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 Zicong, REN Xiangning, ZHU Axing, 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.

基于随机森林算法的耕地质量定级指标体系研究

林子聪[1]，任向宁[1]，朱阿兴[1,2]，赵　鑫[1]，胡月明[1,3,4]
(1 华南农业大学 资源环境学院/华南农业大学 地理信息工程研究所/国土资源部建设用地再开发重点实验室/广东省土地利用与整治重点实验室，广东 广州 501642； 2 威斯康星大学−麦迪逊分校 地理系，威斯康星州 麦迪逊 53706；
3 青海大学 农牧学院，青海 西宁 810016； 4 电子科技大学 资源与环境学院，四川 成都 610054)

摘要:【目的】分析研究区域内的耕地质量差异，优化耕地利用与布局，为耕地保护提供参考依据。【方法】以青海省共和县、都兰县和乌兰县的耕地为研究对象，根据历史及现有文献收集耕地质量的影响因素，采用随机森林算法和相关性分析筛选定级指标并确认权重，通过加权求和法计算定级指数并划分级别，得到定级结果。与常用的特尔菲法定级成果进行比较分析。【结果】随机森林算法得到的变量重要性(*I*)范围在 0.03~11.94，相关性分析结果显示，大部分影响因素间相关性不显著，有 8 个为显著相关，综合 *I* 值和相关性分析结果将 30 个影响因素收敛为 4 个纬度下的 14 个定级指标，其中影响研究区域耕地质量的主要因素为生态系统脆弱性、生长季平均降水和年总太阳辐射量，权重分别为 0.11、0.10 和 0.09，随机森林算法评价结果与实际情况相符。【结论】与常用的特尔菲法比较，随机森林算法稳定性更好，级别指数变幅区间更小，更有利于构建省级空间尺度的耕地级别可比序列。

关键词: 耕地质量评价；定级指标体系；随机森林算法；特尔菲法
中图分类号: X53　　**文献标志码:** A　　**文章编号:** 1001-411X(2020)04-0038-11

Research on the index system of cultivated land quality grading based on random forest algorithm

LIN Zicong[1], REN Xiangning[1], ZHU Axing[1,2], ZHAO Xin[1], HU Yueming[1,3,4]
(1 College of Natural Resources and Environment, South China Agricultural University/Institute of Geographic Information Engineering, South China Agricultural University/Key Laboratory of Construction Land Improvement, Ministry of Land and Resources/Guangdong Provincial Key Laboratory of Land Use and Consolidation, Guangzhou 510642, China; 2 Department of Geography, University of Wisconsin-Madison, Madison 53706, USA;
3 College of Agriculture and Animal Husbandry, Qinghai University, Xining 810016, China; 4 School of Resources and Environment, University of Electronic Science and Technology of China, Chengdu 610054, China)

Abstract: 【Objective】To analyze the difference of cultivated land quality in the study region, optimize the use and layout of cultivated land, and provide a reference for cultivated land protection.【Method】Taking the cultivated land in Gonghe County, Dulan County and Wulan County in Qinghai Province as the research object, the influencing factors of cultivated land quality were collected based on the history and existing literature, and the random forest algorithm and correlation analysis were used to screen the grading indicators and confirm the weight. We calculated the grading index and divided the levels by weighted sum method to get the grading

收稿日期:2019–09–18　　**网络首发时间:**2020–06–15 14:28:59
网络首发地址:https://kns.cnki.net/kcms/detail/44.1110.S.20200615.1135.002.html
作者简介:林子聪 (1994—)，男，硕士研究生，E-mail: linzc306575684@163.com；通信作者：胡月明 (1964—)，男，教授，博士，E-mail: yueminghugis@163.com
基金项目:国家重点研发计划 (2016YFC0501801，2018YFD1100103)；青海省科技计划 (2017-ZJ-730)；广州市科技计划 (201804020034)

result. We compared the results with the grading results of commonly used Delphi method.【Result】The value of variable importance *I* obtained by random forest algorithm ranged from 0.03 to 11.94. Correlation analysis showed that the correlation between most influencing factors was not significant, eight of which were significant correlation. The 14 rating indicators under four dimensions were astringed from 30 influencing factors. The main factors influencing the quality of cultivated land in the study area were ecosystem vulnerability, mean precipitation of growing season and annual solar radiation amount, with the weights of 0.11, 0.10 and 0.09, respectively.【Conclusion】Compared with Delphi method，the random forest algorithm has better stability and smaller level of index variation interval, which is more conducive to construct a comparable sequence of cultivated land levels at provincial spatial scale.

Key words: cultivated land quality evaluation; grading indicators system; random forest algorithm; Delphi method

耕地是特殊的公共资源和最为宝贵的自然资源，是粮食生产的载体，也是保障社会安全及社会可持续发展的物质基础[1]。目前，我国经济发展进入新常态，新型工业化、城镇化建设深入推进，耕地占多补少，占优补劣现象突出，优质耕地正大量流失；同时，由于土壤污染和不合理高强度利用，导致耕地退化，较大地影响了农业的生产力[2-4]。

面对严峻的耕地形势，为摸清耕地资源状况，及时准确地掌握耕地质量现状，把握耕地质量动态，我国开展了耕地质量定级评价工作，即对耕地进行综合评定并划分级别，从而反映因耕地自然质量、现实 (或实际可能的) 利用水平和效益水平的不同所造成的耕地生产力水平的差异[5]。耕地质量评价工作是保护耕地的主要途径，其有利于分析区域内耕地质量差异，能够优化耕地利用与布局，能够为耕地保护提供参考依据，对耕地的占补平衡、占优补优工作提供指导。

自 20 世纪 80 年代末开始，我国许多专家学者对耕地定级的方法不断探究，高中贵等[6] 认为：定级指标体系是其定级结果准确可靠的基础，但在实际操作中存在较多困难；金东海等[7] 通过对以分等成果为基础的“两层七参数法”的定级新方法研究后认为：定级指标的筛选、权重的赋值是定级方法的关键。

