

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局



(10) 国际公布号

WO 2020/258918 A1

(43) 国际公布日

2020 年 12 月 30 日 (30.12.2020)

(51) 国际专利分类号:

G06F 16/00 (2019.01)

(21) 国际申请号:

PCT/CN2020/078258

(22) 国际申请日:

2020 年 3 月 6 日 (06.03.2020)

(25) 申请语言:

中文

(26) 公布语言:

中文

(30) 优先权:

201910550362.0 2019年6月24日 (24.06.2019) CN

(71) 申请人: 中山大学 (SUN YAT-SEN UNIVERSITY)

[CN/CN]; 中国广东省广州市海珠区新港西路135号, Guangdong 510275 (CN)。

(72) 发明人: 赵铜铁钢 (ZHAO, Tongtiegang); 中国广

东省广州市海珠区新港西路135号, Guangdong

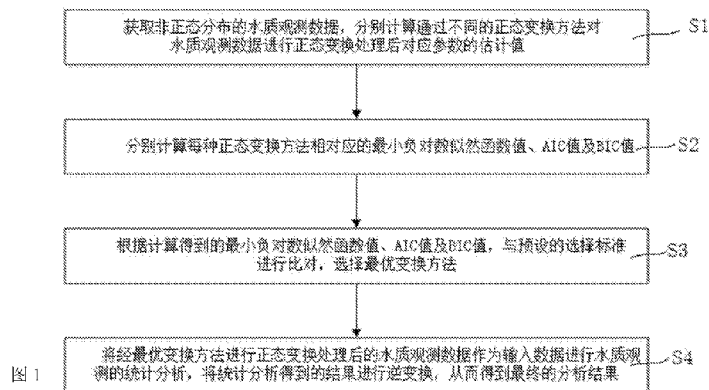
510275 (CN)。 陈浩玲 (CHEN, Haoling); 中国广东省广州市海珠区新港西路135号, Guangdong 510275 (CN)。

(74) 代理人: 广州粤高专利商标代理有限公司 (YOGO PATENT AND TRADEMARK AGENCY LIMITED COMPANY); 中国广东省广州市天河区体育西路中石化大厦B塔4416室, Guangdong 510620 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,

(54) Title: POWER TRANSFORM ANALYSIS METHOD FOR NON-NORMALLY DISTRIBUTED WATER QUALITY OBSERVATION DATA

(54) 发明名称: 一种面向非正态分布水质观测数据的幂变换分析方法



S1 ACQUIRE NON-NORMALLY DISTRIBUTED WATER QUALITY OBSERVATION DATA, AND RESPECTIVELY CALCULATE ESTIMATED VALUES OF CORRESPONDING PARAMETERS AFTER NORMAL TRANSFORMATION PROCESSING IS PERFORMED ON THE WATER QUALITY OBSERVATION DATA BY USING DIFFERENT NORMAL TRANSFORMATION METHODS
S2 RESPECTIVELY CALCULATE THE MINIMUM NEGATIVE LOG LIKELIHOOD FUNCTION VALUE, AIC VALUE AND BIC VALUE CORRESPONDING TO EACH NORMAL TRANSFORMATION METHOD
S3 ACCORDING TO THE CALCULATED MINIMUM NEGATIVE LOG LIKELIHOOD FUNCTION VALUE, AIC VALUE AND BIC VALUE, PERFORM COMPARISON WITH PRESET SELECTION CRITERIA TO SELECT AN OPTIMAL TRANSFORMATION METHOD
S4 USE THE WATER QUALITY OBSERVATION DATA AFTER NORMAL TRANSFORMATION PROCESSING AS INPUT DATA FOR STATISTICAL ANALYSIS OF WATER QUALITY OBSERVATION, AND INVERSELY TRANSFORM THE RESULTS OBTAINED BY STATISTICAL ANALYSIS TO THEREBY OBTAIN A FINAL ANALYSIS RESULT

(57) Abstract: Disclosed is a power transform analysis method for non-normally distributed water quality observation data. First, estimated values of corresponding parameters after normal transformation processing is performed on water quality observation data by using different normal transformation methods are respectively calculated, specific measurement indicators are configured, and calculation and comparison are performed so as to select an optimal normal transformation method according to data characteristics of water quality observation so that the transformed data has better normality. Finally, the water quality observation data after normal

MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US,
UZ, VC, VN, WS, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告 (条约第21条(3))。

transformation processing is performed by using the optimal transformation method is used as input data for statistical analysis of water quality observation, thus improving the effect of analysis. The method of the present invention may enable transformed data to have better normality, thus facilitating further data analysis, and solving problems such as a data transformation effect being poor due to a selected transformation method not being suited to the characteristics of an observation variable of a water plant.

(57) 摘要: 本发明公开了一种面向非正态分布水质观测数据的幂变换分析方法, 首先分别计算通过不同的正态变换方法对水质观测数据进行正态变换处理后对应参数的估计值, 设置具体的衡量指标并进行计算和比对, 从而根据水质观测的数据特征选择最优的正态变换方法, 使得变换后的数据具有更好的正态性, 最后经最优变换方法进行正态变换处理后的水质观测数据作为输入数据进行水质观测的统计分析, 提升了分析的效果。本发明方法能使得变换后的数据具有更好的正态性, 便于进一步的数据分析, 解决了由于选择的变换方法不适应水厂观测变量自身特征导致数据的变换效果差等问题。

一种面向非正态分布水质观测数据的幂变换分析方法

技术领域

本发明涉及环境工程技术领域，尤其涉及一种面向非正态分布水质观测数据的幂变换分析方法。

背景技术

水质观测序列的挖掘和统计分析，往往要求数据呈正态分布，而实际操作中，很多原始水质序列不呈正态分布，需要在不丢失信息的前提下进行数据的正态变换。

目前常用于水质序列的变换方法为对数变换，而水厂实际运行中，一些变量经对数变换后仍是偏态分布，尤其是负偏态数据，经对数变换后反而会增加其偏度。同时，由于污水处理厂进出水观测变量多、序列长且分布不一，单一类型的变换并不适用于所有的观测变量序列；而通过分析人员的主观判断对不同的变换方法进行选择时，由于选择标准不一，难以根据水厂观测变量自身特征选择最合适的变换方法，导致变换后的数据也无法满足常用数据挖掘和统计分析所要求的线性、方差齐性和正态性的要求，利用这些变换后的数据进行实际分析应用时丢失数据的一些重要性质，影响了分析效果。

发明内容

本发明为解决现有的非正态分布水质观测数据进行数据变换时，由于选择的变换方法不适应水厂观测变量自身特征导致数据的变换效果差等问题，提供了一种面向非正态分布水质观测数据的幂变换分析方法。

为实现以上发明目的，而采用的技术手段是：

一种面向非正态分布水质观测数据的幂变换分析方法，包括以下步骤：

S1.获取非正态分布的水质观测数据，分别计算通过不同的正态变换方法对水质观测数据进行正态变换处理后对应参数的估计值，对应参数的估计值包括：进行正态变换处理后，水质观测数据分布的均值、标准差以及变换参数；

S2.对于步骤 S1 中所述的不同的正态变换方法，分别计算每种正态变换方法相对应最小负对数似然函数值、AIC 值及 BIC 值；

S3.根据计算得到的最小负对数似然函数值、AIC 值及 BIC 值，与预设的选

择标准进行比对,根据比对结果从所述不同的正态变换方法中选择得到最优变换方法;

S4.将经所述最优变换方法进行正态变换处理后的水质观测数据作为输入数据进行水质观测的统计分析,将统计分析得到的结果进行逆变换,从而得到最终的分析结果,降低了分析过程的复杂性,提高了分析的准确性。

上述方案中,首先分别计算通过不同的正态变换方法对水质观测数据进行正态变换处理后对应参数的估计值,设置具体的衡量指标并进行计算和比对,从而根据水质观测的数据特征选择最优的正态变换方法,使得变换后的数据具有更好的正态性,最后经最优变换方法进行正态变换处理后的水质观测数据作为输入数据进行水质观测的统计分析,提升了分析的效果。

优选的,步骤 S1 中所述不同的正态变换方法包括有恒等变换、对数变换、Box-Cox 变换及 Yeo-Johnson 变换。

优选的,步骤 S1 中所述的估计参数采用最大似然函数的方法,并采用下山单纯形法进行求解。

优选的,所述步骤 S1 中计算通过 Box-Cox 变换对水质观测数据进行正态变换处理后对应参数的估计值的具体步骤为:

