【统计应用研究】

基于改进的自适应传播模型的农业风险区划分析

谢远涛[1],杨 娟[2],刘皓宇[3]

(1. 对外经济贸易大学 保险学院,北京 100029;
2. 中国科学技术发展战略研究院,北京 100038;3. 德勤中国,北京 100738)

摘要:农业险定价中的核心问题是农业风险区划问题,为了体现农业区划中个体指标的动态发展特征,根据近邻传播改进自适应近邻传播聚类方法对数据进行优化,基于轮廓系数、归属度和吸引度得到最佳聚类中心和几何聚类中心,并将聚类转化为新数据集的聚类问题;选取代表性的棉花为例进行实证分析,通过计算生产、销售、收入、财政等指标进行棉花风险区划实例分析,计算最优棉花风险区划,结果表明对于具有动态特征的数据,本模型具有很好的有效性、实用性和解释性。

关键词:面板数据聚类;近邻传播;自适应近邻传播;聚类中心

中图分类号:C81 文献标志码:A 文章编号:1007-3116(2017)01-0033-08

一、引 言

(一)背景

中国在 2007 年推行政策性农业保险,2015 年农业保险原保险保费收入为 374.7 亿元,仅占总保费收入的 1.56%,农业保险的推广程度和发展水平远不及寿险、意外险和其他类别的产险,农业精算风控还有很远的路要走。

目前,中国没有专门的农业保险定价管理规定,对农业保险定价问题进行明确规定的法规文件散见于《中央财政种植业保险保费补贴管理办法》(财金[2008]26 号)、《中央财政养殖业保险保费补贴管理办法》(财金[2008]27 号)、《财政部关于进一步加大支持力度做好农业保险保费补贴工作的通知》(财金[2012]2 号)、《农业保险条例》(国务院令第 629 号)、《中国保监会关于加强农业保险条款和费率管理的通知》(保监发[2013]25 号)、《关于进一步完善中央财政保费补贴型农业保险产品条款拟订工作的通知》(保监发[2015]25 号)等。

国内的农业保险以补偿物化成本为补偿目标,对补贴险种按"低保障、广覆盖"来确定保障水平,其费率首先由保险公司根据保险责任、保险标的多年平均损失情况、地区风险水平等因素确定,然后由各地财政、农业、林业部门结合本地的财力、风险水平等多方面因素审核确认,并广泛听取农民代表的意见,最后由中国保监会审批或备案。农业保险的费率为单一费率,全县甚至全省采取统一的费率,而这种费率系统未能结合各地区的风险差异和农户的风险差异进行区别对待。这样,一方面在过去 20 年间农业保险承担的平均赔付率达到 120%,与较低水平的保费收入不相称,加之运营和管理费用,中国绝大部分保险公司在农业保险业务上都是亏损的,考虑到中国自然灾害频发,保险公司仅靠保费收入和有限的财政补贴难以承受如此巨大的赔付额;另一方面很多地方农业保险保障水平与直接物化成本之间存在一定差距,2012 年全国农业保险保障水平与

收稿日期:2016-08-19;修复日期:2016-10-24

基金项目:国家自然科学基金项目《风险信息共享背景下的个体风险评估研究》(71303045);对外经济贸易大学"优秀青年学者培育计划"(15YQ09);中国保险学会研究课题《生猪费率厘定系统分析:基于养猪风险管理因子与保险数据的研究》(15HX158);国家社会科学基金重大项目《巨灾保险的精算统计模型及其应用研究》(16ZDA052)

作者简介:谢远涛,男,湖北随州人,经济学博士,副教授,研究方向:非寿险精算与统计模型;
杨 娟,女,湖北武汉人,经济学博士,助理研究员,研究方向:统计模型;
刘皓宇,男,江苏镇江人,应用精算科学硕士,德勤中国精算服务副总监,研究方向:精算和风险管理。

