湖南大學

金融与统计学院

2019-2020 学年春季学期

《质量管理统计学》课程设计报告

题目：	控制图在地震预测中的应用
姓名：	张笑竹
学号：	201618070114
专业班级：	统计学 2016 级 01 班

2020 年 6 月

目录

1. 引言

1.1 背景介绍

地震是地壳快速释放能量过程中造成的震动，期间会产生地震波。据统计，地球上每年约发生 500 多万次地震，其中绝大多数太小或太远以至于人们感觉不到，而真正能对人类造成严重危害的地震大约有一、二十次。危害性的地震常常造成地面的破坏，引起建筑物倒塌、山体滑坡、泥石流和海啸等灾害。例如 1976 年 7 月 28 日发生在河北唐山的里氏 7.8 级地震，造成 24.2 万人死亡，16.4 万人重伤，97%的地面建筑毁坏，一座拥有百万人口的工业城市被夷为平地；再如 2008 年 5 月 12 日，四川汶川的里氏 8.0 级地震，造成 6.9 万人死亡，37.4 万人受伤，直接经济损失 8452 亿元人民币。它们对人类的危害是巨大的。

虽然预测地震目前仍是世界性难题，但迄今科学界普遍认为，有可能反映地震前兆特征的指标不少于 10 个。已经有专业仪器在多个定点实时按秒记录这些指标的数据，期望通过对记录数据的分析研究找到地震的前兆特征。

本文采用的是某地 2013 年 1 月 1 日至 2018 年 6 月 30 日按小时观测的 10 多个指标的数据，以及该地区该时期内已发生地震的时刻、经纬度、震级及震源深度的数据。原始数据来自于中国地震局。通过挖掘数据中隐藏的地震前兆特征，可以度量各个指标对地震发生的敏感程度，并在此基础上构造综合指标，从而对潜在的地震进行预警。

1.2 分析思路

地震的特征之一，是其本身和各项指标的“脉冲式”变化。在地震发生前，各项指标表现得较为平稳；而在地震发生区间内，对地震发生敏感的指标往往会出现大幅度的震动；当地震结束后，这些指标又会恢复到先前的正常水平。基于这种考虑，预测地震的一大关键点，就在于合理地抓取数据“冲击”“跳跃”的片段，并将这些异常片段与地震发生的时刻相互匹配。本文借鉴控制图的思想，观察各项指标数据的变动过程，并利用控制图计算出数据波动的上下限，从而识别出“脉冲”异常值。此外，本文还试图构造“地震波动”的综合指标，并通过过程能力指数反映综合指标的优劣；最后，通过这一综合指标更为简洁地识别潜在的地震。

2. 数据描述和数据处理

2.1 数据描述

本文选取的数据集中，共有 9 个指标，分别是电压、电磁波幅度 EW、电磁波

幅度 NS、地温、水位、气温、气压、水温、气氡。这 9 个指标中有些存在明显的“脉冲”现象，或许与地震有关，有些则存在着明显的季节特性，如水温、气温等。显然，由于地震的发生与季节无关，所以在后续的分析中，应当消除这些变量的季节效应。此外，不乏有些变量存在明显的趋势效应。必须注意到，本文选择控制图作为分析工具，其目的在于识别异常值，而非观察指标中心的移动情况，这与传统情况下利用控制图监测生产过程完全不同；所以，我们就必须保证这些序列是平稳的。后续分析中，变量的趋势效应将被消除。

数据集中的各个观测是按小时记录的，横跨 5 年时间，十分庞大冗杂。考虑到按小时记录的数据会受到天气、气候等外在因素的影响，随机波动较大，因此本文采取了求日平均的方法，消除偶然因素的影响。然后，又对部分缺失值进行填补，并删除了缺失过多、填补确实存在困难的观测，最终得到了 1746 条样本数据。

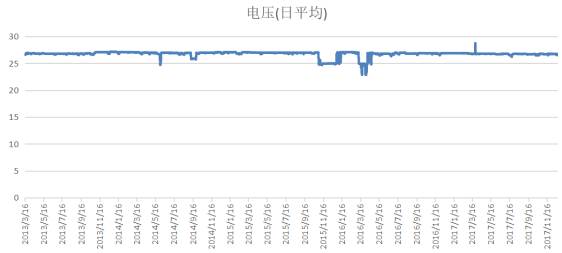
在这 5 年间，该地一共发生了 7 次地震。表 1 展示了这 7 次地震的具体信息。后续的工作，就是要寻找各个指标的变动与地震发生时刻的关系。

表 1　地震发生时间及经纬度

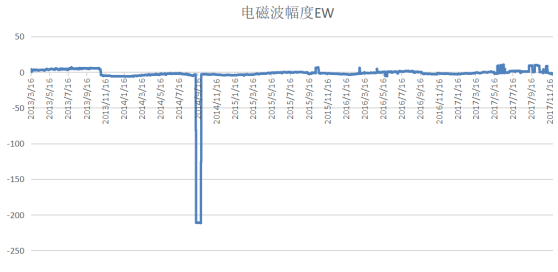
序号	时间	经纬度
1	2013-11-02　15:11	东经 109° 26’　北纬 34° 05’
2	2014-07-26　02:44	东经 109° 05’　北纬 33° 21’
3	2015-03-22　10:51	东经 110° 18’　北纬 35° 06’
4	2015-04-05　01:12	东经 107° 09’　北纬 34° 41’
5	2015-05-15　23:56	东经 107° 32’　北纬 34° 28’
6	2016-07-05　07:37	东经 107° 33’　北纬 34° 19’
7	2017-11-05　07:31	东经 107° 02’　北纬 34° 05’

2.2 数据预处理

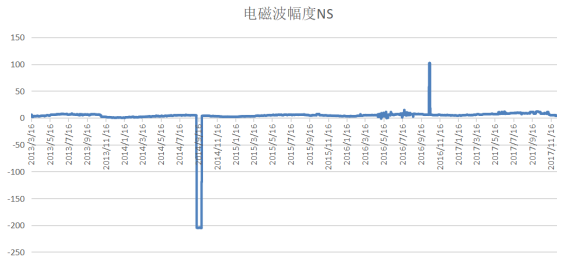
下面，进行数据的预处理。根据 2.1 节中的分析，本文使用控制图的目的，是识别出异常脉冲，因此必须消除数据序列存在的季节性效应和趋势性效应，保证数据序列的相对平稳性。否则，就无法判断超出上下界限的异常值，究竟是由序列内在的季节效应和趋势造成的，还是由与地震有关的“脉冲”造成的。首先，不妨做出这 9 个序列的时间序列图，进行观察。



