Advances in Applied Mathematics 应用数学进展, 2022, 11(9), 6161-6170
Published Online September 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.119649

Hans汉斯

基于ASAMC算法的地质层中氧元素含量的变点问题

张梦琇

石河子大学理学院，新疆 石河子

收稿日期：2022年7月29日；录用日期：2022年8月21日；发布日期：2022年9月1日

摘　要

变点是地质学的热门问题，多用于微量元素检测，具有很强的应用价值。由Kim，J. and Cheon，S.在2010年提出的退化后的随机逼近蒙特卡罗算法(Annealing Stochastic Approximation Monte Carlo，简称ASAMC)；ASAMC算法既可以检测变点个数，又可以检测变点所在的位置。本篇文章基于ASAMC算法，首先对地质层中氧元素含量进行正态性检验，随后，研究不同温度、不同季节下，土壤中氧元素含量的变化情况。最后，利用R软件，我们发现高温多雨的天气会较明显的影响土壤中氧元素的含量。

关键词

变点，ASAMC算法，正态分布均值变点，R软件

A Change-Point Problem of Oxygen Element in Geological Strata Is Based on ASAMC Algorithm

Mengxiu Zhang

College of Science, Shihezi University, Shihezi Xinjiang

Received: Jul. 29th, 2022; accepted: Aug. 21st, 2022; published: Sep. 1st, 2022

Abstract

The change-point is a hot issue in geology, and it is widely used in tracing element detection, and it

文章引用: 张梦琇. 基于 ASAMC 算法的地质层中氧元素含量的变点问题[J]. 应用数学进展, 2022, 11(9): 6161-6170.
DOI: 10.12677/aam.2022.119649

has strong application value. The annealing stochastic approximation Monte Carlo (ASAMC) algorithm proposed by Kim, J. and Cheon, S. in 2010 can detect both the number of change points and the location of change points. In this paper, based on the ASAMC algorithm, the normality test of oxygen content in the geological layer is carried out firstly. Then, the change of oxygen content in the soil under different temperatures and seasons is studied. Last, using R software, we find that the high temperature and rainy weather will obviously affect the oxygen content in the soil.

Keywords

The Change-Point, Annealing Stochastic Approximation Monte Carlo, Change Point of Normal Distribution of Mean, R Software

1. 引言

在大数据背景下，各种分布的极限分布均为正态分布。因此，考察正态分布均值变点模型，对于研究大数据下各种分布的均值变点模型来说都有很大的应用价值，正态分布均值变点已成为解决其他分布均值变点较为普遍的方法。这篇文章基于 ASAMC 算法，首先对地质层中氧元素含量进行正态性检验，随后，利用 R 软件研究不同温度、不同季节下，土壤中氧元素含量的变化情况，我们发现高温多雨的天气会较明显的影响土壤中氧元素的含量。

ASAMC 算法是通过 SAMC 算法演变而来，这种方法最早出现在 Kim，J. and Cheon，S (2010) [1]的文章中，随后，许欢在 2016 年将 ASAMC 算法用于研究气象变化规律[2]，成守尧等在 2022 [3]年采用 Wilcoxon 秩检验法处理变点数据，程夏 2022 [4]年将变点理论与混合分布模型相结合处理数据，文章[5] [6] [7]将变点理论与地质数据相结合，利用变点理论的处理地形、水文等自然现象。本篇利用 ASAMC 算法与气象数据相结合，找出数据链中的变点。本篇文章将 ASAMC 算法用于探寻地质层中氧元素含量变化的成因，利用统计方法处理地质数据，减少了地质调研的时间并结合乌鲁木齐市水磨沟区地质特点，分析氧元素含量与哪些因素有关。

2. 正态分布均值变点模型

变点理论是上世纪五十年代提出的，具有很强的应用价值。变点模型在工业、金融方面应用较广，变点主要分为以下三种，参数变点、概率变点、位置变点，其中位置变点定义为：

假设存在一个数据集，每个数据观测值相互独立，如果在某一时刻，模型中的某个或某些变量突然发生了变化，及存在一个时间点，在该点之前，数据集符合一个分布，在该点之后，数据集符合另一个分布，则该点为该数据集的位置变点[8]。

