

# 空气状况、气象条件、舒适度指数与 人体健康关系及预报反演

孔祥意, 齐 麟, 滕 聪

(山东科技大学 测绘科学与工程学院, 山东 青岛 266590)

**摘 要:**随着社会经济的发展,人体健康问题受到普遍关注,为了人类健康地生活,对人体健康问题的研究及预报尤为重要。呼吸系统和循环系统的健康是人体健康的重要指标,在对北京市 2010 年和 2011 年空气质量、气象、舒适度指数、呼吸系统疾病和循环系统疾病就诊人数数据分析的基础上,使用逐步回归分析法和 BP 神经网络法分别构建就诊人数的预报模型。研究表明,呼吸系统疾病和循环系统疾病就诊人数和部分因素显著相关,且呼吸系统疾病年份不同相关因素不同,循环系统疾病季节不同相关因素不同,对两种预报模型进行对照, BP 神经网络的平均绝对误差更小,平均比逐步回归模型小 15 以上,对人体健康的预报十分有效,从而为人体健康的预报实验提供验证性依据。

**关键词:** 相关分析; 逐步回归分析; BP 神经网络; 预报模型; 人体健康

中图分类号: P208

文献标识码: A

文章编号: 1672-5867(2020)10-0148-05

## Relationship between Air Condition, Weather Condition, Comfort Index and Human Health and Prediction Inversion

KONG Xiangyi, QI Lin, TENG Cong

(College of Geomatics, Shandong University of Science and Technology, Qingdao 266590, China)

**Abstract:** With the development of social economy, human health problems are generally concerned, in order to human health, the study and prediction of human health is particularly important. The health of respiratory system and circulatory system is an important indicator of human health. In the city of Beijing on 2010 and 2011, air quality, weather, comfort index, respiratory diseases and patients of circulatory system diseases on the basis of data analysis, using stepwise regression analysis and BP neural network prediction model of the number of visits were constructed respectively. Research shows that the correlation of respiratory diseases and patients of circulatory system diseases and respiratory system diseases and some factors in different years for different factors related to diseases of the circulatory system in different seasons, different correlation factors, two kinds of prediction model of BP neural network control, the average absolute error is smaller than the average regression model of small is above 15. Prediction on human health is very effective, so as to provide basis for verification of human health prediction experiment.

**Key words:** correlation analysis; stepwise regression analysis; BP neural network; forecasting model; human health

## 0 引 言

随着物质生活、精神生活的提高,人们对人体健康的渴望越来越强烈,健康是人类追求的共同目标。是否患有疾病是人体健康的重要指标,建立人体健康预报模型可以使人们更了解健康问题,对人们的出行提供依据,对实现身心的全面发展、促进人体健康是十分有效的。

过去,关于单一因素和人体健康关系的研究内容较多,多因素的研究基本没有。本文通过分析 15 个气象因素、舒适度指数和空气污染指数(API)与呼吸系统疾病人数和循环系统疾病人数,采用逐步回归分析和较为经典的 BP 神经网络得到就诊人数与各指标的相关关系及预报模型,并对这两种预报模型进行准确性分析,对两种模型进行比较,进而对人体健康问题进行研究。

收稿日期: 2019-05-27

作者简介: 孔祥意(1994-),男,山东烟台人,测绘工程专业硕士研究生,主要研究方向为环境遥感。

1 数据获取与分析

所用数据由人口与健康科学数据共享平台提供。疾病就诊人数是人体健康的有效反应,气象、空气污染指数(API)、舒适度指数数据获取较为简单,有很强的即时性,资料的传输高速可靠,所需要的数据,在制作预报前一般都可以得到,但人体健康数据获取十分困难,采用 2010—2011 年北京市数据对人体健康进行相关因素的分析及预报模型建立对人体健康预测是必要的。

对获取的数据进行初阶分析,共计 730 d 的数据。在气象参数中 20 时—20 时降水量用代号 32 700 代表的数值为微量降水,将其数值赋值为 0.1。在数据中舒适度指数为冷暖感觉形式给出,需对舒适度指数进行分级处理,分级标准见表 1。