定义获取到的非正态分布的水质观测数据序列为 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, λ 为变换参数, $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ 为输出序列;

若 $\mathbf{x}$ 中各项均为正数,则 Box-Cox 变换的函数形式为:

$$y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x_i & \lambda = 0 \end{cases}, \quad i = 1, 2, \dots, n$$

若 $\mathbf{x}$ 中存在 $x_i \leq 0$, 则对整个水质观测数据序列进行平移 ε , 使 $x_i + \varepsilon > 0$, 对应的 Box-Cox 变换的函数形式如下:

$$y_i = \begin{cases} \frac{(x_i + \varepsilon)^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x_i + \varepsilon) & \lambda = 0 \end{cases}, \quad i = 1, 2, \dots, n$$

其中变换参数 λ 通过最大似然法进行估计;

定义经过变换后,水质观测数据服从均值为 μ , 方差为 σ^2 的正态分布,则变换后输出的第 i 个水质观测数据 y_i 的密度为:

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

$\mathbf{x}$ 的密度为:

$$L(\lambda, \mu, \sigma | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\} J(\lambda; \mathbf{x})$$

其中, $J(\lambda; \mathbf{x})$ 为变换的雅可比矩阵:

$$J(\lambda, \mathbf{x}) = \prod_{i=1}^n \left| \frac{dy_i}{dx_i} \right|$$

$$\frac{dy_i}{dx_i} = \begin{cases} x_i^{\lambda-1} & \lambda \neq 0 \\ \frac{1}{x_i} & \lambda = 0 \end{cases}$$

若 $\mathbf{x}$ 中各项均为正数, 获取对数似然函数为:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i)$$

令其中的 $\log \sigma = s$, $\mu / \sigma = v$, 同时去掉常数项 $(-n \log(2\pi) / 2)$, 得到:

$$l_n(\lambda, v, s | \mathbf{x}) = -ns - \frac{1}{2\exp(2s)} \sum_{i=1}^n (y_i - v \exp(s))^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i)$$

对上式的对数似然函数取负值, 然后采用数值法求解使对数似然函数的函数值最小的参数组合, 得到最小负对数似然函数值 $-L$, 则最大似然函数值为 L ;

若 $\mathbf{x}$ 中存在 $x_i \leq 0$, 获取对数似然函数为:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i + \varepsilon)$$

令其中的 $\log \sigma = s$, $\mu / \sigma = v$, 同时去掉常数项 $(-n \log(2\pi) / 2)$, 得到:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i + \varepsilon)$$

对上式的对数似然函数取负值, 然后采用数值法求解使对数似然函数的函数值最小的参数组合, 得到最小负对数似然函数值 $-L$, 则最大似然函数值为 L 。

优选的, 所述步骤 S1 中计算通过 Yeo-Johnson 变换对水质观测数据进行正

态变换处理后对应参数的估计值的具体步骤为:

定义获取到的非正态分布的水质观测数据序列为 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, λ 为变换参数, $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ 为输出序列;

则 Yeo-Johnson 变换的函数形式为:

$$y_i = \begin{cases} \frac{(x_i + 1)^\lambda - 1}{\lambda} & x_i \geq 0, \lambda \neq 0 \\ \log(x_i + 1) & x_i \geq 0, \lambda = 0 \\ -\frac{(-x_i + 1)^{2-\lambda} - 1}{2 - \lambda} & x_i < 0, \lambda \neq 2 \\ -\log(-x_i + 1) & x_i < 0, \lambda = 2 \end{cases}, \quad i = 1, 2, \dots, n$$

其中变换参数 λ 通过最大似然法进行估计;

定义经过变换后, 水质观测数据服从均值为 μ , 方差为 σ^2 的正态分布, 则变换后输出的第 i 个水质观测数据 y_i 的密度为:

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

$\mathbf{x}$ 的密度为:

$$L(\lambda, \beta, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\} J(\lambda; \mathbf{x})$$

其中, $J(\lambda; \mathbf{x})$ 为变换的雅可比矩阵:

$$J(\lambda, \mathbf{x}) = \prod_{i=1}^n \left| \frac{dy_i}{dx_i} \right|$$

$$\frac{dy_i}{dx_i} = \begin{cases} (x_i + 1)^{\lambda-1} & x_i \geq 0, \lambda \neq 0 \\ \frac{1}{x_i + 1} & x_i \geq 0, \lambda = 0 \\ (-x_i + 1)^{1-\lambda} & x_i < 0, \lambda \neq 2 \\ \frac{1}{-x_i + 1} & x_i < 0, \lambda = 2 \end{cases}$$

$$\log \frac{dy_i}{dx_i} = (\lambda - 1) \operatorname{sgn}(x_i) \log(|x_i| + 1) = \begin{cases} (\lambda - 1) \log(x_i + 1) & x \geq 0 \\ (1 - \lambda) \log(-x_i + 1) & x < 0 \end{cases}$$

获取对数似然函数为:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \text{sgn}(x_i) \log(|x_i| + 1)$$

其中， $\text{sgn}(\cdot)$ 为符号函数，当其中的变量 x_i 为正时取值为 1，当其中的变量取值 x_i 为负时为-1，否则取值为 0；

令其中的 $\log \sigma = s$ ， $\mu / \sigma = v$ ，同时去掉常数项 $(-n \log(2\pi) / 2)$ 后，得到：

$$l_n(\lambda, v, s | \mathbf{x}) = -ns - \frac{1}{2 \exp(2s)} \sum_{i=1}^n (y_i - v \exp(s))^2 + (\lambda - 1) \sum_{i=1}^n \text{sgn}(x_i) \log(|x_i| + 1)$$

对上式的对数似然函数取负值，然后采用数值法求解使对数似然函数的函数值最小的参数组合，得到最小负对数似然函数值 $-L$ ，则最大似然函数值为 L 。

优选的，根据计算得到的最小负对数似然函数值、AIC 值及 BIC 值，首先选出最小负对数似然函数值、AIC 值及 BIC 值三者同时低于原始水质观测数据的对应参数值所对应的正态变换方法，否则认为原始的水质观测数据满足正态性假设，不对其进行变换，结束本步骤；

若最小负对数似然函数值、AIC 值及 BIC 值三者同时低于原始水质观测数据的对应参数值所对应的正态变换方法有多个，则其中最低的 BIC 值所对应的正态变换方法为最优变换方法；

其中最小负对数似然函数值表示为 $-L$ ；

AIC 值表示为： $\text{AIC} = 2k - 2 \ln(L)$ ；

其中 k 是估计的参数数量， L 是最大似然函数值；

BIC 值表示为： $\text{BIC} = \ln(n)k - 2 \ln(L)$ ；

其中 k 是估计的参数数量， L 是最大似然函数值， n 为水质观测数据的个数。

与现有技术相比，本发明技术方案的有益效果是：

本发明方法通过对水质观测数据自身的携带信息确定变换参数，设置具体的衡量指标在多种正态变换方法中进行计算和比对，从而根据水质观测的数据特征选择最优的正态变换方法，最后经最优变换方法将序列转入一个服从或近似服从正态分布函数的空间内，得到与原序列相应的新序列，以排除数据序列中可能的非线性、异方差性和非正态性；通过幂变换方法直接对数据进行变换，变换后变量序列相对于原始值的序列不会改变，也就没有改变变量中某个特定值的概率密度，变换过程通过将原序列进行收敛或发散实现变量整体分布的改变。本发明方法能使得变换后的数据具有更好的正态性，便于进一步的数据分析，解决了由于

