基于聚类和分类与回归树的地力等级评价研究①

闫一凡[1,2]，刘建立[1*]，李晓鹏[1]， 张佳宝[1]，赵炳梓[1]

(1 中国科学院南京土壤研究所，南京 210008；2 中国科学院大学，北京 100049)

摘 要：以黄淮海平原粮食主产区河南封丘县为研究区域，利用基于 GIS 的土壤空间和属性数据库，采用聚类分析和分类与回归树(CART)相结合的方法建立了耕地地力评价模型。研究结果表明，基于聚类分析和 CART 的地力评价模型准确度为 93.56%，较单独使用决策树模型的准确度有明显提高；根据耕地地力分级规则，一等地至五等地分别占全县 61 733.3 hm^2 耕地的 28.167%、49.518%、9.389%、5.77% 和 7.156%；地力等级较高的耕地主要分布于封丘西北部，地力较低的区域主要在东南部，由西北向东南地力呈带状递减趋势。本文的研究结果可为当地中低产田及其障碍因子的解析和农田精准管理提供参考依据。

关键词：耕地地力；评价；分类与回归树(CART)；聚类分析；模型

中图分类号：S158.2

耕地地力是指在当前管理水平下由耕地土壤本身特性、自然背景条件和基础设施水平等要素综合构成的耕地的生产能力。耕地地力的高低直接决定了粮食的产量和质量，对其进行全面、客观评价以获得地力的空间分布，是解析中低产田主要障碍因子、开展地力定向培育的前提，对我国科技增粮计划的顺利实施具有重要意义。

目前，国内耕地地力评价仍处于起步阶段，常用的评价方法有经验判断指数和法、层次分析法、模糊综合评价法等[1–2]。已有研究表明，层次分析法需要结合专家打分法(即特尔斐法)，往往会导致评价指标权重系数的主观性太强[3]。而且，由于土壤因素对作物生长的限制作用是渐变的，模糊数学中采用的隶属度函数及其分类也具有一定的局限性[4]；此外，模糊综合评价中可能会丢失许多有用信息，参评因素越多误判的可能性越大。而经验判断指数和法并不适用于有限制因子参与的耕地地力评价[5]。为解决上述问题，不少研究者对地力评价方法作了新的尝试和探索，如王瑞燕等[6]、孔维娜等[7]先后采用人工神经网络(ANN)–产量定量评价模型，即以产量为目标输出，通过人工神经网络训练得到评价模型。由于评价过程中不需确定各因素的权重，消除了前述方法确定权重时的人为影响，增强了评价结果的客观性。段兴武等[8]利用改进后的 PI 模型对东北松嫩黑土区土壤生产力进行评价，该方法所需指标较少，易于操作。孙微微等[9]将决策树模型中的 C4.5 算法应用于广东省的土壤质量等级研究，准确率达到了 96.61%。决策树模型具有直观明了、稳健性高、结果易于理解等优点。与农业部推荐的层次分析与模糊数学相结合的方法相比，决策树模型不仅避免了特尔菲法中的人为因素的影响，而且当评价属性集合发生变化时也无需专家重新确定属性权重的繁琐过程，在地力评价中显示出良好的应用前景。

本文选择位于黄淮海平原粮食主产区的河南封丘县为研究区域，利用前期工作中建立的基于 GIS 的县域土壤空间数据库和属性数据库，采用聚类分析和分类与回归树(classification and regression tree，CART)相结合的方法建立了耕地地力评价模型，评价结果可望为当地农业管理和决策提供参考依据。

1 研究区概况

封丘县位于河南省东北部的新乡市(114°14′ ~ 114°46′E，34°53′ ~ 35°14′N)，海拔 65 ~ 72.5 m。该县东临长垣县，南隔黄河与开封市相望，西靠延津县，北接安阳市滑县。县境南北长 38.2 km，东西宽 48.7 km。面积 1 220.5 km^2，耕地面积 61 733.3 hm^2。主要土壤类型为黄河沉积物发育的潮土，并伴有部分盐土、碱土、沙土和沼泽土的插花分布。封丘县属暖温带大

①基金项目：973 计划课题项目(2011CB100506)和国家自然科学基金项目(41171179、41001127)资助。

* 通讯作者(jlliu@issas.ac.cn)

