中国农业生态文明模式空间分布数据集（2018–2020）

王 曙[1]，诸云强[1,2*]，钱 朗[3]，宋 佳[1,2]，袁 文[1]

1. 中国科学院地理科学与资源研究所，资源与环境信息系统国家重点实验室，北京 100101；
2. 江苏省地理信息协同创新中心，南京 210023；
3. 华南师范大学计算机学院，广州 510631

摘 要：农业生态文明模式是人类合理利用自然环境形成的可模仿、可复制农业发展示范样例。调查农业生态文明模式的空间分布能够揭示农业发展的空间差异、聚集效应和多样性，对农业发展规划、农业生态文明建设和农业可持续发展研究具有重要意义。为此，作者汇集 2018–2020 年间中国农业生态文明模式的新闻报道（包含政府网站、央视网、人民网和新华网），采用自然语言处理等技术，将报道蕴含的农业生态文明模式进行抽取与分类，并进一步解析农业生态文明模式的报道时间和空间位置，最终获得中国地区农业生态文明模式的点状空间分布数据集。数据集详细记录了每项农业生态文明模式的类型、空间位置、报道时间、文本关键词、原始描述、文本来源等信息。数据文件包括.xlsx、.shp 格式 33,440 条记录，数据量为 168 MB（压缩后 21.4 MB）。

关键词：农业生态文明模式；空间分布；中国；新闻报道；2018-2020
DOI: https://doi.org/10.3974/geodp.2021.02.10
CSTR: https://cstr.escience.org.cn/CSTR:20146.14.2021.02.10
数据可用性声明：
本文关联实体数据集已在《全球变化数据仓储电子杂志（中英文）》出版，可获取：
https://doi.org/10.3974/geodb.2021.06.02.V1 或 https://cstr.escience.org.cn/CSTR20146.14.2021.06.02.V1.

1 前言

中国农业生态文明模式是在联合国可持续发展目标下，结合中国自身自然环境和社会经济条件，发展形成的农业发展示范样例[1,2]。这些农业生态模式具有榜样示范作用，能够为当地农业发展提供可参照的路径，进而从发展方向具体指导当地农业种植、农业生产和农业经营等方面，对区域可持续发展具有重大现实意义[3]。然而，中国的国土面积辽阔，南北东西各区域具有不同的农业生产环境、技术环境与市场环境，造就了中国农业生态文明模式空间分布的差异。这种空间分布的差异给区域农业发展模式的参照带来了极大挑战，既无法从宏观上整体了解农业生态文明模式的布局，也无法从微观上探知每类农业生态文

收稿日期：2021-04-22；**修订日期：**2021-06-04；**出版日期：**2021-06-25
基金项目：中国科学院（XDA23100100）；国家自然科学基金（42050101，41771430，41631177，42101467）
***通讯作者：**诸云强 L-6116-2016, 中国科学院地理科学与资源研究所, zhuyq@lreis.ac.cn
数据引用方式：[1] 王曙, 诸云强, 钱朗等. 中国农业生态文明模式空间分布数据集（2018–2020）[J]. 全球变化数据学报, 2021, 5(2): 181–188. https://doi.org/10.3974/geodp.2021.02.10. https://cstr.escience.org.cn/CSTR:20146.14.2021.02.10.
[2] 王曙, 诸云强, 钱朗等. 中国农业生态文明模式及其空间分布数据集（2018–2020）[J/DB/OL]. 全球变化数据仓储电子杂志, 2021. https://doi.org/10.3974/geodb.2021.06.02.V1. https://cstr.escience.org.cn/CSTR:20146.11.2021.06.02.V1.