以往传统的定级指标筛选和赋权中，通常采用专家经验法，该方法主要通过专家根据经验对各个定级指标进行打分，累加各定级指标的分值并根据分值大小确认权重[8]。该方法完全根据专家组的经验判断，主观性较强，其选定的定级指标和确认的权重缺乏定量地分析，定级指标之间存在层级关系，相关性较强，导致定级结果不够客观准确。随着科学技术方法的发展，应用耕地产量[9] 于数学统计与数学模型的客观定量法能够定量测算定级指标的权重，同时可以避免定级指标之间的相关性，降低定级结果的多重共线性问题。因此，客观定量法已经成为必然趋势[10]。随机森林算法基于分类树构建模型，利用分类树与实际值计算误差得到变量重要性，属于客观定量法。

本文以青海省共和县、都兰县及乌兰县为研究区域，应用随机森林算法，建立耕地产量与自然、社会经济、区位等影响因素的有机联系，同时结合因素的相关性分析结果，确定合理的定级指标及定级指标权重，并与传统方法结果进行比较，为耕地定级工作提供科学依据。

1　材料与方法

1.1　研究区域概况

研究区域位于青海省的共和县、都兰县和乌兰县，是国家生态保护与建设的战略要地，是国家乃至全球重要的水源地和生态屏障，是高原生物多样性基因资源的宝库。另外，青海省生态系统较为脆弱，水土流失、荒漠化、沙化面积扩大，湿地萎缩，草场退化等问题突出，而生态环境的自我恢复能力、净化能力越低的地区，越容易影响到耕地的产出水平。

研究区域地处柴达木盆地东部，山脉、盆地、平原各地貌交错，地势北低南高，平均海拔 3 500 m，海拔差最大达 3 076 m。位于大陆腹地，远离海洋，具有典型的高寒大陆性气候特点：年降水少、时空分布不均，蒸发强烈、干燥度大；温度低、温差大、日照长、辐射强；冬春季多大风，灾害性天气频繁。研究区域 2017 年辖 3 县 12 镇 11 乡 242 个行政村，土地总面积 74 187.94 km^2，耕地面积 49 739.29 hm^2；总人口 26.27 万，其中农业人口 16.87 万，集中分布在研究区域内气候较适宜的平原及盆地地区。耕地的主要作物为小麦、青稞及油菜。

根据研究区域耕地的空间分布特征，在研究区域共布设 97 个采样点，通过实地问卷调查，收集到 2015—2017 年耕地产量及对应分布的基础数据。

表 1 研究区域调查样点产量数据

Table 1 Yield information of investigation plot in research area

耕地类型 Type of cultivated land	耕地数量/块 Amount of cultivated land	耕地密度/(块·hm^{-2}) Density of cultivated land	产量/(kg·hm^{-2}) Yield		
			最大值 Max.	最小值 Min.	平均值 Mean
旱地 Rainfed cropland	20	306.26	3 840	1 980	2 671.50
水浇地 Irrigable cropland	77	556.42	9 375	2 100	5 296.35
总计 Total	97	512.78	9 375	1 980	4 755.15

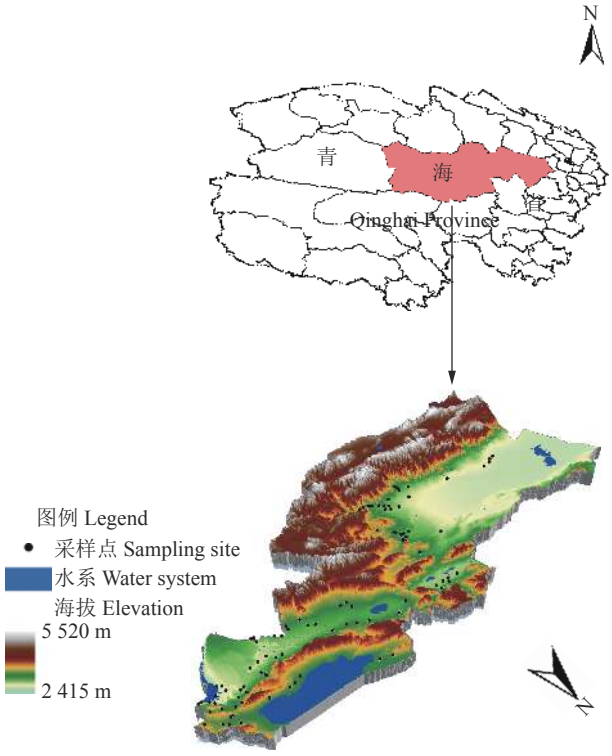


图 1 研究区位置及采样点分布图

Fig. 1 Location of the research area and spatial distribution of the sampling sites

对 3 年的产量数据取平均值，得到各调查样点的年均产量数据，其数据特征及分布如表 1 与图 1 所示。

1.2 数据来源

目前，各种类型耕地质量评价所构建的评价指标体系都以气候因素、地形自然条件因素和土壤物理化学性状因素等为主。根据耕地质量的概念和内涵，影响耕地质量的因子可分为自然因素和社会经济因素两类[11-13]。此外还有一些研究表明，随着高标准农田建设的开展，耕地工程要素对农业生产的便捷性、经济性具有重要影响[14-15]。

因此，本文收集的影响因素分为以下 5 类：1) 自然因素：气候数据，包括温度、降水、太阳辐射等；基础地理数据，主要包括海拔、地形坡度；土壤理化数据；水资源状况数据。2) 生态因素：指研究区域生态系统的脆弱程度，自然灾害等易发程度。3) 社会经济因素：指基础设施、耕作便利程度、土地利用方面数据。4) 区位因素：指城镇、农贸市场、道路等对耕地的影响。5) 工程因素：指高标准农田建设中的土地整治工程。