作者简介：闫一凡(1989—)，女，河南新乡人，博士研究生，主要从事数字农业、土壤水文模型研究。E-mail: yfyan@issas.ac.cn

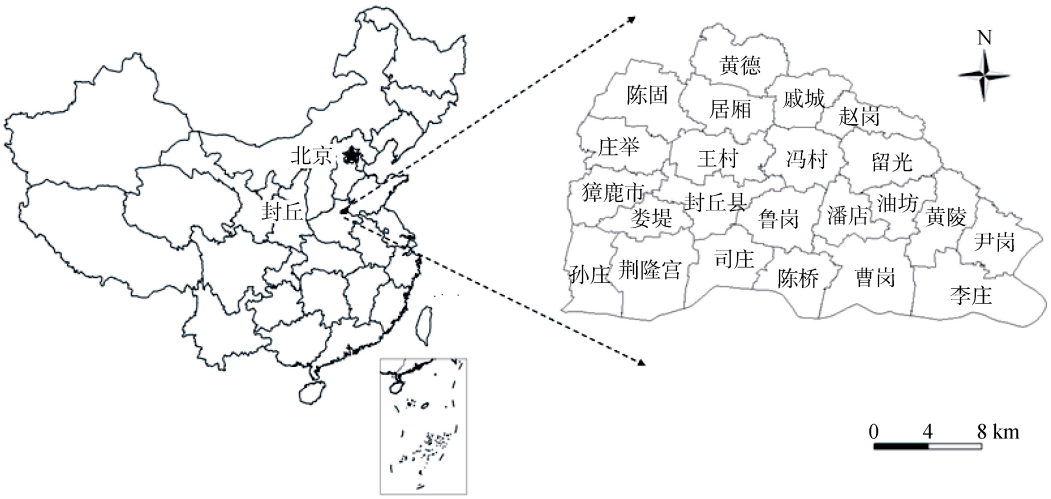

图 1　研究区行政区划图
Fig. 1　Administrative map of researched area

陆性季风气候，降水时空分布不均，干湿季分明，且年际变化大。冬春少雨，土壤易返盐。黄河位于封丘南部边界，因河床高于地面，排水出路不畅。主要粮食作物为冬小麦和夏玉米，此外还盛产大豆、棉花、花生等，为一年两熟或两年三熟制地区。

2　材料与方法

2.1　数据来源

2.1.1　基础图件及专题图资料　　本研究收集了研究区土壤图、土地利用类型图、行政区划图、水系分布图等相关基础图件，利用 ArcGIS9.3 进行了数字化、图形编辑、图幅校正等处理。

2.1.2　土壤数据　　在全县范围内按照 2 km × 2 km 的均匀网格，于 2008—2012 年先后采集了 187 个点位的耕层土样(深度为 0～20 cm)，其中 113 个点同时包括 2008 年和 2009 年作物年产量(冬小麦+夏玉米)数据。土样经风干、研磨和过筛处理后，在实验室内进行分析测试，项目包括：土壤有机质含量(SOM)(重铬酸钾氧化–外加热法测定)、pH、阳离子交换量(CEC)(乙酸铵法测定)、电导率(水土比 5∶1，电导法测定)、全氮(半微量开氏法测定)、全磷(钼锑抗比色法测定)、全钾(氢氧化钠熔融火焰光度法测定)、有效磷(碳酸氢钠法测定)、速效钾(醋酸铵浸提火焰光度法测定)等。

2.2　研究方法

2.2.1　评价单元划分　　划分评价单元是进行地力评价的基础。评价单元选取是否合理，直接关系到评价工作量和结果的准确性。本文根据《全国耕地地力调查与质量评价技术规程》[10]要求，在 GIS 平台支持下对从 1∶50 000 土地利用现状图中提取出地类号为 111、113、114、115 的耕地图斑和 1∶50 000 土地利用类型图进行数字化叠加和综合取舍，形成评价单元图斑。全县共划分为 3 800 个评价单元，同一单元内土壤类型和土地利用类型均相同。

图 2　评价单元划分图
Fig. 2　Assessment units map

2.2.2　评价指标的选取　　我国农业部曾于 2000 年组织专家根据气候和地貌特征等，用穷尽法建立了一套供全国地力评价公用的指标体系。该指标体系中包含了气候、立地条件、剖面性状、耕层理化性质、土壤养分状况、障碍因素、土壤管理 7 类共 64 项指标[11]。由于不同区域的气候、土壤母质、质地、坡面性状、理化性状等各方面都存在着巨大的差异，在实际工作中并不一定将全部指标都纳入考虑，而是根据经验和因地制宜的原则加以筛选。筛选中一般需遵循主导因素原则、差异性原则、稳定性原则、敏感性原则[5]。根据以上原则并结合实际，最终选取土壤有机质、pH、阳离子交换量、电导率、全氮、全磷、全钾、有效磷、速效钾、灌溉保证率这 10 个指标作为封丘县地力评价的指标。