2.1. 正态分布均值变点模型

假设 $\{Y_i, i=(1,2,\cdots,n)\}$ 是服从正态分布的随机变量序列。其中，随机变量序列存在 k 个均值变点，即含有 k 个变点的正态分布均值变点模型如下：

$$Y_i \sim \begin{cases} N\left(\mu_1,\sigma^2\right), i=c_0,2,\cdots,c_1 \\ N\left(\mu_2,\sigma^2\right), i=c_1+1,c_1+2,\cdots,c_2 \\ \quad\vdots \\ N\left(\mu_k,\sigma^2\right), i=c_{k-1}+1,c_{k-1}+2,\cdots,c_k \\ N\left(\mu_{k+1},\sigma^2\right), i=c_k+1,c_k+2,\cdots,c_{k+1} \end{cases}$$

其中 $c_i\left(i=1,2,\cdots,n\right)$ 表示变点位置，其中 $c_0=1$，$c_{k+1}=n$，σ 为已知数，记 $\boldsymbol{\theta}=\left(k,c_0,\cdots,c_{k+1},\mu_1,\cdots,\mu_{k+1}\right)$ 为该模型参数向量，则正态分布均值变点序列似然函数为

$$L\left(Y\mid\theta\right)=\prod_{j=c_0}^{c_1}f_1\left(y_j\mid\mu_1,\sigma^2\right)\cdot\prod_{j=c_1}^{c_2}f_2\left(y_j\mid\mu_2,\sigma^2\right)\cdots\prod_{j=c_k+1}^{c_{k+1}}f_{k+1}\left(y_j\mid\mu_{k+1},\sigma^2\right)$$

其中 $f_i=\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left\{\dfrac{-\left(y-\mu_i\right)^2}{2\sigma^2}\right\}, i=1,\cdots,k+1$。

2.2. 各种分布与正态分布之间的联系

在大数据背景下，样本均值服从的分布可以转成正态分布，因此，正态分布均值变点在大数据背景下有较强的应用价值。

辛钦大数定理[9]：设 $X_1,X_2,\cdots$ 是相互独立，服从同一分布的随机变量序列，且具有数学期望 $E\left(X_k\right)=\mu\left(k=1,2,\cdots\right)$。作前 n 个变量的算术平均 $\dfrac{1}{n}\sum_{k=1}^{n}X_k$，则对于任意 $\varepsilon>0$，有

$$\lim_{n\to\infty}P\left\{\left|\frac{1}{n}\sum_{k=1}^{n}X_k-\mu\right|<\varepsilon\right\}=1。$$

即序列 $\bar{X}=\dfrac{1}{n}\sum_{k=1}^{n}X_k$ 依概率收敛于 μ，$\bar{X}\xrightarrow{P}\mu$。($n$ 趋向于正无穷)

独立同分布的中心极限定理[9]：设随机变量 $X_1,X_2,\cdots,X_n,\cdots$ 相互独立，服从同一分布，其具有数学期望和方差：$E\left(X_k\right)=\mu, D\left(X_k\right)=\sigma^2>0\left(k=1,2,\cdots\right)$，则随机变量之和的标准变化量为

$$Y_n=\frac{\sum_{k=1}^{n}X_k-E\left(\sum_{k=1}^{n}X_k\right)}{\sqrt{D\left(\sum_{k=1}^{n}X_k\right)}}=\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}$$

的分布函数 $F_n\left(x\right)$ 对于任意 x 满足

$$\lim_{n\to\infty}F_n\left(x\right)=\lim_{n\to\infty}P\left\{\frac{\sum_{k=1}^{n}X_k-n\mu}{\sqrt{n}\sigma}\le x\right\}=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{t^2}{2}}\mathrm{d}t=\Phi\left(x\right).$$