直接物化成本之间的平均差额为35%左右(财政部数据),从而难以实现"保障农户灾后恢复生产"这一出发点。

对全球农业保险发达国家的经验进行总结会发现,农业风险区划是农业保险费率厘定的科学基础之一。美国和加拿大等发达国家都实行了农业风险区划,将面临相同风险、种类、发生频率强度以及时间空间分布的农作物,根据受影响程度按照一定原则在地域上加以区别,并将风险相同或相近的地域划分在一起作为同一个风险区而进行风险区域划分,以便于控制和管理风险。农业风险因气象、地质灾害的特点,其风险单位在灾害损失中常表现为时间和空间的相关性,故农业风险具有高度相关性,并且农业巨灾风险明显不符合"理想可保风险"的准则。农业保险作为主要的农业风险管理手段已经显现出诸多劣势和不足之处,而对不同区域的农作物进行风险区划,才是满足保险的风险一致性原则,同时也能避免逆向选择等问题。

农业风险区划本质上是聚类分析。传统聚类分析有两种:一种是最常见的截面数据聚类,可解决空间的依赖性,但是没有办法解决时序上的连贯性;另一种是时间序列聚类,解决了时序上的连贯性,但是没有解决空间上的依赖性。经典文献中聚类分析的结果不能体现这种发展的阶段性,因此不适用于分析具有明显动态特征的面板数据。保险是一个系统长期的风险管控方案,即使2013年的聚类结果和2014年的聚类结果不同,也很难快速调整风险区划,而风险区划允许调整,但是需要时序上具有一定的稳定性。农业数据获取困难,往往需要综合多年的数据进行分析,而使用面板数据进行分析则是最好的解决方案。

针对传统面板数据聚类的一些弊端,本文改进了传统的面板数据聚类方法,通过对粮食单位面积产量变异系数、受灾减产率、农业生产专业化水平、农业生产效率以及抗灾能力等几个反映农业风险的因素进行分析和风险区划,结合风险区划的结果探讨最终农业巨灾保险费率,以望能为中国的农业巨灾风险管理和转移提供一定的建议。

(二)文献综述

中国国内关于风险区划和风险测度的研究已经具有一定积累[1]。庹国柱等对农业保险区划的必要性和理论依据进行了研究,认为划分风险区域的具体指标是作物产量水平、产量变异系数、灾害发生频率和强度指标、气候综合评判值、地理指标、土壤等级、水利设施指标、作物结构以及其他经济技术条件指标等[2];周延等运用粮食单产变异系数、因灾减产强度和地区抗旱能力这三个指标对中国各省份农业生产水平进行了风险区划,并结合粮食单产产量异常波动率,对农业巨灾保险进行了费率厘定[3];于洋基于非参数核密度估计研究了农作物产量保险区域化差别费率厘定[4]。以上分析主要使用了截面数据特征进行分析,但其本身并不是面板数据分析方法。

在面板数据聚类方面,Bonzo等基于概率连接函数定义相似系数,采用改进的自适应模拟退火—遗传算法优化目标函数[5];Nie等将不同时期的观测给予不同权重并构造了距离函数[6];还有学者将单指标面板数据转化为截面数据并进行了聚类分析;任娟等基于形状特征提炼多指标信息,构建了面板数据聚类方法[7];杨娟等基于密度构建了面板数据聚类分析[8]。

以上文献通过提取面板数据的数字特征(均值、方差)、几何特征(位置、形状)及波动特征来构建面板数据的相似性度量,但是聚类结果无法体现个体阶段性发展的动态特征。为了解决这个问题,本文在基于自适应近邻传播算法(Adaptive Affinity Propagation Clustering,ad-AP)的基础上,优化面板数据得到了个体最佳聚类中心并组成了新数据集,再将面板数据聚类问题转化为新数据集的聚类问题,并以棉花为实例进行面板数据聚类分析,从而得到了较好的聚类结果。

本文的创新点是:用ad-AP方法分别计算每个个体的最佳聚类中心,而最佳聚类中心为个体的某些样品能够代表个体的不同发展阶段,并对具有明显动态特征的面板数据进行聚类分析,由此提供了一种新的途径。

二、自适应近邻传播聚类

(一)近邻传播聚类(AP)