明模式形成的自然环境与社会经济机理。鉴于此，调查中国农业生态文明模式的空间分布，对揭示中国每类农业生态文明模式的空间形态与形成机理，乃至进一步规划和调整国家尺度农业发展模式，具有深远影响。

2001 年至 2003 年期间，农业部（现农业农村部）科技司对农业生态文明模式进行过一次全国性调研，其通过行政手段采用自下向上逐级上报的方法，征集得到 370 种生态模式或技术体系，最终经专家组评审公布了中国十大农业模式，包括有北方“四位一体”生态模式、南方“猪-沼-果”生态模式、草地生态恢复与持续利用模式、农林牧复合生态模式、生态种植模式、生态渔业模式、丘陵山区小流域综合治理利用型生态农业模式、设施生态农业和观光生态农业模式[4,5]。这种方法虽然得出了十类中国典型的农业发展模式，但是无法对每一类农业生态文明模式确切空间分布准确定位[6]。这种情况使得在农业模式借鉴过程中，虽然知道典型的农业发展模式，但无法了解每类农业生态文明模式在不同区域的差异，进而无法结合自身自然环境和社会经济条件，选取更加精准的模式参考。

鉴于此，本研究生产了更加精确的中国农业生态文明模式点状空间分布数据，利用典型突出农业生态文明模式会被重点报道的特性，以新闻报道文本为数据源，采用自然语言处理技术、空间位置解析等技术，揭示了每项农业生态文明模式的类型、空间位置、报道时间等信息，构成了中国农业生态文明模式空间分布数据集（2018–2020）。

2 数据集元数据简介

《中国农业生态文明模式及其空间分布数据集（2018–2020）》[7]的名称、作者、地理区域、数据年代、时间分辨率、空间分辨率、数据集组成、数据出版与共享服务平台、数据共享政策等信息见表 1。

3 数据研发方法

3.1 技术路线

中国农业生态文明模式空间分布数据集研发技术路线如图 1 所示，主要包括语料获取和信息抽取两大部分。

3.1.1 语料获取

语料获取包含两个步骤：农业生态文明模式本体构建和农业生态文明模式语料爬取。其中，农业生态文明模式本体构建需要人工通过阅读文献、报告、专著等资料构建农业生态文明模式的分类体系及相关词汇库，参见表 2。农业生态文明模式语料获取来自于新闻网的报道文本。其中，新闻网站具体包含：政府官方网站、央视网、人民网和新华网。政府官方网站选取中华人民共和国农业农村部新闻栏目[1]；央视网为央视网新闻栏目[2]；人民网为人民网新闻搜索栏目[3]；新华网为新华网新闻搜索栏目[4]。

[1] 中华人民共和国农业农村部新闻栏目. http://www.moa.gov.cn/xw/.

[2] 央视网新闻栏目. https://news.cctv.com/.

[3] 人民网新闻搜索栏目. http://search.people.cn/.

[4] 新华网新闻搜索栏目. http://so.xinhuanet.com/.

3.1.2 信息抽取

信息抽取步骤主要是从获取的新闻报道中，解析获取关于农业生态文明模式的报道时间、空间位置、模式描述和模式类型等信息，并将它们关联形成农业生态文明模式的记录，对应的时间信息抽取、空间信息抽取、模式描述抽取等算法描述参见下节。