3. 实例分析

3.1. ASAMC 算法

ASAMC 算法既可以检验正态分布均值变点位置又可以检验变点个数，运用 ASAMC 算法，一般先将数据正态化，将服从其他分布的随机变量序列转化为服从正态分布的随机序列；再利用 ASAMC 算法

检测正态分布均值变点的位置和个数。

3.2. ASAMC 算法步骤[1]

(1) 样本数据正态化

随机产生一组样本数据，记作 $Z=(z_1,z_2,\cdots,z_n)$，检验样本数据 $Z=(z_1,z_2,\cdots,z_n)$ 是否服从正态分布，具体做法如下：画出样本数据的散点图，观察这些点的分布特点，是否在一条直线附近波动，若存在这样的直线，我们就认为该样本数据服从正态分布；反之，样本数据不服从正态分布，要将样本数据进行正态化处理。

(2) 设定一个新的变量 $X_1=(x_1,x_2,\cdots,x_n)$，随机产生 k 个变点，$\left(k\le\left[\frac{n}{10}\right]-1\right)$，从 1 到 n 中随机选取 k 个位置记为 $c_1,c_2,\cdots,c_k$，其中 $1<c_1<c_2<\cdots<c_k<n$，假定，处于变点位置处的 $x_i=1,i=c_1,c_2,\cdots,c_k$，其余位置为 0。

(3) 定义相对样本频率 $\frac{n_i}{n}\cdot 100\%$ 来代替概率，迭代过程具体分为以下步骤：

(3.1) 根据样本数据的对数后验分布概率密度函数 $U(x)=\log p\left(x^k\middle|Z\right)$，我们将样本空间 χ 进行划分，

$$\begin{aligned}&E_1=\{x\in\chi:U(x)<u_1\}\\&E_2=\{x\in\chi:u_1\le U(x)<u_2\}\\&\vdots\\&E_m=\{x:U(x)\ge u_{m-1}\}\end{aligned}$$

分成 m 份，即 ，第 t 次迭代时样本空间记为：$\chi^{(t)}=\bigcup_{i=1}^{\varpi\left(U_{\min}^{(t)}+\Delta\right)}E_i$

(3.2) 根据贝叶斯分析理论，从建议分布 $q(x,y)$ 中随机抽取样本 y

(3.3) 计算接受概率 $\min(1,r)$，其中

$$r=C\frac{\phi(y)q(y,x_t)}{\phi(x_t)q(x_t,y)}$$

其中 $q(x,y)$ 是定义在样本空间 χ 上的函数，C 是与参数无关的常系数

(3.4) 利用 R 软件在 $(0,1)$ 上的均匀分布中抽取随机数 u，

$$x_{t+1}=\begin{cases}y,u\le\min(1,r)\\x_t,u>\min(1,r)\end{cases}$$

(3.5) 多次重复(3.1)~(3.4)，找出不同的 U 和 x 的值。

(4) 若满足 $U(x_{t+1})<U_{\max}^{(t)}$，则 $U_{\max}^{(t+1)}=U(x_{t+1}),x_{\max}^{(t+1)}=x_{t+1}$

(5) 输出 $U_{\max}^{(t+1)},x_{\max}^{(t+1)}$，做出相关数据的频数直方图。

(6) 观察频数直方图，若各组间有显著差别，则认为样本序列存在变点，反之不存在变点。

(7) 若模型存在变点，找出最大的 $U_{\max}^{(t+1)},x_{\max}^{(t+1)}$，$x_{\max}^{(t+1)}$ 的值表示变点的位置。

3.3. 运用 ASAMC 算法检测土壤中氧元素含量

变点理论在地质学中应用广泛。ASAMC 算法既可以检测出正态分布均值变点个数，又可以检测出变点所在的位置。这篇文章将 ASAMC 算法应用于地质层中氧元素含量的变化上，探究地质层中元素含量与哪些因素有关。下面我们以乌鲁木齐水磨沟区新疆师范大学温泉校区的土地为研究对象(数据来源于新疆师范大学地理科学实验室)。

