文章编号：2095-2163(2022)02-0049-05 中图分类号：TP274 文献标志码：A

基于相关性的传感数据分析与处理

罗 宇[1]，李 颖[2]，郝昕宇[3]，杨光松[3]
（1 贵州省广播电视局 645 台，贵阳 550200；2 集美大学 诚毅学院，福建 厦门 361021；
3 集美大学 信息工程学院，福建 厦门，361021）

摘 要：环境监测的无线传感器网络中存在大量监测数据，有效地挖掘数据之间的相关性，可以缩短分析时间，提高分析精度。本文首先介绍了协方差概念和相关性的计算方法；其次，分析了时间序列数据的相关性和多变量的互相关性；最后，对某地的环境监测数据进行了相关性分析。分析结果证明：PM2.5 与空气质量指数存在高度相关性；PM2.5 与时间无相关性；日照与臭氧含量高度相关。

关键词：相关性；时间序列；协方差；互相关

Analysis and processing of sensing data based on correlation

LUO Yu[1], LI Ying[2], HAO Xinyu[3], YANG Guangsong[3]
(1 Radio Transmitting Station 645# of Guizhou Radio and Television Bureau, Guiyang 550200, China;
2 Chengyi Colledge, Jimei University, Xiamen Fujian 361021, China;
3 School of Information Engineering, Jimei University, Xiamen Fujian 361021, China)

【Abstract】 There are a lot of data in wireless sensor networks for environmental monitoring. If we mine the correlation between these data, we can save analysis time and improve the analysis accuracy. In this paper, we first introduce the concept of the covariance and calculation method of the correlation and analyze the correlation of time series data and multivariable. Then the correlation of environmental monitoring data of a city is analyzed. The analysis results show that there is a high correlation between PM2.5 and the air quality index. The correlation of PM2.5 is independent of time. The multivariate correlation analysis indicates that sunshine is highly correlated with ozone content.

【Key words】 relevance; time series; covariance; cross correlation

0 引 言

随着信息技术的快速发展，处理数据的能力不断增强，目前商用的云存储平台已经具有存储大量数据的能力，如何对海量数据进行分析，已成为当前的一个研究热点[1]。为了对环境传感数据进行监测，获取不同时间、不同空间的数据信息，将分布在不同地域的传感器节点，依靠通信协议组网，最终通过特定网关，将获取的数据传输到云平台上，通过分析数据之间所存在的相关性，寻找其固有的规律[2]。

利用数据相关性的检测，可以为监测工作提供精准且全面的数据支持。通过研究数据相关性来制定策略，从而采取相应的处理措施，在环境工程、环境生物学和地球科学等方面得到广泛应用。在水利方面，利用往年的数据可以分析雨季何时来临；在地理方面，可分析出降雨量对土壤成分的影响，预防泥石流的形成[3]；在农业中，可以分析出哪一种变量会影响农作物的产量或者甜度，从而可以用安全的方式增加产量或者提升口感[4]；在环境监测方面，降雨数据、臭氧密度、气温温度等数据之间都存在相关性，其中任何一个量的变化都会引起其他一种或者几种分量的变化[5]。因此，需要对所有变量进行相关性分析，从而发现变量之间的关联关系。

本文主要从协方差、时间序列分析、互相关等方面，讨论相关性的计算、估计方法，并以环境监测数据为例进行相关性分析。

1 相关性及协方差

相关性是指事物之间存在相似的程度[6]。相关关系是指变量之间存在的一种不确定的数量依存关系，即一个变量的数值发生变化时，另一个变量的

基金项目：福建省中青年教师教育科研项目(JT180877)；福建省自然科学基金(2021J01865)。

作者简介：罗 宇(1968-)，男，学士，高级工程师，主要研究方向：通信技术、数据处理；李 颖(1983-)，女，硕士，副教授，主要研究方向：无线通信网络、人工智能。

通讯作者：李 颖 Email：cyxxliying@jmu.edu.cn

收稿日期：2021-12-02

数值也相应地发生变化,变化的数值不是确定的,但在一定的范围内。

协方差是一种用来度量两个随机变量关系的统计量,假设有两类数据 x_i 和 x_j, 可将其视为随机变量,两者之间的关系可以由一个联合概率密度函数 $p(x_i,x_j)$ 来表示,与 $p(x_i,x_j)$ 相关的协方差矩阵 $\boldsymbol{C}_{ij}$ 可定义为式(1)[7]:

$$\boldsymbol{C}_{ij}=\int_{-¥}^{+¥}\int_{-¥}^{+¥}(x_i-\bar{x_i})(x_j-\bar{x_j})p(x_i,x_j)\mathrm{d}x_i\mathrm{d}x_j \quad (1)$$