2.2.3　评价单元赋值　　本文采用克里格法进行评

价单元的属性赋值。由于克里格插值对要求数据呈现严格的正态分布，首先利用 SPSS 对数据进行检验和转换，然后将转换后的数据导入 GS$^+$，进行地统计学插值。相关变量的插值结果见表 1。

表 1 各变量模型拟合参数
Table 1 Parameters fitting of variables

变量	变异函数模型	块金值	基台值	变程	块金系数	R^2
年产量	指数模型	2 000	70 630	0.063	0.028	0.798
SOM	球状模型	10.120	22.190	0.158	0.456	0.833
pH	指数模型	0.080	0.951	0.033	0.084	0.714
CEC	高斯模型	0.068	0.139	0.059	0.489	0.777
电导率	球状模型	794	1 631	0.065	0.487	0.626
全氮	指数模型	0.022	0.045	0.168	0.499	0.744
全磷	球状模型	0.048	0.896	0.024	0.054	0.480
全钾	球状模型	0.037	0.107	0.137	0.346	0.878
有效磷	球状模型	0.021	0.955	0.017	0.022	0.218
速效钾	球状模型	0.096	0.194	0.183	0.497	0.888

按照区域化空间变量相关性分级标准，当块金系数≤25%、25% ~ 75% 和≥75% 时，分别表示变量的空间自相关性为强烈、中等及微弱[12]。当变量空间自相关程度为微弱时，其变异主要由随机变异组成，不适合采用插值方法进行空间赋值。由表 1 可知，产量、pH、全磷和有效磷具有强烈的空间相关性，SOM、CEC、电导率、全氮、全钾、速效钾具有中等的空间自相关，可进行空间插值。根据《全国耕地地力调查与质量评价技术规程》，对上述 10 项评价指标先插值形成栅格化的点位数据，再与评价单元图叠加后赋值给各单元，得到各土壤属性专题图。灌溉保证率的获取参见王良杰等[2]的方法，以耕地距离河流、灌渠的空间距离进行换算(数据未列出)。

2.2.4 分类与回归树(CART)模型 分类与回归树(CART)是 Brieman 等[13] 1984 年提出的数据挖掘工具，属于决策树模型算法的一种。CART 采用二分递归分割技术将观测集(训练样本)进行分类，以使得子集(新样本)达到最大的同质性(即纯度)。本研究采用 Gini 系数作为杂质度量指标。CART的基本原理如下：

若采用 k (k = 1, 2,…,C)来表示类，其中 C 为类的总数目，则矩形 A 的 Gini 不纯度定义为：

$$I(A)=1-\sum_{k=1}^{c}P_k^2 \quad (1)$$

式中：P_k是观测点中属于类 k 的比例。当所有观测点都属于同一类时，当所有类在 A 中以相同概率出现时，$I(A)$最大化为$(C-1)/C$。树节点的分裂准则为：对每个属性均遍历所有可行的二分后，选择使 Gini 系数达到最小的分割方法，作为此节点的分裂标准[14]。

除 Gini 系数外，CART 模型的另一个关键概念是利用独立测试集对训练生长树型的剪枝，亦即认为一个很大的树会过度拟合训练数据，从而导致其仅捕捉到训练集的噪声，而不能反映在将来数据集中可能发生的模式。剪枝过程需要在验证数据集的误分与被剪枝树决策点数目间进行权衡和折中，以得到既可反映数据模式又能排除训练数据噪声的树型。

根据本文的研究目的，将独立的离散型属性变量耕地地力等级按作物年产量(2008 和 2009 年)遵循等差原则分为一等(16 897 ~ 18 795 kg/hm^2)、二等(15 000 ~ 16 897 kg/hm^2)、三等(13 100 ~ 15 000 kg/hm^2)、四等(11 203 ~ 13 100 kg/hm^2)和五等(9 315 ~ 11 203 kg/hm^2)，并作为 CART 模型的输出。表 1 中除年产量外的其他 9 个指标以及灌溉保证率作为输入属性。将数据集按 3∶1 的比例划分为训练集和检验集，在 Clementine 12.0 软件[15]中实现建模。