气候数据是依据马昊翔等[16]制作的青海省生长季均温图、青海省生长季降水图，保广裕等[17]制作的青海省年太阳总辐射量空间分布图获得；基础地理数据是根据中国科学院计算机网络信息中心地理空间数据云平台 (http://www.Gscloud.cn) 获取的 30 m 分辨率数字高程图，在 Arcgis10.2 中提取海拔和坡向获得；依据 2017 年共和县、都兰县和乌兰县耕地质量定级成果，获取有效土层厚度、表层土壤质地、砾石含量、土壤有机质含量、土壤酸碱度、灌溉保证率、灌溉水质量、田块形状、田块大小以及利用现状等耕地质量定级指标值；生态因素依据的是全国主体功能区规划[18]中的自然灾害危险性评价图、生态脆弱性评价图及青海省国土局制作的青海省水土流失状况评价图；依据 2016 年共和县、都兰县和乌兰县的土地利用现状数据库，获取道路、城镇、农贸市场、耕地等分布数据，在 Arcgis10.2 中通过与耕地进行空间分析，获取耕作距离、农田破碎度、农田路网密度、城镇影响度、农贸市场影响度、道路通达度、对外交通便利度等状况数据；工程因素是对 2012—2016 年共和县，都兰县和乌兰县的高标准基本农田整理项目资料进行数据化，得到灌溉排水工程、道路工程、农田防护林工程、土地平整工程的分布及工程建设状况。

影响因素数据类别及描述如表 2 所示。

1.3 分析方法

1.3.1 随机森林算法 随机森林是由 Breiman 等[19]在 2001 年提出的一种基于分类树的算法，具有防止过拟合、模型稳定性强以及易于处理非线性回归等特点，由于其良好的性能表现，在众多领域的问题解决中都取得了不错的效果[20-26]。

表 2　影响耕地质量的因素

Table 2　Factors impacting the quality of cultivated land

类别 Classification	影响因素 Impact factor	类别 Classification	影响因素 Impact factor
自然因素	生长季均温	社会经济因素	林网化程度
Natural factor	Mean temperature of growing season	Socioeconomic factor	Degree of forestation
	生长季降水量		耕作距离
	Mean precipitation of growing season		Cultivation distance
	年总太阳辐射量		农田路网密度
	Annual solar radiation amount		Farmland road network density
	海拔高度		田块形状
	Elevation		Field shape
	地形坡度		田块大小
	Topographic slope		Field size
	有效土层厚度		农田破碎度
	Effective soil thickness		Farmland fragmentation degree
	表层土壤质地		利用现状
	Surface soil texture		Utilization status
	砾石含量	区位因素	城镇影响度
	Gravel content	Location factor	Urban influence degree
	土壤有机质含量		农贸市场影响度
	Soil organic matter content		Agricultural market influence degree
	土壤酸碱度		道路通达度
	Soil pH		Road accessibility degree
	灌溉保证率		对外交通便利度
	Irrigation guarantee rate		External traffic convenience degree
	灌溉水质量	工程因素	灌溉排水工程
	Irrigation water quality	Engineering factor	Irrigation drainage project
生态因素	自然灾害危险性		道路工程
Ecological factor	Natural disaster risk		Road construction project
	生态系统脆弱性		农田防护林工程
	Ecosystem vulnerability		Farmland protective forest project
	水土流失状况		土地平整工程
	Soil erosion condition		Land leveling project

随机森林是基于二进制分割数据解决分类和回归问题的算法[27]。随机森林算法首先采用 Bootstrap 抽样技术从原始数据集中抽取 N 个训练集，每个训练集的大小约为原始数据集的 2/3；然后，为每个训练集分别建立分类回归树，产生由 n 棵分类树组成的森林，在每棵树生长过程中，从全部 M 个特征变量中随机抽选 m 个 ($m \leqslant M$) 特征变量，在这 m 个属性中根据 Gini 系数最小原则选出最优属性进行内部节点分支；最后，集合 n 棵分类树的预测结果，采用投票的方式决定新样本的类别，每次抽样约有 1/3 的数据未被抽中，利用这部分袋外数据 (Out-of-bag，OOB) 进行内部误差估计，产生 OOB 误差[28]。

随机森林算法利用 OOB 误差计算特征变量重要性 (I)：首先，根据袋外数据计算随机森林中每个分类树的袋外误差 (E)；然后，随机改变袋外数据第 j 个特征变量 (X^j) 的值，并计算新的袋外误差 (E_i^j)；最后，变量X^j的重要性 [$I(X^j)$] 表示为

$$I\left(X^j\right)=\frac{1}{n}\sum_{i=1}^{n}(E_i^j-E_i), \tag{1}$$

变量X^j的变化引起的袋外误差增加越大，精度减少的越多，说明该变量越重要[29-30]。但是根据相关的研究成果表明，随机森林算法能够避免剔除重要的变量，因为这些重要变量可能与其他变量有相关性[30]。

随机森林算法通过 R 语言软件平台实现，运行过程中需要至少定义 2 个参数：分类树的数目 (n) 和节点分裂时输入的特征变量个数 (m)。若做分类分析，则 m 设定为变量个数的平方根，回归分析则设定为变量个数的 1/3[31]。本研究影响因素为 30 个，

因此，n=500，m=5，其余参数均根据模型默认值进行设定。

1.3.2 相关性分析 通过 SPSS19.0 平台，对在随机森林模型中模拟的各影响因素进行 Pearson 相关性分析。

1.3.3 定级方法 1) 定级指标量化：根据定级指标分布类型及对定级评价单元的影响方式，对定级指标分为 3 类进行量化。其中：面状指标包括耕作距离、农田破碎度、林网化密度和城镇影响度；线状指标包括道路通达度和农田路网密度；点状指标包括农贸市场影响度和对外交通便利度。面状指标的量化主要通过直接指标或者间接指标进行，线状指标的量化则采用线性衰减法进行。

2) 定级评价单元划分：定级单元是级别划分和质量评定的基本空间单位，单元划分的主要方法为地块法、网格法和叠置法。本文为了保持原始地类图斑的自然属性和形状，与土地利用现状图中的耕地图斑保持一致，更贴近现实耕作情况，采用地块法[32] 进行定级评价单元划分，最终划定定级单元数为 9 219。