Frey等提出的基于聚类中心的近邻传播聚类算法(Affinity Propagation,AP),适用于截面数据聚类分析,聚类结果为族(或类)和相应的聚类中心[9],其思想是将所有的个体作为潜在的聚类中心,通过计算个体之间的实值消息传递,用信息传播方法产生高质量的聚类中心,并且产生相应的族。

截面数据用X_{im}表示,包含N个截面个体和M个指标或变量,其中$X_{im}\in D, i=1,2,\cdots,N, m=1,2,\cdots,M$,任意的截面个体记为个体$i$,指标记为$x_{im}$。定义两种消息:吸引度记为$r(i,k)$,从个体$i$传播到

候选聚类中心 k，表示 k 作为 i 的聚类中心的合适程度；归属度记为 $a(i,k)$，从候选聚类中心 k 传播到个体 i，考虑了其他数据点的支持，表示 i 选择 k 作为聚类中心的合适程度，所有的个体能否作为聚类中心，取决于接受或发送的 $a(i,k)$ 和 $r(i,k)$；变量 C_i 表示个体 i 的聚类中心，$\hat{c}_i=k$ 表示 i 分到了某一族中，k 为 i 的聚类中心；$\hat{c}_k=k$ 表示 k 为聚类中心；$c_i\in K\subseteq\{1,2,\cdots,N\}$，聚类中心的集合为 $K\subset\{1,2,\cdots,N\}$；非聚类中心集合为 $\overline{K}\subseteq\{1,2,\cdots,N\}\backslash K$；相似系数 $s(i,k)$ 表示 k 作为 i 的聚类中心的合适程度，通过最大化相似个体和聚类中心的相似性来得到族，相似系数可以定义为 $s(i,k)=-\|\boldsymbol{x}_i-\boldsymbol{x}_k\|^2$ $(i\neq k)$，也可以用其他的模型来定义；$s(k,k)$ 表示个体 k 成为聚类中心的一个优先条件，令 $s(k,k)=p$；偏向参数 P 需要提前设定，且 $k=1,2,\cdots,N$。

AP 算法的基本步骤如下：

输入：$\{s(i,k)\}$

输出：$\hat{\boldsymbol{c}}=(\hat{c}_1,\cdots,\hat{c}_N)$

步骤一：令 $\forall i,k,\ \ a(i,k)=0$。

步骤二：计算 $a(i,k)$ 和 $r(i,k)$ 直到收敛。

$\forall_i,k$：$a(i,k)$

$$=s(i,k)-\max_{k',k'\neq k}[s(i,k')+a(i,k')]$$

$\forall_i,k$：$a(i,k)$

$$=\begin{cases}\sum_{i,i'\neq i}\max[0,r(i,k')] & k=i\\ \max\left[0,r(k,k)+\sum_{i,i'\notin\{i,k\}}\max[0,r(i,k')]\right] & k\neq i\end{cases}$$

$$\hat{c}_i=\arg\max_k[a(i,k)+r(i,k)]$$

步骤三：重复步骤二，直到结果收敛。

$\forall_i,k,s(k,k)=p$，随着 P 值增大，越多的类代表将成为最终的聚类中心，但是 P 与 AP 聚类的类数不是一一对应关系（Frey 证明了 k 和 P 不是一一对应的关系），因此很难用 P 作为最佳聚类结果的标准。

Bodenhofer 等人根据以上原理编写了 AP 的 R 程序包[10]。AP 方法十分灵活便于自由定制，广泛应用于图像聚类和生物信息领域。具有多阶段发展特征的面板数据，在进行聚类分析问题中体现出了多阶段特征，即可以选择几何聚类中心也可以选择某一样品作为聚类中心，而从解释能力上看选择样品作为聚类中心的聚类结果说服力更强。用 AP 进行聚类分析，其聚类中心为个体的样品能够很好地解释个体的发展与阶段。

基于 AP，本文设计了两种面板数据聚类方案：

第一种，将面板数据看作截面数据，用 AP 进行聚类分析将会产生两类问题：一类是如何将同一个体的样品划分正确，即将同一个体的样品划分到不同的类中，因同一类中包含了多个个体的样品；另一类是如何找到每个个体的最佳聚类中心，因参数的设置不同而聚类中心也不同，从而导致聚类结果不一致并缺乏解释性。