3 结果与分析

3.1 CART 模型的预测效果

为了避免树的过度生长，采用预先剪枝。由于本文数据集记录达 3 800 条，因此将 CART 模型的最大树深度定为 6，分类节点包含数据点少于 20 即可停止划分。剪枝树选择1标准差规则，若发现一颗子树的最小估计误判成本在 1 个标准差内，即停止检索最佳剪枝树，此时的树为最优树[16]。验证选择交叉验证(K = 10)。

为提高模型的学习准确率，本次还引入错分代价，即在进行质量等级的识别时，设置学习发生错误时所花费的代价，预测等级与实际等级差别越大，代价越大。具体划分标准如表 2 所示。

表 2　错分代价表
Table 2　Misclassification cost

实际等级	预测等级				
	一等地	二等地	三等地	四等地	五等地
一等地	0	0.25	0.50	0.75	1.00
二等地	0.25	0	0.25	0.50	0.75
三等地	0.50	0.25	0	0.25	0.50
四等地	0.75	0.50	0.25	0	0.25
五等地	1.00	0.75	0.50	0.25	0

模型运算得到的训练集准确率为 82.97%，测试集的预测准确率为 78.85%，其中二等地的误判较多，在表现评价(performance evaluation)中也可以清楚地看到二等地的分值最低(表 3、表 4)。

3.2　聚类分析与 CART 的结合

由初步建模结果可知模型的预测准确率不高，说明该模型的泛化能力不强，无法完全反映封丘县的地力等级分布情况。为此，我们采用对数据先聚类再建模的方式，以改善模型的预测效果。由于数据量较大，

表 3　符合矩阵
Table 3　Coincidence matrix

测试集	1	2	3	4	5	训练集	1	2	3	4	5
1	190	39	16	3	4	1	504	101	24	3	3
2	13	335	14	2	6	2	41	1076	32	1	20
3	8	28	83	8	24	3	14	73	210	27	54
4	2	7	10	76	18	4	0	11	13	264	43
5	1	0	2	5	99	5	0	2	4	12	275

表 4　表现评价
Table 4　Performance evaluation

地力分等	测试集	训练集
1	1.252	1.383
2	0.788	0.715
3	1.474	1.707
4	1.961	1.987
5	1.806	1.898

本文采用 K-Means 聚类方法获得 K 个聚类，并在每个聚类中心附近抽取 6 条记录构成新的训练样本，然后建立分类回归树模型并检验其合理性。随着 K 值和抽取样本个数的增加，模型预测准度的变化如图 3 所示。

由图 3 可见，当聚类数为 50、抽样样本数为 300 时 CART 模型的预测准确率最高，为 93.56%。预测准确率第二次达到峰值是在聚类数为 80、样本

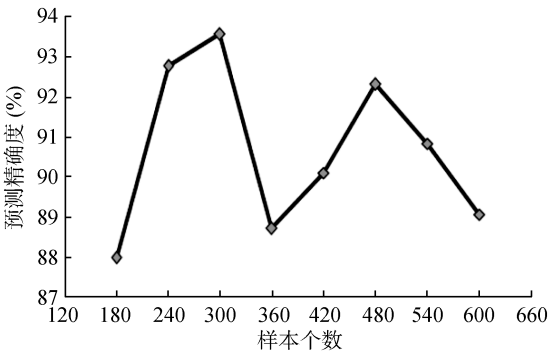

图 3　不同样本个数的预测精度
Fig. 3　Prediction accuracy of different numbers of samples

数为 480 时，此时的预测准确率为 92.23%，略低于前者，且此时抽取样本较多。可以发现，先聚类再 CART 建模的结果准确率明显高于直接利用 CART 建模的准确率(86.08%)，说明建模前利用 K-means 聚类抽样有助于改善模型的预测能力。本文通过抽取 300 个样本来组建训练样本并建立 CART 模型，得到的地力分级结果如图 4 所示。建模中各评价指标的权重如表 5。

