R 语言在地球化学数据趋势面分析中的应用

邱志勇 [1], 邱文虎 [2], 肖永红 [1], 张阳阳 [1], 李 超 [1]

(1. 安徽省地质环境监测总站, 安徽 合肥 230000)
(2. 陆金所 (上海) 科技服务有限公司, 上海 200120)

摘 要: R 语言作为 GNU 系统的一个自由, 免费, 源代码开放的软件, 是一种适合推广应用于统计计算和统计制图的优秀工具. 在地球化学大数据的趋势面分析中借助 R 语言软件, 选择 spatial 库包来进行 kriging 分析和点模式分析; spatial 库中的 surf.gls 方法用最小二乘法来拟合趋势面; 使用 anova 方法比较多个嵌套模型的拟合优度, 实现了趋势面函数模型的最优拟合, 计算与绘图自动完成, 增强了分析的可靠性. 该文以安徽省全椒县中部地区主要商品粮基地 215.86 km^2 范围内 586 个土壤样品测试数据为例, 通过对 Zn, Cu 及 Zn/Cu 比值的趋势面分析, 求得三阶趋势面函数拟合度最优, 三阶趋势面图形与地质环境条件基本吻合, 证实了 R 语言的应用优势.

关键词: R 语言; 趋势面分析; 函数模型; 全椒县; 土壤样品; 锌铜比值

1 引言

趋势面分析法是用于分析地球化学元素含量空间分布和变化规律的一种常用多元统计分析数学方法 [1−4], 求解的趋势面是一个光滑的数学曲面, 它能够集中地反映空间数据在大范围内的变化趋势, 是揭示面状区域上连续或非连续分布的现象空间变化规律的理想工具. 由于传统趋势面分析是经典统计学在点数据进行空间展面上的应用, 属于全局多项式插值, 对整个研究区域用一个多项式进行拟合. 它的缺点在于: 当研究区域范围较大, 地形很复杂时, 需要用高阶多项式拟合以提高精度, 但高阶将增加其计算成本, 工作量巨大.

专门进行数据分析和图形绘制的 R 语言软件环境和编程语言为解决上述问题提供了可能. 利用 R 语言多重包处理空间数据, 选择 spatial 库包来进行 kriging 分析和点模式分析, 通过最小二乘法来拟合趋势面; 同时, 使用 anova 方法比较多个嵌套模型的拟合优度, 进一步提高趋势分析的可靠性. R 语言作为 GNU 系统的一个自由, 免费, 源代码开放的软件, 避免了软件版权问题纠葛, 是一种适合推广应用于统计计算和统计制图的优秀工具.

本文以安徽省全椒县中部地区 215.86 km^2 范围的农田采集 586 个土壤测试样品中 Zn, Cu 及 Zn/Cu 比值的趋势面分析为例, 讨论分析 R 语言在趋势面分析中应用的优越性.

收稿日期: 2018-05-23
资助项目: 中国地质调查局 “土地地球化学调查” 工程 “淮河皖江经济区土地质量地球化学调查” 项目中的子项目 (IHEGDD2016066)

2 趋势面分析法

2.1 分析思路

趋势面分析是在映射的数据通过控制点的地理坐标的多项式展开, 求解最小二乘法确定多项式函数的系数, 确保趋势面的平方偏差的总和最小. 多项式可以扩展到任何期望的程度, 由于舍入误差而存在计算限制, 通过求解一组包含 X, Y 和 Z 值的幂和交叉乘积的联立线性方程来找到未知系数. 一旦求解了系数, 就可以在地图区域内的任何点评估多项式函数. 通过将网格节点的坐标代入多项式, 并计算每个节点的曲面估计, 以此创建节点值的网格矩阵. 相比其它类似的多项式方程的数据近似, 最小二乘法是最优的拟合方法 [5−6].

本分析将映射变量分为两部分, 即趋势和趋势残差. 趋势对应于 “区域特征” 的概念, 而残差代表 “局部特征”, 本次使用趋势面分析是一种用于过滤空间数据的全局方法.