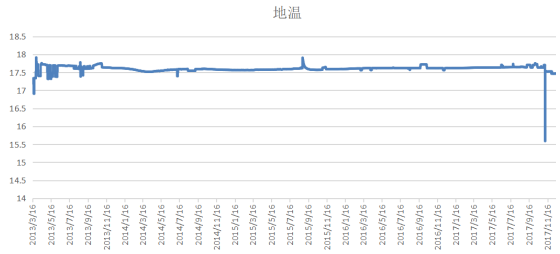
（a）电压



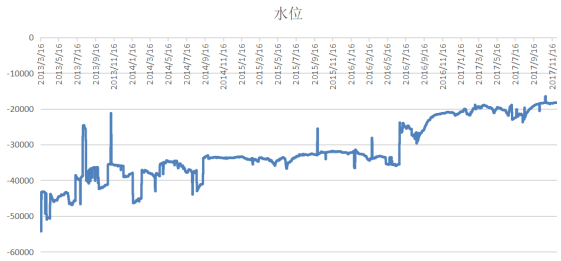
（b）电磁波幅度 EW



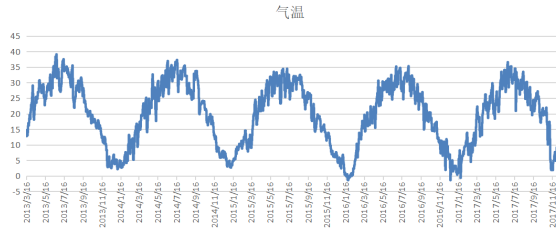
（c）电磁波幅度 NS



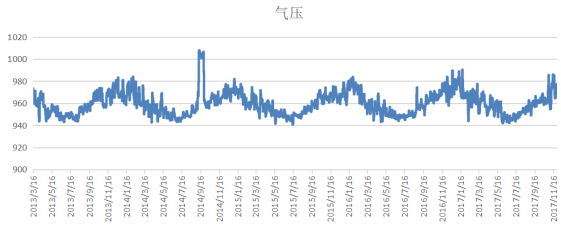
（d）地温



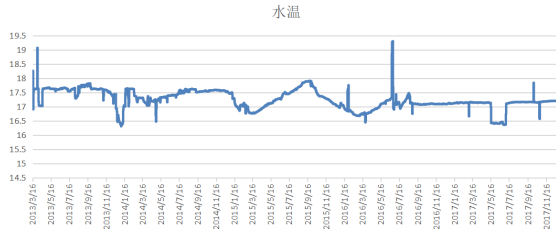
（e）水位



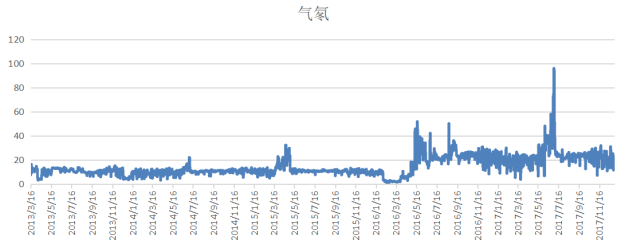
（f）气温



（g）气压



（h）水温



（i）气氢

图 1　　9 个指标的时间序列图

根据图 1，不难发现，电压、电磁波幅度 EW、电磁波幅度 NS、地温、水温这几个指标，除了脉冲的影响，基本呈现平稳状态。水位存在明显的上升趋势，而气氡在序列后部出现的较大的震动。气温、气压这两个指标表现出了明显的周期性，脉冲的影响反而较小。

2.2.1 消除季节效应

在这一部分，首先需要消除气温、气压这两个指标的季节效应。根据时间序列的因素分解理论，所有序列的波动都可以归纳为四个因素的综合影响：长期趋势、循环波动、季节性变化、随机波动，即$x_t = f(T_t, C_t, S_t, I_t)$。考虑到气温和气压的

季节波动幅度不随时间变化，所以采用季节加法模型进行刻画：

$$x_t = T_t + C_t + S_t + I_t.$$

为了分解出季节趋势，首先需要计算时间序列$\{x_t\}$的移动平均：

$$M(x_t) = \frac{x_{t-182} + \cdots + x_t + \cdots + x_{t+182}}{365}.$$

然后，将移动平均$M(x_t)$从原序列中减去：

$$y_t = x_t - M(x_t).$$

接着，应当计算季节指数：

$$S_j = \bar{y}_j - \bar{y}, \qquad j = 1,2\ldots,m.$$

其中，$\bar{y}$为所有y_t的均值。而

$$\bar{y}_j = \frac{\sum_{i=1}^{k} y_{ij}}{k}, \qquad j = 1,2\ldots,m,$$

y_{ij}指的是第j个“季节”的第i个y_t. 最后，将求得的季节指数S_j从原序列x_t中减去，就能够得到消除季节因素的指标值。

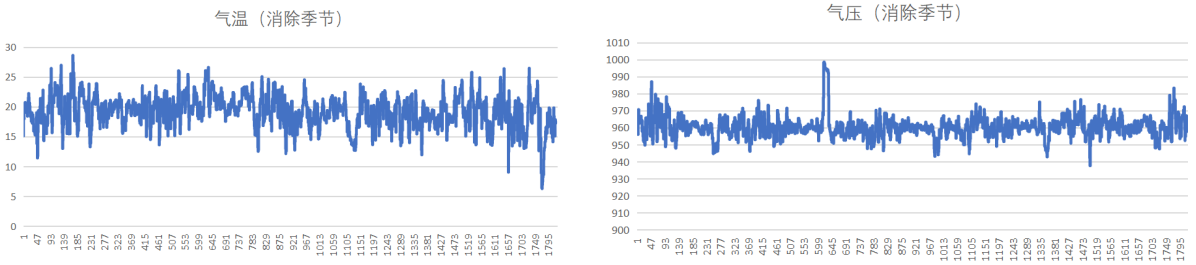


图 2　消除季节效应后的气温和气压

通过图 2，可以清楚地看出季节效应已被消除。时间序列中未被提取的，只有随机波动和脉冲部分。

2.2.1 消除趋势效应

在消除了季节效应后，部分序列（如水位）仍然存在明显的长期趋势。为了消除这一效应，下面采用最常见的差分法进行处理。若记原序列为$\{x_t\}$，则新序列$\{x_t'\}$就是原序列的一阶差分：

$$x_t' = \nabla x_t = x_t - x_{t-1}.$$

经过观察，水位这一指标，经过一阶差分后的确趋于平稳。此外，尽管将原本已经平稳的序列进行再次差分，会出现“过差分”的问题；但是，本文后面将要用到的控制图对于平稳性的要求较为严格，序列稍有趋势就易出界，再加上为了更好地与一阶差分后水位的“变化”这一经济意义对应，不妨将所有其他的指标序列也进行一次差分。经过这样的操作，最终那些稍有幅度但又不显著的波动，就会被抹平，从而留下的都是真正的“脉冲波”。此时，观测值个数从 1746 减少至 1744 个。

