农业资源与环境学报 中文核心期刊 中国科技核心期刊
JOURNAL OF AGRICULTURAL RESOURCES AND ENVIRONMENT
欢迎投稿 http://www.aed.org.cn

耕地质量评价缺失数据填充方法研究

陈宇， 周悟， 胡月明， 谢健文

引用本文:
陈宇, 周悟, 胡月明, 等. 耕地质量评价缺失数据填充方法研究[J]. 农业资源与环境学报, 2021, 38(6): 1132–1141.

在线阅读 View online: https://doi.org/10.13254/j.jare.2021.0201

陈宇, 周悟, 胡月明, 等. 耕地质量评价缺失数据填充方法研究[J]. 农业资源与环境学报, 2021, 38(6): 1132-1141.
CHEN Y, ZHOU W, HU Y M, et al. Research on filling methods of missing data in cultivated land quality evaluation[J]. *Journal of Agricultural Resources and Environment*, 2021, 38(6): 1132-1141.

开放科学 OSID

耕地质量评价缺失数据填充方法研究

陈宇[1], 周悟[1], 胡月明[1,2,3,4,5,6*], 谢健文[1,2,3,4,6]

(1.华南农业大学资源环境学院, 广州 510642; 2.广东省土地信息工程技术研究中心, 广州 510642; 3.广东省土地利用与整治重点实验室, 广州 510642; 4.自然资源部建设用地再开发重点实验室, 广州 510642; 5.青海大学农牧学院, 西宁 810016; 6.青海-广东自然资源监测与评价联合重点实验室, 西宁 810016)

摘 要:在耕地质量数据调查与采集过程中会由于人为、环境等因素造成数据缺失,而目前数据缺失填充方法都存在适用性不足的问题,为完善耕地质量数据库从而提高耕地质量评价精度,对耕地质量评价缺失数据填充方法的研究是十分重要的。本研究以广州市从化区耕地质量数据库为样本集,根据空间相关性和空间分布将数据集划分为空间关联性数据集和非空间关联性数据集,利用多种填充方法对其进行缺失填充模拟,采用十字交叉法进行精度验证。结果表明:选取数据整体异常值比例不足1.2%,且高程、气温、有效锌等25组因素具有空间相关性。对空间关联性数据填充精度最高的是四象最近邻算法,在缺失率20%以下时精度仍高达80%,精度随缺失率增大而降低,其次为K最邻近(KNN)算法、期望最大化法、多重填充法、回归模型算法,四象最近邻算法相较于KNN算法在数据密集时精度更好。对非空间关联性数据填充精度最高的是相似聚集填充算法,在缺失率25%以下时精度超过80%,其次为期望最大化法、多重填充法、回归模型算法。综上,本研究提出的四象最近邻算法和相似聚集填充算法相比其他算法在耕地质量评价缺失数据填充中精度更高,效果更稳定,且实用性更广。

关键词:耕地质量评价;缺失;数据;填充;从化区;精度
中图分类号:F323.211; S158 **文献标志码**:A **文章编号**:2095-6819(2021)06-1132-10 **doi**: 10.13254/j.jare.2021.0201

Research on filling methods of missing data in cultivated land quality evaluation

CHEN Yu[1], ZHOU Wu[1], HU Yueming[1,2,3,4,5,6*], XIE Jianwen[1,2,3,4,6]

(1.College of Natural Resources and Environment, South China Agricultural University, Guangzhou 510642, China; 2. Guangdong Province Engineering Research Center for Land Information Technology, Guangzhou 510642, China; 3. Guangdong Provincial Key Laboratory of Land Use and Consolidation, Guangzhou 510642, China; 4. Key Laboratory of the Ministry of Natural Resources for Construction Land Transformation, Guangzhou 510642, China; 5.College of Agriculture and Animal Husbandry, Qinghai University, Xining 810016, China; 6.Qinghai-Guangdong Joint Key Laboratory of Natural Resources Monitoring and Evaluation, Xining 810016, China)

Abstract: In the process of cultivated land quality data investigation and collection, there will be missing data due to human, environmental, and other factors. However, the current missing data-filling methods have insufficient applicability. In order to improve the cultivated land quality database and evaluation accuracy, it is important to explore missing data-filling methods in cultivated land quality evaluation. In this study, the cultivated land quality database of Conghua District Guangzhou City was used as the sample set. According to the spatial correlation and spatial distribution, the dataset was divided into spatial and non-spatial correlation datasets. Various filling methods were used to simulate the missing data filling, and a cross method was used to verify the accuracy. The results indicated the proportion of total outliers was less than 1.2%, and 25 factors such as elevation, temperature, and available zinc showed spatial correlation. The four-image nearest neighbor algorithm presented the highest filling accuracy for spatial association data, and the accuracy was as high as 80% when the missing rate was less than 20%. The accuracy decreased with the increase in the missing rate. The four-image nearest

收稿日期:2021-04-06 **录用日期**:2021-06-10
作者简介:陈宇(1998—),男,湖北荆州人,硕士研究生,从事土地资源大数据研究。E-mail:yc980718@163.com
***通信作者**:胡月明 E-mail:yueminghugis@163.com
基金项目:国家重点研发计划课题(2020YFD1100204);国家自然科学基金项目(U1901601)
Project supported: National Key R&D Program of China(2020YFD1100204); The National Natural Science Foundation of China(U1901601)

neighbor algorithm was followed by K-nearest neighbor algorithm (KNN), expectation maximization algorithm, multiple interpolation algorithm, and regression model algorithm. The four-image nearest neighbor algorithm showed better accuracy than K-nearest neighbor algorithm when the data was dense. For the non-spatial correlation dataset, the highest filling accuracy was the similar aggregation filling algorithm, which could maintain more than 80% accuracy within 25% of the missing rate, followed by expectation maximization algorithm, multiple interpolation algorithm, and regression model algorithm. To sum up, the four-image nearest neighbor algorithm and the similar aggregation filling algorithm proposed in this study show higher accuracy, more stable effect, and wider practicability than other algorithms for filling missing data in cultivated land quality evaluation.