为了对已有数据进行分析处理[10]，我们进行散点图处理，分析温度，土壤中氧元素含量的变化情况。图 1 表示从 2014 年 1 月到 2017 年 5 月新疆师范大学温泉校区温度变化情况；图 2 表示近年的温度变化导致地质层中氧元素含量变化情况，如下图所示：

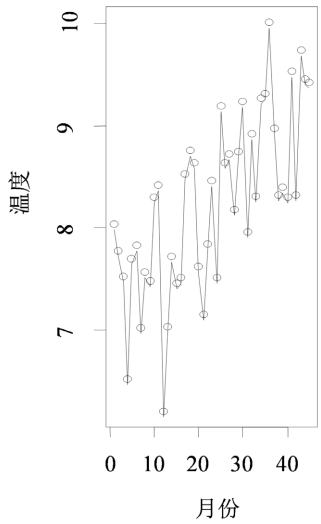


Figure 1. Change of temperature
图 1. 温度变化情况

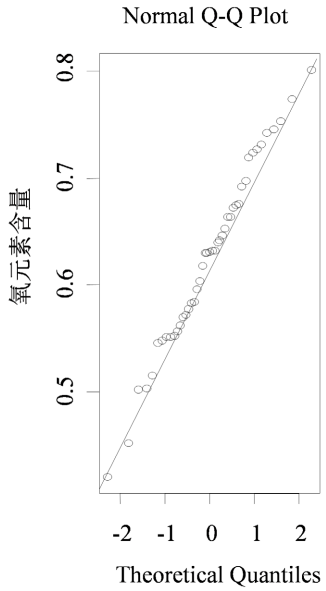


Figure 2. Change of oxygen content
图 2. 氧元素变化情况

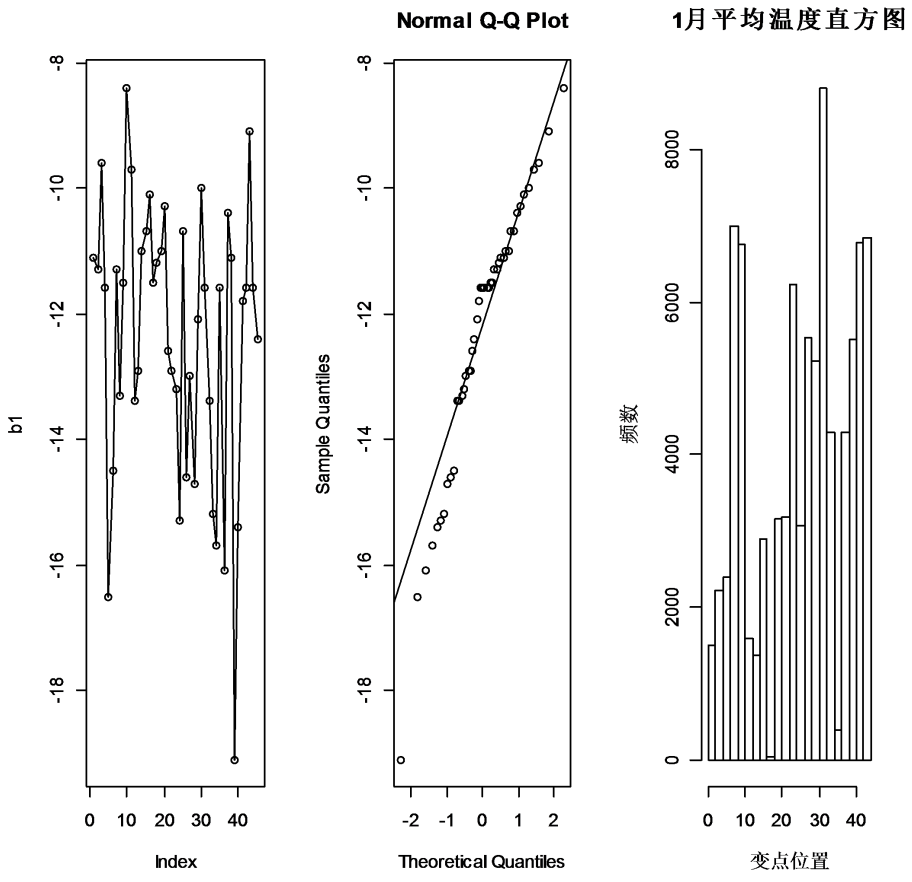
观察图 1、图 2，我们发现地质层中氧元素含量大致呈现正态化，并且地质层中氧元素含量与 2014 年 1 月到 2017 年 5 月温度变化呈正相关，我们以地面为参考平面，测量土壤所在位置的距离(即该土壤层距离地面的最短垂直距离)，随着距离的加深，氧元素的含量在降低。考察温度和土壤中氧元素含量的关系见下表 1：