最后还需要说明的是，序列中有一些数据存在明显的错误。例如，2014 年 9 月 11 日至 9 月 16 日的电磁波幅度 EW，从一直以来所处的区间[-10, 10]直接跌至-211 左右，且这一现象此后从未出现过，其发生时间与地震时间也没有重叠之处。这一记录产生的原因，可能是探测仪器出现故障，也可能是记录人员出现失误。总之，面对这样明显不合理的记录，本文采取的措施是将其删去，并以附近的平均值进行填补，从而得到真正的平稳序列。

3. 用控制图计算有震报准率

在这一节中，本文针对上述每一个平稳序列，通过控制图寻找异常值，进而分析这些异常值和地震发生的关系。在此基础上，本文将度量这 9 个指标对于地震发生的敏感度。

3.1 用控制图寻找异常值

3.1.1 “地震用”控制图合理性的论证

首先，需要说明在此处使用控制图的合理性，以及它与传统生产过程用控制图的区别。

在传统的生产过程中，工人、机器总是以某一个特定值为目标，尽最大的可能使产品参数向这一目标值靠近。因此，产品参数就会服从以这一目标值为中心，上下波动的分布，且往往是正态分布。作为一个简单易行的控制系统，控制图能够有效识别异常波动，包括分布位置的变化、分布标准差的变化、分布形状的变化。控制图以质量特性的均值为中心，分别规定了以 3 倍标准差为半径的上下界限UCL和LCL，一般情况下，超出上下界限的点子就可被视为异常。为了应对第二类错误的发生，常规控制图中还额外增加了一些特别的“判异准则”，用以监测控制中心可能发生的潜在变化。在应用时，既要监测极差（或标准差）的变化，又应当监测均值（或个值）的变化，两个控制图结合使用，都不能出现异常。此外，控制图分为“分析用”控制图和“控制用”控制图，前者用来了解过程，调整失控状态，达到客户要求；而后者被视作一种成型的规范，用来进行后续生产过程的控制。最后，还可以利用过程能力指数C_p，来刻画生产过程的潜在能力；其需要满足的条件至少包括，过程稳定，以及过程质量特性服从正态分布。

然而，在地震预警中，控制图的使用目的，以及它的许多特性都发生了变化。第一，不同于目标性强、控制精密的质量特性，地温、气压等指标都是自然现象，尽管它们在消除了季节效应、趋势效应后趋于平稳，但是它们的波动性仍然不可控制，变化较大。第二，不同于传统控制图“识别异常、消除异常”的目的，在监测

地震相关指标时，控制图的目的仅仅在于“识别异常”，尤其是从平稳“突跃”而后又迅速恢复的异常，却无法消除异常。第三，不同于被人为活动所约束的质量特性数据，自然现象的变化中掺杂着众多偶然因素，因此应当放适当宽控制图的上下界限，本文采用以 6 倍标准差为半径的上下界限UCL和LCL。第四，相比于地震相关指标变化中心的移动，本文更加关注“脉冲式”异常值现象，再加上各个指标已经经过差分，基本平稳，因此本文不再利用八条“判异准则”进行分析。第五，对于一对常用的“单值—移动极差图”，本文将采用不同的分析方法；对于“移动极差图”，只要在UCL之上的点子不超过 5%，就能够接受；对于“单值图”，则不加以控制，只记录超出上下界限的点子个数，用以后续与地震发生时刻进行对比。第六，只要“移动极差图”满足上述条件，计算得到的上下界限就可被确定为“控制用”控制图的界限；如果“移动极差图”存在超过 5%的点子在UCL之上，就仍将其视作“分析用”控制图，留待进一步调试。最后，本文将过程能力指数C_p视作度量综合指标稳定性的指标，C_p越大，说明得到的综合指标越稳定，混杂的不稳定因素越少，因而其出现的“脉冲”异常值就越具有说服力。

综上所述，“生产用”控制图和“地震用”控制图存在较大的差别，并且这一差别的根源来自于两者目的截然不同。那么，对控制图进行如此大幅度的改动，并将其用于地震的监测，究竟是否合理？这一问题，或许可以从地震监测本身的数学模型角度出发解答。事实上，监控地震相关指标的运行，并且合理地截取有效片段，是一件非常困难的事情；如何定义“有效”，如何定义与地震“相关”，到目前为止都无定论。单纯的时间序列分析模型，与地震是否发生这类离散的 0-1 变量不易结合；神经网络等模型则被视作“黑箱”，其中的机制难以窥探。而控制图这一工具，则为我们提供了一个独特的视角，既能巧妙地截取各个时间序列指标的异常片段，又能通过计数的方式，将异常点与地震时刻匹配。因此，我们讨论的并不是如何使得“地震现象”贴近“生产过程”，而应当是如何使得质量管理的模型更具有普适性，在更广泛的问题中发挥其强大的作用。从这一角度而言，此处使用控制图是完全合理的。