Keywords: evaluation of cultivated land quality; missing; data; filling; Conghua District; accuracy

耕地是一种特定的土地，是人类活动的产物，是人类开垦之后用于种植农作物并经常进行耕耘的土地[1]。它是人类所需食物的主要源泉，是农业生产发展的主要物质基础，而耕地关乎粮食安全，粮食安全关乎国家发展与社会稳定[2]。耕地质量评价可准确评估耕地生产力与适宜性，是耕地保护、开发、政策完善等的重要前提[3]。

耕地质量评价数据是对耕地质量产生影响的指标数据集，而耕地质量评价缺失数据即是数据集中部分遗漏、未采集、已知错误的数据。耕地质量评价数据量大、类型众多，在数据获取、输入、传输过程中，存在因人员操作不当、机器故障等原因导致的数据错误与缺失的情况，而数据的错误也是数据缺失的表现形式，进而直接影响数据分析与挖掘，使得评价结果不准确、数据利用不充分[4]。而目前对于缺失数据填充方法已有相关研究，尤其插值法、最近邻填充、回归模型、期望最大化法、多重填充等方法应用相对广泛，但这些方法都存在明显的不足。空间插值法在不同区域不同数据中的最优表现有明显差异，如克里格插值、反距离加权两种方法在不同研究中表现出各自最优，但空间插值法存在方法的选择和结论的不确定性问题[5-7]；最近邻填充法是简单高效且相对高精度的填充算法，但面对不同数据集难以有稳定的填充效果，并且存在K值难以度量的问题[8-10]；回归模型法填充局限性较大，对于数据之间的相关性要求极高，即需要数据存在必然的因果关系，并且根据数据关系构建模型费时费力，修改也极其不易[11-13]；期望最大化法是一种迭代优化过程，执行简单且稳定，逐步寻找最优解，但该算法适用于大样本，且数据集应服从正态分布[9-10,14]；多重填充法是对每个数据缺失值生成多个预测值，与上述方法不同的是该算法表现了数据集原有的不确定性，其随机性强，但运算过程复杂，精度相对较低[15]。

数据的填充能弥补数据自身的缺失或满足应用的需求，如仪器设备测量问题、操作员录入问题、分析问题等会使得数据结果与真实值存在较大差异，最终严重影响耕地质量评价结果[16]。某些数据的直接测量极其复杂或耗时耗力，甚至无法实现，因而需要采用数据填充法，如刘菲等[14]利用相关性因子对森林地林木平均胸径的填充，就是间接运用数据之间的关联性得到所需的数据。目前耕地数据库日益增加，数据规范性、完整性不足的问题愈发突出，导致数据的缺失填充愈发重要；同时对耕地调查评价愈发频繁，评价指标不断丰富，新增指标数据的获取也成为主要问题。

目前数据缺失已是不可避免的现实，而对耕地质量评价数据而言，数据的完整才是耕地质量评价的基础，由于耕地数据的采样极其复杂耗时，所以对于耕地质量评价数据的缺失填充研究迫在眉睫。当前耕地质量评价缺失数据填充没有得到系统地研究，现有的研究基本上只对耕地土壤成分缺失数据进行空间插值填充，为了科学评价耕地质量，保证土地政策和制度的有效推行，必须对当前方法进行合理利用与改进，提出耕地质量评价缺失数据填充方法，提高耕地质量评价结果的精确性和可信度。针对目前耕地质量评价数据缺失现状，本研究对耕地缺失数据的填充方法进行探讨，旨在提高耕地质量评价缺失数据的填充精度，从而完善耕地质量评价数据体系，为今后耕地质量评价等相关研究提供的理论依据，并对填充算法的应用提供更多思路与可能。

1 材料与方法

1.1 研究区概况及数据来源

1.1.1 研究区概况

从化区地处广东省中部、广州市东北部，位于113°17′~114°04′E、23°22′~23°56′N，全区总面积1 984.2 km^2，2019年末人口64.17万。属于亚热带季风气候，年平均气温21.2 ℃，降水充足，河道纵横，水

资源丰富。从化区处于珠江三角洲到粤北山区过渡地带，地势自北向南倾斜，东北高，西南低，地形呈阶梯状。2019年农村人口占比54.89%，而基本农田面积为174.9 km^2，占全区面积不足10%。从化区地理位置、耕地及采样点分布如图1所示。

1.1.2 数据来源

本研究数据主要来源于广东省/广州市统计年鉴、第二次全国土壤调查、广州市基本农田调查、数据挖掘及问卷调查等。根据常用的评价指标发现[17-18]，土壤条件对耕地质量影响最大，而地形、气候虽然在小区域变化不大，但也是影响耕地质量的重要因子。本研究主要选取从化区基本农田数据、土壤重金属数据（76个样点）、样点基础数据（204个样点）等，将其划分为地类地形、土壤条件、气候条件3个方面（表1），共32个指标，5 888条耕地质量评价数据，这些数据充分体现了从化区耕地质量的现状，为耕地质量评价奠定了基础。

1.2 方法与设计

缺失数据填充方法从应用对象上主要分为两大类型，即空间性和非空间性。空间性方法是充分考虑到数据本身存在空间关联性，从而利用自身空间关联特征来通过已知数据对缺失数据进行填充的方法；而非空间性数据之间不存在任何地理关联性，只能寻找与其他数据内部的关联性，利用其关联性对未知数据进行预测填充[17]。而对于耕地质量评价数据而言，其自身的复杂多样性决定了单一方法无法解决，因此本研究在缺失数据填充方法基础上进行改进后对耕地质量评价缺失数据进行填充，并与传统方法进行精度比较。

表1 耕地质量评价指标

Table 1 Cultivated land quality evaluation index

类别 Category	影响因素 Influence factors
地类地形	坡度、高程
土壤条件	微生物活性、总有机碳、pH值、土壤有机质、重金属含量（镉等5种）、砂石含量（粗、细、粉）、黏粒、土壤养分元素（13种）
气候条件	年日照时数、年降水量、湿度、温度

1.2.1 空间相关性分析

空间自相关分析是检验具有空间属性的要素是否对相邻空间点属性值产生影响，所以空间相关性分析必须对其属性的空间位置和属性值进行统计。目前对空间相关性分析的方法较多，最常用的是