表 1 舒适度指数分级表  
Tab.1 Comfort index scale

寒冷	冷	微冷	凉	凉爽	舒适	微热	热	炎热	酷热
-4	-3	-2	-1	-1	0	1	2	3	4

对患病就诊人数进行时间序列<sup>[5]</sup>显示如图 1 所示。从图 1 中可以看出 2011 年呼吸系统疾病日就诊人数比 2010 年明显增多,总人数接近 2010 年的两倍,2011 年呼吸系统日就诊人数最低值基本持平于 2010 年最高值,为了确保有效性,故对呼吸系统日就诊人数分为 2010 年和 2011 年分别进行研究,得到呼吸系统日就诊人数预报模型。从图 1 可以看出循环系统日就诊人数高峰期在冬季,其他季节人数基本稳定,故对循环系统的就诊人数预报模型分为冬季(12、1、2 月)和其他季节两部分进行<sup>[6]</sup>。

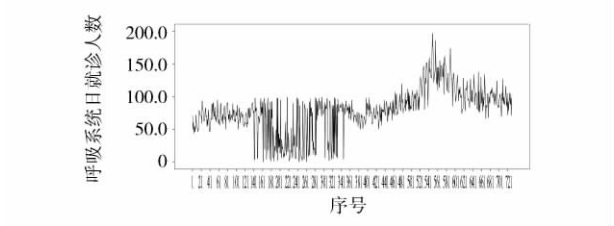


图 1 2010—2011 呼吸系统日就诊人数走势图  
Fig.1 Trend chart of daily visits to respiratory system in 2010—2011

数据进行处理,提取出进入模型中的变量因子,防止日就诊人数不相关的变量进入到模型中,对全部参数进行相关性分析,相关分析是指按照变量的现实数据,计算出表示自变量和因变量间相关程度和性质的统计量——相关系数,并进行显著性检验<sup>[7]</sup>。采用双变量相关分析方法,能得到每个参数与日就诊人数之间的相关性,从而得出与日就诊人数显著相关的参数。

相关系数的查验一般采取 F 检验法和 t 检验法,对相关系数  $r$  的显著性进行检验。F 检验的计算公式为:

$$F = \frac{r^2}{(1 - r^2)/(n - 2)}, df_1 = 1, df_2 = n - 2 \quad (1)$$

t 检验的计算公式为:

$$t = \frac{r}{s_r}, df = n - 2 \quad (2)$$

式中,  $s_r = \sqrt{(1 - r^2)/(n - 2)}$ , 叫作相关系数的标准误。

我们得到用来进行建立模型的符合 F 检验和 t 检验的显著性变量,2010 年呼吸系统日就诊人数的显著性变量:平均气压、日最高气压、日最低气压、平均气温、日最高气温、日最低气温、平均相对湿度、最小相对湿度、平均风速、最大风速、舒适度指数在置信水度(双侧)为 0.01 时显著相关,极大风速在置信水度(双侧)为 0.05 时显著相关;2011 年比 2010 年多一个 20 时到 20 时降水量,全部在置信水度(双侧)为 0.01 时显著相关。冬季循环系统就诊人数的显著性变量:平均相对湿度、最小相对湿度、日照实数在置信水度(双侧)为 0.01 时显著相关,20 时到 20 时降水量、最大风速、最大风速的方向在置信水度(双侧)为 0.05 时显著相关;其他季节有日最高气温、20 时到 20 时间降水量、最大风速、舒适度指数在置信水度(双侧)0.05 时显著相关,极大风速、日照时数在置信水度(双侧)为 0.01 时显著相关。

2 研究方法与技术路线

2.1 逐步回归预报模型建立与分析

在实际问题中,影响因变量  $Y$  的因素(自变量  $X$ )可能有很多,通常我们希望从中挑选出有显著影响的自变量,剔除那些影响不显著的变量。在回归分析中,当自变量的数量比较大的时候,普遍采用逐步回归的方法来解决实际问题,进行相关分析得到的显著性变量仍然较多,需要进一步筛选。逐步回归法就是将变量再次筛选得到更优的回归变量。逐步回归方法在参数变量非常多的情况下能够筛选出与因变量最相关的参数,对于回归方程的建立是非常有利的,避免了方程中引入太多的无关变量<sup>[9]</sup>。