3.1.2 各项指标的控制图

根据上一节的论证，我们对“地震用”控制图做出如下规定：

（1）UCL和LCL距离中心线 6 倍标准差。

（2）采用“单值图—移动极差图”。

（3）观察“移动极差图”，只有不超过 5%的数据在UCL之上时，才进行第（4）步；否则，移除部分极端值，直到满足上述条件。

（4）无论第（3）步中是否移除了极端值，在绘制“单值图”时，所有的数据

都应当包含在内。

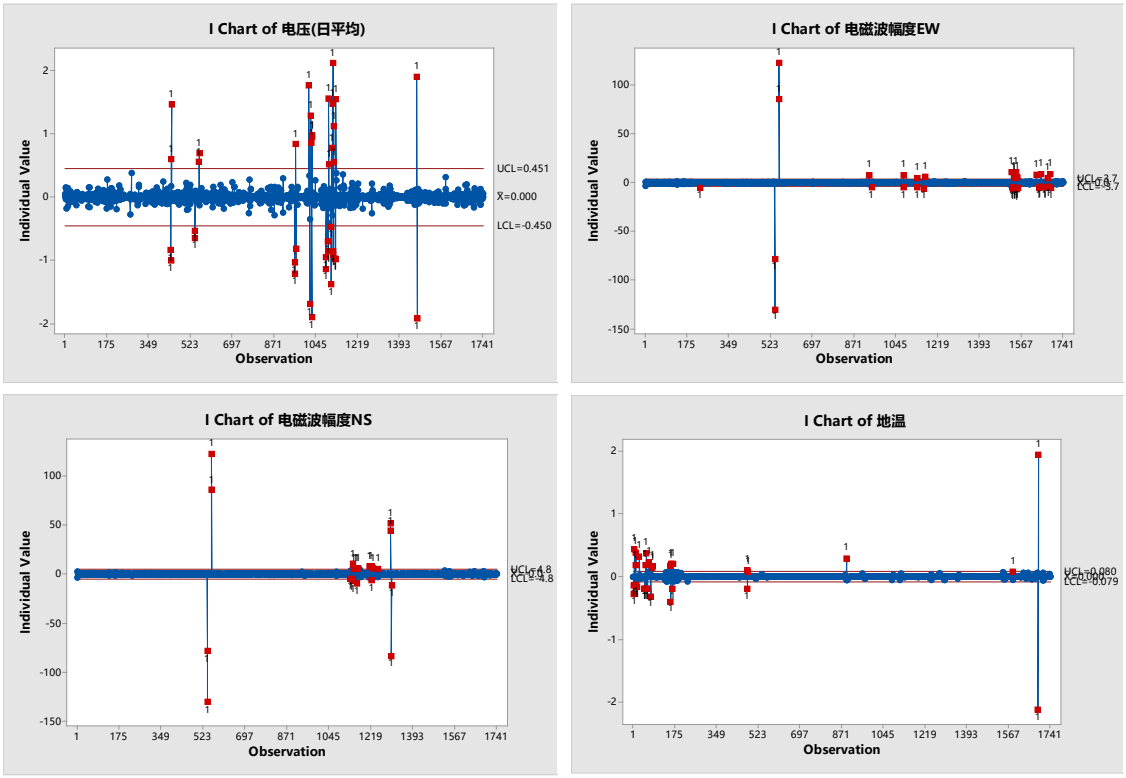
（5）计算“单值图”中超过上下界限的点子编号，留待后续分析。

表 2 9 个指标“单值—移动极差图”的超限情况

指标	电压	电磁波幅度 EW	电磁波幅度 NS	地温	水位
MR 超限比例（%）	2.92	2.81	2.47	3.15	5.79/ 4.93
单值超限个数	39	40	22	32	63/ 108
指标	气温	气压	水温	气氡	
MR 超限比例（%）	0.23	0.29	4.53	1.83	
单值超限个数	2	2	50	11	

表 2 展示了 9 个指标“单值—移动极差图”的超限情况。不难发现，除了水位，其他指标的移动极差图中，超过上限*UCL*的数据均在 5%以内；而水位指标原本的 MR 超限比例大于 5%，因此移除部分极端值，进行调整，最终得到新的 MR 超限比例（%）为 4.93，符合要求。此外，表 2 还给出了 9 个指标的单值图中超限观测个数。