3) 定级指数与级别划分：定级指数计算采用加权求和法[33]。耕地级别一般根据单元定级指数值进行划分，通常采用等间距法、数轴法或总分频率曲线法进行土地级别的划分。本文采用等间距法对单元定级指数进行级别划分。

1.3.4 特尔菲法 目前耕地质量定级评价工作中，应用特尔菲法确认定级指标体系是比较常用的方法，本文将研究成果与采用特尔菲法的定级结果进行对比分析，尝试分析随机森林算法的优势与不足。本文采用的特尔菲法的定级指标体系如表 3 所示，共计 30 个指标作为影响因素。

表 3 采用特尔菲法的定级指标体系

Table 3 Grading indicators system using Delphi method

一级 Primary level		二级 Secondary level		定级 Grading level	
一级指标 Primary indicator	权重 Weight	二级指标 Secondary indicator	权重 Weight	定级指标 Grading indicator	权重 Weight
生态因素 Ecological factor	0.19	生态状况 Ecological condition	0.19	生态系统脆弱性 Ecosystem vulnerability	0.02
				自然灾害危险性 Natural disaster risk	0.08
				水土流失状况 Soil erosion condition	0.09
自然因素 Natural factor	0.48	气候状况 Climate condition	0.16	生长季平均温度 Mean temperature of growing season	0.06
				生长季平均降水 Mean precipitation of growing season	0.08
				年总太阳辐射量 Annual solar radiation amount	0.02
		地形状况 Terrain condition	0.08	海拔高度 Elevation	0.03
				地形坡度 Topographic slope	0.04
				砾石含量 Gravel content	0.01
		土壤条件 Soil condition	0.22	有效土层厚度 Effective soil thickness	0.06
				表层土壤质地 Surface soil texture	0.06
				土壤酸碱度 Soil pH	0.03
				土壤有机质含量 Soil organic matter content	0.07
		水资源状况 Water resources condition	0.02	灌溉水质量 Irrigation water quality	0.02
社会经济因素 Socioeconomic factor	0.09	基础设施条件 Infrastructure condition	0.04	农田路网密度 Farmland road network density	0.01
				林网化程度 Degree of forestation	0.01
				农田破碎度 Farmland fragmentation degree	0.01
				灌溉保证率 Irrigation guarantee rate	0.01

续表 3　Continued table 3

一级 Primary level		二级 Secondary level		定级 Grading level	
一级指标 Primary indicator	权重 Weight	二级指标 Secondary indicator	权重 Weight	定级指标 Grading indicator	权重 Weight
		耕作条件 Cultivating condition	0.03	耕作距离 Cultivation distance	0.01
				田块形状 Field shape	0.01
				田块大小 Field size	0.01
		土地利用状况 Land use status	0.02	利用现状 Utilization status	0.02
区位因素 Location factor	0.20	区位条件 Locational condition	0.11	城镇影响度 Urban influence degree	0.06
				农贸市场影响度 Agricultural market influence degree	0.05
		交通条件 Traffic condition	0.09	道路通达度 Road accessibility degree	0.05
				对外交通便利度 External traffic convenience degree	0.04
工程因素 Engineering factor	0.04	工程建设状况 Construction condition	0.04	灌溉排水工程 Irrigation drainage project	0.01
				农田防护林工程 Farmland protective forest project	0.01
				道路工程 Road construction project	0.01
				土地平整工程 Land leveling project	0.01

2　结果与分析

2.1　随机森林算法分析结果

以 97 个产量采样点 2015—2017 年的平均标准产量作为因变量，30 个影响因素作为自变量建立随机森林回归模型，并对 30 个影响因素进行重要性分析。模型的预测产量与实际产量的拟合度 (R^2) 达到 79.47，反映出模型的拟合度较好。

表 4 列出了随机森林模型分析的变量重要性 I 值。30 个影响因素的变量重要性 I 值在 0.03~

表 4　随机森林算法对影响因素的变量重要性 (*I*) 排序

Table 4　Variable importance (*I*) ranking of impact factors by random forest algorithm

影响因素 Impact factor	I	影响因素 Impact factor	I
生态系统脆弱性 Ecosystem vulnerability	11.94	灌溉保证率 Irrigation guarantee rate	4.74
生长季降水量 Mean precipitation of growing season	10.63	有效土层厚度 Effective soil thickness	4.65
自然灾害危险性 Natural disaster risk	10.01	灌溉排水工程 Irrigation drainage project	4.58
年总太阳辐射 Annual solar radiation	9.08	表层土壤质地 Surface soil texture	4.50
土壤酸碱度 pH	8.54	海拔高度 Elevation	3.85
灌溉水质量 Irrigation water quality	7.94	农田路网密度 Farmland road network density	3.85
对外交通便利度 External traffic convenience	7.60	道路工程 Road construction project	3.63
地形坡度 Slope	7.47	利用现状 Utilization status	3.61
生长季均温 Mean temperature of growing season	6.82	砾石含量 Gravel content	3.38
农田破碎度 Farmland fragmentation	6.75	农田防护林工程 Protective forest project	2.85
城镇影响度 Urban influence	6.14	林网化程度 Degree of forestation	2.62
农贸市场影响度 Agricultural market influence	6.10	道路通达度 Road accessibility	2.57
土壤有机质含量 Soil organic matter	5.79	土地平整工程 Land leveling project	2.01
水土流失状况 Soil erosion condition	5.73	田块大小 Field size	1.02
耕作距离 Cultivation distance	5.70	田块形状 Field shape	0.03

11.94，其中生态系统脆弱性、生长季平均降水和自然灾害危险性的重要性较强，*I* 值分别为 11.94、10.63 和 10.01。年总太阳辐射量、土壤酸碱度、灌溉水质量、对外交通便利度、地形坡度、生长季平均温度、农田破碎度、城镇影响度、农贸市场影响度、土壤有机质含量、水土流失状况和耕作距离的 *I* 值在 5.70~9.08，其他 15 个影响因素的 *I* 值在 4.74 以下。