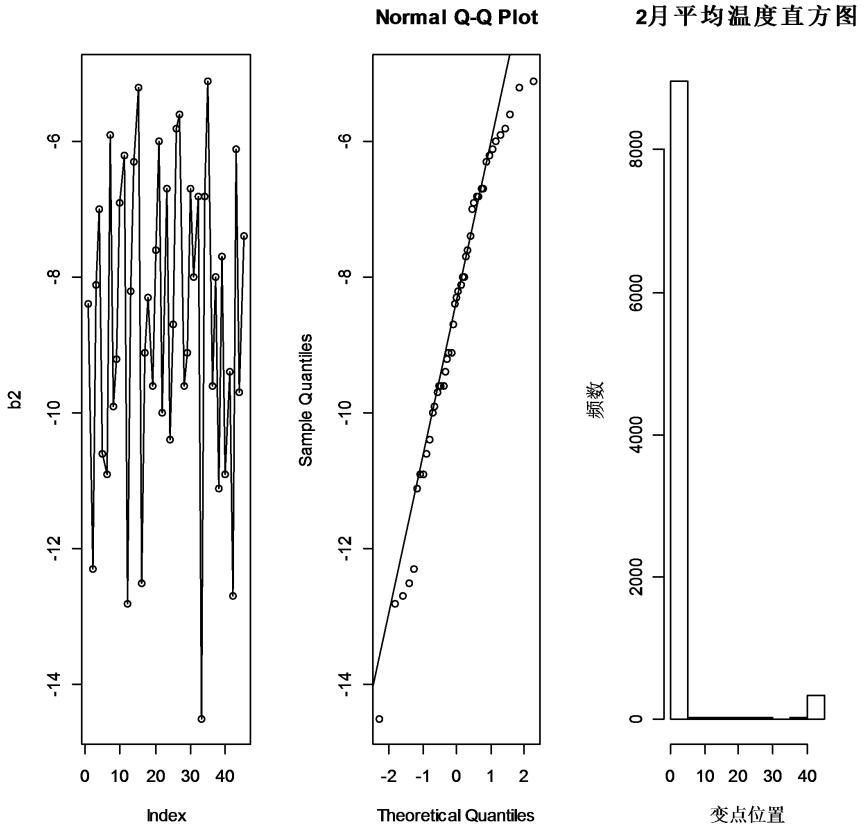
Table 1. Change of oxygen content
表 1. 氧元素变化情况

