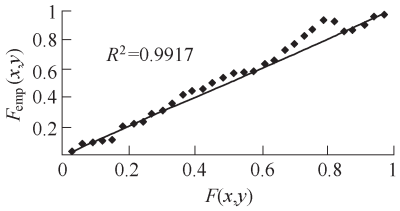

图7　Gumbel-Hougaard函数水量和氨氮联合观测点的经验分布和理论分布的比较

对其进行进一步的拟合优度检验，可采用离差平方和最小准则(OLS)来评价Copula方法的有效性，并选取OLS最小的Copula作为联结函数。OLS[3]的计算公式为：

$$OLS=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(p_{e_i}-p_i\right)^2} \tag{5}$$

式中：p_{e_i}、p_i分别为经验频率和理论频率；i为样本序号。

计算得到Clayton Copula函数的*OLS*分别为0.028 51和0.124 6，Gumbel-Hougard Copula函数的*OLS*分别为0.016 33和0.050 42，可知Gumbel-Hougard Copula函数分布拟合精度比较高，选择Gumbel-Hougard Copula函数来构造蚌埠闸的水量水质联合分布函数。

3个变量的Gumbel-Hougard Copula联合分布的相关系数的平方R^2分别是0.998 5和0.991 7，这说明经验点据与所选取的Gumbel-Hougard Copula联结函数的分布情况拟合较好，说明选取的二维Copula函数是合理的，可用来分析水量水质联合分布问题。

图8、图9分别描述了水量、高锰酸盐指数、氨氮指数3个变量之间的理论联合分布[12]，从图中我们可以更加清楚的了解联合分布函数。

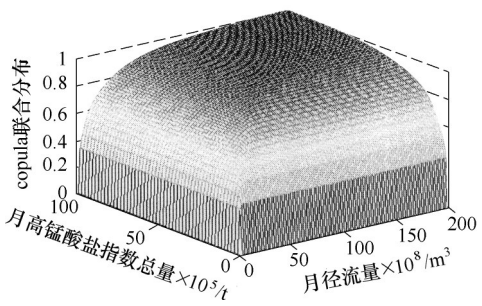

图8　水量和高锰酸盐指数的联合分布图

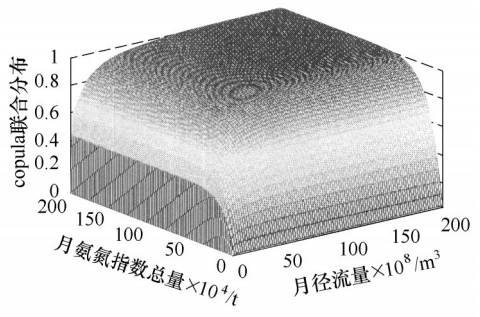

图9　水量和氨氮的联合分布图

3.2 三维Copula联合分布

3.2.1 三维边缘分布 三维Copula联合分布建立所用的方法和资料和二维Copula联合分布一样，其概率密度拟合图与二维边缘分布的一样，这里不再赘述。以下得到的是三维边缘分布的参数值：

表3 三维边缘分布参数估计值

边缘分布参数估计值	α	β	a_0
月径流量	0.336	0.043	0.698
月高锰酸盐指数总量	0.336	0.076	0.505
月氨氮总量	0.406	0.017	0.432

3.2.2 三维联合分布 在水量水质联合分布的推求中，有3个变量需要考虑，分别是水量、氨氮指数、高锰酸盐指数，本文选择三维Gumbel-Hougaard Copula函数来构造水量水质的联合分布。它又分为对称型和不对称型两种[9]，下面是它们的表达式：

三维对称Gumbel-Hougaard Copula：

$$C(u_1,\ u_2,\ u_3)=\exp\left\{-\left[(-\ln u_1)^{\theta}+(-\ln u_2)^{\theta}+(-\ln u_3)^{\theta}\right]^{1/\theta}\right\},\ \theta\geqslant 1 \tag{6}$$

式中：θ为表达相关性的参数，当$\theta=1$时，为变量之间相互独立的情形，此时$C(u_1,\ u_2,\ u_3)=u_1u_2u_3$。

三维非对称Gumbel-Hougaard Copula：

$$C(u_1,\ u_2,\ u_3)=\exp\left\{-\left(\left[(-\ln u_1)^{\theta_2}+(-\ln u_2)^{\theta_2}\right]^{\theta_1/\theta_2}+(-\ln u_3)^{\theta_1}\right)^{1/\theta_1}\right\},\ \theta_1,\ \theta_2\geqslant 1 \tag{7}$$