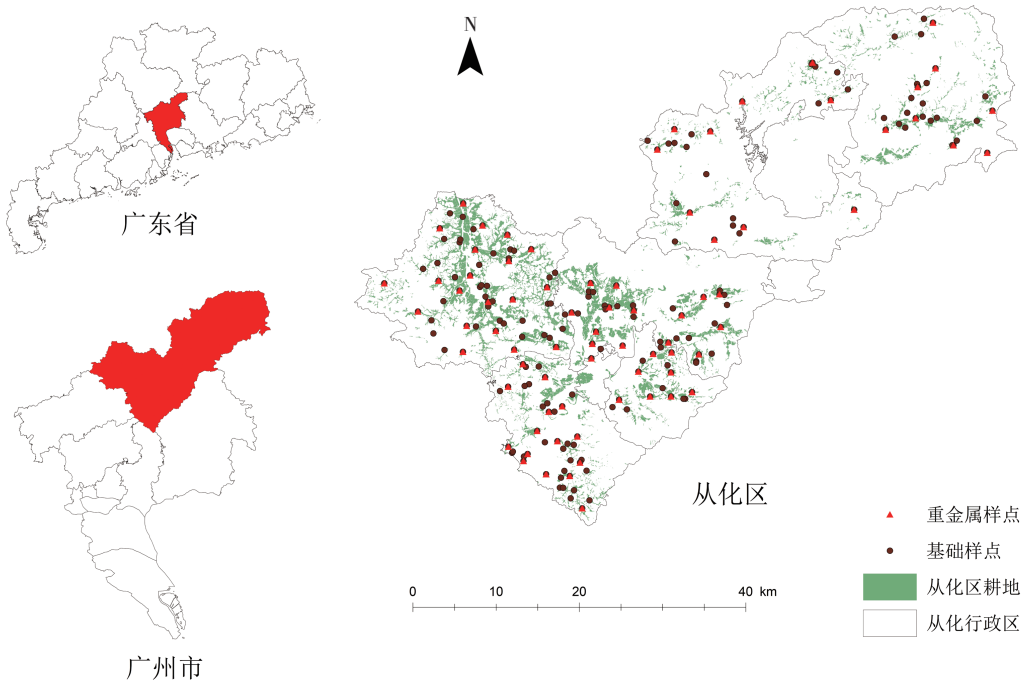

图1 从化区地理位置、耕地及样点分布图

Figure 1 Geographical location, cultivated land and samples distribution of Conghua District

Moran′s I指数，当I>0时，为正相关；I=0时不相关；I<0为负相关。具体计算见公式(1)[19]：

$$I=\frac{n\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n}W_{ij}\right)\sum_{i=1}^{n}(x_i-\bar{x})^2}(i\neq j) \quad (1)$$

式中：x_i和x_j分别为i和j所在位置的属性值；$\bar{x}$为该元素属性平均值；W_{ij}为权重。

1.2.2 填充方法

缺失数据填充方法研究已近百年，方法在不断被提出与完善，目前已提出的方法有回归模型填充、期望最大化填充(Expectation maximization，EM)、多重填充(Multiple imputation，MI)、K最邻近填充(K-nearest neigbor，KNN)、空间插值、神经网络、随机森林等，本研究选取常用的几种缺失数据填充方法进行简单介绍并提出改进方法。

(1)常用填充方法

回归模型填充是通过对自变量与因变量之间的关系进行建模预测[5-7]。该方法对于数据之间的相关性要求极高，即需要数据存在必然的因果关系。因此主要用于分析结果数据预测，多用于时间序列预测法。

KNN填充是利用欧氏距离度量与当前数据最相似的K条记录，然后用这K条记录在当前属性出现频率最高的值进行填充或者利用这K条记录对缺失位置的属性利用距离的归一化进行加权填充[20-22]。该方法多用于空间样点数据的填充，与常用空间插值反距离权重插值原理相同，该插值方法常应用于土壤数据、气候数据的填充，不同之处在于前者是对已知点缺失值的填充，后者是对未知点数据的预测[23-24]。

多重填充方法是对每个数据缺失值生成多个预测值，呈现缺失数据的不确定性；每个值都用来填充数据集中的缺失值，产生若干个完整数据集合；再利用相同的方法对多个数据集进行分析，筛选出最优解[15]。

期望最大化法是一种迭代算法，由两步组成：第一步是求出期望，第二步则是将随机参数进行极大化。先给随机变量一个初始值，求出模型中各个参数的估计值，然后再利用新估计出的模型对该随机变量进行估计，如此反复迭代，直至模型收敛为止[9-10,14]。

(2)四象最近邻填充

四象最近邻填充是在KNN填充的基础上进行改进，由于KNN法是直接筛选出最近的K个对象，有可能会存在K个对象都趋向于一方的现象，导致最终的填充结果有较大偏差，所以针对该方法的不足进行改进，提出四象最近邻填充方法[23]。四象最近邻填充方法是针对某个对象属性缺失值，在数据样本中寻找该对象每个象限中最邻近的n个对象，并利用其对应属性进行反距离加权运算，最终结果为该对象缺失值的预测值。该方法既弥补了KNN的不足，也避免了K值选择的困难。具体过程如下：

①距离度量的确定：计算出所有耕地数据对象的属性距离，用于衡量两两之间的影响程度。本研究采用目前最常用的距离度量算法——欧式距离。

$$d_{ab}=\sqrt{\sum_{i=1}^{2}(x_{ia}-x_{ib})^2} \qquad i=1,2 \quad (2)$$

式中：d_{ab}为对象a和对象b之间的度量距离，m；x_{ia}表示第a个对象的第i维坐标，m；x_{ib}表示第b个对象的第i维坐标，m；i代表对象数据维度(本研究耕地数据为二维)；a和b代表某个数据对象。

②邻近筛选：对缺失数据对象点周边其他对象进行逐一象限筛选，对存在对象的每个象限选择n(n≤3)个对象用来填充缺失数据，n过大会导致距离太远，从而关联性降低，对于周边对象少的n取值为1，保证数据具有较高的关联性。