基本思想:逐个引入自变量。逐步引入对因变量显著影响的因子,不显著的因子剔除,最终得到的方程中有对  $Y$  影响显著的变量,又没有对  $Y$  影响不显著的变量<sup>[10]</sup>。

逐步回归的技术路线如图 2 所示。基本步骤:首先给出引入变量的显著性水平  $\alpha_{in}$  和剔除变量的显著性水平  $\alpha_{out}$ ,然后按图 2 筛选自变量。采用逐步回归方法为向前引入法,及时将参数变量一个一个逐步引入到回归方程中,每引入变量后对变量进行检验,检验合格则引入该变量,不合格将其剔除。

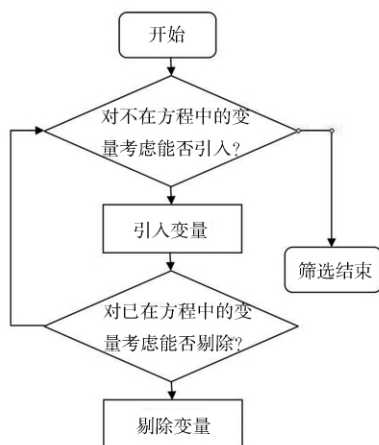


图2 逐步回归的技术路线

Fig.2 The technical route of gradual return

## 2.2 模型建立与检验

通过相关分析得到这些显著性变量后,将 2010—2012 年的显著性变量分别和呼吸系统日就诊人数和循环系统日就诊人数数据进行逐步回归,使用其中 75% 的数据进行试验,25% 的数据用来模型验证<sup>[11]</sup>,从而筛选出最终进入回归方程的变量,得到人体健康预报模型。

通过相关性分析得到 2010 年与呼吸系统就诊人数显著相关的参数有 12 个,将 12 个参数 75% 的数据与呼吸系统日就诊人数进行逐步回归分析,从而筛选出最终进入回归方程的变量共有 3 个,得到预报模型,进行逐步回归后我们得到的逐步分析结果见表 2。

表2 模型摘要

Tab.2 Model summary

模型	R	R 平方	调整后 R 平方	标准估算的错误	更改统计量				
					R 方变化	F 更改	df1	df2	显著性 F 更改
1	0.416a	0.173	0.170	26.296 3	0.173	56.623	1	271	0.000
2	0.197b	0.197	0.191	25.963 3	0.024	7.966	1	270	0.005
3	0.460c	0.211	0.203	25.772 4	0.015	5.016	1	269	0.026

a. 预测变量: 舒适度指数( 常量)  
b. 预测变量: 舒适度指数, 最大风速( 常量)  
c. 预测变量: 舒适度指数, 最大风速, 平均气温( 常量)

由表 2 可以看出第 3 个模型才是最终结果,所以本文只给出了最终模型的系数表。表中列出了终极回归方程的常数项和各个自变量对应的非标准化系数、非标准化系数、t 值和显著性程度。

综合分析结果可得,用逐步回归方法最终筛选出 3 个参数,分别为舒适度指数、最大风速、平均气温,求得的最优多元回归方程式为:

2010 年呼吸系统日就诊人数 =  $(-11.206) \times \text{舒适度指数} + 0.224 \times \text{最大风速} + 0.129 \times \text{平均气温} + 27.671$

2011 年与呼吸系统日就诊人数显著相关的参数有 13 个,通过与 2010 年同样方法最终筛选出 3 个参数,分别为舒适度指数、最大风速、平均相对湿度,求得的最优多元回归方程式为:

2011 年呼吸系统日就诊人数 =  $9.519 \times \text{舒适度指数} + (-0.148) \times \text{最大风速} + 0.131 \times \text{平均相对湿度} + 97.580$

冬季循环系统日就诊人数显著相关的参数有 6 个,通过上述同样方法最终筛选出 2 个参数,分别为平均相对湿度和最大风速的方向,求得的最优多元回归方程式为:

冬季循环系统日就诊人数 =  $(-1.723) \times \text{平均相对湿度} + 3.307 \times \text{最大风速的方向} + 329.776$

其他季节循环系统日就诊人数 =  $(-0.334) \times \text{极大风速} + (0.201) \times \text{日照时数} + 227.5$

## 2.3 神经网络方法模型建立

### 2.3.1 人工神经网络的原理与方法

神经网络运行的基本处理单位和运行基础是神经元,一般把构成神经元的模型概括为 3 种要素:

1) 神经元功能函数。它是连接输入神经与输出神经的中间部分,也称为激活函数。激活函数有简单线性函数、阈值函数、Sigmoid 函数。在已经熟练和广泛使用的神经网络中,一般都是用 S 型函数作为激励函数,如 Logistic 函数,定义如下:

$$\varphi(v) = 1 / (1 + \exp(-av)) \quad (3)$$

其中,  $a$  是 sigmoid 函数的倾斜参数。

2) 加权和函数其对象是输入信号。这个操作步骤是将输入信号转换成一个线性组合。

3) 突触或连接链,一般两个靠近的神经元都是紧密相连地靠它连接的。根据连接情况,把神经网络分成前馈神经网络和反馈神经网络两种形式。

### 2.3.2 BP 误差反传神经网络

Back Propagation 神经网络(简称 BP 神经网络)是应用最成熟和普及的神经网络,它属于多层前馈网络。BP 神经网络的结构如图 3 所示。

图 3 表现的神经网络结构中,含有两个隐含层,即一个输入层和一个输出层,并且网络之间都是全部相连接

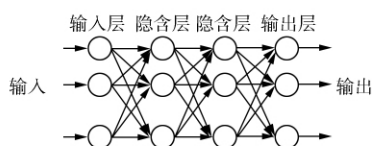


图 3 具有两个隐藏层的神经网络

Fig.3 Neural network with two hidden layers

的,就是说任意层上的神经元与它之前神经元上的节点全部连接,输入信号从输入层到隐藏层再到输出层,方向上镶嵌的。

一般的 BP 神经网络都是用经验来设计网络结构、激励函数、学习算法以及所使用的样本。在进行模型预测时,一般把样本分成训练集和测试集两个部分,比例关系一般为 3:1,样本的选择要有一定的代表性,隐藏层结构的设计通常由经验来决定,一般来说单隐藏层结构能够解决大多数的问题,隐藏层节点数的设计方法一般通过经验公式算出:

$$m = \sqrt{n + l} + \alpha \text{ 或 } m = \sqrt{nl} \quad (4)$$

式中, $m$  为隐藏层节点数; $n$  为输入层节点数; $l$  为输出层节点数; $\alpha$  为调节常数,在 1—10 之间<sup>[15]</sup>。

### 3 研究结果与分析

#### 3.1 回归预报模型检验结果

为了分析模型预测的有效性,用 75% 的数据进行逐步回归得到回归方程对剩余的 25% 没有参与建模过程的测试数据集进行预测,预测效果如图 4 至图 7 所示。

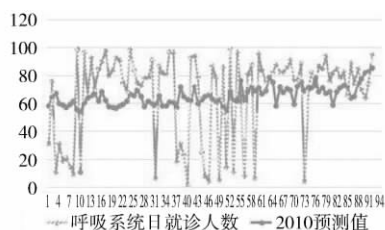


图 4 2010 年呼吸系统日就诊人数逐步回归模型预测效果

Fig.4 Prediction effect of the stepwise regression model for the number of daily visits to the respiratory system in 2010

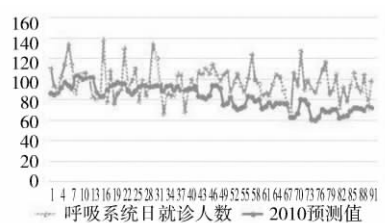


图 5 2011 年呼吸系统日就诊人数逐步回归模型预测效果

Fig.5 Predictive effect of the stepwise regression model of the number of daily visits to the respiratory system in 2011