2.2 基本原理

假设 $Z_i(x_i, y_i)$ 表示所要分析的地理要素的实际观测值, 即特征值. 其中 (x_i, y_i) 为所要研究的地理区域内各调查点的坐标值. 把测试值 $Z_i(x_i, y_i)$ 的变化分解成” 区域特征” 和” 局部特征” 两个部分, 即:

$$Z_i(x_i, y_i) = f(x_i, y_i) + \varepsilon_i. \tag{1}$$

式中 $f(x_i, y_i)$ 为趋势值, 表示由整个区域因素决定的部分, 反映了在较大区域范围内 Z 随着 x 和 y 变化的特点; 而 ε_i 为剩余值 (残差值), 表示由局部区域因素和随机因素决定的部分, 反映了在局部区域范围内 Z 有异于一般规律变化的情况和随机性的干扰所造成的偏差.

趋势面分析的核心就是从实际值出发推算趋势面, 使得残差平方和趋于最小. 以此来估计趋势面参数, 是在最小二乘法意义下的趋势面拟合.

2.3 分析步骤

用回归方法求得趋势值和剩余值, 即根据已知数据 $Z_i(x_i, y_i)$ 的一个回归方程 $f(x_i, y_i)$, 使得如下残差的平方之和

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left[Z_i(x_i, y_i) - f(x_i, y_i) \right]^2. \tag{2}$$

达到极小, 根据测试数据值利用最小二乘法拟合曲面.

建立趋势面回归方程. 在趋势面分析中, 通常选择多项式作为回归方程.

一阶趋势面函数:

$$f_1(x, y) = b_0 + b_1 x + b_2 y. \tag{3}$$

二阶趋势面函数:

$$f_2(x, y) = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2. \tag{4}$$

三阶趋势面函数:

$$f_3(x, y) = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2 + b_6 x^3 + b_7 x^2 y + b_8 xy^2 + b_9 y^3. \tag{5}$$

n 阶趋势面函数:

$$f_n(x, y) = b_0 + b_1 x + b_2 y + b_3 x^2 + b_4 xy + b_5 y^2 + \cdots + b_p y^n. \tag{6}$$

其中, $p = \frac{1}{2}(n+1)(n+2) - 1$

上述式中: $f(x,y)$ – 各坐标确定的相应位置的趋势变化值; x,y – 相应的横坐标与纵坐标值; b – 系数; n – 阶数.

趋势面分析拟合程度与回归模型的效果直接相关, 本文采用 AIC 信息准则和 P 值校验, 检验趋势面拟合程度. 趋势面方程不是阶次越高越好, 随着拟合次数的增加, 其通用性和预测性也就越低, 计算也越复杂. 一般结合实际情况综合考虑, 选择适当阶次的趋势面方程. 本研究采用 R 语言以及相关软件包建立趋势面方程并绘制等值线图.

3 R 语言应用

3.1 应用原理

R 语言具有处理空间和时空数据的高级功能, 是用于数据分析和图形绘制, 并被广泛使用的编程语言和软件环境具有无与伦比的优势条件 [7−8].

R 语言中处理空间数据的多种包, 包括 sp 包, maptools 包, gstat 包 [9]. 本应用选择其中的 spatial 包 [10] 来进行 kriging 分析和点模式分析; spatial 包中的 surf.gls 方法通过最小二乘法来拟合趋势面; 同时利用 anova 方法比较多个嵌套模型的拟合优度.

3.2 计算步骤

根据采集的数据, 使用 R 语言 spatial 软件包中的 surf.gls 方法来生成趋势面模型. 使用 prmat 法生成克里金 (Kriging) 曲面, 使用 semat 法生成预测标准差曲面 [10].

```
surf.gls(np, covmod, x, y, z, nx=1000,··· )
prmat(obj, xl, xu, yl, yu, n)
semat(obj, xl, xu, yl, yu, n)
```

surf.gls 方法中 np 为需要生成的趋势面阶数, covmod 为采用的协方差函数. 将采样数据中 “横坐标” 作为 x, “纵坐标” 作为 y, “Zn/Cu” 作为 z, 编写 R 代码, 分别拟合次数 1 至 6 阶的多项式曲面.

变量 obj 为 surf.gls 方法生成的对象, 使用 prmat 方法生成曲面. 参数 xl,xu 分别为在 x 轴上的绘图范围; 参数 yl,yu 为在 y 轴上的绘图范围; 参数 n 表示在绘图区域中使用的网格大小.

方法 semat 的参数与 prmat 类似, 它将 surf.gls 方法生成的 obj 对象生成预测标准差曲面. 参数 xl, xu 分别为在 x 轴上的绘图范围; 参数 yl, yu 为在 y 轴上的绘图范围; 参数 n 表示在绘图区域中使用的网格大小. 上述方法求得 x,y,z 值后, 通过 contour 方法绘图.