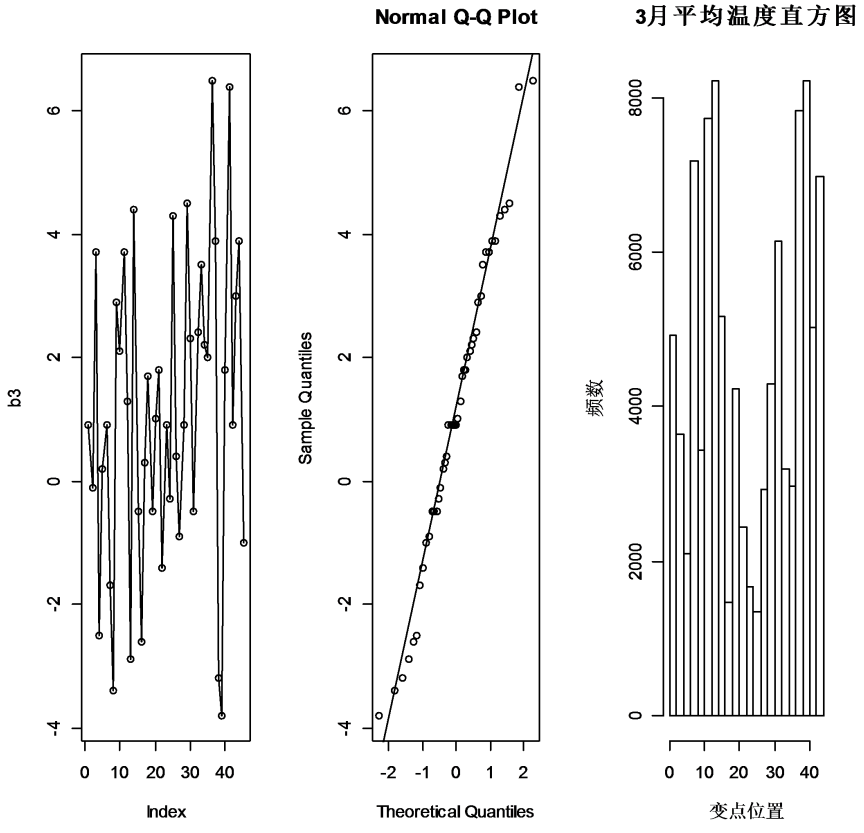
No.	The number of change-points	Change patterns	Negative log-posterior
1	1	24	0.782
2	2	(23, 24)	0.752
3	2	(24, 25)	0.652
4	1	23	0.655
5	1	25	0.6447
6	3	(23, 24, 25)	0.6577
7	1	21	0.5648
8	2	(22, 23)	0.556
9	1	26	0.5567
10	1	20	0.5822