③权重分配：采用距离权重反比，根据缺失对象与样本点对象的距离进行加权度量，一般取值权重与距离平方成反比。具体计算表达式见公式(3)：

$$w_{ak}=\frac{\left(\frac{1}{d_{ak}}\right)^2}{\sum_{k=1}^{4n}\left(\frac{1}{d_{ak}}\right)^2} \qquad k=1,2,\cdots,4n \quad (3)$$

式中：w_{ak}为对象k对对象a的影响权重系数；d_{ak}为对象a和对象k之间的度量距离，m；k为缺失数据对象筛选出的第k个对象。

④缺失填充：根据缺失对象筛选出的样本对象对应属性值与权重系数计算缺失填充值。存在的特殊分类数据先将其转换为数值数据，直接选取重复率最高的进行填充。具体计算表达式见公式(4)：

$$T=\sum_{k=1}^{4n}w_{ak}\times v_k \qquad k=1,2,\cdots,4n \quad (4)$$

式中：T为缺失填充值；v_k是第k个对象对应的属性值。

(3)相似聚集填充

相似聚集填充是将数据集划分为完整数据集和缺失数据集，通过对完整数据集内部数据自身相似关联性进行分析，通过不断迭代运算计算出数据对象间的相似性，最终利用缺失数据集中已知数据和对象相

似性结果预测缺失数据集中缺失值。该方法具体步骤如下：

①数值归一化：由于耕地数据类型众多、数据量大，数据会因为属性值范围不一、文本数据、离散数据等原因，导致数据不同属性产生影响的不平衡性，所以需要将所有数据属性值归到相同数值范围内，将文本数据转换为数值数据，使所有属性影响相同。为简化归一结果，一般都选择[0，1]。数值归一化过程具体计算见公式(5)、(6)：

$$C_i=(a_i-\bar{a})/S \quad i=1,2\cdots,n \tag{5}$$

$$\text{其中}\ \bar{a}=(\sum_{i=1}^{n}a_i)/n,S=\sqrt[n-1]{\sum_{i=1}^{n}(a_i-\bar{a})^2}$$

式中：$\bar{a}$为属性值的平均值；a_i为该属性第i个属性值；n为该属性中属性值的个数；S为该属性的标准差；C_i为数据格式化该属性第i个属性值。

$$D_i=(C_i-U_{\min})/(U_{\max}-U_{\min}) \tag{6}$$

式中：数据集$U=\{C_1,C_2,\cdots,C_n\}$，$U_{\max}$和$U_{\min}$是表示该属性数据集的最大值和最小值；D_i为归一化处理后该属性中第i个属性值。

②相似度度量：计算完整数据集中对象之间的相似度，连续变量相似度计算见公式(7)，离散变量相同为1，否则为0；构建相似度矩阵S。再通过构建吸引度矩阵X和归属度矩阵G(初始值为0)不断迭代直到聚集中心不变后停止，确定最终对象相似度矩阵[25]。

$$S_{jj'}=[n-\frac{\sum_{i=1}^{n}(a_{ij}-a_{ij'})^2}{n}] \tag{7}$$

$$x_{ij}=s_{ij}-\max\{g_{ij'}+s_{ij'}\} \tag{8}$$

$$g_{ij}=\min\{0,x_{jj}+\sum\max\{0,x_{i'j}\}\} \tag{9}$$

$$g_{jj}=\sum\max\{0,x_{i'j}\} \tag{10}$$

式中：a_{ij}为对象j的第i个属性的值；s_{ij}为第i和第j的对象之间的相似度；x_{ij}为第i和j的对象之间的吸引度；g_{ij}为第i和j的对象之间的归属度；i'和j'均表示非i和非j；当$g_{jj}+x_{jj}>0$时，迭代停止，此时与对象相似度最高的为该对象的聚集中心。

③缺失值填充：选择与缺失值对象最高相似度的k个对象作为参考值，如果其中对象也存在对应缺失值，即向下寻找下一个相似度最接近的对象。权重确定方法选择距离权重反比，具体计算同公式(2)；再通过权重和已知样品数值计算缺失值，计算式同公式(3)；对离散数据选择重复率最高的作为预测值。

1.2.3 实验设计

由于耕地数据覆盖面广、类型众多、结果复杂、数据量大、数据采集周期长等原因，对耕地质量评价缺失数据的研究较少，本研究在原有填充算法不足的前提下，提出四象最近邻和相似聚集填充方法较以往填充方法的优势。提出的两种方法是针对耕地质量评价数据结构特征，具有针对性，所以该方法在本研究的适用性较好。为验证其方法的精度并与其他填充方法比较，利用Python 3.7和SPSS 26进行数据处理和精度计算，具体过程如下。

(1)缺失处理：为验证数据填充方法的精度，选取真实完整的数据进行实验。首先使用正态分布对数据异常值进行剔除，避免数据填充过程中数据异常值影响过大，导致填充精度过低。利用空间相关性和空间分布图分析将数据集划分为空间数据集和非空间数据集；再对空间数据集中数据除去坐标数据外随机删除1%、5%、10%、15%、20%数据信息，用于模拟缺失数据集，采用四象最近邻填充方法和其余传统填充方法进行填充；对非空间数据集中随机选取5%、10%、15%、20%、25%属性因素，在其中随机删除部分属性信息，模拟缺失数据集，采用相似聚集填充方法和其余传统填充方法进行填充。

(2)精度检验：由于数据对方法的适应能力不同，为了避免偶然性，每次试验都得出不同的精度，一般取多次结果的精度平均值对模型方法精度进行估计，本研究取10次计算结果的平均值为最终精度。精度采用预测值与真实值相关系数计算，具体见公式(11)：

$$Q=\sum_{i=1}^{n}\left[1-\frac{(X-X')^2}{X^2}\right]/n \quad i=1,2,\cdots,n \tag{11}$$