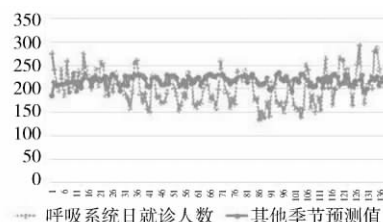


图 6 冬季循环系统日就诊人数逐步回归模型预测效果

Fig.6 Predictive effect of the stepwise regression model for the number of daily visits to the circulatory system in winter

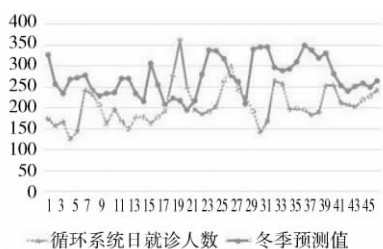


图 7 其他季节循环系统日就诊人数逐步回归模型预测效果

Fig.7 Predictive effect of the stepwise regression model of the number of daily visits to the circulatory system in other seasons

图中可以看出虚线条为观测数据,实线条为预测数据,由于疾病就诊人数存在很大不确定性,故人体健康预报模型的建立有一定的难度。2010 年验证数据就诊人数总数为 6 155,预报模型总数为 6 045,2011 年验证数据就诊人数总数为 8 911,预报模型总数为 7 581,冬季循环系统疾病验证数据就诊人数总数为 9 495,预报模型总数为 12 559,其他季节循环系统疾病验证数据就诊人数总数为 27 751,预报模型总数为 29 772。由此可以看出预报模型对每一天的预报相关性不高,但总人数具有高度相关性,并且变化趋势具有一定的相似性,所以通过逐步回归的方法得到的模型,在一定程度上对就诊人数的预测也是有意义的,但效果并不是太好。

#### 3.2 BP 神经网络预报模型建立与分析结果

在用神经网络模型预报之前,首先要做的是将网络结构设计好。在上一步用逐步回归分析时,已经通过相关分析挑选显著性相关的参数。用显著性相关参数和日就诊人数来确定节点数,输出只有一个节点是日就诊人数,输入节点数就是相关参数个数<sup>[16]</sup>。迭代次数设置为 3 000 次,参数学习率定为 0.1,Sigmoid 参数设置为 0.95,动态参数为 0.7,其他参数 Matlab 环境默认即可。2010 和 2011 年呼吸系统日就诊人数训练数据为 274 个、测试数据为 93 个,冬季循环系统日就诊人数训练数据为 134 个、测试数据为 46 个,其他季节训练数据为 412 个、测试数据为 138 个。利用 Matlab 软件对 BP 神经网络进行编程实现,获得 BP 神经网络预报模型如图 8 至图 11 所示。

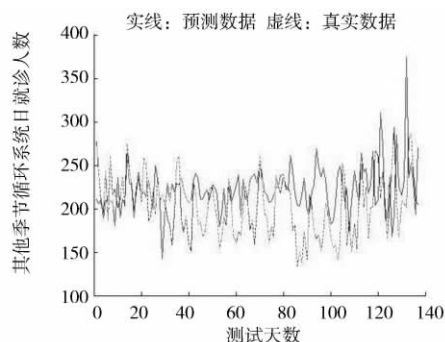


图 8 2010 年呼吸系统日就诊人数神经网络预测图  
Fig.8 Neural network forecast of the number of daily visits to the respiratory system in 2010

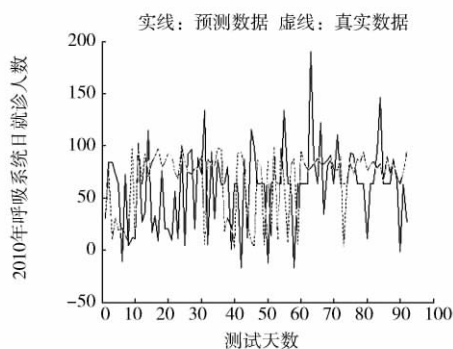


图 9 2011 年呼吸系统日就诊人数神经网络预测  
Fig.9 Neural network prediction of the number of daily visits to the respiratory system in 2011