4 应用实例

以 “环巢湖地区 (全椒县)1:5 万土地质量地球化学调查” 项目为例, 对全椒县中部地区 215.86 km^2 范围的耕地采集 586 个土壤测试样品中 Zn, Cu 及 Zn/Cu 值进行趋势面分析. 该项目的目的任务是查明区内土壤化学元素及化合物的分布与分配特征, 评价土地营养元素及有益元素丰缺程度, 有毒有害元素的污染程度, 以及对农业有益有害元素异常分布范围和存在的主要土地质量问题, 为当地土地资源合理开发利用, 保护和建设高标准基本农田提供依据.

4.1 地质环境背景条件

工作区是位于安徽省全椒县中部的石沛镇, 六镇镇, 二郎口镇等三个镇的主要商品粮基地. 其中, 北部的石沛镇耕地与南部两镇耕地不连续分布, 地面标高一般 10~56m, 合计耕地面积约 130km^2.

工作区地处北亚热带季风气候区, 多年平均降水量 1017.6mm, 属长江左岸支流: 滁河流域. 地貌类型主要为丘陵东南前缘的岗坡地与河漫滩, 地表岩性以第四系松散岩类为主.

土壤母质类型分为 4 类, 河流冲积物成土母质约占 7.31%, 晚更新世黄土成土母质约占 50.38%, 红色碎屑岩类成土母质约占 30.14%, 碳酸盐类成土母质约占 12.18%.

4.2 样品采集测试

耕地土壤采样点均匀布局, 采用 GPS 确定采样点位坐标. 采样密度, 采样深度等工作方法符合设计要求. 样品代表性较好, 实物样品齐全. 样品加工, 处理流程符合要求. 防污染措施得当, 质量管理体系运行有效. 调查项目野外工作质量顺利通过专家评审验收.

样品测试由安徽省地质实验研究所承担. 该所实验室拥有生态地球化学调查样品测试资格 (五十二种元素) 和全国地下水污染调查评价样品测试资格证.

4.3 数据列表

对 586 个土壤测试样品中 Zn, Cu 及 Zn/Cu 值进行趋势面分析. 数据格式见如表 1, 分析数据共 586 组.

表 1 锌, 铜及锌铜比值格式表

横坐标	纵坐标	Zn(mg/kg)	Cu(mg/kg)	Zn/Cu
606140	565140	79.00	29.40	2.687074865
608183	564061	59.00	26.30	2.243346073
609219	564488	69.20	28.90	2.394463594
607385	564211	71.40	29.70	2.404040394

4.4 生成趋势面

利用上述方法构建趋势面模型, 自动绘制 Zn/Cu 比值等值线图. R 语言中最多可以得到 6 阶多项式曲面, 1 至 6 阶的多项式曲面模型如下:

```
> library(spatial)
> data.gls1 <-surf.gls(1, covmod=expcov, data, d=0.7)
> data.gls2 <-surf.gls(2, covmod=expcov, data, d=0.7)
> data.gls3 <-surf.gls(3, covmod=expcov, data, d=0.7)
> data.gls4 <-surf.gls(4, covmod=expcov, data, d=0.7)
> data.gls5 <-surf.gls(5, covmod=expcov, data, d=0.7)
> data.gls6 <-surf.gls(6, covmod=expcov, data, d=0.7)
```

其中, data 为包含 586 个采样数据的 x, y, z 值的数据结构. 以 data.gls3 为例, 生成 3 阶曲面图与标准差图过程如下:

```
> GLS <-data.gls3
> krig <-prmat(GLS, data.ls$rx[1], data.ls$rx[2],data.ls$ry[1], data.ls$ry[2], 100)
> par(mfrow=c(1,2))
```

```
> contour(krig)
> points(data$x, data$y, cex=0.5)
> SEs <- semat(GLS, data.ls$rx[1], data.ls$rx[2], data.ls$ry[1], data.ls$ry[2], 100)
> contour(SEs)
> points(data$x, data$y, cex=0.5)
```