2.2 相关性分析结果

对 30 个影响因素进行 Pearson 相关分析，结果显示，大部分影响因素间相关性不显著。其中有 8 个影响因素为显著相关 (表 5)：自然灾害危险性与生态系统脆弱性的 Pearson 相关系数 r=0.957，为显著正相关；表层土壤质地与砾石含量的相关系数 r=0.790；灌溉排水工程、道路工程、农田防护林工程和土地平整工程之间 r>0.850，结果表明，在该研究区域，以上影响因素之间存在较强的相关性，会对定级结果造成多重共线性的问题。

在面对多重共线性问题上，最常用做法是保留重要解释变量，去掉次要或可替代解释变量。然而，相关性分析虽然可以判断影响因素之间的相关性，但是无法识别其中重要的解释变量。因此，本文结合随机森林的变量重要性的分析结果进行了解释变量的识别。

表 5 显著相关影响因素的相关系数[1)]

Table 5 Correlation coefficients matrix of significantly related impact factors

影响因素 Impact factor	*Y*1	*Y*2	*Y*3	*Y*4	*Y*5	*Y*6	*Y*7	*Y*8
*Y*1	1							
*Y*2	0.790**	1						
*Y*3	0.254*	0.310**	1					
*Y*4	−0.278*	−0.150*	0.957**	1				
*Y*5	0.132*	0.119*	0.243*	−0.251*	1			
*Y*6	0.142*	0.156*	0.222*	−0.258*	0.856**	1		
*Y*7	0.140*	0.177*	0.278*	−0.262*	0.905**	0.986**	1	
*Y*8	0.190*	0.104*	0.189*	−0.247*	0.851**	0.969**	0.982**	1

1)*Y*1：表层土壤质地；*Y*2：砾石含量；*Y*3：自然灾害危险性；*Y*4：生态系统脆弱性；*Y*5：灌溉排水工程；*Y*6：道路工程；*Y*7：农田防护林工程；*Y*8：土地平整工程；“*”和“**”分别表示在0.05和0.01水平显著相关(Pearson法)

1)*Y*1: Surface soil texture; *Y*2: Gravel content; *Y*3: Natural disaster risk; *Y*4: Ecosystem vulnerability; *Y*5: Irrigation drainage project; *Y*6: Road construction project; *Y*7: Farmland protective forest project; *Y*8: Land leveling project; “*” and “**” indicate significant correlation at 0.05 and 0.01 levels, respectively(Pearson method)

2.3 定级指标体系

对重要性 *I* 值进行标准化后的研究区域的定级指标体系如表 6 所示，这套耕地质量定级指标体系的一级指标权重序列为自然因素>区位因素>生态因素>社会经济因素，其中自然因素权重最大，为 0.53；区位因素和生态因素分别为 0.19 和 0.16，社会经济因素权重最小，为 0.12。该套耕地质量定级指标体系包含了生态环境状况、气候状况、地形状况、土壤条件、水资源状况、基础设施条件、耕作便利条件及区位、交通等方面，选取的指标都不同程度地对研究区域耕地质量有所影响，其权重与影响程度比较相符，能够比较合理地评价研究区域的耕地质量。

2.4 随机森林算法和特尔菲法定级结果的比较

从随机森林算法和特尔菲法定级结果 (图 2) 可知，随机森林算法和特尔菲法计算的耕地质量定级结果具有比较相似的空间分布情况，即共和县东部及乌兰县中部耕地级别较高，共和县西部及南部比较低，但局部地区定级结果存在较大差异。由于研究区域耕地分布较为零散，难以从总体空间和数量上对比 2 种方法定级结果的差异，故在研究区内设置能够穿越最多耕地斑块的 2 条典型样带，对随机森林算法及特尔菲法的定级结果进行比较观察，即“东北−西南”样带及“西北−

表 6　耕地质量定级指标体系

Table 6　The index system of cultivated land quality grading

一级 Primary level		二级 Secondary level		定级 Grading level		
一级指标 Primary indicator	权重 Weight	二级指标 Secondary indicator	权重 Weight	定级指标 Grading indicator	*I*	权重 Weight
生态因素 Ecological factor	0.16	生态状况 Ecological condition	0.16	生态系统脆弱性 Ecosystem vulnerability	11.94	0.11
				水土流失状况 Soil erosion condition	5.73	0.05
自然因素 Natural factor	0.53	气候状况 Climate condition	0.25	生长季平均温度 Mean temperature of growing season	6.82	0.06
				生长季平均降水 Mean precipitation of growing season	10.63	0.10
				年总太阳辐射量 Annual solar radiation amount	9.08	0.09
		地形状况 Terrain condition	0.07	地形坡度 Topographic slope	7.47	0.07
		土壤条件 Soil condition	0.14	土壤有机质含量 Soil organic matter content	5.79	0.08
				土壤酸碱度 pH	8.54	0.06
		水资源状况 Water resources condition	0.07	灌溉水质量 Irrigation water quality	7.94	0.07
社会经济因素 Socioeconomic factor	0.12	基础设施条件 Infrastructure condition	0.06	农田破碎度 Farmland fragmentation degree	6.75	0.06
		耕作便利条件 Cultivating condition	0.06	耕作距离 Cultivation distance	5.70	0.06
区位因素 Location factor	0.19	区位条件 Locational condition	0.12	城镇影响度 Urban influence degree	6.14	0.06
				农贸市场影响度 Agricultural market influence degree	6.10	0.06
		交通条件 Traffic condition	0.07	对外交通便利度 External traffic convenience degree	7.60	0.07