式中：θ_1、θ_2为表达相关性的参数，当$\theta_1=1$时，$C(u_1,\ u_2,\ u_3)=u_3C(u_1,\ u_2)$，当$\theta_2=1$时，$C(u_1,\ u_2,\ u_3)=C(u_1u_2,\ u_3)$，当$\theta_1=\theta_2=\theta$时，上面不对称型三维Copula函数即变成了对称型三维Copula函数。$\theta_1=\theta_2=1$时，为变量之间相互独立的情形，此时$C(u_1,\ u_2,\ u_3)=u_1u_2u_3$。

(1)三维联合分布经验频率。水量、氨氮指数、高锰酸盐指数分别用X、Y和Z来描述，在实测数据中，把$(X_1,\ X_2,\ \cdots,\ X_n)$，$(Y_1,\ Y_2,\ \cdots,\ Y_n)$和$(Z_1,\ Z_2,\ \cdots,\ Z_n)$按照系列$X$的升序排列，挑选排好次序数据中的$X_i\leqslant X_j$，$Y_i\leqslant Y_j$，$Z_i\leqslant Z_j(i<j=1,\ \cdots,\ n)$的数据对，通过这些数据对来计算联合分布函数的经验频率值。计算公式为：

$$p(X_1\leqslant x_1,\ X_2\leqslant x_2,\ X_3\leqslant x_3)=\frac{\sum_{m=1}^{i}\sum_{n=1}^{i}\sum_{p=1}^{i}N_{mnp}-0.44}{n+0.12} \tag{8}$$

式中：P为$X\leqslant x_i$，$Y\leqslant y_i$，$Z\leqslant z_i$的三维联合概率值；N_{mnp}为$X\leqslant x_i$，$Y\leqslant y_i$，$Z\leqslant z_i$的序号，n为总的数据对数目。

(2)三维联合分布理论频率。首先计算出Copula函数的关联性参数θ的值，由上面的边际分布的参数估计值估计得到X和Y的kendall秩相关系数$\tau_{XY}=0.220$；Y和Z的kendall秩相关系数$\tau_{YZ}=0$，说明氨氮指数和高锰酸盐指数之间是独立的，是不相关的；X和Z的kendall秩相关系数$\tau_{XZ}=0.816$。按照秩相关系数τ和θ之间的关系，可以计算得知，对称型三维Gumbel-Hougaard Copula的参数$\theta=1/3\left[1/(1-\tau_{XY})+1/(1-\tau_{XZ})+1/(1-\tau_{YZ})\right]\approx 2.577$；非对称型三维Gumbel-Hougaard Copula的参数$\theta_1=1/2\left[1/(1-\tau_{XY})+1/(1-\tau_{YZ})\right]\approx 1.141$，$\theta_2=1/(1-\tau_{XZ})\approx 5.448$。

3.2.3 三维联合分布结果分析 将对称型三维Gumbel-Hougaard Copula函数和非对称型三维Gumbel-Hougaard Copula经验联合分布累积频率和理论联合分布累积频率点绘得到图10、图11：

计算得对称型三维和非对称型三维Archimedean Copula的离差平方和最小准则*OLS*值分别为0.052和0.162 8，表明对称型三维Archimedean Copula建立的联合分布比非对称型三维Archimedean Copula的更合理。说明对称型三维Archimedean Copula构造的联合分布比非对称型三维Archimedean Copula

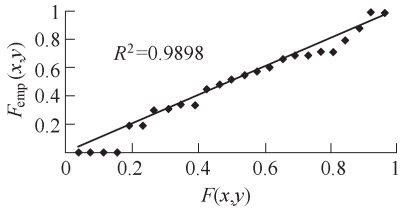

图10　对称型三维Copula函数联合观测点的经验分布和理论分布的比较

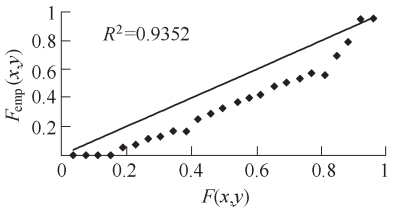

图11　非对称型三维Copula函数联合观测点的经验分布和理论分布的比较

要更加满足实际情况。对称型与非对称型三维Gumbel-Hougaard Copula函数的经验和理论累积频率的相关性系数R^2分别为0.989 8和0.935 2，同样也可以看出用对称型三维Copula函数来构造水量水质之间的联合分布更合理。