其中, $\bar{x_i}$ 和 $\bar{x_j}$ 分别是 x_i 和 x_j 均值。

通常,可以通过观测数据构造的近似概率密度函数方块图来估计 $\boldsymbol{C}_{ij}$。协方差估算的散点图,如图 1 所示,可将 (x_i,x_j) 平面划分为许多小的方格,按照 s 编号。每个方格 s 的面积为 $\Delta x_i\Delta x_j$, 其中心坐标为 $(x_i(s),x_j(s))$。于是可得式(2):

$$p(x_i,x_j)\Delta x_i\Delta x_j\approx\frac{N_S}{N} \quad (2)$$

其中, N 表示平面中数据对的总数, N_S 表示方格中数据对的数量(即互相关的数目对)。

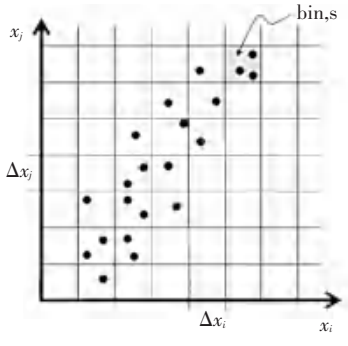


图 1 协方差估算的散点图

Fig. 1 Scatter plot of covariance estimation

综合考虑(1)式和(2)式,可得 $\boldsymbol{C}_{ij}$ 的近似计算公式(3):

$$C_{ij}\approx\frac{1}{N}\sum_s[x_i^{(s)}-\bar{x_i}][x_j^{(s)}-\bar{x_j}]N_s \quad (3)$$

进一步进行规一化处理,将方格大小缩小,使其每个方格中至多有一个数据对 ($Ns=0$ or $Ns=1$), 于是可得式(4):

$$\boldsymbol{C}_{ij}\approx\frac{1}{N}\sum_{k=1}^{N}[x_i^{(k)}-\bar{x_i}][x_j^{(k)}-\bar{x_j}]N_s \quad (4)$$

当数据表现出一定程度的相关性时,协方差是非 0 的,但其实际数值取决于数据量。通过方差乘积的平方根进行缩放,可将范围标准化为 1,式(5)。

$$\boldsymbol{R}_{ij}=\frac{\boldsymbol{C}_{ij}}{\sqrt{C_{ii}C_{jj}}} \quad (5)$$

$\boldsymbol{R}$ 被称为相关系数矩阵,其元素称为相关系数。当通过某个数据集对其进行估计时,被称为样本相关系数。

2 时间序列的数据相关性

式(1)的协方差矩阵 $\boldsymbol{C}_{ij}$ 可以用于量化联合概率密度函数的相关度,亦可用于描述时间序列的相关度。将式(1)展开,可得式(6)。

$$\boldsymbol{C}_{ij}=\int_{-¥}^{+¥}\int_{-¥}^{+¥}(x_i-\bar{x_i})(x_j-\bar{x_j})p(x_i,x_j)\mathrm{d}x_i\mathrm{d}x_j=$$
$$\int_{-¥}^{+¥}\int_{-¥}^{+¥}x_ix_jp(x_i,x_j)\mathrm{d}x_i\mathrm{d}x_j-2\bar{x}+\bar{x}^2=A_{ij}-\bar{x}^2 \quad (6)$$

其中, $\boldsymbol{A}_{ij}$ 被称为自相关矩阵,式(7)。

$$\boldsymbol{A}_{ij}=\int_{-¥}^{+¥}\int_{-¥}^{+¥}x_ix_jp(x_i,x_j)\mathrm{d}x_i\mathrm{d}x_j \quad (7)$$

如果时间序列是平稳的,其统计特性将不随时间变化,时间序列的均值也与时间无关。当均值 $\bar{x_i}$ 等于 0 时, $\boldsymbol{A}_{ij}$ 就等于时间序列的协方差矩阵 $\boldsymbol{C}_{ij}$。