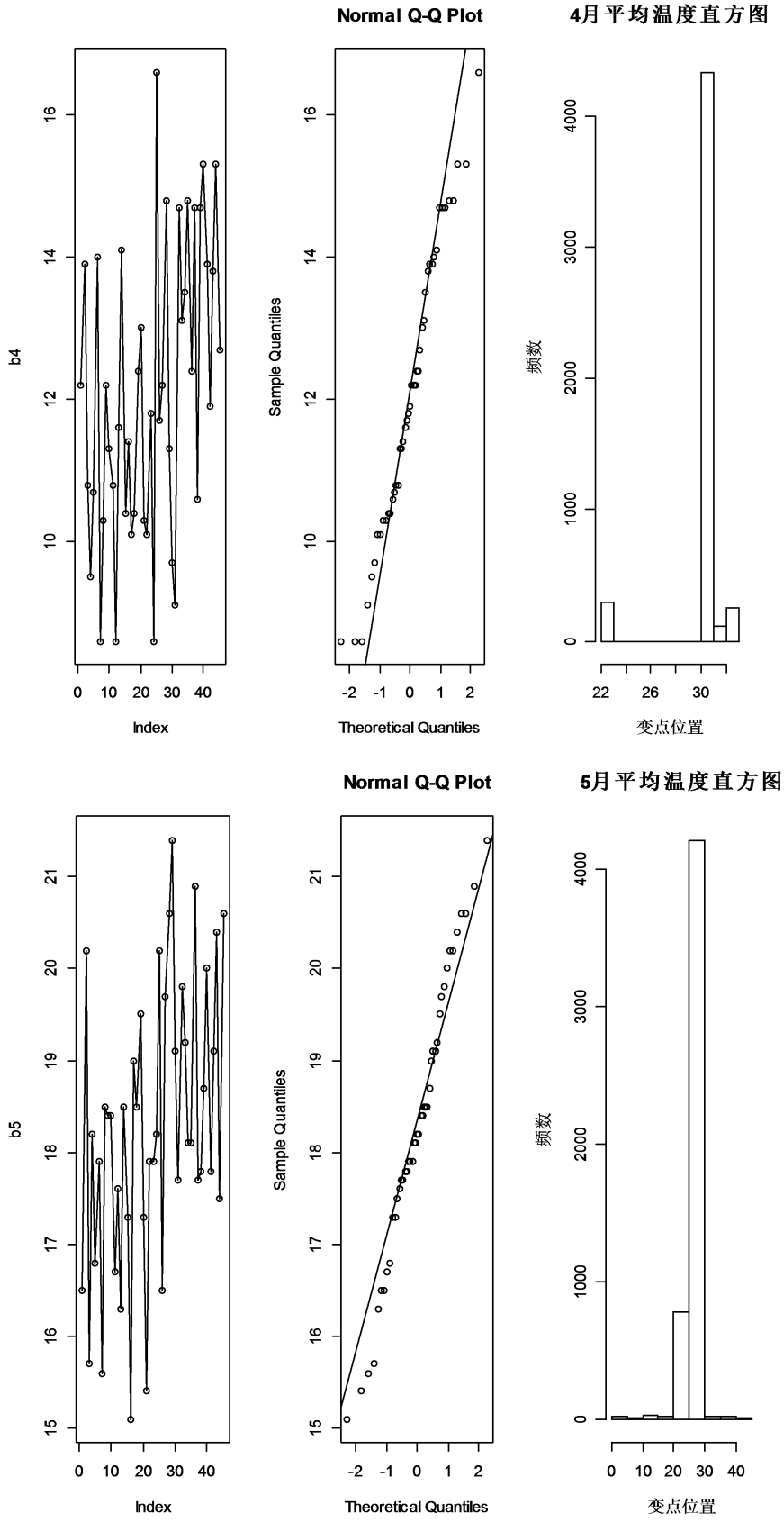
通过对表 1 的研究，地质层中氧元素含量与月平均温度的关系，通过对数据的研究，我们发现数据中存在变点，大多数据中只含有一个变点，因此，我们发现变点位置大致为 24，即年，即 2016 年 6 月月平均温度对土壤中氧元素的含量的影响发生明显的变化，高温会使得土壤中氧元素含量增加。

下面研究季节对土壤中氧元素含量的影响，我们将研究重点放在春夏两个季节，见图 3。




Normal Q-Q Plot
2月平均温度直方图
b2
Index
Sample Quantiles
Theoretical Quantiles
频数
变点位置


Normal Q-Q Plot
3月平均温度直方图
b3
Index
Sample Quantiles
Theoretical Quantiles
频数
变点位置


Normal Q-Q Plot
4月平均温度直方图
b4
Index
Sample Quantiles
Theoretical Quantiles
频数
变点位置
Normal Q-Q Plot
5月平均温度直方图
b5
Index
Sample Quantiles
Theoretical Quantiles
频数
变点位置