4　结论

本文采用Copula联结函数，构建了蚌埠闸的水量和水质的二维和三维联合分布函数，分析了蚌埠闸的水量水质联合分布频率，结果表明蚌埠闸的Copula函数水量水质联合分布的经验累积频率和理论累积频率的一致性是很高的，用Copula函数构造水量水质联合分布的拟合精度是满意的。对三维Copula函数来说，对称型三维Copula函数比非对称型三维Copula函数更合理，Copula方法是构建联合分布的一种有效方法，用Copula方法描述水量水质之间的联合分布关系是可行的。在水质方面的研究中，虽然很多Copula函数可以用来计算联合分布，但是阿基米德Copula函数是进行水质分析研究较合适的函数。目前关于水量水质联合分布的研究还很少，所以今后在这方面的研究还需要继续加强，并进一步应用于水量水质综合管理的风险分析之中。

参　考　文　献：

[1] 熊立华，郭生练，肖义，等．Copula联结函数在多变量水文频率分析中的应用[J]．武汉大学学报(工学版)，2005，38(6)：16-19

[2] 许月萍，李佳，曹飞凤，等．Copula在水文极限事件分析中的应用[J]．浙江大学学报(工学版)，2008，42(7)：1119-1122．

[3] 莫淑红，沈冰，张晓伟，等．基于Copula函数的河川径流丰枯遭遇分析[J]．西北农林科技大学学报(自然科学版)，2009，37(6)：131-136．

[4] 闫宝伟，郭生练，刘攀，等．基于Copula函数的径流随机模拟[J]．四川大学学报(工程科学版)，2010，42(1)：5-9．

[5] Sklar A．Fonctions de repartition an dimensions et leurs marges[J]．Publication de l'Institut de Statistique de l'Universite de Paris，1959，8：229-231．

[6] 张娜，郭生练，闫宝伟，等．Copula函数在分期设计洪水中的应用研究[J]．水文，2008，28(5)：28-32．

[7] Joe H．Asymptotic efficiency of the two-stage estimation method for Copula-based models[J]．J．Multivariate Anal，2005；94：401-419．

[8] 韦艳华，张世英．Copula理论及其在金融分析上的应用[M]．北京：清华大学出版社，2008．

[9] 肖义．基于Copula函数的多变量水文分析计算研究[D]．武汉：武汉大学，2007．

[10] 郭生练，闫宝伟，肖义，等．Copula函数在多变量水文分析计算中的应用及研究进展[J]．水文，2008，28(3)：1-7．

[11] Zhang L．Multivariate Hydrological Frequency Analysis and Risk Mapping[D]．PhD thesis，Department of Agricultural and Mechanical College，Louisiana State University，USA，2005．

[12] 闫宝伟，郭生练，陈璐，等．长江和清江洪水遭遇风险分析[J]．水利学报，2010，41(5)：553-559．

[13] 钟波, 张鹏 . Copula选择方法[J]. 重庆工学院学报(自然科学), 2009, 23(5): 155-160 .

Jointed distribution function of water quality and water quantity based on Copula

ZHANG Xiang[1], RAN Qi-xiang[1], XIA Jun[2], SONG Xing-yuan[1]

(1. *Wuhan University*, *Wuhan* 430072, *China*;

2. *Key Laboratory of Water Cycle and Related land Surface Processes*, *IGSNRR*, *CAS*, *Beijing* 100101, *China*)

Abstract: The Copula is applied to perform multivariate frequency analysis of water quality and water quantity. The definition, its main classifications, methods of parameter estimation and goodness-of-fit tests of Copula function are introduced. The application of Copula as the bivariate and trivariate (which divide into symmetric and asymmetric) jointed distribution functions for water quality and water quantity on a hydrological gage station was illustrated. The result shows that it is feasible to apply Copula function to describe the jointed distribution function of water quality and water quantity.

Key words: Copula function; water environment; jointed distribution of water quality and water quantity; multivariate

(责任编辑: 韩 昆)