最后，将 9 个单值控制图展示如下。



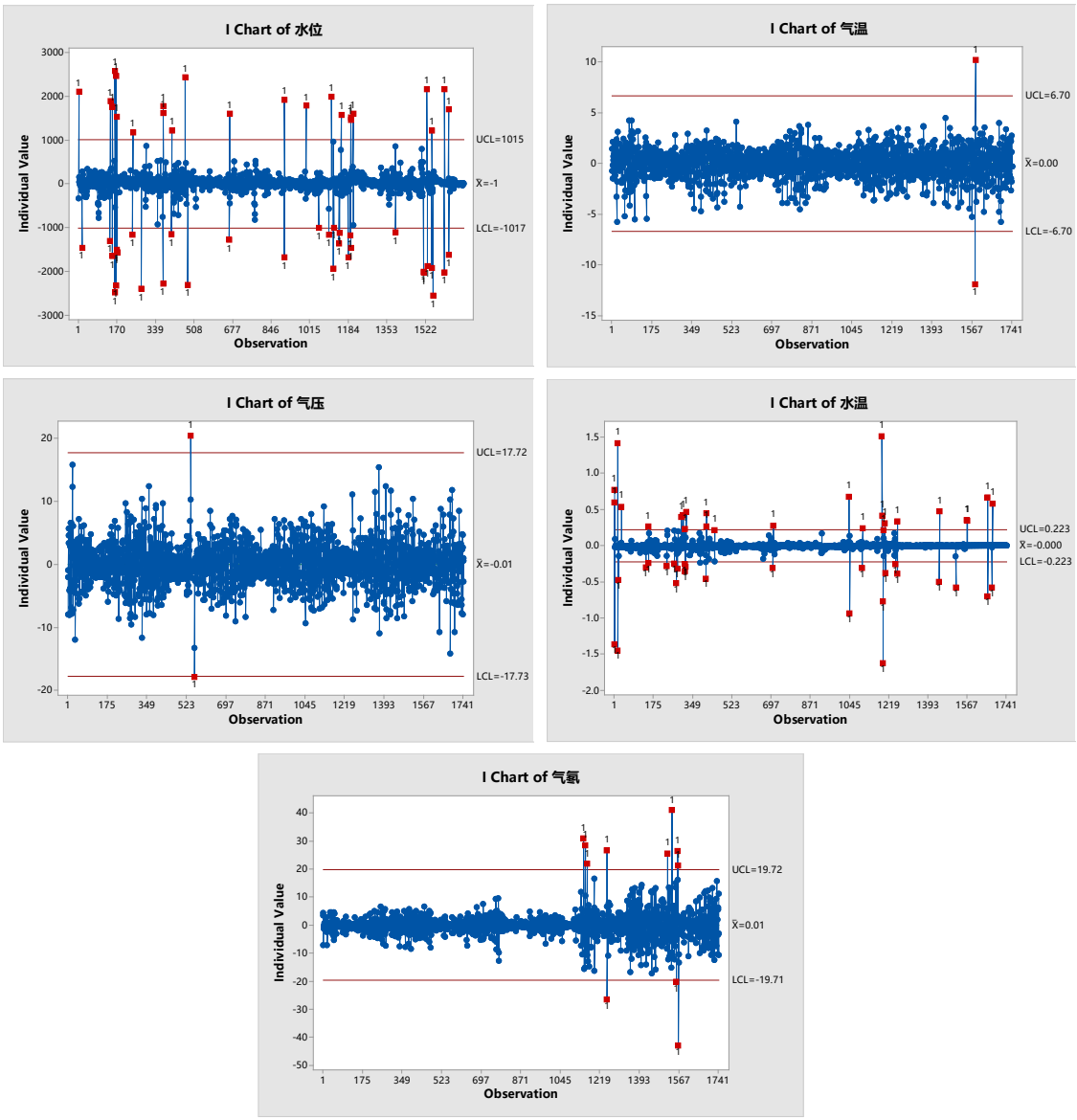


图 3　　9 个指标的单值控制图

3.2 有震报准率

为了寻找单值图中超限的观测与地震发生时刻的相关关系，下面，参考朱俊杰等（2017），构造“有震报准率”这一衡量标准：

$$\text{有震报准率} = \frac{\#(\text{预测有效数据})}{\#(\text{异常数据})}$$

其中，“异常数据”指的就是上一节单值图中超限的观测数据；“预测有效的数据”，指的是单值图中超限，并且与 7 次地震中的某一次相隔不超过 30 天的观测数据。有震报准率反映了各个指标在地震发生前后一个月内的活跃变动与地震联系的紧密程度。有震报准率越大，说明异常活动能够更多地指向潜在地震；反之，则说明指标的异常变动与地震关系不大。

表 3 展示了 9 个指标“有震报准率”的计算过程和最终结果。经过对比，电磁波幅度 NS 和电磁波幅度 EW 是两个对地震最为敏感的变量，而气温、气压则对地震完全不敏感，起不到任何预测的作用。

表 3　9 个指标的“有震报准率”概况

指标	电压	电磁波幅度 EW	电磁波幅度 NS	地温	水位
#（异常数据）	39	40	22	32	108
#（预测有效数据）	3	13	13	5	25
有震报准率（%）	7.69	32.50	59.09	15.63	23.15
有震报准率排名	7	2	1	6	5
指标	气温	气压	水温	气氡	
#（异常数据）	2	2	50	11	
#（预测有效数据）	0	0	13	3	
有震报准率（%）	0.00	0.00	26.00	27.27	
有震报准率排名	8.5	8.5	4	3	

举例来说，电磁波幅度 EW 一共出现了 40 个异常数据，其中的 13 个处于地震发生前后 30 天内；特别地，2 个异常数据捕捉到了第一次地震，4 个数据捕捉到了第五、六次地震，7 个数据捕捉到了第七次地震。电磁波幅度 NS 一共出现了 22 个异常数据，而其中所有的 13 个发生在地震前后 30 天内的数据，捕捉的都是第五、六次地震。其他的指标也可以进行类似的分析。从这一角度来看，尽管有些指标的“有震报准率”较高，它们却难以捕捉所有的地震。如果想要尽可能多地捕捉所有的地震，就不得不构建综合指标，充分考虑每一种可能性，发掘潜在的地震。

4. 综合指标的构建

在这一节中，本文将依据上述分析得到的“有震报准率”，构建用来监测潜在地震的综合指标。构建这一指标的方法可以有多种，得到的指标性能也有较大差异。因此，下面首先会讨论两种构建方法，然后利用过程能力指数比较它们，最后利用性能较好的那一个综合指标进行分析。

4.1 两种方法

无论采用的是哪种方法，为了保持量纲的一致性，首先都需要对数据进行标准化。对于第i个观测的第j个指标值x_{ij}，标准化后的取值$\tilde{x}_{ij}$为：

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}.$$

其中，$\bar{x}_j$为第j个指标的样本均值，s_j为第j个指标的样本标准差。

4.1.1 主成分分析法

构建综合指标，最常用的方法是主成分分析法。其原理是对指标的协方差矩阵进行分解，经过线性变化，把多个指标浓缩为 2-3 个主成分，进而得到这些主成分的加权平均值，形成综合指标。在本文中，根据表 3，不难发现电压的有震报准率较小，在 10%以下，而气温、气压的有震报准率则直接为零。因此，将上述三个指标移出分析范围，只考虑排名前六的指标。

将电磁波幅度 EW、电磁波幅度 NS、地温、水位、水温、气氡的标准化指标值分别记为$\tilde{x}_{\cdot 1}, \tilde{x}_{\cdot 2}, \ldots, \tilde{x}_{\cdot 6}$，将它们组成的数据集记为$\boldsymbol{X}$。然后计算$\boldsymbol{X}$的协方差矩阵$\Sigma_{6\times 6}$。对$\Sigma_{6\times 6}$进行谱分解，计算特征值$\lambda_1, \lambda_2, \ldots, \lambda_6$和特征向量$U_1, U_2, \ldots, U_6$。最后根据特征值大于 1 的原则，提取了 3 个主成分。根据特征值$(\lambda_1, \lambda_2, \lambda_3)^{\mathrm{T}}$，计算得到前 3 个主成分的方差占比为$(0.3185, 0.2148, 0.168)^{\mathrm{T}}$，经过归一化，得到$\boldsymbol{\lambda} = (0.4542, 0.3063, 0.2395)^{\mathrm{T}}$。线性变换矩阵为：

$$\boldsymbol{U} = \begin{array}{c} \\ \tilde{x}_{\cdot 1} \\ \tilde{x}_{\cdot 2} \\ \tilde{x}_{\cdot 3} \\ \tilde{x}_{\cdot 4} \\ \tilde{x}_{\cdot 5} \\ \tilde{x}_{\cdot 6} \end{array} \begin{array}{c} \begin{array}{ccc} C_1 & C_2 & C_3 \end{array} \\ \begin{bmatrix} 0.500 & 0.119 & 0.007 \\ 0.500 & 0.125 & -0.011 \\ 0.015 & -0.042 & 0.883 \\ -0.139 & 0.578 & 0.030 \\ -0.064 & -0.604 & -0.182 \\ 0.011 & 0.214 & -0.422 \end{bmatrix} \end{array}.$$

因此，主成分矩阵为$\boldsymbol{C}_{1744\times 3} = (C_1, C_2, C_3) = \boldsymbol{XU}$. 进一步可以得到综合指标$\boldsymbol{I}^{(1)}_{1744\times 1} = \boldsymbol{C\lambda}$。为了直接表示从数据集到综合指标的线性变换，记$\boldsymbol{w^{(1)}} = \boldsymbol{U\lambda}$，从而有$\boldsymbol{I}^{(1)}_{1744\times 1} = \boldsymbol{Xw^{(1)}}$。