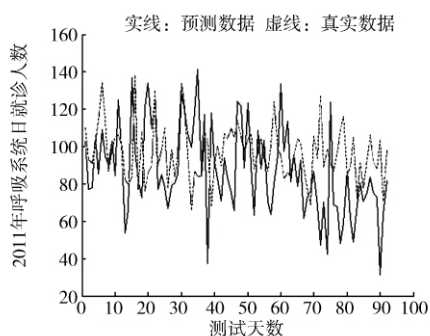


图 10 冬季循环系统日就诊人数神经网络预测  
Fig.10 Neural network prediction of the number of daily doctors in winter circulatory system

从图中可以看出,神经网络对日就诊人数的预测效果整体上要比逐步回归分析的预测效果要好,并且误差更小;对呼吸系统疾病日就诊人数的预测效果整体要比循环系统日就诊人数的预测更好,这也说明呼吸系统疾病预测较为简单;对其他季节循环系统日就诊人数的预测效果要比冬季的预测更好,这说明冬季的影响因素不稳定,影响系数较小。

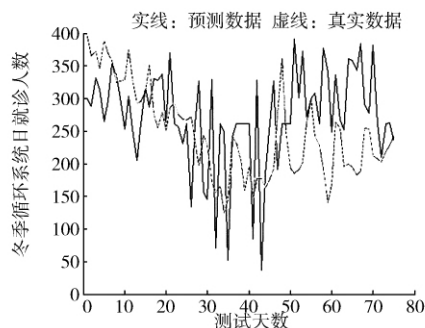


图 11 其他季节循环系统日就诊人数神经网络预测图  
Fig.11 Neural network prediction diagram of daily visits to circulatory system in other season

2010 年呼吸系统日就诊人数用神经网络预测的平均绝对误差为 39.4,2011 年平均绝对误差为 22.2,季循环系统日就诊人数平均绝对误差为 42.4,其他季节平均绝对误差为 32.6。2010 年呼吸系统日就诊人数用逐步回归分析得到的预报平均绝对误差为 87.3,2011 年平均绝对误差为 62.4,冬季循环系统日就诊人数平均绝对误差为 77.4,BP 神经网络方法使预测结果精度显著提高。

#### 4 结束语

本文通过使用逐步回归分析和 BP 神经网络两种方法,分别对呼吸系统日就诊人数、循环系统日就诊人数和气象数据、空气质量和舒适度指数的关系进行分析,得到显著性指标,并建立日就诊人数的预报模型。对两种方法进行比较。得到以下结论:

1) 呼吸系统日就诊人数和年份有关,2011 年比 2010 年明显增加;循环系统日就诊人数和季节有关,冬季比其他季节日就诊人数明显增加。

2) 所有数据中,只有部分数据和日就诊人数显著相关,并且每一个模型建立所用的显著性相关参数并不相同。

#### 参考文献:

- [1] 王祖承,陈正洪.冷空气对武汉市人群呼吸道和心脑血管疾病的影响[J].湖北预防医学杂志,2001,12(1): 15-16.
- [2] 高少华,冯皓.气象条件对呼吸道疾病影响的关系分析[J].陕西气象,2005(4): 25-26.
- [3] HAJAT S, HAINES A. Associations of cold temperature with GP consultations for respiratory and cardiovascular disease amongst the elderly in London [J]. journal of Epidemiology, 2002, 31(4): 825-830.
- [4] BRAGA A L F, ZANOBETTI A, Schwartz J. The effect of weather on respiratory and cardiovascular deaths in 12 US cities [J]. Environmental Health Perspectives, 2002, 110(9): 859.