由表 5 可知，10 个参评指标中的灌溉保证率及速效磷没有出现，这是由于 CART 的优点之一是在构建最优树的过程中会自动选择最优化分的变量，而权重极小的变量则被剔除。这两个指标没有权重输出说明其权重极小，对评价结果影响极小，可以忽略。反之，耕层土壤的电导率和全磷含量权重最大，对封丘地力等级的影响较为显著。由图 4 可以更直观地看出，电导率 >91.43 μS/cm 时地力等级较低，为三等至五等地，且以四等地为主。一等和二等地均分布在电导率 <91.43 μS/cm 的分枝中，且以二等地为主。在电导率<91.43 μS/cm 分枝中，全磷含量<0.665 g/kg 的分枝地力等级较低，主要集中于三、四、五等地，且五等地比例最高；在全磷含量>0.665 g/kg 的分枝地力等级较高，主要集中于一、二等地，且二等地比例最高。其他层次的分等规则也可按上述方法确定。图 4 中各参评指标单位：电导率(μS/cm)，全磷(g/kg)，阳离子交换量(mmol/kg)，全氮(g/kg)，全钾(g/kg)，有机质(g/kg)，速效钾(mg/kg)。

3.3　封丘县域耕地地力空间分布

采用前文建立的分类与回归树模型，对封丘县耕

图 4 地力等级评价结果
Fig. 4 Result of productivity grading

表 5 各参评指标的权重
Table 5 Weights of indexes

电导率	全磷	阳离子交换量	全氮	全钾	有机质	速效钾	pH
0.372	0.215	0.111	0.099	0.068	0.064	0.056	0.001 5

地地力的空间分布情况作了评价，结果表明：全县 61 733.3 hm^2 耕地中，一等地为 17 386.7 hm^2(占总面积的 28.17%)，二等地 30 573.3 hm^2(49.52%)，三等地 5 793.3 hm^2 (9.39%)，四等地 3 560 hm^2(5.77%)，五等地 4 420 hm^2 (7.16%)。由此可知，一等地和二等地合计占全县耕地的 77.69%，而地力较低的四等地、五等地合计占 12.93%，说明封丘县总体地力水平较高，但仍有近 8 000 hm^2 的中低产田需要开展地力培育。

由封丘县地力等级空间分布图(图 5)可知，地力等级较高的耕地主要分布于封丘县西北部，地力较低的区域则主要集中于东南部，且由西北向东南方向地力水平呈带状递减趋势，这一结果与实地调查和测产的数据是一致的。导致东南部地力较差的原因主要在于，黄河流经封丘县东南边界，此区域的土质较砂，土体结构不良，黏粒及有机质含量较低，保水保肥性差。此外，通过分类准则也可知道该区域的耕层电导率普遍较高，因此存在次生盐渍化的可能。

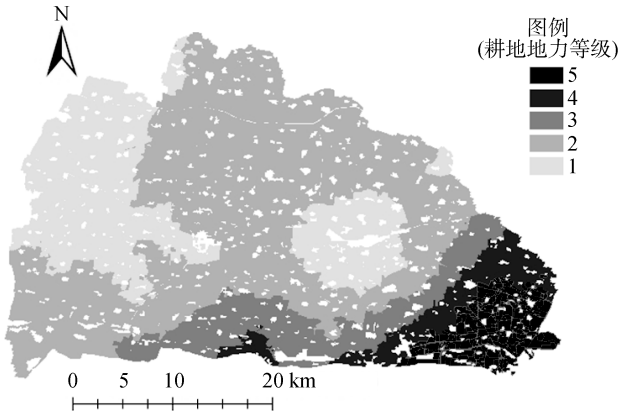

图 5 耕地地力等级空间分布图
Fig. 5 Spatial distribution of farmland productivity grading

4 结论

(1) 与单纯采用 CART 建模相比，采用聚类分析抽取训练样本然后与 CART 结合建模，既可减少学习样本空间、节省建模时间，又可有效提高模型预测精度，显示出较好的应用前景。

(2) 当聚类数为 50、抽样样本数为 300 时，CART 模型的预测准确率最高(达到 93.56%)。据此建立了封丘县域耕地地力的分级规则，结果表明：一等地至五等地分别占全县 61 733.3 hm^2 耕地的28.167%、49.518%、9.389%、5.77% 和 7.156%。

(3) 地力等级较高的耕地主要分布于封丘县西北部，地力较低的区域则主要集中于东南部，且由西北向东南方向地力水平呈带状递减趋势。

(4) 由于数字化的地图多是二次普查时期的，而补充采样是近几年的，两者之间存在偏差，这也是影响预测结果准确度的重要原因之一。

参考文献：

[1] 周艺红, 熊东红, 杨忠, 何毓荣, 曾云英. 长江上游典型地区基于 SOTER 数据库的耕地地力评价[J]. 土壤通报, 2005, 36(2): 145–148