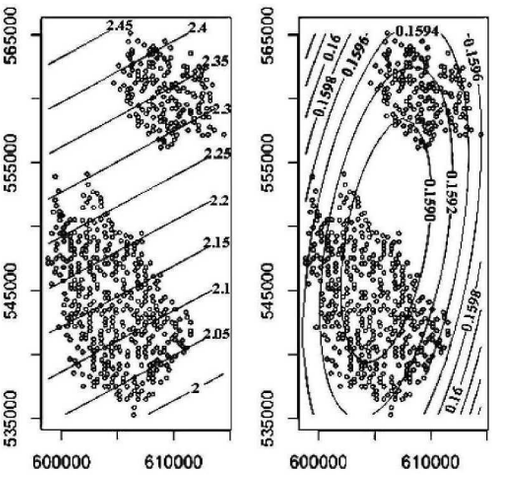

图 1 一阶趋势面图 (左) 及标准差图 (右)

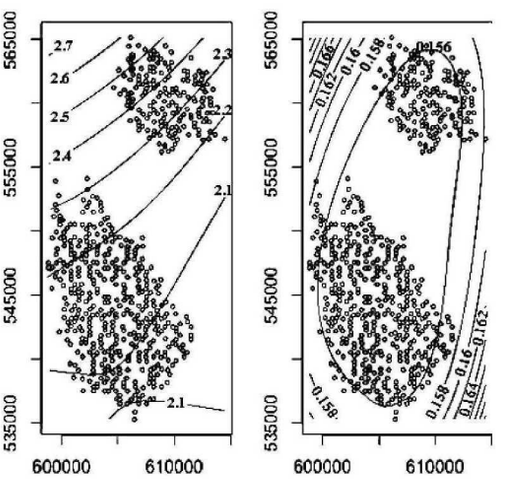

图 2 二阶趋势面图 (左) 及标准差图 (右)

各阶趋势面及标准差如图 1, 2, 3, 4, 5, 6 所示.

4.5 数据分析

随着阶数的提高, 模型的残差越小. 从图中可以看出, 一阶曲面是一个光滑的平面, 采样点相对均匀的分布在各趋势区间内; 而当阶数提高到 5 至 6 阶时, 所有在采样点都分布在一个底部平缓, 中部略微隆起的曲面内底部.

使用 summary 方法可以获得每个模型的 AIC 值:

```
> summary(ata.gls1)
Analysis of Variance Table
Model: surf.gls(np=1, covmod=expcov, x=data, d=0.7)
...
AIC: (df=3) -2150.745
...
```

每个模型的 AIC 值如表 2 所示: 一般情况下 AIC 值最小的模型是最佳的, 本次分析中每组模型的 AIC 差距不大. 希望尽量使模型简单化, 通过 anova 方法来比较多个模型的拟合优度, 随着阶数的增加, 低阶模型的项完全包含在高阶模型中.

通过方差分析 (表 3) 对以上模型进行检验. 低阶模型的项完全包含在高阶模型中, 阶数升到 3 阶时 F 值增加明显, P 值足够小, 表明 3 阶模型十分显著, 在统计意义上关联显著, 三阶趋势面函数拟合度最优, 锌铜比值分布规律与土壤类型分布对比基本吻合 (见图 7), 向下游方向土壤锌含量呈现明显减少的趋势. 3 阶曲面 3D 示意图如图 8 所示.

表 2　模型拟合度对比表

阶数	自由度	AIC 值
1	3	–2150.745
2	6	–2176.712
3	10	–2255.803
4	15	–2257.35
5	21	–2285.64
6	28	–2301.63

表 3　方差分析表

阶数	Res.Df	Res.Sum Sq	Df	Sum Sq	F 值	P 值
1	582	14.656				
2	579	13.877	3	0.77940	10.8399	6.153e-07
3	575	11.957	4	1.91948	23.0756	< 2.2e-16
4	570	11.724	5	0.23371	2.2726	0.0460883
5	564	10.943	6	0.78025	6.7020	7.222e-07
6	557	10.397	7	0.54688	4.1856	0.0001653

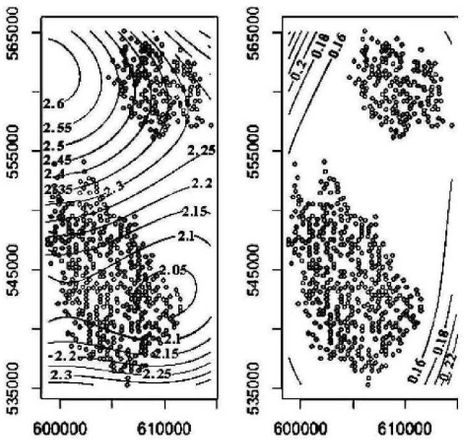

图 3　三阶趋势面图 (左) 及标准差图 (右)