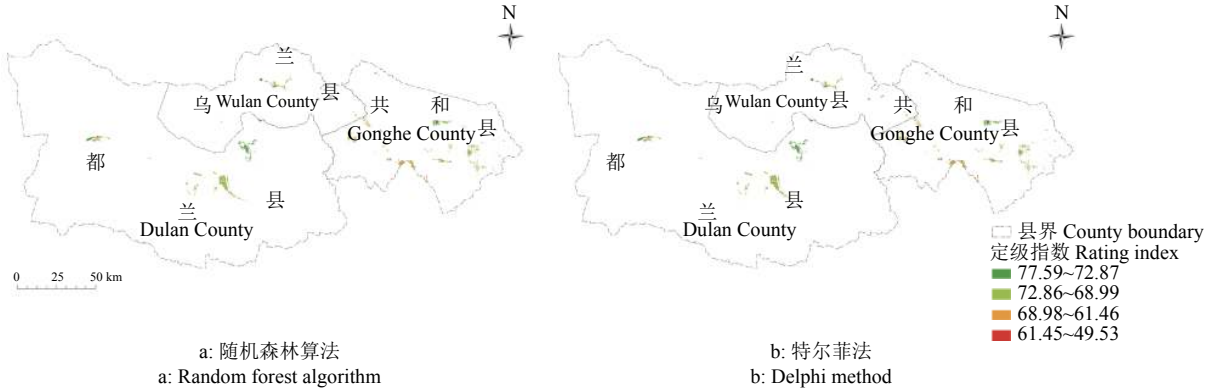
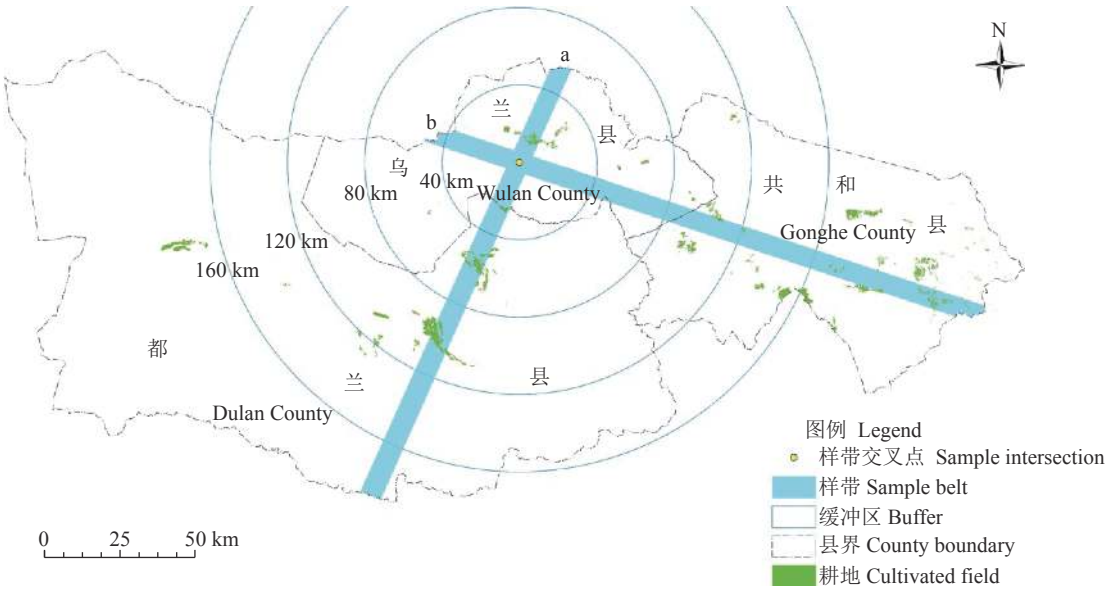


图 2　研究区域 2 种方法的定级结果及空间分布图

Fig. 2　The grading result and spatial distribution map of two methods in the research area

东南”样带，样带宽度为 5 km(图 3)。

从对样带的定级结果 (图 4) 可知，随机森林算法和特尔菲法的定级结果空间变化趋势基本一致的趋同性强，对同一地块的级别高低判断基本一致，都能够较好地体现出耕地级别的空间变异性。另外，随机森林算法更加稳定，级别变化程度小，更



a: 东北–西南样带; b: 西北–东南样带

a: Sample belt from northeast to southwest; b: Sample belt from northwest to southeast

图 3 研究区域 2 条样带位置图

Fig. 3 Position of two sample belts in the study area

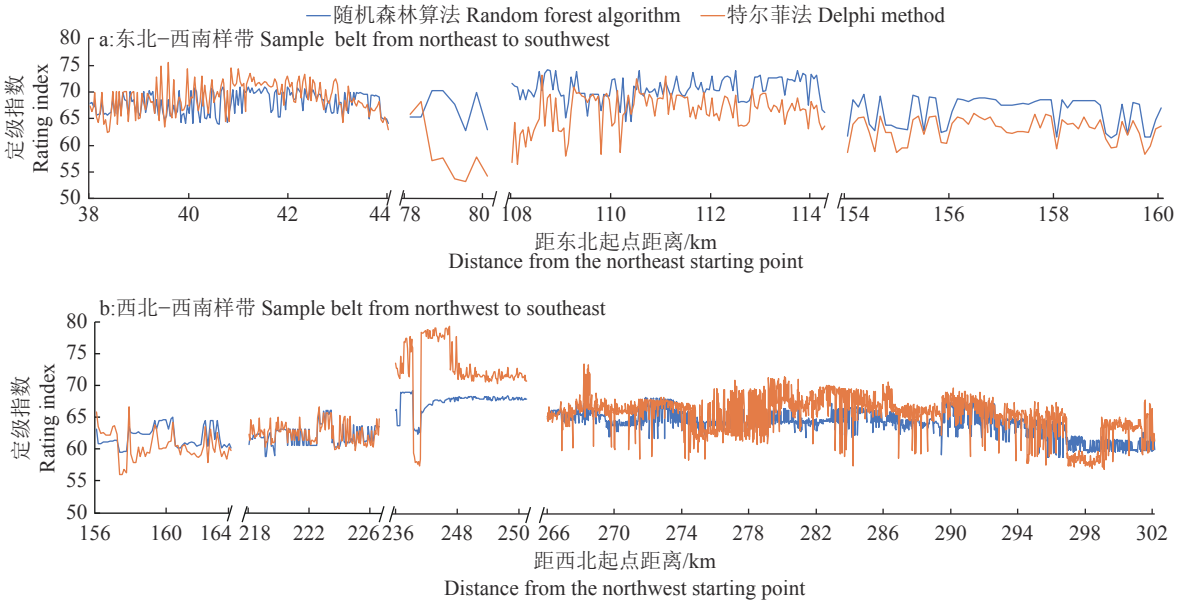


图 4 随机森林算法和特尔菲法对 2 条样带定级结果的对比

Fig. 4 Comparison of grading results of random forest method and Delphi method to two sample belts