表 1　《中国农业生态文明模式及其空间分布数据集（2018–2020）》[7]元数据简表

条　目	描　述
数据集名称	中国农业生态文明模式及其空间分布数据集(2018–2020)
数据集短名	CEApatterns_2018-2020
作者信息	王曙, 中国科学院地理科学与资源研究所, wangshu@igsnrr.ac.cn 诸云强 L-6116-2016, 中国科学院地理科学与资源研究所, zhuyq@lries.ac.cn 钱朗, 华南师范大学, 2018022623@m.scnu.edu.cn 宋佳, 中国科学院地理科学与资源研究所, songj@igsnrr.ac.cn 袁文, 中国科学院地理科学与资源研究所, yuanwen@igsnrr.ac.cn
地理区域	中国
数据年代	2018–2020
时间分辨率	1 天
空间分辨率	100 m
数据格式	.xlsx、. shp
数据量	168 MB（压缩后 21.4 MB）
数据集组成	由 33,440 条农业生态文明模式记录构成
基金项目	中国科学院（XDA23100100）；国家自然科学基金（42050101，41631177）
出版与共享服务平台	全球变化科学研究数据出版系统 http://www.geodoi.ac.cn
地址	北京市朝阳区大屯路甲 11 号 100101，中国科学院地理科学与资源研究所
数据共享政策	全球变化科学研究数据出版系统的“数据”包括元数据（中英文）、通过《全球变化数据仓储电子杂志（中英文）》发表的实体数据集和通过《全球变化数据学报（中英文）》发表的数据论文。其共享政策如下：（1）“数据”以最便利的方式通过互联网系统免费向全社会开放，用户免费浏览、免费下载；（2）最终用户使用“数据”需要按照引用格式在参考文献或适当的位置标注数据来源；（3）增值服务用户或以任何形式散发和传播（包括通过计算机服务器）“数据”的用户需要与《全球变化数据学报（中英文）》编辑部签署书面协议，获得许可；（4）摘取“数据”中的部分记录创作新数据的作者需要遵循 10% 引用原则，即从本数据集中摘取的数据记录少于新数据集总记录量的 10%，同时需要对摘取的数据记录标注数据来源[8]
数据和论文检索系统	DOI, CSTR, Crossref, DCI, CSCD, CNKI, SciEngine, WDS/ISC, GEOSS

3.2 算法原理

数据集研发涉及核心算法包括以下几个部分：时间信息抽取算法、空间信息抽取算法、模式抽取算法和农业生态文明模式记录聚合算法。

（1）时间信息抽取算法

时间信息抽取是指从新闻报道中获取农业生态文明模式的报道时间。在数据集研发中，因为农业生态文明模式的报道时间在四类新闻网站中具有固定表达形式，所以其获取可利用互联网 XPATH 解析语句从 HTML 文本中直接获取，中华人民共和国农业农村部新闻栏

目、央视新闻网站为央视网新闻栏目、人民网新闻搜索栏目和新华网新闻搜索栏目的XPATH语句分别为(1)、(2)、(3)和(4)。

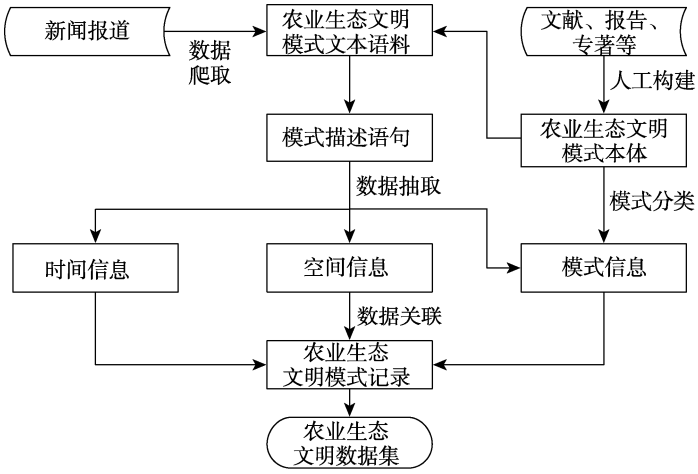

图1 中国农业生态文明模式空间分布数据集研发技术路线图

表2 农业生态文明模式分类体系.

一级	二级
生态种植业模式	林粮间作模式、林药间作模式、林菜间作模式、林苗间作模式、林菌间作模式、林草间作模式、林花间作模式、林果间作模式、菌草间作模式
	时间差间套模式、空间差间套模式、养分差间套模式
	水肥一体化模式、膜下滴灌模式、抗旱保苗模式、农业贮藏模式、节水农业模式、节肥农业模式、精准施肥模式、设施农业模式、小农模式、种植修复模式
	稻鱼类模式、稻畜类模式、林草畜模式、果园畜养模式、沙地散养模式、种养加模式、稻鱼鸭类模式、畜沼果模式、多元立体循环种养、秸秆循环种养、种养结合大类、种养循环模式
生态养殖业模式	发酵床养殖模式、粪便还田模式、粪便资源化利用模式、沼气集中利用模式、
	水禽水产模式、二段式养殖模式、鸡猪模式、错时养殖模式、养殖加工模式、循环养殖模式、放养散养模式、设施养殖模式、养殖修复治理模式
创新性生态农业发展模式	微生物农业模式、物联网精准农业模式、光伏农业模式、工业化种养殖模式、全链条管理种养模式、高品质种养模式、吨粮模式
	互联网农产品销售模式、农业众筹模式、订单认养模式、共享农业模式
	生态园区综合体模式、庄园采摘模式、科技农业园区模式