第二种，将面板数据拆分为截面数据，分别进行聚类分析，聚类中心组成新数据集，将面板数据聚类转化为新数据集的聚类问题。由于 AP 中参数的设置不同而聚类中心也不同，最终面板数据的聚类结果也不一致并缺乏解释性。

因此，虽然 AP 能够计算出聚类中心，但聚类结果不一致，并缺乏解释性，故需对 AP 进行改进。

（二）自适应近邻传播聚类（ad-AP）

王开军用自适应的方法改进了 AP 算法，解决了如何产生最佳聚类结果的问题[11]。

ad-AP 通过自适应方法，计算 AP 所有可能的聚类结果，然后用轮廓系数求得最优聚类结果。AP 聚类结果的类的个数 k 依赖于偏向参数 P 给定的数据集，AP 算法中偏向参数 P 与聚类结果个数 k 不是一一对应的，因此 ad-AP 通过扫描 P 的参数空间来搜索聚类个数空间可得到一系列聚类结果，其中采用的技术有自适应扫描技术、下降步幅选择技术、确定扫描区间技术、扫描加速技术。

聚类有效性技术是评价聚类结果质量的有效方法，对于某个聚类算法产生的一系列聚类结果，用聚类有效性指标找到最优的聚类结果。Kaufman 等提出的轮廓系数，同时考虑了类间的可分性和类内部的紧密性，轮廓系数对聚类结构有良好的评价能力，ad-AP 用轮廓系数来求得最佳聚类结果[12]108－117。

三、基于 ad-AP 面板数据聚类设计

面板数据记为 $X_{i(t)m}$，表示具有 N 个个体连续观测 T 个时间，并包括 M 个指标或变量的面板数据，其中 $i=1,2,\cdots,N$、$t=1,2,\cdots,T$、$m=1,2,\cdots,N$。任意的个体称为个体 i，其样品记为 $i(t)$，第 m 个指标记为 x_m，所有指标记为 $\boldsymbol{x}_i$。例如北京 2011 为样品，记为 $i(2011)$，第 m 个指标记为 $x_{i(2011)m}$，所有指标记为 $\boldsymbol{x}_{i(2011)}$。

基于 ad-AP，本文设计了面板数据的聚类方法如下：

用 ad-AP 分别处理每一个个体，得到个体 i 的最佳聚类结果，对应的最佳聚类中心用 $\{i(t^*)\}$ 表示，最佳聚类中心的数目用 K_i^* 表示；所有个体的最佳聚类中心组成一个新的数据集，用 X^* 表示，并将 X^* 作为截面数据，选择合适的截面数据聚类方法进行分析。基于 ad-AP，本文将面板数据的聚类问题转化为数据集 X^* 的聚类问题。

本文设计的面板数据聚类方法有如下特点：

第一，本文用 ad-AP 分别计算个体的最佳聚类中心，对于个体 i 最佳聚类中心 $\{i(t^*)\}$ 为其样品，而不是其他个体的样品；第二，由于 ad-AP 解决了最佳聚类结果的问题，使得个体 i 的最佳聚类中心 $\{i(t^*)\}$ 是唯一确定的，保证了数据集 X^* 的唯一性。如果用 AP 计算个体的聚类中心则不能确保数据集 X^* 的唯一性，将导致最终的聚类结果随着 X^* 的变化而变化；第三，每个个体至少存在两个最佳聚类中心[①]，即 $K_i^* \in K[2, \sqrt{T_i})$，$T_i$ 为个体样品的个数，因此 X^* 为非平衡面板数据；第四，将 X^* 作为截面数据进行聚类分析，其结果可能为同一个体的最佳聚类中心被划分到不同的族，这样的结果说明该个体的指标具有明显的动态特征，体现了各个变化阶段和其他个体的关系。同时，还存在这样的情况，即同一个体的最佳聚类中心被划分为同一族，说明该个体指标的动态特征不明显，虽然指标发生了变化，但仍然处于同一族的变化范围内。