式中：X为真实值；X'为预测值；n为填充个数；Q为填充精度。

2 结果与讨论

2.1 数据统计结果

由于采集的数据会存在少量异常值，需对所有数据进行正态分布检验，本研究取置信区间为$(\bar{x}-3\sigma,\bar{x}+3\sigma)$，将置信区间外的属性值划为异常值，数据检验结果(表2)表明，32组属性数据基本符合正态分布，异常值比例均小于3.5%，平均异常值比例仅为1.2%。

2.2 空间相关性分析

利用ArcMap10.2的空间自相关(Moran′s I)工具对32组属性数据进行空间相关性检验，Moran′s I指数取值范围为[-0.261 9，0.652 1]，其中具有空间正相

关的因素有高程、气温等25个,具有空间负相关的因素有全氮、粉砂粒等7个,具体相关性统计结果见表3。

虽然空间自相关分析较为客观,但为避免偶然性,本研究再利用ArcMap10.2生成空间分布图,进一步分析数据是否具有聚集相关性[26],部分空间分布图如图2所示。

由图2可以看出:从化区西南部海拔低、东北部海拔较高,具有明显的空间分布差异性;pH值基本呈现西南部偏低、东部较高、北部居中,也具有明显的空间分布差异性;而全氮含量分布不存在明显的规律和特征;微生物含量呈现与海拔高度相反的趋势,西南部含量高,东北部含量低,具有显著的空间分布差异性。而气候条件中气温与高程分布特征基本相似,东北部山区气温偏低,西南部平原气温偏高;降水及湿度与地形特征具有较大关联性,降水量相对较高的地区分布在东北部山区南坡和西南地区。数据空间分布结果分析与空间自相关分析整体基本一致,根据最终分析结果将32组数据集分为空间性数据和非空间性数据。

表2 从化区数据统计结果

Table 2 Statistical results of Conghua District

样点数 Points	因素 Factors	最小值 Min	最大值 Max	平均值 Mean	标准差 SD	变异系数 CV	置信区间 Confidence interval	异常个数 Number of exceptions
204（基础样点）	高程/m	15.00	550.00	100.74	106.00	1.05	[0,418.73)	4
	坡度/(°)	0.31	28.57	6.75	4.98	0.74	[0,21.70)	3
	有效磷/(mg·kg^{-1})	4.60	140.80	43.32	26.30	0.61	[0,122.23)	3
	速效钾/(mg·kg^{-1})	2.00	350.00	77.39	59.38	0.77	[0,255.52)	3
	有效硅/(mg·kg^{-1})	7.55	170.58	63.53	33.38	0.53	[0,163.67)	2
	有效硫/(mg·kg^{-1})	2.25	164.26	20.24	22.69	1.12	[0,88.32)	5
	有效钙/(cmol·kg^{-1})	0.10	47.50	2.47	3.74	1.51	[0,13.70)	2
	有效铁/(mg·kg^{-1})	13.20	2 456.90	320.42	276.95	0.86	[0,1 151.26)	3
	有效硼/(mg·kg^{-1})	0.01	0.30	0.10	0.07	0.64	[0,0.30)	0
	有效锰/(mg·kg^{-1})	0.50	125.50	13.08	13.21	1.01	[0,52.71)	3
	有效钼/(mg·kg^{-1})	0.01	0.27	0.09	0.04	0.49	[0,0.23)	2
	全氮/(mg·kg^{-1})	284.00	2 140.00	846.50	403.18	0.48	[0,2 056.05)	2
	有效镁/(cmol·kg^{-1})	0.10	1.10	0.33	0.20	0.59	[0,0.93)	5
	有效铜/(mg·kg^{-1})	0.05	17.20	1.83	2.42	1.32	[0,9.08)	7
	有效锌/(mg·kg^{-1})	0.16	24.82	2.96	3.49	1.18	[0,13.41)	7
	微生物活性/(10^5 cfu·g^{-1})	0.94	10.52	2.72	1.43	0.53	[0,7.00)	5
	土壤有机质/(g·kg^{-1})	4.89	68.90	22.24	8.85	0.40	[0,48.78)	1
	总有机碳/(g·kg^{-1})	0.44	6.78	1.78	0.94	0.53	[0,4.59)	5
	粗砂粒/%	0.91	81.34	22.02	19.81	0.90	[0,81.45)	0
	细沙粒/%	2.86	64.60	29.30	13.99	0.48	[0,71.27)	0
	粉沙粒/%	4.23	42.56	22.14	8.68	0.39	[0,48.18)	0
	黏粒/%	2.05	85.27	26.54	17.37	0.65	[0,78.63)	4
	pH值	4.20	8.20	5.74	0.52	0.09	(4.19,7.30)	2
	气温/℃	18.75	23.65	21.70	1.04	0.05	(18.58,24.82)	0
	年降水量/mL	1 747.50	2 947.30	2 243.48	256.43	0.11	(1 474.19,3 012.77)	0
	湿度/%	73.00	84.00	78.55	2.24	0.03	(71.82,85.28)	0
	年日照时数/h	1 225.30	1 686.50	1 449.80	80.60	0.06	(1 208.00,1 691.60)	0
76（重金属样点）	总镉/(mg·kg^{-1})	0.04	0.74	0.17	0.09	0.53	[0,0.44)	1
	总汞/(mg·kg^{-1})	0.07	0.47	0.17	0.07	0.41	[0,0.38)	2
	总砷/(mg·kg^{-1})	1.90	98.11	13.01	15.32	1.18	[0,58.98)	1
	总铅/(mg·kg^{-1})	10.50	111.91	55.06	19.05	0.35	[0,112.21)	0
	总铬/(mg·kg^{-1})	17.90	240.20	81.46	45.60	0.56	[0,218.27)	1