$$\mathrm{XPATH}_{\text{农业农村部}} = //\mathrm{span}[@\mathrm{class}="\mathrm{fbsj}"] \tag{1}$$

$$\mathrm{XPATH}_{\text{央视新闻}} = //\mathrm{div}[@\mathrm{class}="\mathrm{src-tim}"]//\mathrm{span}[@\mathrm{class}="\mathrm{tim}"] \tag{2}$$

$$\mathrm{XPATH}_{\text{人民网}} = //\mathrm{span}[@\mathrm{class}="\mathrm{tip-pubtime}"] \tag{3}$$

$$\mathrm{XPATH}_{\text{新华网}} = //\mathrm{div}[@\mathrm{class}="\mathrm{easynews}"]//\mathrm{div}[@\mathrm{class}="\mathrm{newstime}"]//\mathrm{span}] \tag{4}$$

（2）空间信息抽取算法

空间信息抽取是指从新闻报道中获取农业生态文明模式的空间位置信息。数据集研发

过程中，首先使用 NLPIR 工具集[5]识别地名，例如识别出“庄浪县以发展绿色……”句中的地名“庄浪县”；然后利用百度地理编码服务，解析地名对应的经纬度信息[6]。需要注意的是，研发数据集的解析精度参数设置为 100 m，坐标系为百度坐标（BD09）。

（3）模式抽取算法

模式抽取是从新闻报道文本中获取农业生态文明模式的描述文字。数据集采用基于正则表达式的模式匹配方法，其正则表达式分为两类：有触发词类和无触发词类。有触发词类是利用触发词等文本字符特征抽取对应的农业生态文明模式，其正则表达式例如“采取了{0,1}“((.)+)”(.)+模式”；无触发词类是利用模式描述的结构特征抽取对应的农业生态文明模式，其正则表达式例如“(“([\u4e00-\u9fa5]+)(—([\u4e00-\u9fa5]+))+”)”。具体规则集合参见数据集研发的开源代码[7]。

（4）农业生态文明模式记录聚合算法

农业生态文明模式记录聚合指将抽取的离散时间信息、空间信息和模式信息进行关联，构建形成农业生态文明模式记录。聚合算法的基本原理是句内、段内和上下文中语义描述具有连贯性，因而可以将句子内部的时间、空间和模式信息进行关联，缺省信息按照句、段、篇章顺序依次填补。

4　数据结果与验证

4.1　数据集组成

数据集的.xlsx 文件由 33,440 条农业生态文明模式记录构成，每条记录包含 22 个字段（表 3）。

表 3　数据集中的记录字段表

序号	字　段	序号	字　段
1	ID（序号）	12	MODE_TYPE_LEVEL_1_ZH（农业生态文明模式一级分类_中文）
2	DATASOURCE_ZH（数据来源_中文）	13	MODE_TYPE_LEVEL_1_EN（农业生态文明模式一级分类_英文）
3	DATASOURCE_EN（数据来源_英文）	14	MODE_TYPE_LEVEL_2_ZH(农业生态文明模式二级分类_中文)
4	URL（文本链接）	15	MODE_TYPE_LEVEL_2_ EN(农业生态文明模式二级分类_英文)
5	TITLE_ZH（文本标题_中文）	16	EXTRACT_MODE_ZH (抽取原始农业模式描述_中文)
6	TITLE_EN（文本标题_英文）	17	EXTRACT_MODE_ EN (抽取原始农业模式描述_英文)
7	REPORT_DATE（报道时间）	18	KEYWORDS_ZH (文本描述关键词_中文)
8	LOCATION_ZH（地名位置_中文）	19	KEYWORDS_ EN (文本描述关键词_英文)
9	LOCATION_ EN（地名位置_英文）	20	CONTENT (文本正文内容),
10	LNG（经度）	21	SHORT_SENTENCE (描述农业模式子句)
11	LAT（纬度）	22	LONG_SENTENCE (描述农业模式长句)