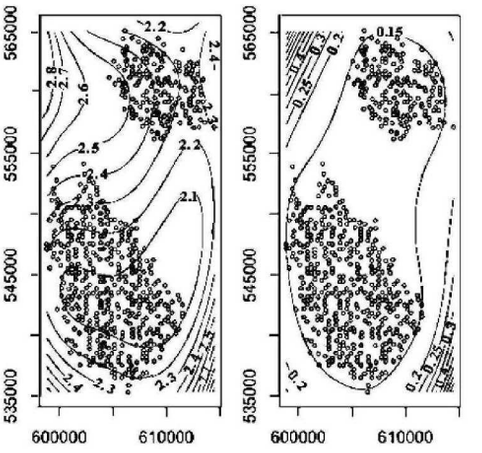

图 4　四阶趋势面图 (左) 及标准差图 (右)

```
> anova(data.gls1, data.gls2, data.gls3, data.gls4,data.gls5,data.gls6)
```

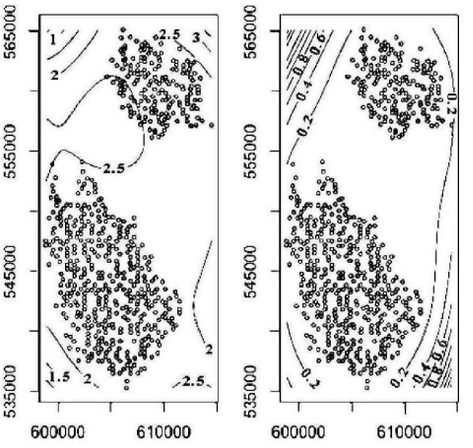

图 5　五阶趋势面图 (左) 及标准差图 (右)

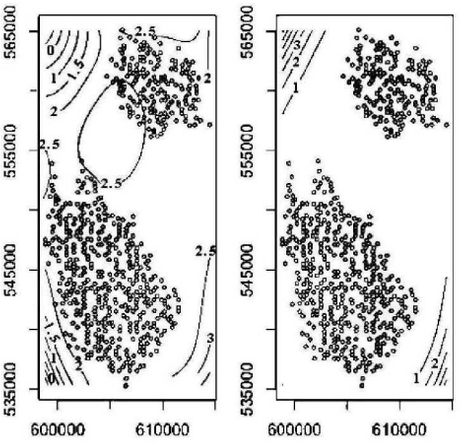

图 6　六阶趋势面图 (左) 及标准差图 (右)

5　讨论

1) 在地形较复杂的工作区域, 需要实现完全定量化的分析评价, 建立高阶趋势函数模型仍然是一种选择, 传统计算分析工作量较大, 难以获得高效率的解决方案.

2) R 语言为解决高阶趋势函数模型的自动计算分析和绘图提供了可能. R 语言的 spatial 包中的多项式插值法引入了概率模型, 认为一个统计模型不可能完全精确地得出预测值. 进行预测时, 应该给出预测值的误差. Kriging 插值是一个最优的无偏估计法. 获得预测图并不要求数据呈正态分布, 当数据呈正态分布时, kriging 插值法将是无偏估计法中效果最好的一种方法.

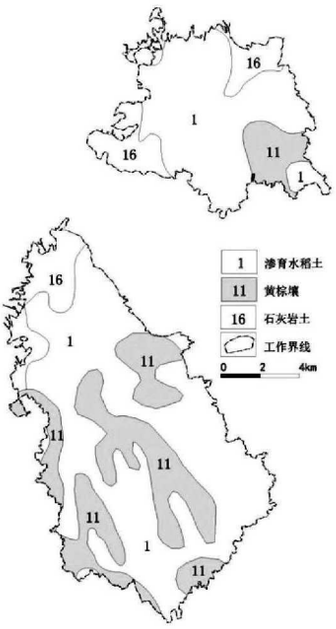

图 7 工作区土壤类型分布图

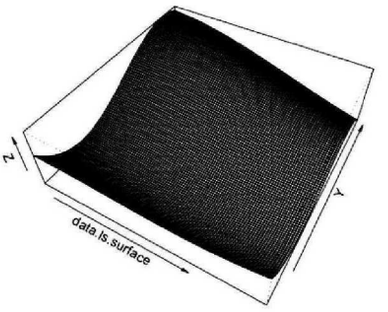

图 8 3 阶曲面 3d 示意图

3) 作为自由, 免费, 源代码开放的 R 语言软件, 不存在软件版权纠葛, 它是一个用于统计计算和统计制图的优秀工具.