选择的变换方法不适应水厂观测变量自身特征导致数据的变换效果差等问题。

附图说明

图 1 为本发明方法的总流程图。

图 2 为本发明中使用的 Box-Cox 变换方法在不同参数下的变换效果图。

图 3 为本发明中使用的 Yeo-Johnson 变换方法在不同参数下的变换效果图。

图 4 为实施例 2 中原始水质观测序列的 Q-Q 图。

图 5 为实施例 2 中经 Box-Cox 变换后水质观测序列的 Q-Q 图。

图 6 为实施例 2 中经 Yeo-Johnson 变换后水质观测序列的 Q-Q 图。

图 7 为实施例 2 中经对数变换后水质观测序列的 Q-Q 图。

图 8 为实施例 2 中原始水质观测序列的示意图。

图 9 为实施例 2 中经对数变换后水质观测序列的示意图。

图 10 为实施例 2 中原始水质观测数据的分布图。

图 11 为实施例 2 中经对数变换后水质观测数据分布图。

图 12 为实施例 2 中逆变换后的水质观测数据与原始水质观测数据的关系图。

图 13 为实施例 2 中自回归统计分析结果经逆变换所得序列与原始水质观测序列对比示意图。

具体实施方式

附图仅用于示例性说明，不能理解为对本专利的限制；

为了更好说明本实施例，附图某些部件会有省略、放大或缩小，并不代表实际产品的尺寸；

对于本领域技术人员来说，附图中某些公知结构及其说明可能省略是可以理解的。

下面结合附图和实施例对本发明的技术方案做进一步的说明。

实施例 1

一种面向非正态分布水质观测数据的幂变换分析方法，包括以下步骤：

S1. 获取非正态分布的水质观测数据，分别计算通过不同的正态变换方法对水质观测数据进行正态变换处理后对应参数的估计值，对应参数的估计值包括：进行正态变换处理后，水质观测数据分布的均值、标准差以及变换参数；其中本实施例 1 中，不同的正态变换方法包括有恒等变换、对数变换、Box-Cox 变换及 Yeo-Johnson 变换；其中估计参数采用最大似然函数的方法，并采用下山单纯形

法进行求解；

对于 Box-Cox 变换，步骤 S1 中计算通过 Box-Cox 变换对水质观测数据进行正态变换处理后对应参数的估计值的具体步骤为：

定义获取到的非正态分布的水质观测数据序列为 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ ， λ 为变换参数， $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ 为输出序列；

若 $\mathbf{x}$ 中各项均为正数，则 Box-Cox 变换的函数形式为：

$$y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x_i & \lambda = 0 \end{cases}, \quad i = 1, 2, \dots, n$$

若 $\mathbf{x}$ 中存在 $x_i \leq 0$ ，则对整个水质观测数据序列进行平移 ε ，使 $x_i + \varepsilon > 0$ ，对应的 Box-Cox 变换的函数形式如下：

$$y_i = \begin{cases} \frac{(x_i + \varepsilon)^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x_i + \varepsilon) & \lambda = 0 \end{cases}, \quad i = 1, 2, \dots, n$$

其中变换参数 λ 通过最大似然法进行估计；

定义经过变换后，水质观测数据服从均值为 μ ，方差为 σ^2 的正态分布，则变换后输出的第 i 个水质观测数据 y_i 的密度为：

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

$\mathbf{x}$ 的密度为：

$$L(\lambda, \mu, \sigma | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\} J(\lambda; \mathbf{x})$$

其中， $J(\lambda; \mathbf{x})$ 为变换的雅可比矩阵：

$$J(\lambda, \mathbf{x}) = \prod_{i=1}^n \left| \frac{dy_i}{dx_i} \right|$$

$$\frac{dy_i}{dx_i} = \begin{cases} x_i^{\lambda-1} & \lambda \neq 0 \\ \frac{1}{x_i} & \lambda = 0 \end{cases}$$

若 $\mathbf{x}$ 中各项均为正数，获取对数似然函数为：

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i)$$

令其中的 $\log \sigma = s$, $\mu / \sigma = v$, 同时去掉常数项 $(-n \log(2\pi) / 2)$, 得到:

$$l_n(\lambda, v, s | \mathbf{x}) = -ns - \frac{1}{2\exp(2s)} \sum_{i=1}^n (y_i - v \exp(s))^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i)$$

对上式的对数似然函数取负值, 然后采用数值法求解使对数似然函数的函数值最小的参数组合, 得到最小负对数似然函数值 $-L$, 则最大似然函数值为 L ;

若 $\mathbf{x}$ 中存在 $x_i \leq 0$, 获取对数似然函数为:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i + \varepsilon)$$

令其中的 $\log \sigma = s$, $\mu / \sigma = v$, 同时去掉常数项 $(-n \log(2\pi) / 2)$, 得到:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i + \varepsilon)$$

对上式的对数似然函数取负值, 然后采用数值法求解使对数似然函数的函数值最小的参数组合, 得到最小负对数似然函数值 $-L$, 则最大似然函数值为 L 。

对于 Yeo-Johnson 变换, 步骤 S1 中计算通过 Yeo-Johnson 变换对水质观测数据进行正态变换处理后对应参数的估计值的具体步骤为:

定义获取到的非正态分布的水质观测数据序列为 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, λ 为变换参数, $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ 为输出序列;

则 Yeo-Johnson 变换的函数形式为:

$$y_i = \begin{cases} \frac{(x_i + 1)^\lambda - 1}{\lambda} & x_i \geq 0, \lambda \neq 0 \\ \log(x_i + 1) & x_i \geq 0, \lambda = 0 \\ -\frac{(-x_i + 1)^{2-\lambda} - 1}{2 - \lambda} & x_i < 0, \lambda \neq 2 \\ -\log(-x_i + 1) & x_i < 0, \lambda = 2 \end{cases}, \quad i = 1, 2, \dots, n$$

其中变换参数 λ 通过最大似然法进行估计;

定义经过变换后, 水质观测数据服从均值为 μ , 方差为 σ^2 的正态分布, 则变换后输出的第 i 个水质观测数据 y_i 的密度为:

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

$\mathbf{x}$ 的密度为:

$$L(\lambda, \beta, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \right\} J(\lambda; \mathbf{x})$$

其中, $J(\lambda; \mathbf{x})$ 为变换的雅可比矩阵:

$$J(\lambda, \mathbf{x}) = \prod_{i=1}^n \left| \frac{dy_i}{dx_i} \right|$$

$$\frac{dy_i}{dx_i} = \begin{cases} (x_i + 1)^{\lambda-1} & x_i \geq 0, \lambda \neq 0 \\ \frac{1}{x_i + 1} & x_i \geq 0, \lambda = 0 \\ (-x_i + 1)^{1-\lambda} & x_i < 0, \lambda \neq 2 \\ \frac{1}{-x_i + 1} & x_i < 0, \lambda = 2 \end{cases}$$

$$\log \frac{dy_i}{dx_i} = (\lambda - 1) \operatorname{sgn}(x_i) \log(|x_i| + 1) = \begin{cases} (\lambda - 1) \log(x_i + 1) & x \geq 0 \\ (1 - \lambda) \log(-x_i + 1) & x < 0 \end{cases}$$

获取对数似然函数为:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \operatorname{sgn}(x_i) \log(|x_i| + 1)$$

其中, $\operatorname{sgn}(\cdot)$ 为符号函数, 当其中的变量 x_i 为正时取值为 1, 当其中的变量取值 x_i 为负时为-1, 否则取值为 0;

令其中的 $\log \sigma = s$, $\mu / \sigma = v$, 同时去掉常数项 $(-n \log(2\pi) / 2)$ 后, 得到:

$$l_n(\lambda, v, s | \mathbf{x}) = -ns - \frac{1}{2 \exp(2s)} \sum_{i=1}^n (y_i - v \exp(s))^2 + (\lambda - 1) \sum_{i=1}^n \operatorname{sgn}(x_i) \log(|x_i| + 1)$$

对上式的对数似然函数取负值, 然后采用数值法求解使对数似然函数的函数值最小的参数组合, 得到最小负对数似然函数值 $-L$, 则最大似然函数值为 L 。

S2. 对于步骤 S1 中所述的不同的正态变换方法, 分别计算每种正态变换方法相对应的最小负对数似然函数值、AIC 值及 BIC 值;

S3. 根据计算得到的最小负对数似然函数值、AIC 值及 BIC 值, 与预设的选择标准进行比对, 根据比对结果从所述不同的正态变换方法中选择得到最优变换方法; 具体如下:

根据计算得到的最小负对数似然函数值、AIC 值及 BIC 值, 首先选出最小

负对数似然函数值、AIC 值及 BIC 值三者同时低于原始水质观测数据的对应参数值所对应的正态变换方法，否则认为原始的水质观测数据满足正态性假设，不对其进行变换，结束本步骤；

若最小负对数似然函数值、AIC 值及 BIC 值三者同时低于原始水质观测数据的对应参数值所对应的正态变换方法有多个，则其中最低的 BIC 值所对应的正态变换方法为最优变换方法；