用类似式(2)~式(4)的方法,可以用散点图的形式求式(7)的近似积分,于是可求得 $\boldsymbol{A}_{ij}$ 的近似值为[5]:

$$A_{i,k+i-1}\approx\frac{1}{N-|k-1|}\sum_s x_i^{(s)}x_{k+i-1}^{(s)}N_s=$$
$$\frac{1}{N-|k-1|}\sum_{i=1}^{N-k+1}x_ix_{k+i-1}=\frac{a_k}{N-|k-1|} \quad (8)$$

其中, $\boldsymbol{a}_k$ 为在时间差 $\tau=k-1$ 时的自相关,式(9)。

$$\boldsymbol{a}_k=\sum_{i=1}^{N-k+1}x_ix_{k+i-1} \quad (9)$$

由 $\boldsymbol{a}_k$ 构成的列向量 $\boldsymbol{a}$ 称为时间序列的自相关。由于 A 是对称的,所以时间间隔为正的自相关等于时间间隔为负的自相关, 即当 $k=|i-j|+1$ 时, $\boldsymbol{A}_{ij}=\boldsymbol{a}_k$。

当均值 $\bar{x_i}$ 随时间变化时,其相关的程度将取决于测量时间以及均值之间的时间差。在一个时间序列中相邻的样本通常是高度相关的,因此是可预测的,例如某条河流的流量,某区域的有害气体含量等。

3 互相关

自相关研究的问题是从相同变量的时间序列中间隔时延 t 的样本;而互相关研究是不同变量的时间序列中间隔时延 t 的样本。例如降水 u 和河水流

量 v 的时间序列，在降水量高的时候，可以预计河水流量也会很大。但由于河水流动需要时间，因此，当降水时间序列相对于水流时间序列时间间隔一定时间时，降水时间序列与水流时间序列的相关性最大。已知变量 u 和 v，定义互相关性为其概率密度函数 $p(u_i, v_j)$，分别为时间序列 u 的第 i 个样本，和时间序列 v 的第 j 个样本。

可以将自相关的计算，类推到计算互相关 c_k，式(10)

$$c_k = \sum_i u_i v_{k+i-1} \tag{10}$$

互相关可用如式(11)的卷积形式进行计算。

$$c(t) = \int_{-¥}^{+¥} u(\tau) v(t+\tau) \mathrm{d}\tau \tag{11}$$

与自相关不同的是，互相关在时间间隔上是不对称的。$v(t)$ 和 $u(t)$ 的互相关性是 $u(t)$ 和 $v(t)$ 的互相关的时间反转。

4　数据处理与分析

在现实生活中，不同事物之间存在大量的因果关系，通过发掘这些相互关系，可以获得一些有用的信息，帮助做出正确的判断，有助于科学的预测，从而防患于未然。

分析北京市 2017 年一整年的空气质量数据，见表 1[8]。主要基于协方差、自相关、互相关进行分析。

表 1　北京 2017 年环境监测数据

Tab. 1　Environmental monitoring data of a year in Beijing

序号	AQI 指数	PM2.5 指数	PM10 指数	SO_2浓度	NO_2浓度	CO 浓度	O_3 浓度	空气质量状况	质量等级
1	451	425	493	9	114	5.942	4	严重污染	六级
2	212	161	313	8	82	3.308	12	重度污染	五级
3	126	86	200	9	66	2.058	16	轻度污染	三级
4	78	57	0	8	64	1.683	17	良	二级
5	65	46	0	9	62	1.567	16	良	二级
6	178	52	303	12	59	1.792	20	中度污染	四级
7	256	68	388	13	56	1.975	28	重度污染	五级
8	178	101	304	12	64	2.417	34	中度污染	四级
9	273	222	222	11	120	4.217	5	重度污染	五级
……	……	……	……	……	……	……	……	……	……

4.1　基于协方差的数据分析

空气质量数据包含大气中一些污染物的含量，如：PM2.5(细颗粒物)、PM10(可吸入颗粒物)、SO_2(二氧化硫)、NO_2(二氧化氮)、CO(一氧化碳)、CO_2(二氧化碳)、O_3(臭氧)。空气质量的衡量标准是空气质量指数(Air Quality Index，AQI)，选取其中 5 种污染物做相关协方差分析，相关系数矩阵如图 2 所示，横轴和纵轴分别表示这几种因素之间的相关系数，颜色越深，表示相关性越强。可见，从左上至右下的对角线元素都均为黑色，因为每种因素与自身完全相关，与 AQI 最相关的因素是 PM2.5，其次是 PM10、CO、NO_2，SO_2与其相关性较小。