4.1.2 简单权重法

主成分分析法尽管相对复杂，理论完善，但是在有些情况下，它的效果却不尽人意。为了应对这种可能性，本文额外给出一种备选的简单方案，即通过加权平均求出综合指标。注意到，在 3.2 节中，我们计算了各个指标的有震报准率，以反映这些指标对地震是否发生的敏感程度。有震报准率越高，对应的指标就越敏感，在综合指标中也就应当获得更多的重视；反之则反是。基于这种考虑，不妨将六个指标的有震报准率向量进行归一化，并将其直接作为权重，记为

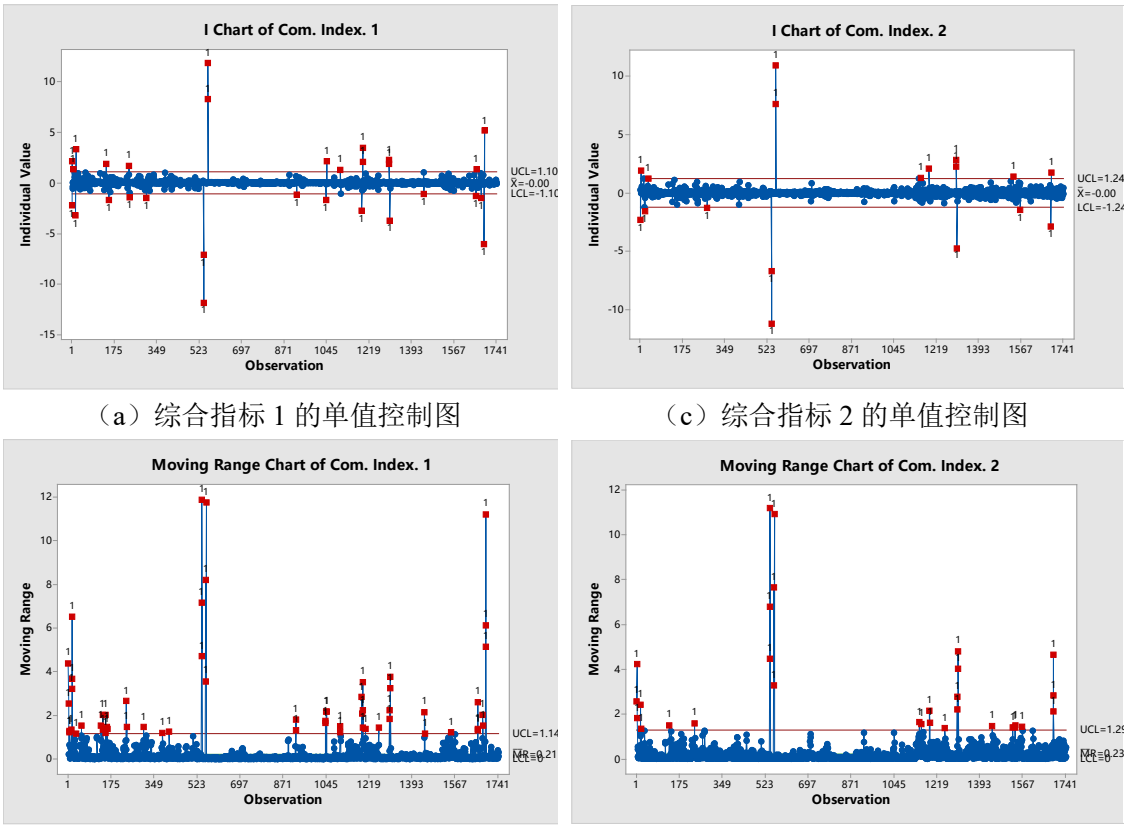
$$\boldsymbol{w^{(2)}} = (0.1770, 0.3218, 0.0851, 0.1261, 0.1416, 0.1485)^{\mathrm{T}}.$$

因此，综合指标就可以写作$\boldsymbol{I}^{(2)}_{1744\times 1} = \boldsymbol{Xw^{(2)}}$。

4.2 综合指标的比较

按照 4.1 节中所述方法，可以得到两个综合指标$\boldsymbol{I}^{(1)}$和$\boldsymbol{I}^{(2)}$。画出它们的控制

图，如图 4 所示。经过进一步统计，得知综合指标 1 的 MR 超限比例为 3.33%，综合指标 2 的 MR 超限比例为 1.72%，均小于规定的 5%临界值，所以可以进行进一步分析。



（a）综合指标 1 的单值控制图　　（c）综合指标 2 的单值控制图

（b）综合指标 1 的移动极差控制图　　（d）综合指标 2 的移动极差控制图

图 4　两个综合指标的“单值-移动极差”控制图

为了比较综合指标 1 和综合指标 2 的优劣，下面引入过程能力指数C_p作为衡量指标。如3.1.1 节所述，C_p能够刻画生产过程在给定的顾客要求之内波动的能力，即过程质量水平满足过客要求的能力。它的基本要求至少包括过程稳定，以及过程质量特性服从正态分布。然而，由于“地震用”控制图的根本目的是监测相关指标的变化，而非进行调整控制，再加上相关指标都是自然现象，不受人类活动限制；因此，“过程稳定”这一条件就被放宽至指标的 MR 超限比例不超过 5%。尽管所讨论的六个指标并不完全服从正态分布，但是在经过一系列的去季节、去趋势处理后，它们全部都呈现对称的单峰分布，处理后的指标分布与正态分布较为接近。经过进一步的标准化，指标的中心全部移动至 0 点附近。

过程能力指数的计算公式为：

$$C_p = \frac{(USL - LSL)}{12\hat{\sigma}} - \frac{|M - \hat{\mu}|}{6\hat{\sigma}}.$$

对于给定的上下界限USL和LSL，C_p越大，就意味着综合指标的标准差越小，混杂的不稳定因素就越少。在越稳定的序列中，出现的“脉冲”异常值就越具有说服力，

所以“地震用”过程能力指数与传统的“生产用”过程能力指数一样，都遵循越大越好的原则。

那么，C_p公式中的USL、M和LSL该如何确定呢？在 3.1.2 节中，我们已经绘制出了 9 个指标的控制图，其中“个值控制图”中的UCL、CL和LCL，为各个指标规定了上下界限和目标值。不同于传统过程能力指数中的顾客要求，这样的界限是完全由数据驱动的。事实上，对水温、水位这类自然指标规定上下界限和目标值，本身就要通过数据统计的方法，人为的限制并无意义。因此，不妨将上述UCL、CL和LCL直接视作不同自然指标的USL、M和LSL，分别按照“主成分分析法”和“简单权重法”进行线性组合，最终求得综合指标的USL、M和LSL。