有利于构建省级空间尺度的耕地级别可比序列。

在西北–东南样带第 220 km 处，随机森林算法的定级结果与特尔菲法比较相似。该位置地形平整，土壤质地为壤土，有效土层厚度大于 1 m，土壤有机质含量较高，但生长季降水及温度条件较差，距离共和县城区较远，区位条件较差。表明尽管随机森林算法比特尔菲法的定级指标数量较少，但定级结果由于权重的大小调整，未影响定级结果的准确性，随机森林算法能够有效地表现研究区域耕地级别的空间变异性。

其次，在东北–西南样带第 110 km 处，随机森林算法的定级结果显著高于特尔菲法的定级结果。此处耕地位于柴达木盆地的东部，年总太阳辐射量为 6 735 MJ·m^{-2}，生长季平均气温为 11~12 ℃，生态系统脆弱性类型为略脆弱型。在研究区域中该位置耕地的太阳辐射量较大，生长季气温高，生态环境比较稳定，是青海省著名的春小麦高产区。表明随机森林算法能够通过变量重要性值大小的设置，充分体现以上定级指标对耕地质量的正向影响作用，有效识别出研究区域的优质高产耕地，修正研究区

域中高值低估的状况。另外，在西北−东南样带第290 km 位置，随机森林算法的定级结果显著低于特尔菲法。该区域靠近城镇及国道，生长季温度为10~11 ℃，生长季温度较高，降水比较充足，区位和交通条件优越。然而该区域位于研究区域中人口总量最多的共和县，耕地开发程度较高，生态系统脆弱性类型为一般脆弱型，人与自然矛盾比较突出，脆弱的生态环境限制了耕地的耕作；另外，该地区的太阳辐射量为研究区较低的区域，较大地限制了耕地作物的产量。表明由于特尔菲法定级指标体系的定级指标过多，各定级指标权重值的差异较小，无法突出一些重要的定级指标，定级结果易受到多影响因素的综合影响，造成低值高估的现象。总之，随机森林算法稳定性更好，级别指数变幅区间更小，更有利于构建省级空间尺度的耕地级别可比序列。

3　讨论与结论

随机森林算法计算得到耕地质量定级指标体系，对该体系的研究结果表明，自然因素是对研究区域耕地影响最大的方面，本文从自然因素的定级指标分别讨论其对研究区域耕地质量的影响。

生长季平均降水和灌溉水质量这 2 个定级指标描述了耕地灌溉情况，也是影响研究区域耕地质量的重要因素。年降水不足，灌溉条件较差，均直接制约着青海省农作物的单产。青海省气候干燥，蒸发量远大于降水量，使农作物的需水量剧增。当降水充足、灌溉水质量高时，将直接影响作物吸收的水分、土壤养分的质量[34]。年太阳辐射量的权重为0.09，排权重的第 3 位，对研究区域耕地质量影响较大。青海省尤其是研究区域中位于柴达木盆地的部分，太阳辐射量仅次于西藏，是青海省发展农业的重要优势，其对于小麦的茎秆发育、扬花、灌浆的作用十分明显，直接影响耕地的产量[35]。青海地区属于高原大陆性气候，易形成土壤盐渍化[36]，土壤pH 也是影响耕地产量的因素之一。地形坡度对地区水热条件的能量交换起重要作用，同时直接影响土壤的形成和植被的生长发育。生长季平均温度对生物的正常发育和生长起着决定性的作用，决定农作物的耕作制度。土壤有机质提供植物生长发育所需要的养分，其含量能影响耕地的产出。

本研究首先根据研究区域相关部门的研究成果、相关文献及实地农业调查，整理出影响研究区域耕地质量的 30 个影响因素，并收集了研究区域近 3 年的耕地产量数据。其次，对影响因素进行相关性分析，并将耕地产量与该 30 个影响因素建立随机森林回归模型。通过综合影响因素的相关性分析结果及随机森林变量重要性结果，将 30 个影响因素筛选为 14 个定级指标，构建了本研究区域的耕地质量定级指标体系，并对研究区域的耕地进行了质量定级。最后将定级结果与定级工作常用的特尔菲法成果进行对比分析，得出以下结论：1) 将随机森林算法与相关性分析结合，构建了定级指标体系。指标体系同时涵盖生态因素、自然因素、社会经济因素和区位因素共 4 方面，既包含了大尺度上生态、气候和区位影响因素的指标，同时又考虑精细到地块尺度的土地属性，结果可以比较全面地评价研究区域耕地的质量。2) 通过对随机森林算法和特尔菲法定级结果的比较，2 种方法定级结果的趋同性强，对同一地块的级别高低判断基本一致，都能够较好地体现出耕地级别的空间变异性；随机森林算法稳定性更好，级别指数变幅区间更小，更有利于构建省级空间尺度的耕地级别可比序列，为随机森林算法的应用及耕地质量定级指标体系的构建提供了新的依据。

此外，本研究的部分呈面状分布指标如生态系统脆弱性，自然灾害危险性等，无法有效表示各程度影响的渐变过程。同时，生态系统的变化比较剧烈和频繁，如何使指标精确定量化是评价耕地质量需要解决的重要问题，有待今后进一步的研究探索。

参考文献：

[1] 吴大放, 刘艳艳, 董玉祥, 等. 我国耕地数量、质量与空间变化研究综述[J]. 热带地理, 2010, 30(2): 108-113.

[2] 温良友, 孔祥斌, 辛芸娜, 等. 对耕地质量内涵的再认识[J]. 中国农业大学学报, 2019, 24(3): 156-164.