根据表 1，进一步绘出 AQI 与 PM2.5 的相关指数，如图 3 所示，两个因素呈现正相关的趋势，利用式(5)，可计算出 PM2.5 与 AQ1 相关系数 $R = 0.99$，证明 PM2.5 与 AQI 具有高度相关性。

由此说明，若想改变空气质量指数，治理 PM2.5 最有成效，因为其相关性最大，降低 PM2.5 指数可以有效的改变空气质量；改变 PM10 在空气中的含量，也可以提升空气质量。

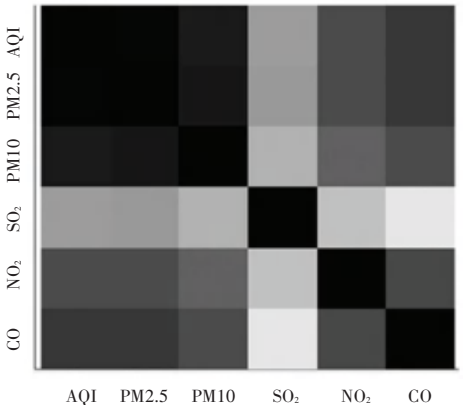


图 2　北京市空气质量数据集相关系数绝对值矩阵

Fig. 2　Absolute value matrix of correlation coefficient of Beijing air quality dataset

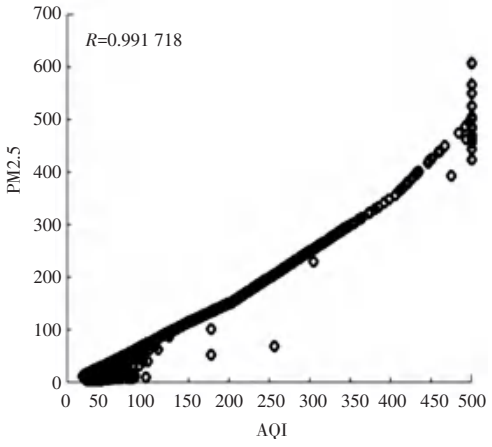


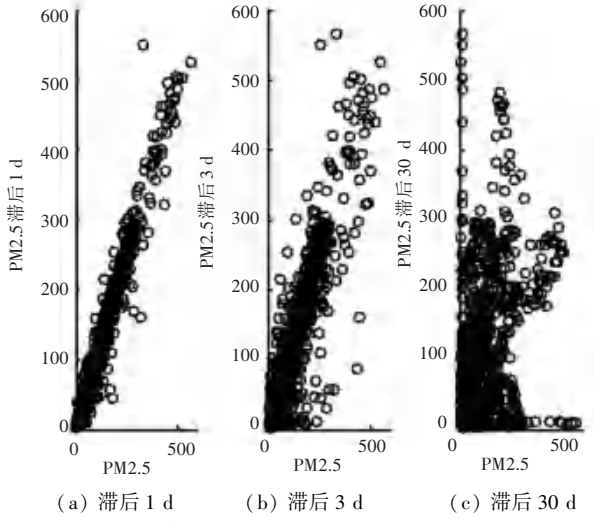
图 3 PM2.5 和 AQI 的相关性

Fig. 3 Correlation between PM2.5 and AQI

4.2 基于自相关的数据分析

PM2.5 在时间序列上自身的变化,时间间隔越大则自相关越小。根据表 1,取不同时刻的 PM2.5 的指数值,可得空气中 PM2.5 的指数与时间间隔的关系,如图 4 所示。图 4(a)~(c)分别是时间间隔为 1 d、3 d、30 d 的自相关函数,横轴为 PM2.5 的含量,纵轴为滞后一段时间后的 PM2.5 含量。

如果把空气在 t_i 时刻 PM2.5 含量记为 d_i, 在 t_j 时刻的含量记为 d_j, 那么其联合概率密度函数为 $p(d_i,d_j)$, 可以预计那个 d_i 和 d_j 在何处有很强的正相关关系,当时间间隔 $\Delta t_i t_j$ 很小时,其相关性很强,短期时间关联度很高,比如昨天的 PM2.5 与今天的 PM2.5 差不多,如图 4(a)所示;当测量值的时间间隔大时,其 PM2.5 的相关性变得越来越小,如图 4(b)为间隔 3 天的情况;在一定时间间隔(如 1 个月左右的时间)后,基本不相关,如图 4(c)。