(下转第 155 页)

$v/v_1$ 。因此,草地在地图中的长度乘以经验值后即为矢量模型中的运算路径长度。

## 2 实验结果与讨论

### 2.1 实验准备

我们采用谷歌发布的最新卫星地图,截取了某区域部分卫星图,如图 2 所示,面积约为  $26 \text{ hm}^2$ 。卫星照片比较清晰,显示区域内路线比较复杂,具有规划意义。而且分布有两片水系,树木覆盖率较高,满足实验要求的特殊地形条件,能够对改进后最短路径算法进行规划检验。

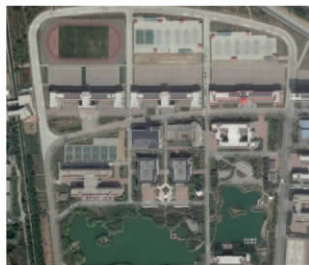


图 2 某区域卫星图

Fig.2 Satellite map of certain area

在卫星图片上加入节点和路线的信息,程序可以自主添加并命名节点和路线,节点之间路线的长度可以通过程序算法计算出来,也可以根据实地测量得到的数据进行路线长度赋值。添加完毕后,会自动生成一幅卫星矢量电子地图。接下来最短路径规划以此为基础进行。

### 2.2 行动路径规划系统操作流程

打开卫星矢量电子地图后,选择行动方式。选定后,对地图进行预处理,预处理的目的是根据选定的方式,对地图矢量模型进行调整、改正。而后输入起始点和终点,开始计算两个目标点最小的代价。

### 2.3 规划结果

在不指定交通工具时,进行单一的最短路径规划,这种情况下,不会考虑特殊的地形和地物,比如水系、林地等,仅在道路网模型中进行规划,所规划的路径都是可通行的无差别道路。选择起始点  $m$ ,终结点  $q_1$ ,规划的结果如图 3 所示,得到最短路径为  $m \geq l = o \geq p \geq z \geq e_1 \geq f_1 \geq p_1 \geq q_1$ 。

如果选择交通方式为徒步,在所选择地图中,有一片草地,在草地上步行可以通行。选择起始点为  $a$ ,终结点为  $r_1$ ,规划结果如图 4 所示,得到最短路径为  $a \geq z_1 \geq g \geq r_1$ 。



图 3 单一模式下路径规划结果

Fig.3 Path planning results in single mode



图 4 步行路径规划结果

Fig.4 Path planning result under walking conditions

## 3 结束语

本文着眼于提高野外测绘工作的效率,利用最短路径算法,结合我们可以获取的卫星图像进行了快速路径规划策略的研究。实验结果表明,这种方法可以准确快速地规划出达到目的地的可行最短路径并计算距离,同时相较于其他方法,本方法具有成本低、实时性好、准确性高的优点。

### 参考文献:

- [1] CORMEN THOMAS H, LEISERSON CHARLES E, RIVEST Ronald L, et al. Introduction to algorithms (second edition) [M]. Cambridge: The MIT Press and McGraw-Hill Higher, 2017.
- [2] 任刚, 王伟. 交通规划中的动态路网及其模型研究 [J]. 公路交通科技, 2002, 12, 19(6): 108-111, 122.
- [3] 王永明. 军事地理信息系统研究和应用 [D]. 北京: 清华大学, 1996.

[编辑: 任亚茹]

(上接第 152 页)

- [5] 吴占福, 马旭平, 李亚奎. 统计分析软件 SPSS 介绍 [J]. 河北北方学院学报(自然科学版), 2006, 22(6): 67-73.
- [6] 张梅, 陈玉光, 李杜敏, 等. 上呼吸道疾病与气象条件的关系及预报 [J]. 辽宁气象, 2004, 20(1): 32-33.
- [7] 王颖. 实验设计与 SPSS 应用 [M]. 北京: 化学工业出版社, 2013.
- [8] 高惠璇. 应用多元统计分析 [M]. 北京: 北京大学出版社, 2005.

- [9] 易丹辉. 统计预测: 方法与应用 [M]. 北京: 中国统计出版社, 2001.
- [10] 高惠璇. 应用多元统计分析 [M]. 北京: 北京大学出版社, 2005.
- [11] 崔泳琳, 何史林, 周忠玉, 等. 气象与环境因子影响心脑血管疾病急诊数预测模型研究 [J]. 环境与健康杂志, 2013, 30(1): 48-52.

[编辑: 张 曦]