[2] 王良杰, 赵玉国, 郭敏, 张甘霖. 基于 GIS 与模糊数学的县级耕地地力质量评价研究[J]. 土壤, 2010, 42(1): 131–135

[3] Duan XW, Xie Y, Feng YJ, Yin SQ. Study on the method of soil productivity assessment in black soil region of Northeast, China[J]. Agricultural Sciences in China, 2009, 8(4): 472–481

[4] 魏善沛, 章景, 王凯. 粗糙集与 SVM 的组合算法在人工林地力评价中的应用[J]. 中南林业科技大学学报, 2013, 33(5): 1–5

[5] 吴鹏飞, 孙先明, 龚素华, 刘洪斌. 耕地地力评价可持续研究发展方向探讨[J]. 土壤, 2011, 43(6): 876–882

[6] 王瑞燕, 赵庚星, 陈丽丽. 基于 ANN-产量的耕地地力定量评价模型及其应用[J]. 农业工程学报, 2008, 24(1): 113–118

[7] 孔维娜，李跃进，李双异，裴久渤，汪景宽. 人工神经网络产量定量评价模型在县域耕地地力评价中的应用[J]. 国土与自然资源研究, 2012(2): 30–32
[8] 段兴武, 谢云, 张玉平, 刘冰. PI 模型在东北松嫩黑土区土壤生产力评价中的应用[J]. 中国农学通报, 2010, 26(8): 179–188
[9] 孙微微, 胡月明, 刘兴才, 薛月菊. 基于决策树的土壤质量等级研究[J]. 华南农业大学学报, 2005, 26(3): 108–110
[10] 中华人民共和国农业部. 耕地地力调查与质量评价技术规程(NY/T1634-2008)[S]. 北京: 中国标准出版局, 2008
[11] 田有国, 辛景树, 栗铁申. 耕地地力评价指南[M]. 北京: 中国农业科技出版社, 2006
[12] Combardella CA, Mooman TB, Novak JM. Field-scale variability of soil properties in central soils[J]. Soil Science Society of America Journal, 1994: 1 501–1 511
[13] Breiman L, Friedman J, Olshen RA, Stone C. Classification and regression trees[M]. Monterey: Wadsworth& Brooks/ Cole Advanced Books & Software, 1984: 232–234
[14] 李大峰, 罗林开, 岑涌. 基于 PCA 与分类回归树的疾病诊断应用研究[J]. 计算机与数字工程, 2007, 35(5): 184–188
[15] SPSS Inc. Clementine12.0 用户指南[DB/OL]. http://www.spss.com. 2007
[16] Feldman D. Mortgage default: Classification tree analysis[J]. The Journal of Real Estate Finance and Economics, 2005, 30(4): 369–396

Assessment of Farmland Productivity with Cluster Analysis and Classification and Regression Tree (CART)

YAN Yi-fan[1,2], LIU Jian-li[1*], LI Xiao-peng[1], ZHANG Jia-bao[1], ZHAO Bing-zi[1]

(1 *Institute of Soil Science*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China*; 2 *University of Chinese Academy of Sciences*, *Beijing* 100049)

Abstract: In the present study, a combination of classification and regression tree (CART) and cluster analysis method was applied in assessing the farmland productivity in Fengqiu County, Henan Province, based on county-level soil spatial and attributive databases. The results indicated that the prediction accuracy of the proposed combination model was considerably improved (to 93.56%) as compared to that by CART approach alone. According to the resulting grading rules, the first, second and third grade farmland accounts for 28.167%, 49.518% and 9.389% of the total area of farmland in this county, respectively; while the fourth and fifth grade farmland accounts for only 5.77% and 7.156% of the farmland, respectively. The higher grading land was mainly distributed in the northwest of Fengqiu County, while the lower grading land was mainly located in the southeast region. There was also an obvious banded decreasing trend of farmland productivity extending from the northwest to the southeast. The results of this paper may help to analyze the spatial distribution of the middle to low-yield fields and limiting factors for improving grain yield, and may also provide references for decision-making on regional farmland management.

Key words: Farmland productivity, Assessment, Classification and regression tree (CART), Cluster analysis, Model