(a) 滞后 1 d (b) 滞后 3 d (c) 滞后 30 d

图 4 空气中 PM2.5 的含量与时间间隔的关系

Fig. 4 Relationship between PM2.5 and time interval

4.3 基于互相关的数据分析

互相关是表示两个变量之间相似性的一个度量,通过与已知变量比较,来寻找未知变量中的特性。利用互相关性分析臭氧和日照的关系。

平流层中的臭氧,能够吸收紫外线,保护地球表面免受太阳紫外线的照射。但对流层中的臭氧是雾霾的主要成分,对人体健康有害,并导致的 AQI 指数降低。

利用半个月的数据,仅包含 4 列数据,时间(d)、臭氧(ppb)、太阳辐射(W/m^2)和气温(c)。

将半月的日照数据(单位为 W/m^2)和臭氧变化,在同一地点按对应的时间进行统计,如图 5 所示。可见两者都表现出明显的周期性,随着日照的强度增大,臭氧浓度也会增多,这是因为在温度较高、日照相对较强时,大气中的氮氧化物和挥发性有机化合物经紫外线照射发生光化学反应,生成臭氧。随着时间序列的变化,这两个变量所反映出来的相关性成正相关性,只要日照强度高,臭氧浓度就会增多。另一方面,从图 5 亦可以观察到,臭氧峰值比日照峰值延迟了几 d(见垂直虚线)。