4) 由于对整个工作区域用一个多项式进行拟合, 存在一定的局限性. 当工作区域范围很大. 地形很复杂时, 需要用多阶至 n 阶多项式拟合以提高精度, 可能误差仍然较大; 需要配合其它方法, 互相印证, 合理选择模型.

5) 建立大区域地球化学数据高阶趋势函数模型时, 建议考虑地质环境背景条件, 以地形地质单元划分趋势分析区块为宜.

6 结语

通过对安徽省全椒县中部地区 586 个土壤样品元素的趋势面分析, 利用 R 语言强大的计算分析功能, 求解出最佳拟合函数, 得到最佳分析结论, 表明在实际应用中 R 语言具有高效、实用的优势. 各项分析计算步骤作为命令文件保存, 今后针对同类型数据分析时, 仅需更改数据源文件后, 在 R 语言中调用相应函数即可自动完成所有计算和绘图工作, 适合推广应用.

参考文献

[1] 赵丹, 赵华甫, 饶杰等. 基于趋势面的耕地质量空间分异特征及影响因素 [J]. 水土保持研究, 2015, 22(6): 219-223.

[2] 胡永定, 李炬, 王石凡等. 徐州市北郊农田土壤重金属污染趋势面分析及其容量研究 [J]. 生态与农村环境学报, 1993, 9(4): 34-38+63.
[3] 李随民, 姚书振, 韩玉丑. Surfer 软件中利用趋势面方法圈定化探异常 [J]. 地质与勘探, 2007, 43(2): 72-75.
[4] Wang H, Zuo R. A comparative study of trend surface analysis and spectrum-area multifractal model to identify geochemical anomalies[J]. Journal of Geochemical Exploration, 2015, 155: 84-90.
[5] 武松. SPSS 统计分析大全 [M]. 清华大学出版社, 2014: 38-286.
[6] 王远飞, 何洪林. 空间数据分析方法 [M]. 北京: 科学出版社, 2007: 58-226.
[7] Bivand R, Gebhardt A. Implementing functions for spatial statistical analysis using the language[J]. Journal of Geographical Systems, 2000, 2(3): 307-317.
[8] TORGO L. Data mining with R: learning with case studies[M]. CRC press, 2016: 55-362.
[9] Bivand R. Implementing spatial data analysis software tools in R[J]. Geographical Analysis, 2006, 38(1): 23-40.
[10] Ripley B, Bivand R, Venables W, et al. Package<spatial>[J]. 2015. http: // www.stats.ox.ac.uk/pub / MASS4/.

Application of R Language in Trend Surface Analysis of Geochemical Data

QIU Zhi-yong[1], QIU Wen-hu[2], XIAO Yong-hong[1], ZHANG Yang-yang[1], LI Chao[1]

(1. General Station of Geo-Environment Monitoring of Anhui Province, Hefei 230000, China)
(2. Lujinsuo (Shanghai) Science and Technology Service Co, Ltd., Shanghai 200120, China)

Abstract: As a free, free, open-source software for the GNU system, the R language is an excellent tool for popularizing statistical computing and statistical drawing. In the trend analysis of geochemical big data, R software is used to select the spatial library package to perform kriging analysis and point pattern analysis; the surf.gls method in the spatial library uses the least squares method to fit the trend surface; use anova method to compare The goodness of fit of multiple nested models realizes the optimal fitting of the trend surface function model, and the calculation and drawing are automatically completed, which enhances the reliability of the analysis. This paper takes 586 soil sample data from the main commercial grain base in the central area of Quanjiao County in Anhui Province as an example. Through the trend surface analysis of Zn, Cu and Zn/Cu ratios, the third-order trend surface function fitting is obtained. The degree is optimal, and the third-order trend surface graph is basically consistent with the geological environment conditions, confirming the application advantages of R language.

Keywords: R language; trend analysis; function model; Quanjiao County; soil samples; Zinc-Copper ratio