定义估计参数时采用最大似然函数的方法获得的最大似然函数值为 L ，

则负对数似然函数值表示为 $-L$

AIC 值表示为： $AIC = 2k - 2\ln(L)$

其中 k 是估计的参数数量， L 是最大似然函数值；

BIC 值表示为： $BIC = \ln(n)k - 2\ln(L)$

其中 k 是估计的参数数量， L 是最大似然函数值， n 为水质观测数据的个数。

S4. 将经所述最优变换方法进行正态变换处理后的水质观测数据作为输入数据进行水质观测的统计分析，将统计分析得到的结果进行逆变换，从而得到最终的分析结果；记统计分析后获得的序列为 $\mathbf{z} = \{z_1, \dots, z_m\}$ ，其逆变换序列为 $\hat{\mathbf{x}} = \{\hat{x}_1, \dots, \hat{x}_m\}$ 。

其中对于经 Box-Cox 变换后的水质观测数据的逆变换形式如下：

对于 Box-Cox 单参数变换，即变换时没有对水质观测数据进行平移：

$$\hat{x}_i = \begin{cases} (z_i \lambda + 1)^{1/\lambda} & \lambda \neq 0 \\ \exp(z_i) & \lambda = 0 \end{cases}$$

其中 $\hat{x}_i$ 为逆变换后的水质数据， z_i 为经 Box-Cox 变换后的水质数据， λ 为变换参数；

对于 Box-Cox 双参数变换，即变换时对整个水质观测数据序列平移了 ε ：

$$\hat{x}_i = \begin{cases} (z_i \lambda + 1)^{1/\lambda} - \varepsilon & \lambda \neq 0 \\ \exp(z_i) - \varepsilon & \lambda = 0 \end{cases}$$

其中 $\hat{x}_i$ 为逆变换后的水质数据， z_i 为经 Box-Cox 变换后的水质数据， λ 为变换参数；

其中对于经 Yeo-Johnson 变换后的水质观测数据的逆变换形式如下：

$$\hat{x}_i = \begin{cases} (z_i \lambda + 1)^{1/\lambda} - 1 & y_i \geq 0, \lambda \neq 0 \\ \exp(z_i) - 1 & y_i \geq 0, \lambda = 0 \\ 1 - [1 - z_i(2 - \lambda)]^{1/(2-\lambda)} & y_i < 0, \lambda \neq 2 \\ 1 - \exp(z_i) & y_i < 0, \lambda = 2 \end{cases}$$

其中 $\hat{x}_i$ 为逆变换后的水质数据, z_i 为经 Yeo-Johnson 变换后的水质数据, λ 为变换参数。

不同参数的 Box-Cox 变换及 Yeo-Johnson 变换效果分别如图 2、3 所示, 两者对变量偏度的改变明显, 在一定程度上甚至会改变偏移的方向。本发明采取对观测变量同时进行不同变换, 选取合适的变换方法, 获得便于实际统计分析计算的变换结果。

实施例 2

本实施例 2 基于实施例 1 的方法进行实验, 以某污水处理厂入水化学需氧量 (COD) 日观测序列为实验数据, 该水质观测序列长度为 655 日, 相关统计参数如表 1 所示;

观测日数	均值	标准差	pvalue	
			正态分布	对数正态分布
655	398.41	182.96	0.00	0.53

表 1

根据表 1 对数正态分布及正态分布 K-S 检验 pvalue, 可以认为该序列服从对数正态分布, 作为输入的非正态分布的水质观测数据水观测序列, 分别对这些数据进行恒等、Box-Cox、Yeo-Johnson 和对数变换参数估计, 获得不同的变换结果。另外本实施例 2 还进一步根据变换结果绘制 quantile-quantile 图 (Q-Q 图), Q-Q 图采用图形的方法鉴别样本数据是否近似于正态分布, 通过 Q-Q 图可以比较直观的获取数据分布信息, 其主要用于辅助判断变换效果。Q-Q 图上的点 (x, y) 反映出其中一个样本数据的经验分布的分位数和正态分布的相同分位数, 若 Q-Q 图上的点近似一条对角直线, 则可认为数据点呈正态分布。本实施例 2 变换后水质观测序列的 Q-Q 图如图 4-7 所示, 变换参数估计结果如表 2 所示; 其中表格中的 Log 项指对数变换, nllf 指负对数似然函数值。

Parameter	Identity	Log	Box-Cox	Yeo-Johnson
-----------	----------	-----	---------	-------------

λ	/	/	0.57	0.04
μ	398.41	5.89	50.91	6.72
σ	182.82	0.45	14.06	0.58
nllf	3,739.07	3,663.61	3,687.02	3,663.46
AIC	7,478.14	7,327.23	7,376.05	7,328.92
BIC	7,478.14	7,327.23	7,380.53	7,333.41

表 2

由图 4-7 和表 2 可以得出,相比于原始的水质观测数据,即表格中的 Identity 项,经 Box-Cox、Yeo-Johnson 和对数变换后,数据的负对数似然函数值、AIC 和 BIC 值均有所降低,其中对数变换的 BIC 值是所有变换方法中最低的,为 7327.23,故采用对数变换作为最优变换方法对水质观测数据进行变换。

采用对数变换作为最优变换方法,对原始水质观测序列进行变换,得到的变换序列如图 9 所示,与图 8 的原始水质观测序列相比,整体波动更为平稳。通过图 10 和图 11 得到,变换后水质观测序列分布正态性得到明显提高,直接对变换序列进行逆变换后,绘制原始水质观测序列与逆变换序列散点图,如图 12 所示,逆变换后的数据与原始数据完全一致,说明变换过程并不会丢失原始数据信息。对变换后序列进行时间序列自回归分析,发现该水厂 COD 序列存在显著的自相关性,对该拟合结果进行逆变换,所得逆变换序列与原始序列对比结果如图 13 所示;从图 13 可以看到,自回归拟合序列结果基本消除了原始序列中存在的噪声,同时能够较好地概括原始序列的存在的变化趋势,利用该逆变换序列,可以进一步预测该水厂未来入水 COD 的变化趋势。

附图中描述位置关系的用语仅用于示例性说明,不能理解为对本专利的限制;

显然,本发明的上述实施例仅仅是为清楚地说明本发明所作的举例,而并非是对本发明的实施方式的限定。对于所属领域的普通技术人员来说,在上述说明的基础上还可以做出其它不同形式的变化或变动。这里无需也无法对所有的实施方式予以穷举。凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等,均应包含在本发明权利要求的保护范围之内。

权 利 要 求 书

1.一种面向非正态分布水质观测数据的幂变换分析方法，其特征在于：包括以下步骤：

S1.获取非正态分布的水质观测数据，分别计算通过不同的正态变换方法对水质观测数据进行正态变换处理后对应参数的估计值，对应参数的估计值包括：进行正态变换处理后，水质观测数据分布的均值、标准差以及变换参数；

S2.对于步骤 S1 中所述的不同的正态变换方法，分别计算每种正态变换方法相对应的最小负对数似然函数值、AIC 值及 BIC 值；

S3.根据计算得到的最小负对数似然函数值、AIC 值及 BIC 值，与预设的选择标准进行比对，根据比对结果从所述不同的正态变换方法中选择得到最优变换方法；

S4.将经所述最优变换方法进行正态变换处理后的水质观测数据作为输入数据进行水质观测的统计分析，将统计分析得到的结果进行逆变换，从而得到最终的分析结果。

2.根据权利要求 1 所述的面向非正态分布水质观测数据的幂变换分析方法，其特征在于，步骤 S1 中所述不同的正态变换方法包括有恒等变换、对数变换、Box-Cox 变换及 Yeo-Johnson 变换。

3.根据权利要求 2 所述的面向非正态分布水质观测数据的幂变换分析方法，其特征在于，步骤 S1 中所述的估计参数采用最大似然函数的方法，并采用下山单纯形法进行求解。

4.根据权利要求 3 所述的面向非正态分布水质观测数据的幂变换分析方法，其特征在于，所述步骤 S1 中计算通过 Box-Cox 变换对水质观测数据进行正态变换处理后对应参数的估计值的具体步骤为：

定义获取到的非正态分布的水质观测数据序列为 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$ ， λ 为变换参数， $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ 为输出序列；

若 $\mathbf{x}$ 中各项均为正数，则 Box-Cox 变换的函数形式为：

$$y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log x_i & \lambda = 0 \end{cases}, \quad i = 1, 2, \dots, n$$