表3 Moran′s *I*指数统计结果

Table 3 Statistical results of Moran′s *I* index

因素 Factors	Moran′s *I*	因素 Factors	Moran′s *I*	因素 Factors	Moran′s *I*	因素 Factors	Moran′s *I*
高程	0.541 7	有效硼	0.194 4	土壤有机质	0.026 6	年降水量	0.261 2
坡度	0.115 6	有效锰	0.003 1	总有机碳	0.045 6	湿度	0.132 1
有效磷	0.039 0	有效钼	−0.084 3	粗砂粒	0.022 7	年日照时数	−0.101 2
速效钾	0.040 7	全氮	−0.261 9	细砂粒	0.141 5	总镉	−0.083 1
有效硅	0.217 6	有效镁	0.125 6	粉砂粒	−0.097 2	总汞	−0.060 9
有效硫	0.052 6	有效铜	0.121 2	黏粒	0.094 1	总砷	0.006 0
有效钙	0.031 6	有效锌	0.250 7	pH值	0.190 7	总铅	−0.005 4
有效铁	0.035 7	微生物含量	0.051 4	气温	0.652 1	总铬	0.081 8

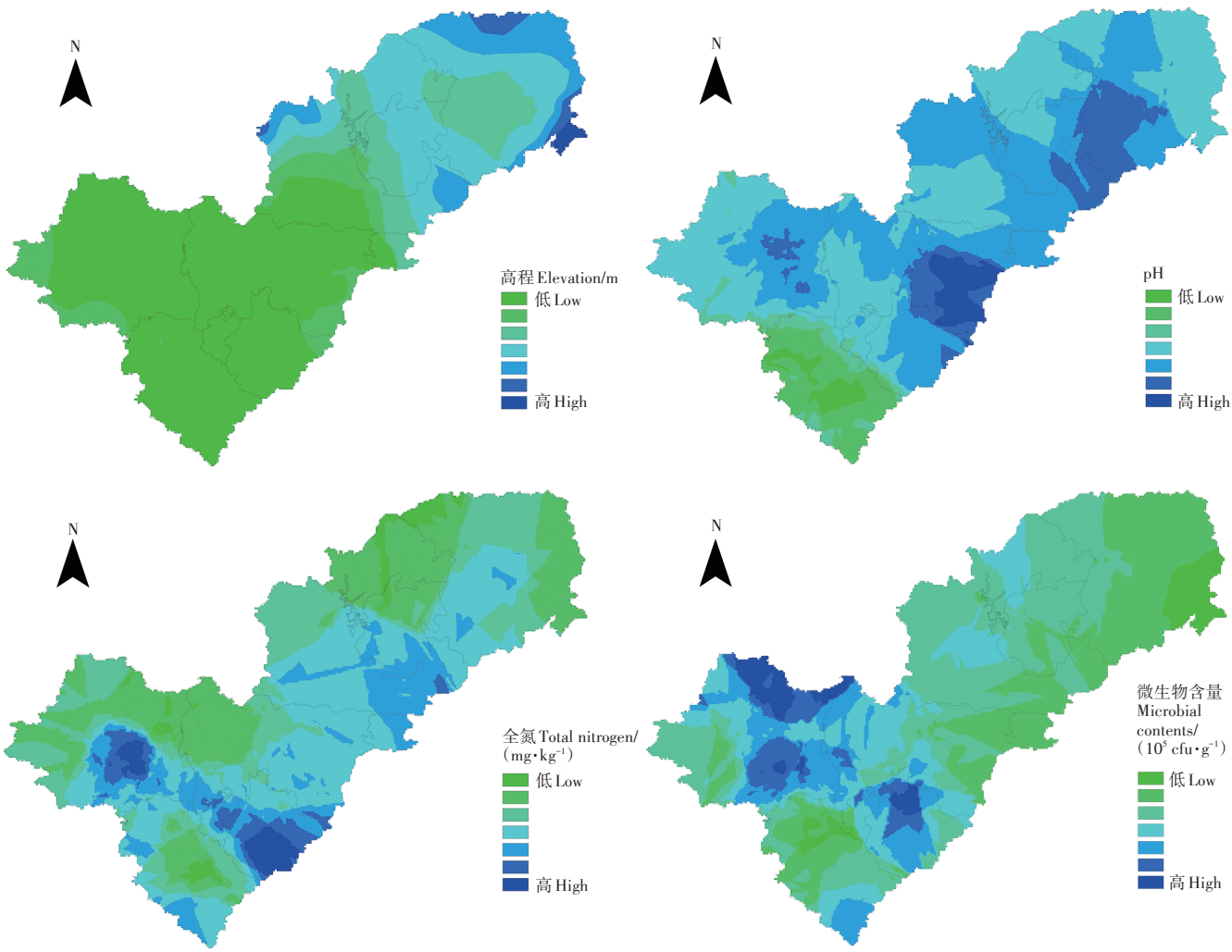

图2 高程、pH值、全氮、微生物含量空间分布图

Figure 2 Elevation, pH value, total nitrogen and microbial content spatial distribution map

2.3 空间性数据填充精度评价

根据空间相关性分析得出耕地质量评价数据中的空间性数据，如高程、气温、有效锌等25组数据，并对上述空间性数据采用回归模型法、KNN法、期望最大化填充法、多重填充算法及四象最近邻填充法进行缺失填充，并计算不同填充方法不同缺失率下的填充精度（表4）。

从表4可以得出，所有填充方法的填充精度均随

表4 不同缺失率下空间性数据各填充方法的填充精度比较(%)

Table 4 Comparison of filling accuracy of different filling methods for spatial data with different missing rate(%)

填充方法 Filling method	缺失率阈值 Missing rate threshold	1.0%	5.0%	10.0%	15.0%	20.0%
四象最近邻	20%	92.6	88.2	85.4	82.5	80.4
KNN	20%	90.2	85.6	82.3	81.6	78.8
期望最大化法	15%	85.3	81.4	76.5	75.6	72.2
多重填充	10%	80.9	75.0	70.4	62.8	58.7
回归模型	15%	75.6	72.9	69.8	66.4	57.7