[5] NLPIR 工具集. http://ictclas.nlpir.org/.

[6] 解析地名对应的经纬度信息，其地理编码服务网址为 http://api.map.baidu.com/geocoding/v3/?address=庄浪县&output=json&ak=ak&callback=showLocation.

[7] 开源代码网址. https://github.com/shuwang8951/EcoCivMdl.

数据集的.shp 文件将.xlsx 文件中的数据记录，在空间上以点数据模型进行存储。

4.2 数据结果

数据集共包含 72 类农业生态文明模式，数量最多的前十类农业生态文明模式为种养结合模式、畜沼果模式、稻鱼模式、生态园区模式、农业+互联网模式、立体循环种养模式、畜粪资源利用模式、水肥一体化模式、秸秆循环利用模式、林草畜模式。以种养结合模式为例，其点状空间分布如图 2 所示，图中每一个点代表一次种养结合农业生态文明模式的出现。为更加清晰地展现种养结合模式在中国的空间分布，可以将种养结合模式的点状数据经过核密度计算，其核密度空间分布如图 3 所示。其中，黑色三角形表示种养结合农业生态文明模式的出现区域，黑色三角形越大表示该地区种养结合模式被报道的频次越多。由此可见，数据集经过数据处理和可视化能够清晰揭示农业生态文明模式的空间分布。例如图 3 揭示了我国种养结合农业生态文明模式分别在吉林中部、宁夏北部、山东北部、湖北南部、四川中东部等地区形成群聚效应。