若 $\mathbf{x}$ 中存在 $x_i \leq 0$ ，则对整个水质观测数据序列进行平移 ε ，使 $x_i + \varepsilon > 0$ ，对应的 Box-Cox 变换的函数形式如下：

$$y_i = \begin{cases} \frac{(x_i + \varepsilon)^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log(x_i + \varepsilon) & \lambda = 0 \end{cases}, \quad i = 1, 2, \dots, n$$

其中变换参数 λ 通过最大似然法进行估计；

定义经过变换后，水质观测数据服从均值为 μ ，方差为 σ^2 的正态分布，则变换后输出的第 i 个水质观测数据 y_i 的密度为：

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

$\mathbf{x}$ 的密度为：

$$L(\lambda, \mu, \sigma | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2}\right\} J(\lambda; \mathbf{x})$$

其中， $J(\lambda; \mathbf{x})$ 为变换的雅可比矩阵：

$$J(\lambda, \mathbf{x}) = \prod_{i=1}^n \left| \frac{dy_i}{dx_i} \right|$$

$$\frac{dy_i}{dx_i} = \begin{cases} x_i^{\lambda-1} & \lambda \neq 0 \\ \frac{1}{x_i} & \lambda = 0 \end{cases}$$

若 $\mathbf{x}$ 中各项均为正数，获取对数似然函数为：

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i)$$

令其中的 $\log \sigma = s$ ， $\mu / \sigma = v$ ，同时去掉常数项 $(-n \log(2\pi) / 2)$ ，得到：

$$l_n(\lambda, v, s | \mathbf{x}) = -ns - \frac{1}{2\exp(2s)} \sum_{i=1}^n (y_i - v \exp(s))^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i)$$

对上式的对数似然函数取负值，然后采用数值法求解使对数似然函数的函数值最小的参数组合，得到最小负对数似然函数值 $-L$ ，则最大似然函数值为 L ；

若 $\mathbf{x}$ 中存在 $x_i \leq 0$ ，获取对数似然函数为：

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i + \varepsilon)$$

令其中的 $\log \sigma = s$, $\mu / \sigma = v$, 同时去掉常数项 $(-n \log(2\pi) / 2)$, 得到:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \log(x_i + \varepsilon)$$

对上式的对数似然函数取负值, 然后采用数值法求解使对数似然函数的函数值最小的参数组合, 得到最小负对数似然函数值 $-L$, 则最大似然函数值为 L 。

5. 根据权利要求 3 所述的面向非正态分布水质观测数据的幂变换分析方法, 其特征在于, 所述步骤 S1 中计算通过 Yeo-Johnson 变换对水质观测数据进行正态变换处理后对应参数的估计值的具体步骤为:

定义获取到的非正态分布的水质观测数据序列为 $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$, λ 为变换参数, $\mathbf{y} = \{y_1, y_2, \dots, y_n\}$ 为输出序列;

则 Yeo-Johnson 变换的函数形式为:

$$y_i = \begin{cases} \frac{(x_i + 1)^\lambda - 1}{\lambda} & x_i \geq 0, \lambda \neq 0 \\ \log(x_i + 1) & x_i \geq 0, \lambda = 0 \\ -\frac{(-x_i + 1)^{2-\lambda} - 1}{2-\lambda} & x_i < 0, \lambda \neq 2 \\ -\log(-x_i + 1) & x_i < 0, \lambda = 2 \end{cases}, \quad i = 1, 2, \dots, n$$

其中变换参数 λ 通过最大似然法进行估计;

定义经过变换后, 水质观测数据服从均值为 μ , 方差为 σ^2 的正态分布, 则变换后输出的第 i 个水质观测数据 y_i 的密度为:

$$f(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right]$$

$\mathbf{x}$ 的密度为:

$$L(\lambda, \beta, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{\sum_{i=1}^n (y_i - \mu)^2}{2\sigma^2} \right\} J(\lambda; \mathbf{x})$$

其中, $J(\lambda; \mathbf{x})$ 为变换的雅可比矩阵:

$$J(\lambda, \mathbf{x}) = \prod_{i=1}^n \left| \frac{dy_i}{dx_i} \right|$$

$$\frac{dy_i}{dx_i} = \begin{cases} (x_i + 1)^{\lambda-1} & x_i \geq 0, \lambda \neq 0 \\ \frac{1}{x_i + 1} & x_i \geq 0, \lambda = 0 \\ (-x_i + 1)^{1-\lambda} & x_i < 0, \lambda \neq 2 \\ \frac{1}{-x_i + 1} & x_i < 0, \lambda = 2 \end{cases}$$

$$\log \frac{dy_i}{dx_i} = (\lambda - 1) \operatorname{sgn}(x_i) \log(|x_i| + 1) = \begin{cases} (\lambda - 1) \log(x_i + 1) & x \geq 0 \\ (1 - \lambda) \log(-x_i + 1) & x < 0 \end{cases}$$

获取对数似然函数为:

$$l_n(\lambda, \mu, \sigma | \mathbf{x}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + (\lambda - 1) \sum_{i=1}^n \operatorname{sgn}(x_i) \log(|x_i| + 1)$$

其中, $\operatorname{sgn}(\cdot)$ 为符号函数, 当其中的变量 x_i 为正时取值为 1, 当其中的变量取值 x_i 为负时为-1, 否则取值为 0;

令其中的 $\log \sigma = s$, $\mu / \sigma = v$, 同时去掉常数项 $(-n \log(2\pi) / 2)$ 后, 得到:

$$l_n(\lambda, v, s | \mathbf{x}) = -ns - \frac{1}{2 \exp(2s)} \sum_{i=1}^n (y_i - v \exp(s))^2 + (\lambda - 1) \sum_{i=1}^n \operatorname{sgn}(x_i) \log(|x_i| + 1)$$

对上式的对数似然函数取负值, 然后采用数值法求解使对数似然函数的函数值最小的参数组合, 得到最小负对数似然函数值 $-L$, 则最大似然函数值为 L 。

6. 根据权利要求 1~5 任一项所述的面向非正态分布水质观测数据的幂变换分析方法, 其特征在于, 根据计算得到的最小负对数似然函数值、AIC 值及 BIC 值, 首先选出最小负对数似然函数值、AIC 值及 BIC 值三者同时低于原始水质观测数据的对应参数值所对应的正态变换方法, 否则认为原始的水质观测数据满足正态性假设, 不对其进行变换, 结束本步骤;

若最小负对数似然函数值、AIC 值及 BIC 值三者同时低于原始水质观测数据的对应参数值所对应的正态变换方法有多个, 则其中最低的 BIC 值所对应的正态变换方法为最优变换方法;

其中最小负对数似然函数值表示为 $-L$;

AIC 值表示为: $\text{AIC} = 2k - 2 \ln(L)$;

其中 k 是估计的参数数量, L 是最大似然函数值;

BIC 值表示为: $BIC = \ln(n)k - 2 \ln(L)$;