根据面板数据的聚类目的选择聚类方法，如果聚类目的是分析个体动态特征，可将 X^* 看作截面数据，可用截面数据聚类方法进行分析；如果要分析个体的平均发展情况，计算每个个体的几何聚类中心 $\{i(\bar{t}^*)\}$，指标 $x_{i(\bar{t}^*)m} = \overline{x}_{i(t^*)m}$，所有个体的几何聚类中心组成截面数据 $\overline{X}^*$，可进行截面数据的聚类分析。

基于 ad-AP 的面板数据聚类分析步骤为：

输入：面板数据

输出：聚类结果

步骤一：用 ad-AP 方法计算每个个体的最佳聚类结果，对应的聚类中心记为 $\{i(t^*)\}$，所有个体的最佳聚类中心组成数据集 X^*。

步骤二：如需分析个体的动态特征，可将 X^* 进行截面数据聚类分析得到聚类结果；如需分析个体动态特征的平均水平，可计算 $\{i(t^*)\}$ 的几何中心 $\{i(\bar{t}^*)\}$，组成截面数据 $\overline{X}^*$，再进行截面数据聚类分析得到聚类结果。

四、农业区划分析

（一）测度指标的选取与定义

中国是世界上棉花产量最大的国家，棉花具有重要的经济价值，也容易受到自然灾害的影响。本文选取各省份 2010—2015 年棉花生产、产量数据[②]、农民人均收入、财政收入等指标，原始数据来源均为《中国统计年鉴》(2010－2015)及国家统计局网站数据库。

1. 棉花单产变异系数。该指标是自然灾害和人为因素影响的综合体现，测度的是棉花单位面积产量的年度波动幅度。该值越大，则说明该省份棉花生产和产量越不稳定，风险越大；反之，则说明棉花生产越稳定，风险越小，其公式为：

$$CV_i = \frac{\sqrt{\sum (Y_{it} - \overline{Y}_{it})^2 / (t-1)}}{\overline{Y}_{it}}$$

其中 CV_i 为 i 省棉花单产变异系数，Y_{it} 为 i 省第 t 年棉花单位面积产量。

2. 棉花种植规模指数。该指标是衡量棉花种植风险的重要指标，反映该省份的棉花种植规模与全国平均水平的比率。一般来说，种植棉花的规模越大，即种植面积占总播种面积比率越高，发生灾害后棉花发生损失的风险和损失额就越大，即：

$$SI_{ic} = \frac{PS_{ic}/PS_c}{PS_i/PS}$$

其中 SI_{ic} 为 i 省棉花种植规模指数，PS_{ic}/PS_c 是 i 省棉花种植面积占全国棉花种植面积的比率，PS_i/PS 是 i 省农作物总播种面积与全国农作物总播种面积的比率。因此，$SI_{ic} > 1$ 表示 i 省在棉花种植规模上大于全国平均水平；反之，$SI_{ic} < 1$ 表示 i 省棉花种植规模相对较小。

3. 棉花单产效率指数。棉花单产产量与全国平均单产相比能反映出该省份棉花种植的成功率和效率，同时也能间接反映出棉花种植的条件质量情况，如土地质量、肥料、气象情况等，即：

$$EI_{ic} = \frac{AP_{ic}}{AP_c}$$

其中 EI_{ic} 代表 i 省棉花单产效率，为 i 省棉花单产与全国棉花单产之比，此指数能反映出与全国平均水平相比该省份棉花生产效率所处的水平。

①ad-AP 中聚类数目 K 的扫描区间为 $(2, \sqrt{N})$，N 为个体数目。

②不包括不生产棉花的黑龙江、广东、海南、青海、西藏自治区以及台湾省、香港特别行政区。

4. 公共财政收入比重。政府的公共财政收入决定其对农业生产的补贴和支持的能力，它是当地政府为农村基础建设投资、农民种植农作物补贴和灾害预防准备的重要金融财政基础，同时对于灾害应急和灾后处理方面都有着巨大的影响，与农作物生产风险水平有着密切的联系，此指标的公式为：