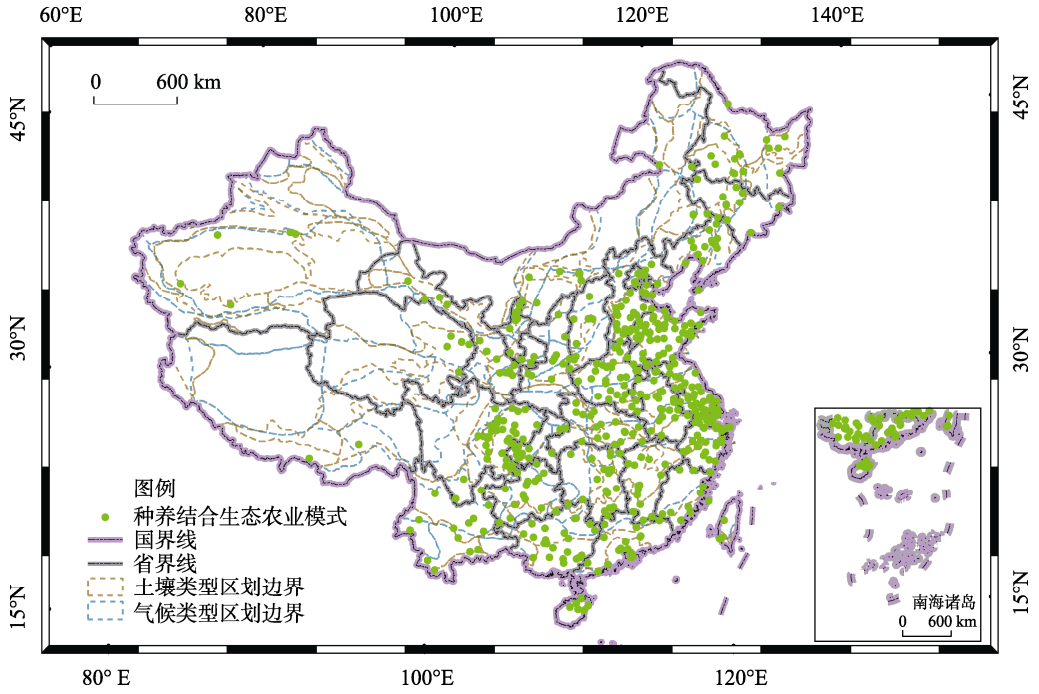

图 2 种养结合农业生态文明模式的点状空间分布图

（依据审图号 GS(2020)4630 号的标准地图制作）

4.3 数据结果验证

作者从两方面对数据集结果进行验证：第一，信息抽取过程，以检验信息抽取过程中的准确性；第二，模式抽取结果，以检验数据集中农业生态文明模式的覆盖度。

针对信息抽取过程的检验，本文从数据集中随机选取 150 条记录，对新闻报道中提及的时间、空间和模式进行人工标注，对比人工标注结果与机器自动抽取结果，得到数据集生产过程中的准确率，参见表 4。

为验证数据集中农业生态文明模式的覆盖度，本文选取一种农业生态文明模式（生态园区模式），参考当前休闲农业园区的典型试点区域，将数据集结果与典型试点区域对比，统计研发数据集的覆盖度。其中，当前休闲农业园区的典型试点区域由两部分构成：其一，农业农村部公布的全国农村创业创新园区（基地）中休闲旅游相关园区[9]，共计 47 个；第二，休闲农业资源相关文献中的典型示范园区[10]，共计 54 个。数据集中生态园区模式在两类园区列表中的覆盖度参见表 5，县级和市级园区的平均覆盖度分别为 87.13%和 92.08%。