着缺失率的上升逐渐降低,空间性数据中填充算法的整体精度表现为:四象最近邻填充>KNN填充>期望最大化填充>多重填充>回归模型填充。四象最近邻填充算法的数据填充精度最高,在1.0%缺失率时填充精度高达92.6%;而KNN算法在缺失率较低时,精度略低于四象最近邻填充算法,随着缺失率的不断提高,两种填充方法的精度逐渐趋于接近,主要原因是缺失率较高时,四象最近邻方法筛选各象限邻近点愈发靠远,寻找较远点导致关联性较低从而降低了填充精度。其他三种算法中期望最大化法精度相对较高,并且随着缺失率的提高精度降幅较为平缓;多重填充法在缺失率为1.0%时精度超过80%,而随着缺失率上升精度急剧下降;回归模型填充算法的精度普遍较低,在缺失率15%以下填充精度趋于稳定,而缺失率为20%时精度快速下滑。期望最大化法填充、多重填充和回归模型填充三种方法的精度相对较低可能是由于数据具有空间相关性,而这几种方法并没有对数据内部关联性进行分析,而只是运用数据值进行分析预测。

综上所述,对于耕地质量评价空间性数据,本研究提出的四象最近邻填充算法在精度上相对突出并稳定,整体上优于其他方法。

2.4 非空间性数据填充精度评价

在耕地质量评价数据中,非空间性数据包括全氮、粉砂粒等7组因素,对该数据类型采取非空间性填充方法进行数据缺失填充,采用回归模型填充、多重填充、期望最大化法填充、相似聚集填充,对非空间性缺失数据进行不同缺失率下的精度计算,结果见表5。

由表5可知:随着数据缺失率的提高,四种数据填充算法的精度都有所降低。而在这些算法中,相似聚集填充算法精度最高,在缺失率为5%~10%时,数据填充精度超过90%,主要原因是该方法集聚关联因素而避免了不同类型因素之间的相互影响。并且该算法在缺失率25%以下时,算法的精度均平稳下降,而期望最大化法填充、多重填充和回归模型填充在缺失率达到15%时精度降幅明显加快,而多重填充和回归模型填充算法在整体上的填充精度较低,即使在缺失率为5%时的精度也仅为80%左右,所以相似聚集填充算法比较稳定,且在缺失率较高时仍然能保持较好的填充精度。综上所述,相似聚集填充算法对本研究中耕地质量评价非空间关联性数据缺失填充具有优势,在精度上明显优于其他填充算法,集中表现了其精度高、稳定性强的特点。

3 结论

本研究以广州市从化区耕地质量评价数据为样本数据集,采用多种数据缺失填充方法进行分析,对数据进行空间相关性分析,并对缺失数据进行填补,结论如下:

(1)从化区耕地质量评价数据基本服从正态分布,异常数据较少,32组数据中有25组具有空间自相关性。

(2)对空间关联性数据填充精度最高的方法是四象最近邻算法,在缺失率20%以下时精度均高达

表5 不同缺失率下非空间性数据各填充方法的填充精度比较(%)

Table 5 Comparison of filling accuracy of different filling methods for non-spatial data with different missing rate(%)

填充方法 Filling method	缺失率阈值 Missing rate threshold	5.0%	10.0%	15.0%	20.0%	25.0%
相似聚集	20%	94.5	91.1	87.5	85.1	81.6
期望最大化法	15%	90.4	88.2	87.3	82.5	75.7
多重填充	15%	82.1	78.2	74.4	66.2	61.1
回归模型	15%	78.0	75.8	73.5	67.4	59.8

80%，精度随缺失率增大而降低，其次为KNN算法、期望最大化法、多重填充法、回归模型法。

（3）对非空间关联性数据填充精度最高的是相似聚集填充法，在缺失率25%以下时可保持80%以上的高精度，其次为期望最大化法、多重填充法、回归模型法。

（4）本研究提出的四象最近邻算法和相似聚集填充算法不仅在相同缺失率情况下精度更高，同时缺失率阈值范围更广，说明其方法的实用性更强。

综上，本研究提出的四象最近邻填充方法和相似聚集填充方法对耕地质量评价缺失数据填充的精度较其他方法有较大提升，并且更加适用于耕地领域。下一步将进行不同研究区的验证研究，来证实本研究提出方法的实用性和可靠性。

参考文献：

[1] 沈仁芳, 陈美军, 孔祥斌, 等. 耕地质量的概念和评价与管理对策[J]. 土壤学报, 2012, 49(6): 1210-1217. SHEN R F, CHEN M J, KONG X B, et al. Conception and evaluation of quality of arable land and strategies for its management[J]. *Acta Pedologica Sinica*, 2012, 49(6): 1210-1217.

[2] 成升魁, 李云云, 刘晓洁, 等. 关于新时代我国粮食安全观的思考[J]. 自然资源学报, 2018, 33(6): 911-926. CHENG S K, LI Y Y, LIU X J, et al. Thoughts on food security in China in the new period[J]. *Journal of Natural Resources*, 2018, 33(6): 911-926.

[3] WANG Z, WANG L M, XU R N, et al. GIS and RS based assessment of cultivated land quality of Shandong Province[J]. *Procedia Environment Sciences*, 2012, 12: 823-830.

[4] 刘思谦. 不完全数据填充算法的研究与应用[D]. 大连: 大连理工大学, 2017. LIU S Q. Research and application of incomplete data imputation algorithm[D]. Dalian: Dalian University of Technology, 2017.

[5] 李新, 程国栋, 卢玲. 空间内插方法比较[J]. 地球科学进展, 2000, 15(3): 260-265. LI X, CHENG G D, LU L. Comparison of spatial interpolation methods[J]. *Advances in Earth Science*, 2000, 15(3): 260-265.

[6] 朱求安, 张万昌, 余钧辉. 基于GIS的空间插值方法研究[J]. 江西师范大学学报(自然科学版), 2004, 28(2): 183-188. ZHU Q A, ZHANG W C, YU J H. The spatial interpolations in GIS[J]. *Journal of Jiangxi Normal University (Natural Sciences Edition)*, 2004, 28(2): 183-188.

[7] 林忠辉, 莫兴国, 李宏轩, 等. 中国陆地区域气象要素的空间插值[J]. 地理学报, 2002, 57(1): 47-56. LIN Z H, MO X G, LI H X, et al. Comparison of three spatial interpolation methods for climate variables in China[J]. *Acta Geographica Sinica*, 2002, 57(1): 47-56.

[8] TROYANSKAYA O, CANTOR M, SHERLOCK G, et al. Missing value estimation methods for DNA microarrays[J]. *Bioinformatics*, 2001, 17(6): 520-525.