[3] 张超, 乔敏, 郧文聚, 等. 耕地数量、质量、生态三位一体综合监管体系研究[J]. 农业机械学报, 2017, 48(1): 1-6.

[4] 刘兴华, 孙鹏举, 刘学录. 甘肃省临夏县耕地资源社会保障价值测算[J]. 干旱区资源与环境, 2013, 27(1): 53-57.

[5] 中华人民共和国国土资源部. 农用地定级规程: GB/T 28405—2012 [S]. 北京: 中国标准出版社, 2012.

[6] 高中贵, 彭补拙. 我国农用地分等定级研究综述[J]. 经济地理, 2004, 24(4): 514-519.

[7] 金东海, 许皞, 秦文利. 基于分等成果的农用地定级新方法:两层七参数法[J]. 中国土地科学, 2004, 18(6): 34-39.

[8] 鲁明星, 贺立源, 吴礼树. 我国耕地地力评价研究进展[J]. 生态环境, 2006, 8(4): 866-871.

[9] 冯超. 中国谷物产出的"面积− 质量"导向因素分析[J]. 干旱区资源与环境, 2015, 29(8): 7-13.

[10] 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.
[11] 张凤荣, 安萍莉, 王军艳, 等. 耕地分等中的土壤质量指标体系与分等方法[J]. 资源科学, 2002, 24(2): 71-75.
[12] 付国珍, 摆万奇. 耕地质量评价研究进展及发展趋势[J]. 资源科学, 2015, 35(2): 226-236.
[13] 盛艳, 姚云峰, 秦富仓, 等. 基于 GIS 的耕地地力等级划分研究[J]. 干旱区资源与环境, 2014, 28(6): 27-32.
[14] 马瑞明, 马仁会, 韩冬梅, 等. 基于多层级指标的省域耕地质量评价体系构建[J]. 农业工程学报, 2018, 34(16): 249-257.
[15] 杜国明, 刘彦随, 于凤荣, 等. 耕地质量观的演变与再认识[J]. 农业工程学报, 2016, 32(14): 243-249.
[16] 马昊翔, 陈长成, 宋英强, 等. 青海省近 10 年草地植被覆盖动态变化及其驱动因素分析[J]. 水土保持研究, 2018, 25(6): 137-145.
[17] 保广裕, 张静, 周丹, 等. 青海省太阳辐射强度时空变化特征分析[J]. 冰川冻土, 2017, 39(3): 563-571.
[18] 樊杰. 中国主体功能区划方案[J]. 地理学报, 2015, 70(2): 186-201.
[19] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
[20] CHEN X W, LIU M. Prediction of protein-protein interactions using random decision forest framework[J]. Bioinformatics, 2005, 21(24): 4394-4400.
[21] WARD M M, PAJEVIC S, DREYFUSS J, et al. Short-term prediction of mortality in patients with systemic lupus erythematosus: Classification of outcomes using random forests[J]. Arthrit Rheumat, 2006, 55(1): 74-80.
[22] OPARIN I, GLEMBEK O, BURGET L, et al. Morphological random forests for language modeling of inflectional languages[C/OL]//IEEE. 2008 IEEE Spoken Language Technology Workshop. Goa: IEEE, 2008: 189-192. [2019-08-25]. https://www.infona.pl/resource/bwmeta1.element.ieee-art-000004777872/tab/summary. doi: 10.1109/SLT.2008.4777872.
[23] ZHANG M, ZHANG H, WU P, et al. Prediction of soil organic carbon in an intensively managed reclamation zone of eastern China: A comparison of multiple linear regressions and the random forest model[J]. Sci Total Environ, 2017, 592: 704-713.
[24] 方匡南, 朱建平, 谢邦昌. 基于随机森林方法的基金收益率方向预测与交易策略研究[J]. 经济经纬, 2010, 27(2): 61-65.
[25] 董师师, 黄哲学. 随机森林理论浅析[J]. 集成技术, 2013, 2(1): 1-7.
[26] 张雷, 王琳琳, 张旭东, 等. 随机森林算法基本思想及其在生态学中的应用：以云南松分布模拟为例[J]. 生态学报, 2014, 24(3): 650-659.
[27] 刘斌, 郭星, 朱宇恩. 基于随机森林模型的土壤重金属源解析：以晋中盆地为例[J]. 干旱区资源与环境, 2019, 33(1): 106-111.
[28] 马玥, 姜琦刚, 孟治国, 等. 基于随机森林算法的农耕区土地利用分类研究[J]. 农业机械学报, 2016, 47(1): 297-303.
[29] ZHU Z, WOODCOCK C E, ROGAN J, et al. Assessment of spectral, polarimetric, temporal, and spatial dimensions for urban and peri-urban land cover classification using Landsat and SAR data[J]. Rem Sens Environ, 2012, 117: 72-82.
[30] VAN BEIJMA S, COMBER A, LAMB A. Random forest classification of salt marsh vegetation habitats using quad-polarimetric airborne SAR, elevation and optical RS data[J]. Rem Sens Environ, 2014, 149: 118-129.
[31] LIAW A, WIENER M. Classification and regression by random forest[J]. R News, 2002, 2(3): 18-22.
[32] 刘欢, 吴克宁, 宋文, 等. 耕地质量定级方法改进研究:以农安县为例[J]. 北京师范大学学报(自然科学版), 2018, 54(3): 315-320.
[33] 赵璐, 郑新奇, 闫弘文, 等. 基于地统计学的县域农用地定级方法[J]. 农业工程学报, 2008, 24(S1): 99-103.
[34] 黄居茂. 青海省农作物生产发展的科学技术探讨[J]. 青海农林科技, 1984, 14(4): 18-26.
[35] 朱文江, 康素珍. 柴达木盆地春小麦高产的气候因素[J]. 中国农业科学, 1978, 19(2): 51-56.
[36] 张玮, 李江. 青海省菜田盐渍化形成及治理[J]. 青海农技推广, 2015, 20(2): 32-33.

【责任编辑 李晓卉】