其中 k 是估计的参数数量, L 是最大似然函数值, n 为水质观测数据的个数。

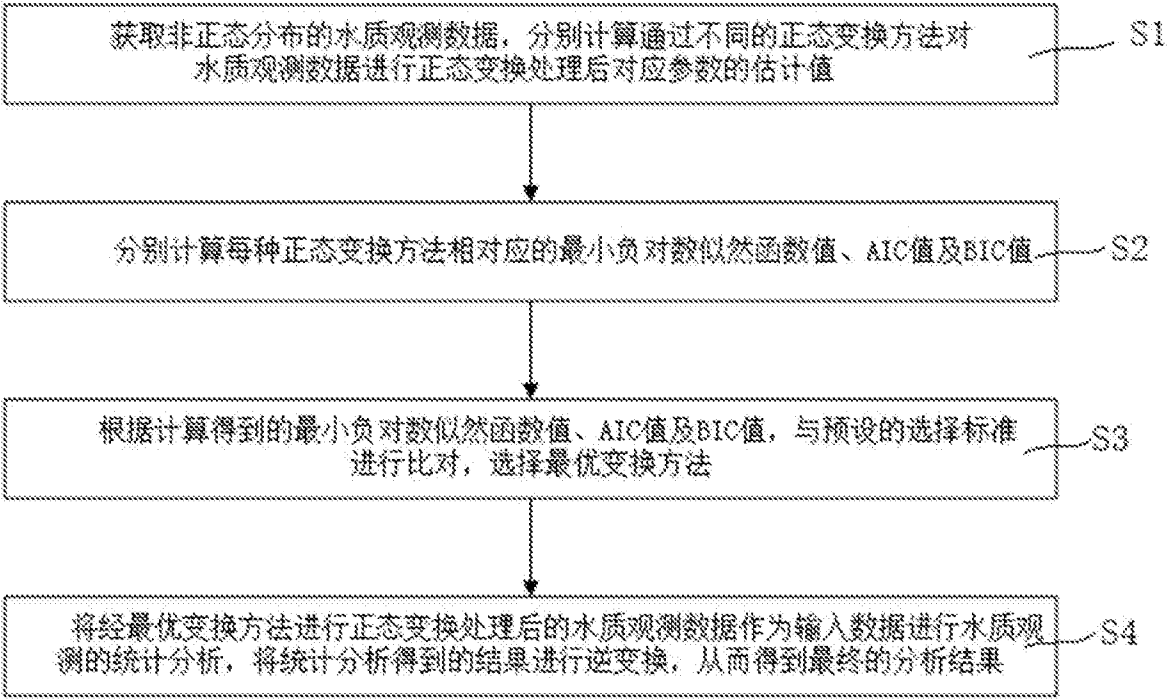


图 1

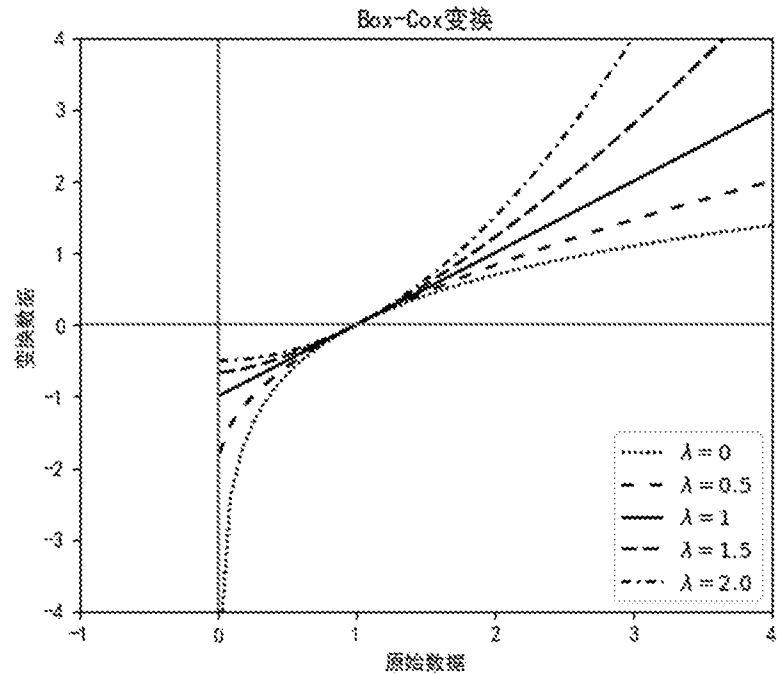


图 2

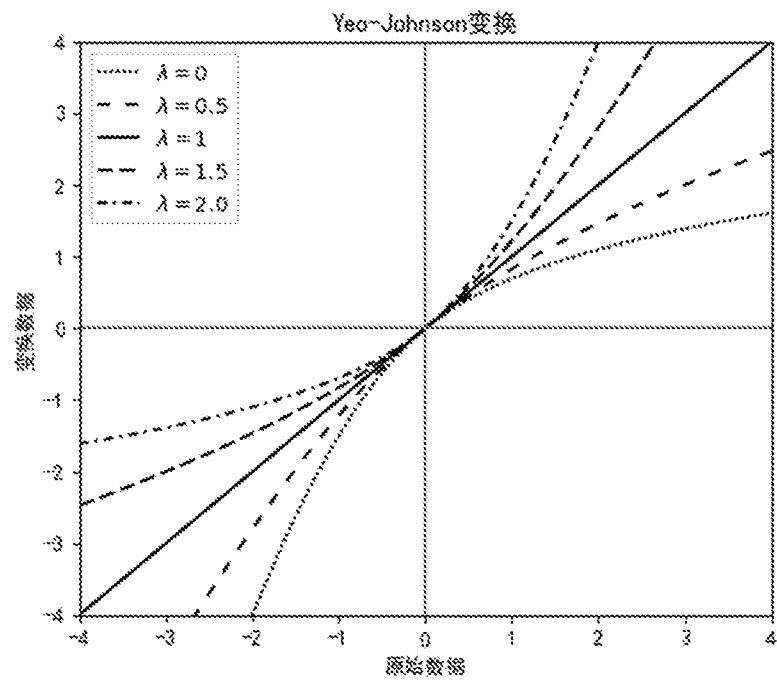


图 3

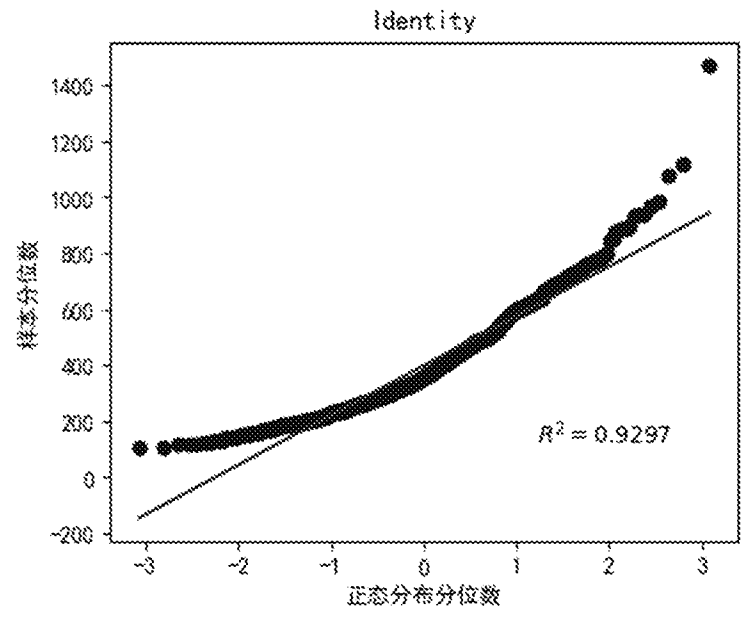


图 4

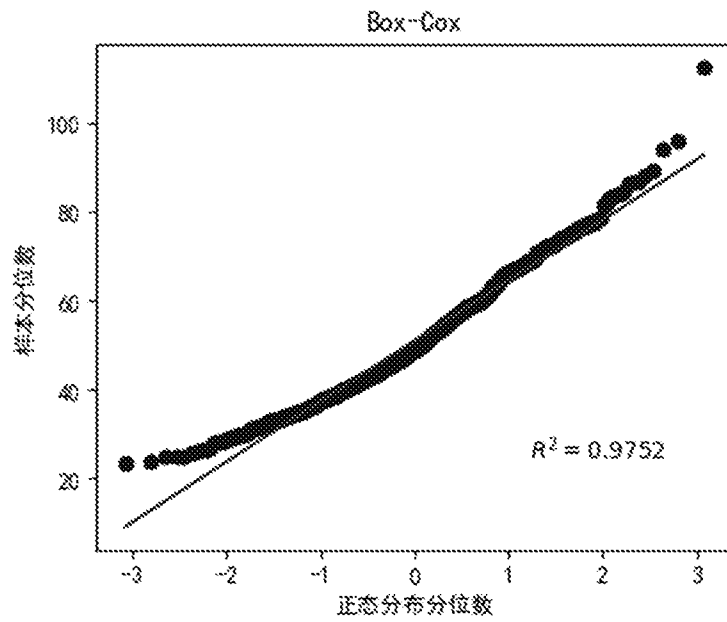


图 5

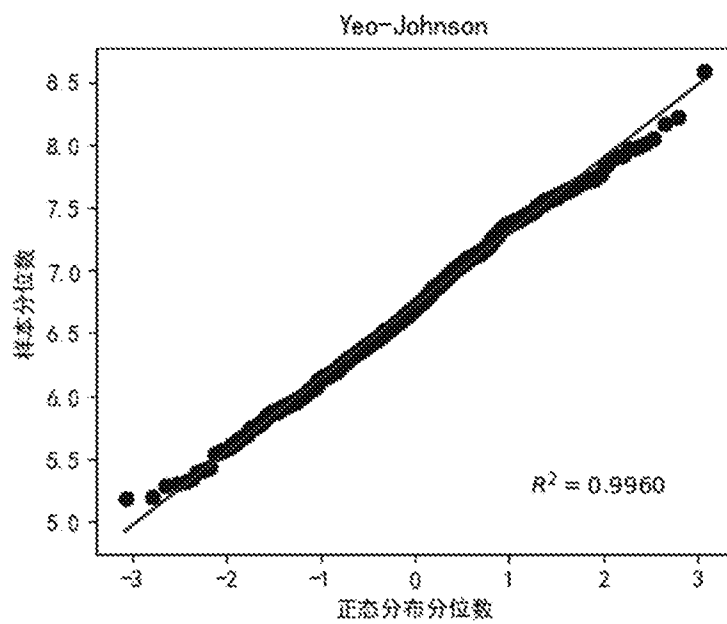


图 6

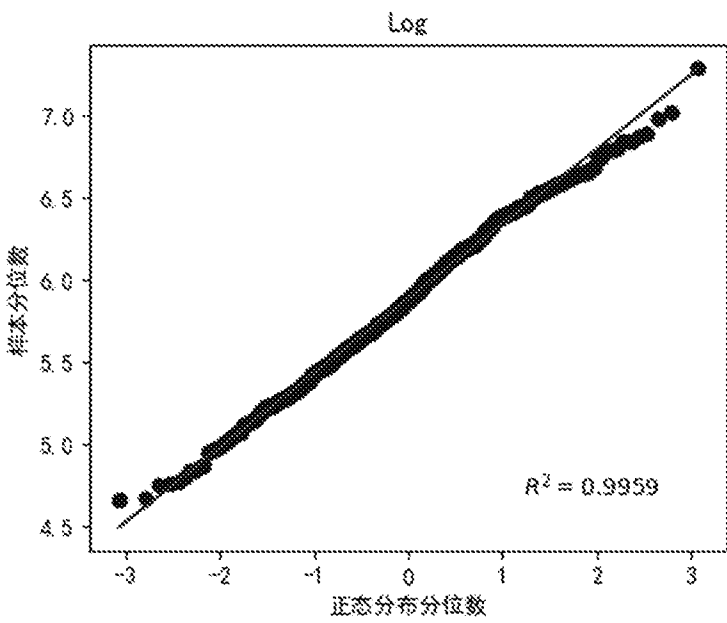


图 7

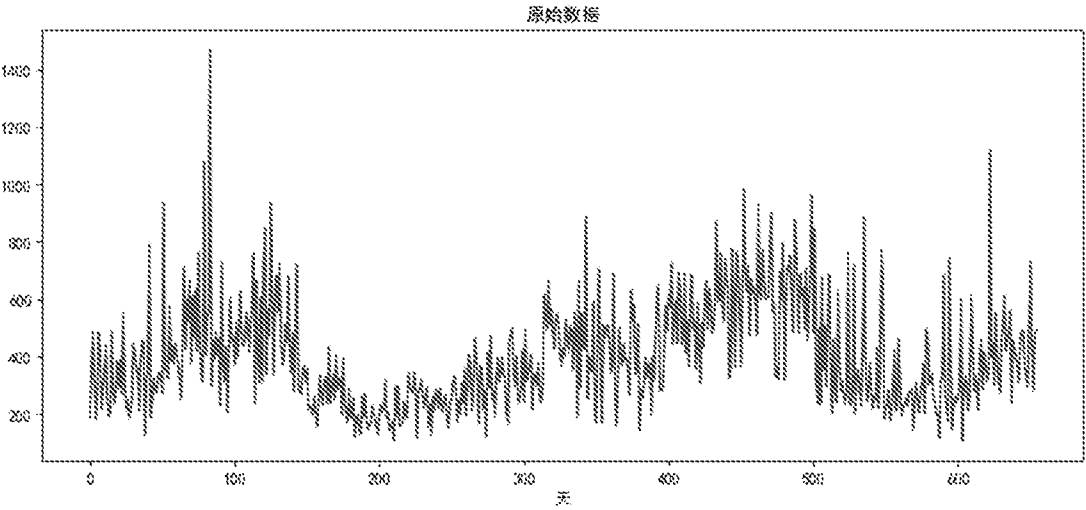


图 8

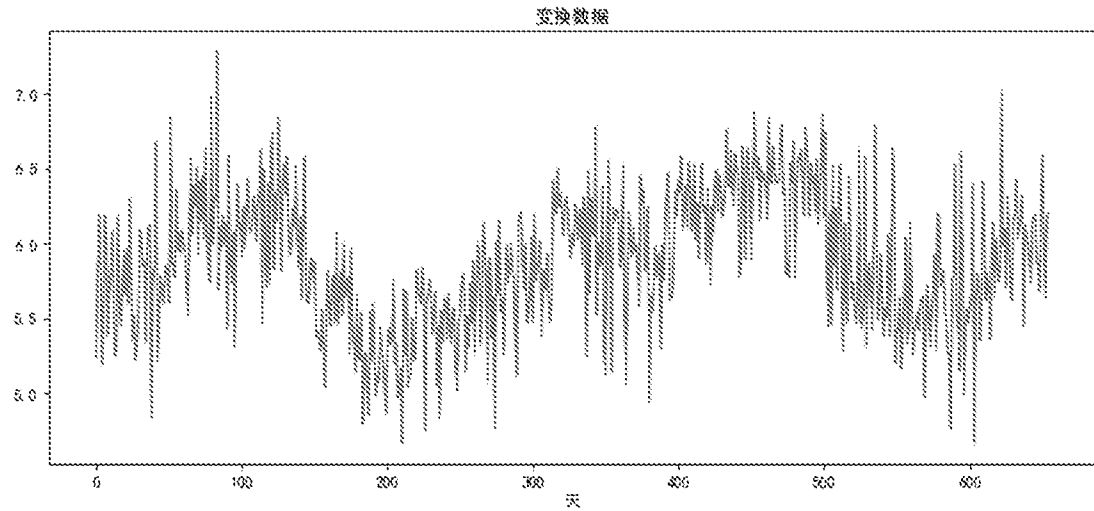


图 9

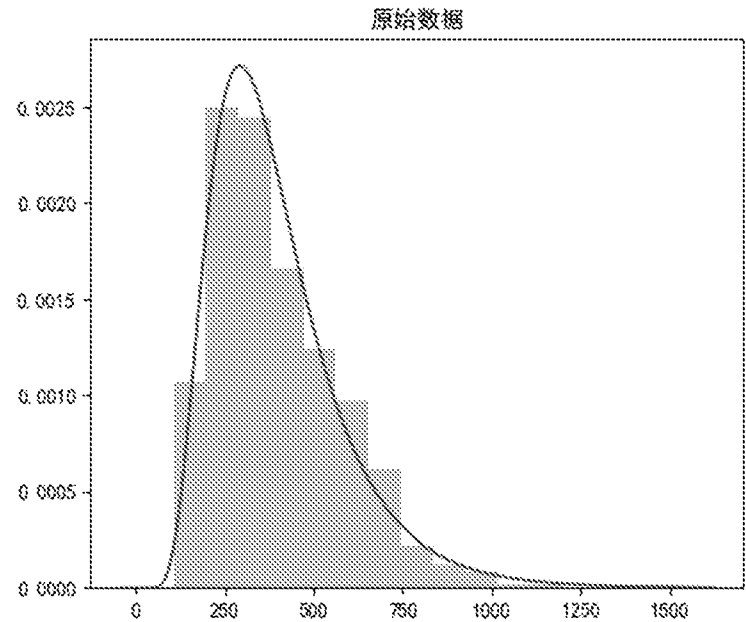


图 10

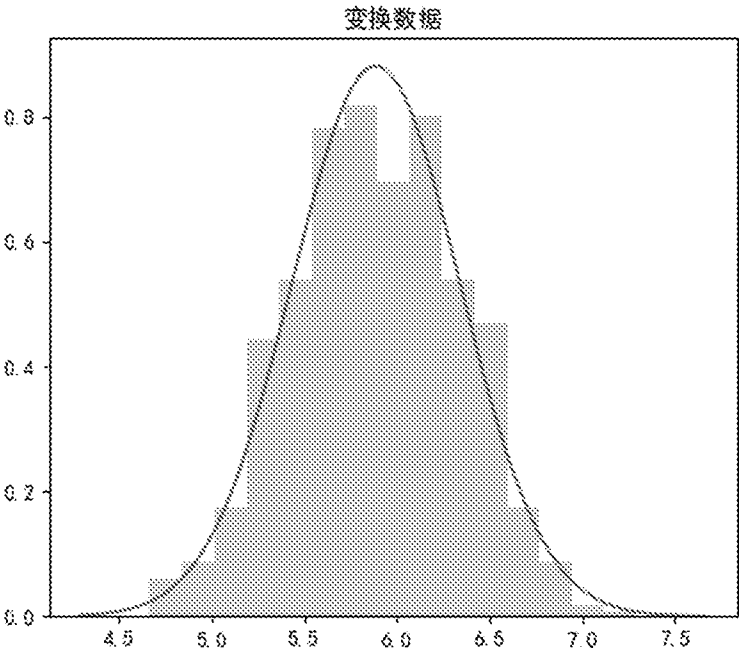


图 11

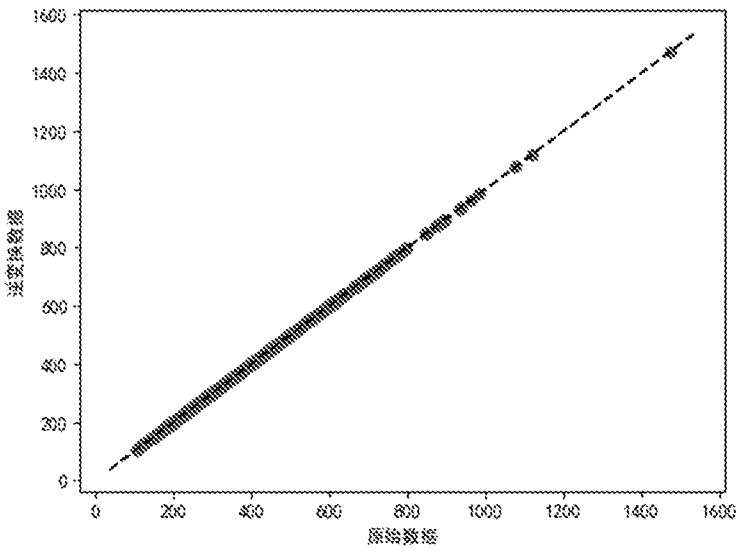


图 12

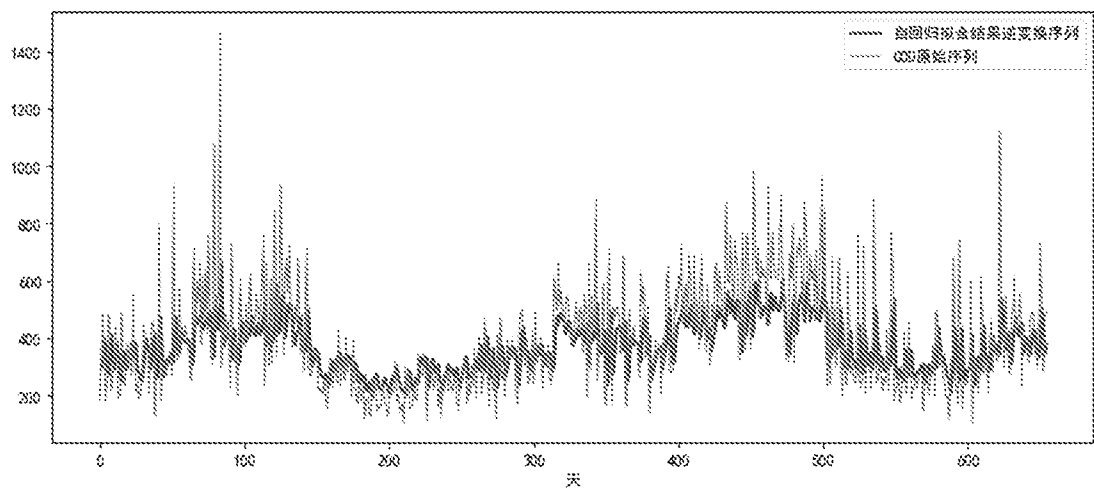


图 13

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2020/078258

A. CLASSIFICATION OF SUBJECT MATTER

G06F 16/00(2019.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

DWPI; SIPOABS; CNABS; CNKI; CNTXT: 正态分布, 非正态分布, 最优, 转换, normal distribution, non normal distribution, optimal, transformation

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN 110309199 A (SUN YAT-SEN UNIVERSITY) 08 October 2019 (2019-10-08) Claims 1-6	1-6
Y	CN 104899419 A (TSINGHUA UNIVERSITY) 09 September 2015 (2015-09-09) description, paragraphs 47-100	1-6
Y	CN 102855757 A (ZHEJIANG UNIVERSITY) 02 January 2013 (2013-01-02) description, paragraphs 53-79	1-6
A	US 2009157351 A1 (XEROX CORP.) 18 June 2009 (2009-06-18) entire document	1-6