$$PFR_{it}=\frac{FP_{it}}{FP_t}$$

其中 PFR_{it} 为 i 省 t 年公共财政收入占全国总公共财政收入的比重。

5. 农民人均纯收入水平。农业风险区划不仅需要考虑农作物所面临的各种风险，同样也需要考虑当地的风险承受能力，包括当地的经济发展情况和政府的财政能力。农民人均纯收入水平越高，潜在的由农业风险造成的收入损失部分就越小，同时农民对于农作物种植损失的敏感性就越低。一般来说，人均收入水平越高，农民的风险厌恶程度就越低，进而影响其进行各种保险及风险防范措施的运用。用 $PCNIL_{it}$ 表示 i 省 t 年的农民人均纯收入与全国水平之比，即：

$$PCNIL_{it}=\frac{PCNI_{it}}{PCNI_t}$$

式中 $PCNI_{it}$ 为 i 省 t 年农民人均纯收入，$PCNI_t$ 为 t 年全国农民纯收入。

6. 地区生产总值（GDP）比重。地区生产总值也是风险区划中的一个重要指标，反映的是一个地区的经济发展状况和财富总值，而此指标会影响农业生产的方方面面，比如相关技术发展水平和相应政策的制定实施。地区生产总值越高，一般来说抵御巨灾风险的能力就越强，其公式为：

$$PGDP_{it}=\frac{GDP_{it}}{GDP_t}$$

其中 GDP_{it} 是 i 省 t 年地区生产总值，GDP_t 是 t 年中国国民生产总值。

7. 农业固定资产投资比重。在农业固定资产方面的投资，是提高农业生产状况和标准与达成农业经济长期稳定快速发展的基础。农业固定资产投资一般来自政府的财政拨款或者农民自身的财富积累，其投资包括对农业生产基础设施的维修、更新和改进、生产加工设备的购买、厂房的建设和生产技术的创新等，而这些投资能大大增加农业生产的效率和稳定性并提高农民的工作效率。此指标如何影响风险区划则还有待进一步分析，其公式为：

$$PAIFA_{it}=\frac{ALFA_{it}}{ALFA_t}$$

其中 $ALFA_t$ 为 i 省 t 年的农业固定资产投资金额，$ALFA_t$ 为 t 年全国农业固定资产投资金额。

上述指标中，计算棉花单产变异系数的原始数据为 2010—2015 年中国各省份棉花单位面积产量，其他 6 个指标则是由 2010—2015 年间相关数据计算所得出的平均值。运用这些指标进行因子分析和聚类分析，可对各省份棉花生产进行风险区划。为了抵消单位不同造成的影响，采用指数功效函数方法。

（二）聚类中心的选择

为了验证本文设计的基于 ad-AP 的面板数据聚类方法，2010—2015 年中每个省市的发展都呈现了多阶段性特点。按照基于 ad-AP 面板数据聚类方法和步骤，用 R 软件编写了相应的面板数据聚类程序，分别计算最佳聚类中心 X^*、几何聚类中心 $\overline{X}^*$ 以及相应的聚类结果和经济学意义。

1. 最佳聚类中心。如果聚类的目的是查看个体发展的动态特征，采用 ad-AP 方法计算每个个体的最佳聚类结果。根据 2010—2015 年每个省市的相关指标，用 ad-AP 计算每个省市的最佳聚类结果，如表 1 所示。每个省市均对应两个聚类中心，说明 31 个省市的发展均可分为两个阶段，北京的最佳聚类中心为其对应的样品北京 2010 年和北京 2014 年，上海对应的最佳聚类中心为样品上海 2011 年和上海 2014 年。

表 1　31 个省市的最佳聚类中心表

最佳聚类中心	省市
2010、2014	北京、陕西、山东、新疆、广东、辽宁、山西、重庆、宁夏、甘肃、江苏、青海、西藏
2011、2014	上海、四川、浙江、天津、黑龙江、湖南、贵州、河北、安徽、吉林、江西、河南、湖北、云南
2009、2014	海南、福建、内蒙古
2010、2015	广西

数据集 X^* 由 ad-AP 方法计算得到，将 X^* 看作截面数据，本文使用 AP 方法对 X^* 进行聚类分析，8 个聚类中心分别为：北京 2014、江苏 2014、湖南 2011、上海 2011、贵州 2014、湖北 2014、辽宁 2014、宁夏 2014。