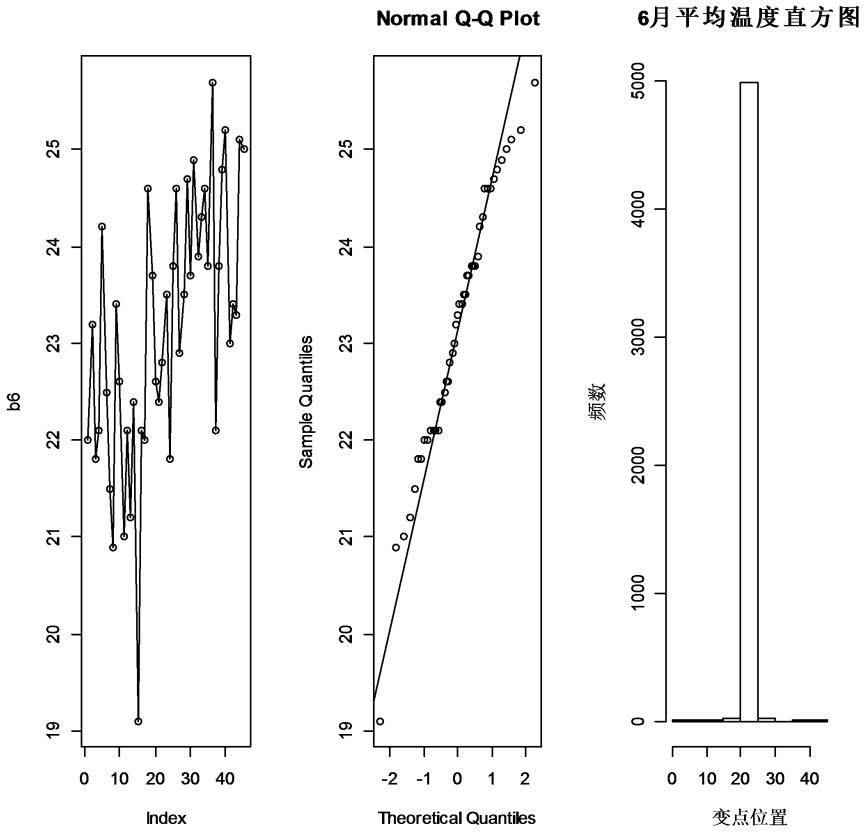


Figure 3. Changes of oxygen content in soil in spring and summer
图 3. 春夏两季土壤中氧元素含量变化情况

根据图 3 中的最大对数概率值可以得到土壤中氧元素含量变化情况，其中夏季氧元素含量明显增加，即变点一般产生在夏季，由于夏季高温炎热，降水量大。因此，我们可以发现温度和季节都会影响土壤中氧元素的含量。

这篇文章通过 ASAMC 方法检验出的乌鲁木齐水磨沟区的新疆师范大学温泉校区的土地中氧元素含量变化情况。

新疆乌鲁木齐的夏季高温、多雨，使得土壤中氧元素含量升高，符合新疆气候特点，分析图 2、图 3 我们发现，影响氧元素含量变化与新疆夏季气候有关，因此利用 ASAMC 算法处理数据具有较强的实用价值。

基金项目

石河子大学自主立项科研项目(ZZZC202032B)。

参考文献

[1] Kim, J. and Cheon, S. (2010) Bayesian Multiple Change-Point Estimation with Annealing Stochastic Approximation Monte Carlo. *Computational Statistics*, **25**, 215-239. https://doi.org/10.1007/s00180-009-0172-x

[2] 许欢. 基于 ASAMC 算法的气象数据多变点估计[D]: [硕士学位论文]. 合肥: 合肥工业大学, 2016.

[3] 成守尧, 陈占寿, 娘毛措, 汪肖阳. 一类长记忆时间序列趋势项变点的 Wilcoxon 秩检验[J]. 浙江大学学报(理学版), 2022, 49(4): 427-434.

[4] 程甜, 夏志明. 带变点的混合模型的统计推断与算法设计[J]. 应用概率统计, 2022, 38(3): 439-453.

[5] 宁婷, 崔伟, 马晓勇. 基于均值变点法提取地形起伏度的影响因素分析——以黄河流域(山西段)为例[J]. 测绘通报, 2022(2): 159-163.

[6] 聂启阳. 基于双侧均值变点法的数字河网阈值划定——以南苕溪流域为例[J]. 绿色科技, 2022, 24(10): 250-254.

[7] 张清杰, 黄领梅. 基于斜率单变点法的小理河流域退水规律分析[J]. 水电能源科学, 2022, 40(1): 21-24.

[8] 陈希孺. 变点统计分析简介[J]. 数理统计与管理, 1991, 10(2): 52-59.

[9] 盛骤, 谢式千, 潘承毅. 概率论与数理统计教程[M]. 北京: 高等教育出版社, 2015.

[10] 肖枝洪, 朱强. 统计模拟及其 R 实现[M]. 武汉: 武汉大学出版社, 2010.