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

12 May 2020

Date of mailing of the international search report

12 June 2020

Name and mailing address of the ISA/CN

China National Intellectual Property Administration (ISA/ CN)
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2020/078258

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	110309199	A	08 October 2019	None			
CN	104899419	A	09 September 2015	None			
CN	102855757	A	02 January 2013	CN	102855757	B	21 May 2014
US	2009157351	A1	18 June 2009	US	7925460	B2	12 April 2011

国际检索报告

国际申请号

PCT/CN2020/078258

A. 主题的分类 G06F 16/00 (2019.01) i 按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类																	
B. 检索领域 检索的最低限度文献 (标明分类系统和分类号) G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用)) DWPI; SIPOABS; CNABS; CNKI; CNTXT: 正态分布, 非正态分布, 最优, 转换, normal distribution, non normal distribution, optimal, transformation																	
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>PX</td> <td>CN 110309199 A (中山大学) 2019年 10月 8日 (2019 - 10 - 08) 权利要求1-6</td> <td>1-6</td> </tr> <tr> <td>Y</td> <td>CN 104899419 A (清华大学) 2015年 9月 9日 (2015 - 09 - 09) 说明书第47-100段</td> <td>1-6</td> </tr> <tr> <td>Y</td> <td>CN 102855757 A (浙江大学) 2013年 1月 2日 (2013 - 01 - 02) 说明书第53-79段</td> <td>1-6</td> </tr> <tr> <td>A</td> <td>US 2009157351 A1 (XEROX CORP) 2009年 6月 18日 (2009 - 06 - 