表 4　6 个指标的上下界限（标准化）

指标	电磁波幅度 EW	电磁波幅度 NS	地温
USL	0.7140	0.8268	1.0499
M	0.0000	0.0000	0.0000
LSL	-0.7140	-0.8268	-1.0367
指标	水位	水温	气氡
USL	1.0378	2.0036	4.6124
M	0.0000	0.0000	0.0000
LSL	-1.0378	-2.0036	-4.6147

表 4 给出了 6 个指标USL、M和LSL的标准化取值。其中标准化的方法，与x_{ij}经过标准化得到$\tilde{x}_{ij}$的方法一致。如果将表 4 写成一个矩阵：

$$\boldsymbol{S}=\begin{array}{c} \\ USL \\ T \\ LSL \end{array}\begin{array}{c} \begin{array}{cccccc} \tilde{x}_{\cdot 1} & \tilde{x}_{\cdot 2} & \tilde{x}_{\cdot 3} & \tilde{x}_{\cdot 4} & \tilde{x}_{\cdot 5} & \tilde{x}_{\cdot 6} \end{array} \\ \begin{bmatrix} 0.71 & 0.83 & 1.05 & 1.04 & 2.00 & 4.61 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ -0.71 & -0.83 & -1.04 & 1.04 & -2.00 & -4.61 \end{bmatrix} \end{array},$$

那么综合指标 1 的$(USL, M, LSL)_{(1)}{}^{\mathrm{T}} = \boldsymbol{S}\boldsymbol{w}^{(\mathbf{1})} = (0.1613, 0, -0.1613)$，综合指标 2 的$(USL, M, LSL)_{(2)}{}^{\mathrm{T}} = \boldsymbol{S}\boldsymbol{w}^{(\mathbf{2})} = (1.581249, 0,\ 1.581249)$。

最后，计算两个综合指标的过程能力指数。由于数据进行了标准化，因此$\hat{\mu} = 0$，从而，过程能力指数的计算公式退化为

$$C_p = \frac{(USL - LSL)}{12\hat{\sigma}},$$

其中的

$$\hat{\sigma} = \frac{\overline{R}}{d_2} = \frac{\overline{R}}{1.128} = 0.8865\overline{R},$$

而$\overline{R}$就是“移动极差图”的中心线。事实上，C_p的分子分母同时除以 2，那么它的分子就是向量$(USL, M, LSL)_{(2)}{}^{\mathrm{T}}$的第一个分量，而它的分母就是“个值控制图”中的$UCL$。从而，

$$C_p^{(1)} = \frac{0.1613}{1.10} = 0.146,$$
$$C_p^{(2)} = \frac{1.581249}{1.24} = 1.275.$$

可以发现，$C_p^{(1)}$和$C_p^{(2)}$的差别非常大。前者低于 0.2，能力过小，在地震预测的背景下，这一数字说明综合指标 1 非常不稳定，其中参杂了很多不稳定因素，因此它所识别出的异常值，究竟是由其他不稳定因素造成，还是由其内在的、与地震有关的因素造成，就令人怀疑。而后者在 1.2 至 1.3 的区间内，能力较为充足，说明对于给定的上下界限，综合指标 2 较为稳定，因此它所识别出的异常值就使人信服。

为什么利用主成分分析法和简单权重法计算得到的指标差别如此之大？究其原因，使用主成分分析方法，需要满足许多条件。首先，各个指标应尽量接近正态分布；其次，变量之间应当存在部分高度相关。但是，所分析的六个指标，尽管呈现对称单峰分布，却不完全满足正态性，从而使得分析结果存在偏差。更重要的是，除了电磁波幅度 EW 和电磁波幅度 NS 高度相关以外，其他变量之间的两两相关系数都很小（见表 5），在这种情况下，仅仅选择 3 个主成分，显然会导致信息提取不充分。如果仅仅将电磁波幅度 EW 和电磁波幅度 NS 进行降维，保持其他互不相关的指标，即选择 5 个主成分，那么主成分分析就和简单的加权平均法差别不大了。因此，在这一案例中，简单权重法的效果更好。

表 5 相关系数矩阵

	电磁波幅度 EW	电磁波幅度 NS	地温	水位	水温	气氡
电磁波幅度 EW	1.000	0.879	0.015	-0.108	0.018	-0.007
电磁波幅度 NS	0.879	1.000	0.005	-0.106	0.015	0.010
地温	0.015	0.005	1.000	-0.031	-0.011	-0.014
水位	-0.108	-0.106	-0.031	1.000	-0.295	0.063
水温	0.018	0.015	-0.011	-0.295	1.000	-0.041
气氡	-0.007	0.010	-0.014	0.063	-0.041	1.000

4.3 利用综合指标分析

根据上一节的分析，综合指标 2 更加稳定，因此后续的分析中，将采用综合指标 2。首先，观察图 4，对超限的观测进行记录，共发现 18 个异常数据，分别为 2，3，20，33，277，541，542，558，559，1156，1190，1302，1303，1306，1537，1566，1692，1693。在这 18 个异常数据中，共有 4 个处于地震发生前后 30 天以内，有震报准率为 22%，在各项指标中处于中位水平。由于只有 4 个“有效数据”，并且所有地震发生前后 30 天的日期并集为 5 段不相交区间；所以，根据鸽巢原理，至少存在 1 次地震没有被预报，这也说明了综合指标仍然存在局限性。因此，在监

测地震相关指标变化时，应当将多个指标综合起来进行观察，单独依靠任意一个指标都是不合理的。

当然，为了克服上述局限性，考虑到许多指标存在较长的先行期或滞后期，可以将“地震发生前后 30 天”这一范围放宽至“地震发生前后 60 天”。这样，“有效数据”个数就上升至 7 个，“有震报准率”为 38.89%，并且基本每一次地震都能够被这些“有效数据”识别。如果将“60 天”的范围进一步放宽，地震识别准确性还将进一步上升。