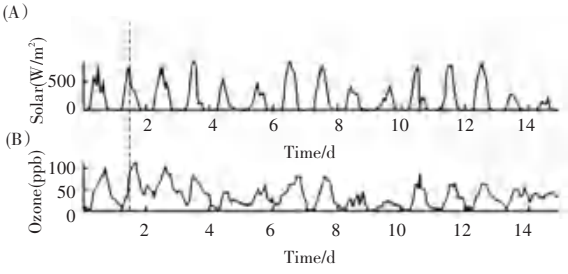


图 5 日照与臭氧含量的相关图

Fig. 5 Correlation between sunshine and ozone content

进一步将两个时间序列相互关联,可得出滞后的时间间隔约为 3 d,如图 6 所示。

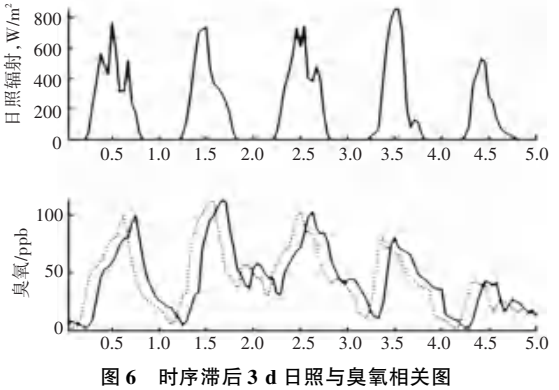


图 6 时序滞后 3 d 日照与臭氧相关图

Fig. 6 Correlation between sunshine and ozone with time lag of 3 d

滞后时间为 3 d,日照与臭氧互相关性,如图 7 所示。互相关系数达到了最大值,约为 3.5×10^{-6}。

因为光照有一个过程，随着光照的增加，臭氧含量也逐渐增加到最大值，所以两者之间的相关函数也相应地呈现出最大值。

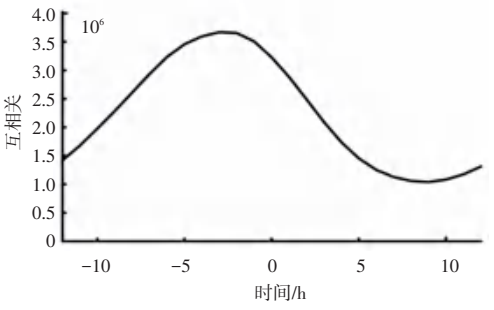


图 7　日照与臭氧互相关性

Fig. 7　Cross correlation between sunshine and ozone content

5　结束语

在环境监测中，传感数据之间存在相关性，充分挖掘这些相关性，有助于分析影响环境的各种因素，从而准确、高效地采取措施。这些相关性方法，还可以广泛应用在灾害预测等方面。本文利用协方差分析变量之间的相关性，对 PM2.5 指数、O_3浓度等与空气质量指数 AQI 的相关性进行分析，得出 PM2.5 是主要影响因素，从而解决提升空气质量的问题；利用自相关系数在同一过程、不同时刻的相互关系，分析了北京市的 PM2.5 的短期自相关性与长期自相关性，研究 PM2.5 随时间情况；利用互相关函数来分析时间序列，对日照与臭氧之间的互相关性进行研究，结果说明，二者之间在一定时间间隔上存在相关性。由于数据收集的局限性，本文数据样本还不够丰富，对一些相关性问题还需要数据样本的支撑。大数据时代的到来，数据相关性分析日益重要，下一步将考虑传感器网络中的数据补全等相关性问题，以进一步提高分析效率，节省传输能量。

参考文献

[1] CHEN M, MAO S, LIU Y. Big data: A survey[J]. Mobile networks and applications, 2014, 19(2): 171-209.

[2] FANG X, BATE I. Issues of using wireless sensor network to monitor urban air quality[C]//Proceedings of the First ACM International Workshop on the Engineering of Reliable, Robust, and Secure Embedded Wireless Sensing Systems. Delft, Netherlands, 2017: 32-39.

[3] MAHMOOD R, LITTELL A, HUBBARD K G, et al. Observed data-based assessment of relationships among soil moisture at various depths, precipitation, and temperature[J]. Applied Geography, 2012, 34(1): 255-264.

[4] FLIS B, ZIMNOCH-GUZOWSKA E, MA N KOWSKI D. Correlations among Yield, Taste, Tuber Characteristics and Mineral[J]. Journal of Agricultural Science, 2012, 4(7): 625-637.

[5] 陆星家. 宁波市空气质量与环境变量的偏相关性分析[J]. 中国资源综合利用, 2018, 36(8): 155-158.

[6] Wolfgang Karl Härdle, Léopold Simar. Applied multivariate statistical analysis[M]. Upper Saddle River: Pearson Education Asia Ltd, 2008: 17-18.

[7] Menke, William, Joshua Menke. Environmental data analysis with Matlab[M]. Salt Lake City: Academic Press, 2016: 168-169.

[8] 王杰.空气质量历史数据查询[EB/OL].(2012-10-23)[2021-08-26]. https://www.aqistudy.cn/history data/.

(上接第 48 页)

抗攻击效果，但本文算法在透明性相差不大情况下，抗攻击效果略胜一筹；

(6)本文算法在低通滤波，*JPEG* 压缩攻击后提取效果和 DWT 变换，Contourlet 变换相差不大的情况下，做到了在椒盐噪声攻击下提取效果的提升。

4　结束语

本文提出了一种基于 DWT-SVD 鲁棒水印算法，在传统离散小波变换基础上，进行奇异值分解，利用图像奇异值良好的稳定性，提高图像的抗攻击性；在图像奇异值基础上，进行奇偶量化嵌入和提取水印，可实现盲提取；为了进一步提高抗攻击性，在图像的各个频段通过加权融合提取水印。仿真和攻击实验结果表明，与其它算法进行比较分析，该算法的鲁棒性有一定提高。

参考文献

[1] 郑玉平. 数字水印技术在数字版权保护中的应用[D]. 北京理工大学, 2015.

[2] 夏煜, 郎荣玲, 戴冠中, 黄殿中, 钱思进. 基于数字图像 LSB 嵌入的检测算法[J]. 计算机工程, 2004(4): 10-11, 161.

[3] 李文娜, 孔祥勇, 高立群, 等. 基于条带波变换的图像水印算法[J]. 东北大学学报(自然科学版), 2013, 34(8): 1086-1090.

[4] 燕鲲鹏. 基于 DWT 变换的分块数字图像水印算法研究[D]. 西安: 西北大学, 2017.

[5] 吴静. 基于 Contourlet 变换的图像数字水印算法研究[D]. 贵州: 贵州师范大学, 2014.

[6] 李磊. 基于 DCT 变换和 SVD 变换的数字水印技术[J]. 电脑知识与技术, 2019, 15(30): 197-199.

[7] 汤永利, 张亚萍, 高玉龙, 等. 基于 DWT-SVD 压缩量化的数字图像盲水印算法[J]. 重庆邮电大学学报(自然科学版), 2018, 30(2): 265-271.