18) 全文</td> <td>1-6</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	PX	CN 110309199 A (中山大学) 2019年 10月 8日 (2019 - 10 - 08) 权利要求1-6	1-6	Y	CN 104899419 A (清华大学) 2015年 9月 9日 (2015 - 09 - 09) 说明书第47-100段	1-6	Y	CN 102855757 A (浙江大学) 2013年 1月 2日 (2013 - 01 - 02) 说明书第53-79段	1-6	A	US 2009157351 A1 (XEROX CORP) 2009年 6月 18日 (2009 - 06 - 18) 全文	1-6
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求															
PX	CN 110309199 A (中山大学) 2019年 10月 8日 (2019 - 10 - 08) 权利要求1-6	1-6															
Y	CN 104899419 A (清华大学) 2015年 9月 9日 (2015 - 09 - 09) 说明书第47-100段	1-6															
Y	CN 102855757 A (浙江大学) 2013年 1月 2日 (2013 - 01 - 02) 说明书第53-79段	1-6															
A	US 2009157351 A1 (XEROX CORP) 2009年 6月 18日 (2009 - 06 - 18) 全文	1-6															
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																	
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																	
国际检索实际完成的日期 2020年 5月 12日		国际检索报告邮寄日期 2020年 6月 12日															
ISA/CN的名称和邮寄地址 中国国家知识产权局 (ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		授权官员 王艳坤 电话号码 86-(10)-62412330															

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2020/078258

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
CN	110309199	A	2019年 10月 8日	无	
CN	104899419	A	2015年 9月 9日	无	
CN	102855757	A	2013年 1月 2日	CN 102855757	B 2014年 5月 21日
US	2009157351	A1	2009年 6月 18日	US 7925460	B2 2011年 4月 12日