5. 结论和展望

本文利用控制图的思想，建立了监测地震相关指标变动的模型。尽管对“生产用”控制图进行了较大的改动，但是通过详细的论文，本文说明了控制图这一工具的普适性和强大作用。通过构建“有震报准率”这一指标，本文将识别得到的异常值与地震发生时刻进行关联，排除了三个对地震基本不敏感的指标。然后，本文选用了两种方法构建综合指标，不仅要求综合指标保持较高的“有震报准率”，还对它提出了识别每一次地震的要求。利用过程能力指数对两个综合指标进行对比，最终选用了更加稳定的后者进行后续分析。得到的指标基本实现了提出的两个要求，能够较为精准地识别地震的发生。

但是，本文也存在着问题和改进的空间。首先，本文没有试图区分先行指标、同步指标、滞后指标，这就给地震的预测、综合指标的外推带来了较大的障碍。其次，本文“有效数据”定义中的时间范围，是地震发生前后 30 天或 60 天；但是当异常值的出现晚于地震发生时，它的识别意义就已经不大了。从这一角度来说，寻找更加合理的“有效数据”的定义，是实现预测性能提升的关键。

综上，本文将较为严格、精密的“生产用”控制图推广至自然现象的监测中，是一次富有创新性的、较有意义的尝试。地震的发生变化莫测、难以捉摸，希望这一尝试能够对地震的监测工作有所启发，也希望质量管理中的控制图能够被推广至更多的场景中，在更广泛的领域中大放异彩。

参考文献

[1] 朱俊杰. 基于多元统计的地震数据分析与预测研究[C]. 中国自动化学会控制理论专业委员会.第 36 届中国控制会议论文集（D）. 中国自动化学会控制理论专业委员会:中国自动化学会控制理论专业委员会, 2017: 730-735.

[2] 周纪芗, 茆诗松. 质量管理统计方法, 第二版[M]. 北京: 中国统计出版社, 2008.

[3] 何晓群. 多元统计分析, 第四版[M]. 北京: 中国人民大学出版社, 2015.

附 录

附录中将展示部分 R 语言代码。第一部分为计算“有震报准率”的函数以及示例。

```r
## Part 1.
## define the function
Anorm <- function(X)
{
  sum <- 0
  sum <- sum + sum(X >= 201 & X <= 261)
  sum <- sum + sum(X >= 466 & X <= 526)
  sum <- sum + sum(X >= 704 & X <= 798)
  sum <- sum + sum(X >= 1123 & X <= 1234)
  sum <- sum + sum(X >= 1660 & X <= 1720)
  lt <- list("#Abnormal" = length(X), "#Seismic" = sum, "ratio"=sum/length(X))

  return(lt)
}

## start data analysis
setwd('C:\\Users\\小竹子\\Documents\\Quality Control Statistics\\Lab 1')
dat <- read.csv("异常数据.csv")
for(i in 1:ncol(dat))
{
  X <- dat[,i]
  X <- X[!is.na(X)]
  result <- Anorm(X)
  print(colnames(dat)[i])
  print(result)
  print("===========")
}

# [1] "电压"
# $`#Abnormal`
# [1] 39
#
# $`#Seismic`
# [1] 3
#
# $ratio
# [1] 0.07692308
#
# [1] "==========="
# [1] "EW"
# $`#Abnormal`
# [1] 40
#
# $`#Seismic`
# [1] 13
#
# $ratio
# [1] 0.325
#
# [1] "==========="
# [1] "NS"
# $`#Abnormal`
```

```
52.     # [1] 22
53.     #
54.     # $`#Seismic`
55.     # [1] 13
56.     #
57.     # $ratio
58.     # [1] 0.5909091
59.     #
60.     # [1] "==========="
61.     # [1] "地温"
62.     # $`#Abnormal`
63.     # [1] 32
64.     #
65.     # $`#Seismic`
66.     # [1] 5
67.     #
68.     # $ratio
69.     # [1] 0.15625
70.     #
71.     # [1] "==========="
72.     # [1] "水位"
73.     # $`#Abnormal`
74.     # [1] 108
75.     #
76.     # $`#Seismic`
77.     # [1] 25
78.     #
79.     # $ratio
80.     # [1] 0.2314815
81.     #
82.     # [1] "==========="
83.     # [1] "气温"
84.     # $`#Abnormal`
85.     # [1] 2
86.     #
87.     # $`#Seismic`
88.     # [1] 0
89.     #
90.     # $ratio
91.     # [1] 0
92.     #
93.     # [1] "==========="
94.     # [1] "气压"
95.     # $`#Abnormal`
96.     # [1] 2
97.     #
98.     # $`#Seismic`
99.     # [1] 0
100.    #
101.    # $ratio
102.    # [1] 0
103.    #
104.    # [1] "==========="
105.    # [1] "水温"
106.    # $`#Abnormal`
107.    # [1] 50
108.    #
109.    # $`#Seismic`
```

```
# [1] 13
#
# $ratio
# [1] 0.26
#
# [1] "==========="
# [1] "气氡"
# $`#Abnormal`
# [1] 11
#
# $`#Seismic`
# [1] 3
#
# $ratio
# [1] 0.2727273
#
# [1] "==========="
```

第二部分为计算综合指标 1 及综合指标 2 的 USL、LSL 代码。

```
## Part 2.
## Comprehensive index 1
## A is the USL/LSL matrix, whose rows are USL/LSL, and cols are the variables

## B is the coefficients matrix
A <- matrix(c(0.7140, 0.8268, 1.0499, 1.0378, 2.0036, 4.6124,
              -0.7140, -0.8268, -1.0499, -1.0378, -2.0036, -4.6124),
            2, byrow = T)
B <- matrix(c(0.5, 0.119, 0.007, 0.5, 0.125, -0.011,
              0.015, -0.042, 0.883, -0.139, 0.578, 0.03,
              0.064, -0.604, -0.182, -0.011, 0.214, -0.422),6,3,byrow=T)
## l is the weight vector
l <- c(0.4542, 0.3063, 0.2395)
A %*% B %*% l

# USL and LSL of comprehensive index 1 are:
#            [,1]
# [1,]  0.1613136
# [2,] -0.1613136

## -----------------

## Comprehensive index 2
## a is the weights/ coefficients vector
a <- c(0.176979792, 0.321781441, 0.085086439,
       0.126053983, 0.141583834, 0.148514511)
## b is the USL vector
b <- c(0.7140, 0.8268, 1.0499, 1.0378, 2.0036, 4.6124)
sum(a*b)

# The USL of comprehensive index 2 is: (the LSL is the opposite)
# [1] 1.581249
```