[9] 严远亭, 吴亚亚, 赵姝, 等. 构造性覆盖下不完整数据修正填充方法[J]. 智能系统学报, 2019, 14(6): 1225-1232. YAN Y T, WU Y Y, ZHAO S, et al. Improving missing data recovery with a constructive covering algorithm[J]. *CAAI Transactions on Intelligent Systems*, 2019, 14(6): 1225-1232.

[10] 花琳琳. 不同缺失值处理技术的模拟比较[D]. 郑州: 郑州大学, 2012. HUA L L. Simulated comparison of different filling methods in missing values[D]. Zhengzhou: Zhengzhou University, 2012.

[11] 谢花林, 李波. 基于Logistic回归模型的农牧交错区土地利用变化驱动力分析——以内蒙古翁牛特旗为例[J]. 地理研究, 2008, 27(2): 294-304. XIE H L, LI B. Driving forces analysis of land-use pattern changes based on logistic regression model in the farming-pastoral zone: A case study of Ongiud Banner, Inner Mongolia[J]. *Geographical Research*, 2008, 27(2): 294-304.

[12] 邓银燕. 缺失数据的填充方法研究及实证分析[D]. 西安: 西北大学, 2010. DENG Y Y. Study on the filling method of missing data and empirical analysis[D]. Xi′an: Northwest University, 2010.

[13] HANSEN B E. Regression kink with an unknown threshold[J]. *Journal of Business & Economic Statistics*, 2017, 35(2): 228-240.

[14] 刘菲, 李明阳, 刘雅楠, 等. 森林资源抽样调查缺失数据填充方法[J]. 林业资源管理, 2018(6): 130-137. LIU F, LI M Y, LIU Y N, et al. Filling method for missing data of forest resource sampling investigation[J]. *Forest Resources Management*, 2018(6): 130-137.

[15] 申宁宁, 房瑞玲, 高宇钊, 等. 纵向研究缺失数据多重填补及混合效应模型分析[J]. 中国药物与临床, 2015, 15(7): 901-905. SHEN N N, FANG R L, GAO Y Z, et al. Using multiple imputation and mixed-effects model on missing data: A longitudinal study[J]. *Chinese Remedies and Clinics*, 2015, 15(7): 901-905.

[16] 万义良. 空间数据质量检查与评估理论研究[D]. 武汉: 武汉大学, 2015. WAN Y L. Research on the theory for spatial data quality inspection and assessment[D]. Wuhan: Wuhan University, 2015.

[17] 邱小倩, 胡月明, 朱阿兴, 等. 基于关联规则的耕地质量评价数据检错方法研究——以广州市为例[J]. 中国土地科学, 2020, 34(3): 75-83. QIU X Q, HU Y M, ZHU A X, et al. Research on associated rule-based error checking method on assessment index database of cultivated land quality: A case study on Guangzhou City[J]. *China Land Science*, 2020, 34(3): 75-83.

[18] 林子聪, 任向宁, 朱阿兴, 等. 基于随机森林算法的耕地质量定级指标体系研究[J]. 华南农业大学学报, 2020, 41(4): 38-48. LIN Z C, REN X N, ZHU A X, et al. Research on the index system of cultivated land quality grading based on random forest algorithm[J]. *Journal of South China Agricultural University*, 2020, 41(4): 38-48.

[19] 邱炳文, 王钦敏, 陈崇成, 等. 福建省土地利用多尺度空间自相关分析[J]. 自然资源学报, 2007, 22(2): 311-320. QIU B W, WANG Q M, CHEN C C, et al. Spatial autocorrelation analysis of multi-scale land use in Fujian Province[J]. *Journal of Natural Resources*, 2007, 22(2): 311-320.

[20] 赵地, 李光强, 李晶晶. 空间不完备数据及其填补方法研究[J]. 西部探矿工程, 2009, 21(1): 137-140. ZHAO D, LI G Q, LI J J. Incomplete data of space and the resume methods on these data[J]. *China Exploration Engineering*, 2009, 21(1): 137-140.

[21] 邱英, 冯春雨, 谢锋云, 等. 基于K邻近算法的转向架构架状态识别研究[J]. 测控技术, 2019, 38(8): 48-53. QIU Y, FENG C Y,

XIE F Y, et al. State recognition of bogie frame based on K-nearest neighbor algorithm[J]. *Measurement and Control Technology*, 2019, 38(8):48-53.

[22] 黄樑昌. KNN填充算法的分析和改进研究[D]. 桂林:广西师范大学, 2010. HUANG L C. The analysis and improvement research of KNN-imputation algorithm[D]. Guilin: Guangxi Normal University, 2010.

[23] 赵业婷. 基于GIS的陕西省关中地区耕地土壤养分空间特征及其变化研究[D]. 杨凌:西北农林科技大学, 2015. ZHAO Y T. Spatial characteristics and changes of soil nutrients in cultivated land of Guanzhong region in Shaanxi Province based on GIS[D]. Yangling: Northwest A&F University, 2015.

[24] 张灏, 王娇, 郑新奇. 针对地质云钻孔数据的空间插值方法选择[J]. 矿山测量, 2020, 48(3):12-16. ZHANG H, WANG J, ZHENG X Q. Selection of spatial interpolation method for geological cloud drilling data[J]. *Mine Surveying*, 2020, 48(3):12-16.

[25] 冷泳林, 陈志奎, 张清辰, 等. 不完整大数据的分布式聚类填充算法[J]. 计算机工程, 2015, 41(5):19-25. LENG Y L, CHEN Z K, ZHANG Q C, et al. Distributed clustering and filling algorithm of incomplete big data[J]. *Computer Engineering*, 2015, 41(5):19-25.

[26] 胡克林, 张凤荣, 吕贻忠, 等. 北京市大兴区土壤重金属含量的空间分布特征[J]. 环境科学学报, 2004, 24(3):463-468. HU K L, ZHANG F R, LÜ Y Z, et al. Spatial distribution of concentrations of soil heavy metals in Daxing County, Beijing[J]. *Acta Scientiae Circumstantiae*, 2004, 24(3):463-468.