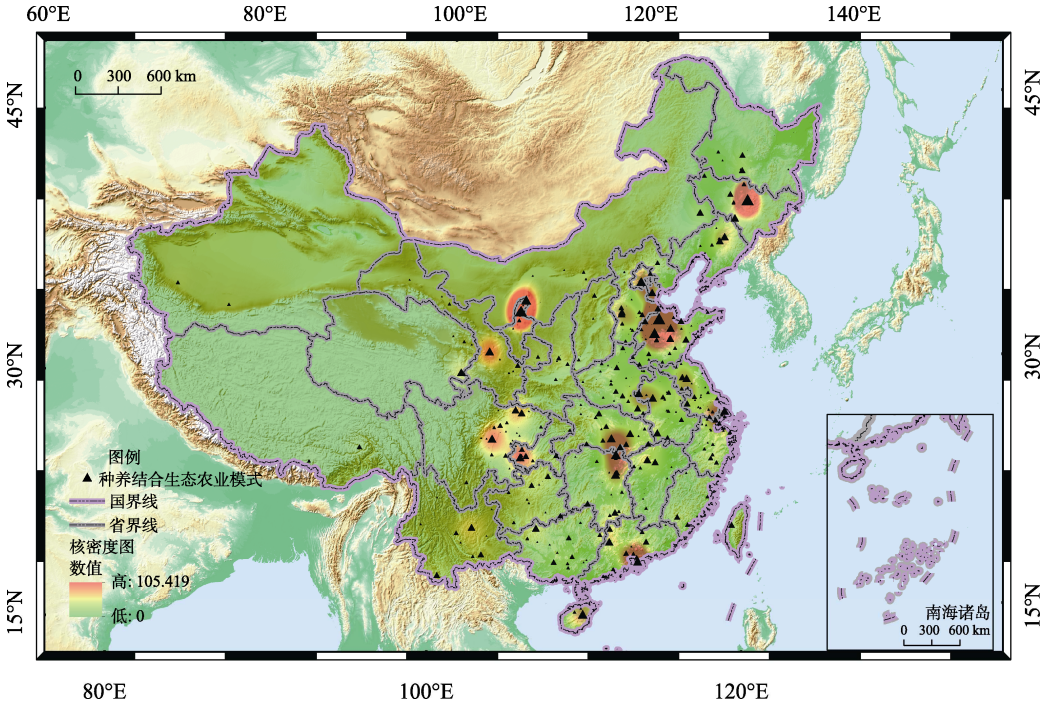

图 3　种养结合农业生态文明模式的核密度图

（依据审图号 GS(2020)4630 号的标准地图制作）

表 4　农业生态文明模式信息抽取准确率统计表

抽取类型	抽样记录数	人工检验错误数	准确率
时间信息	150	0	100%
空间位置	150	7	95.3%
模式描述	150	8	94.7%

表 5　生态园区模式记录在典型试点区域的覆盖度统计表

对照园区	县级覆盖度	市级覆盖度
全国农村创业创新园区（基地）	87.03%	92.59%
休闲农业/观光农业典型示范园区	87.23%	91.49%
平均	87.13%	92.08%

5 结论

为揭示中国农业生态文明模式的空间分布情况，作者汇集了2018–2020年间中国农业生态文明模式的新闻报道，采用自然语言处理等技术，将报道蕴含的农业生态文明模式进行抽取分类，解析获取农业生态文明模式的相关报道时间和空间位置，研发了中国农业生态文明模式的点状空间分布数据集。本数据集从宏观的全国尺度揭示中国农业生态文明模式的空间分布情况，能够作为基础数据支撑农业生态文明模式孕育环境、形成机理和发展规划等各类研究，并且能够为区域农业模式的发展方向提供参照，具有现实指导意义。

作者分工：诸云强对数据集的开发做了总体设计；王曙和钱朗采集和处理了农业生态文明模式新闻报道数据；王曙、宋佳和袁文设计了模型和算法；王曙做了数据验证；王曙撰写了数据论文。

利益冲突声明：本研究不存在研究者以及与公开研究成果有关的利益冲突。

参考文献

[1] Xu, C. Comparative study of Chinese ecological agriculture and sustainable agriculture [J]. *International Journal of Sustainable Development* & *World Ecology*, 2004, 11(1): 54–62.

[2] 尹昌斌, 程磊磊, 杨晓梅等. 生态文明型的农业可持续发展路径选择[J]. 中国农业资源与区划, 2015, 36(1): 15–21.

[3] 刘宗超, 贾卫列. 生态文明理念与模式[M]. 北京: 化学工业出版社, 2015: 82–87.

[4] 农业部科技司. 中国生态农业十大模式和技术[J]. 农业环境与发展, 2003(1): 16.

[5] Li, M., Zhang, Y., Xu, M., *et al*. China eco-wisdom: a review of sustainability of agricultural heritage systems on aquatic-ecological conservation [J]. *Sustainability*, 2020, 12(1): 60.

[6] Wang, X. M. Study on the problems of Chinese organic agriculture development history and present situation [C]. International Conference on Advanced Educational Technology and Information Engineering (AETIE). Beijing, 2015: 984–989.

[7] 王曙, 诸云强, 钱朗等. 中国农业生态文明模式及其空间分布数据集（2018–2020）[J/DB/OL]. 全球变化数据仓储电子杂志, 2021. https://doi.org/10.3974/geodb.2021.06.02.V1. https://cstr.escience.org.cn/CSTR:20146.11.2021.06.02.V1.

[8] 全球变化科学研究数据出版系统. 全球变化科学研究数据共享政策[OL]. https://doi.org/10.3974/dp.policy.2014.05 (2017年更新).

[9] 王甫园, 王开泳, 陈田. 国家级休闲农业园区的分布、类型与优化布局[J]. 农业现代化研究, 2016, 37(6): 1035–1044.

[10] 包乌兰托亚. 我国休闲农业资源开发与产业化发展研究[D]. 青岛: 中国海洋大学, 2013: 175–177.