如果按照 8 个聚类中心将省市分为 8 个族，结果显得十分琐碎。考虑划分为 5 个簇，通过 AP 聚类，上海的两个最佳聚类中心上海 2011 和上海 2014 被划分为族 II，表明虽然上海的发展经历了两个阶段，但从全国范围看来上海的发展水平明显区别于其他大部分省市，因此被划分为同一族。

2.几何聚类中心。如果聚类目的是查看个体阶段性特征的平均水平，需要计算每个个体的几何聚类中心，组成一个截面数据，再使用 AP 对该截面数据进行聚类分析。聚类结果的树状图如图 1 所示，总共有 6 个聚类中心，分别为北京、天津、江苏、湖南、广西、云南，每一族包含的省市如表 2 所示。

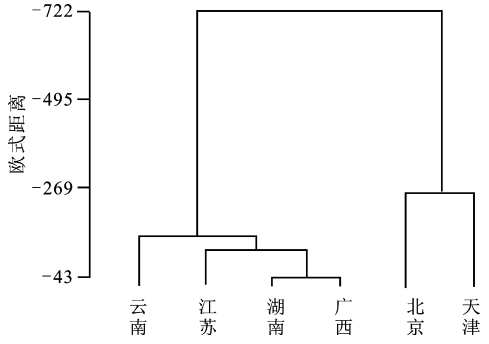

图 1　几何聚类中心的 AP 聚类树状图

虽然每个省市都经历了两个发展阶段，但是基于几何聚类中心的聚类结果不能体现这种变化，只能给出这两个阶段的平均水平。表 3 中按照发展指数两个阶段的平均水平，北京和上海划分为族 I，与表 2 的结果对比，北京 2014 年单独分为一个族，说明与北京 2010 年的阶段相比，北京 2014 年这个阶段的发展指数明显提高，体现出了北京发展指数的阶段性特征。

表 2　基于几何聚类中心的 AP 聚类结果表

族	聚类中心	省市
I	北京	北京、上海
II	新疆	新疆
III	江苏	江苏、浙江
IV	山东	山东、河南
V	宁夏	宁夏 、贵州、云南、广西、重庆、吉林、福建
VI	陕西	其他

(三)聚类结果

为了对各省份的棉花生产和种植进行风险区划，采用因子分析得出的 3 个公共因子为指标进行聚类分析，以便更好地划分不同的风险区域，以确定相关风险区域的风险等级，进而为巨灾债券的定价提供依据。

通过简要分析这 3 个指标的含义和代表的农业生产、支持实际能力，将因子 $F1$ 定义为“金融支持能力因子”，此因子能有效减少棉花产量损失的风险；因子 $F2$ 与 SI(棉花种植规模指数)和 EI(棉花单产效率指数)有着显著的关联，这两个指标也反映了棉花种植所面临的风险，其指标越大，一般就意味着棉花生产所面临的风险就越大，故将因子 $F2$ 定义为“风险暴露因子”；最后一个因子 $F3$ 与 PAIFA(农业固定资产投资比重)有着极大的相关系数，同时也与 PGDP 有一定的关联，故将此因子定义为“农业生产水平因子”，此因子越大，就意味着该地区有更大的风险承担能力和较小的农业生产风险损失概率。

表 3　各省份棉花生产综合风险得分表

省份	排名	聚类	省份	排名	聚类
北 京	22	I 区	山 东	24	II 区
天 津	4	III 区	河 南	20	II 区
河 北	15	III 区	湖 北	9	III 区
山 西	6	III 区	湖 南	11	III 区
内蒙古	8	III 区	广 西	16	III 区
辽 宁	18	III 区	重 庆	17	III 区
吉 林	10	III 区	四 川	19	III 区
上 海	23	I 区	贵 州	14	III 区
江 苏	26	II 区	云 南	12	III 区
浙 江	25	II 区	陕 西	5	III 区
安 徽	13	III 区	甘 肃	2	III 区
福 建	21	III 区	宁 夏	7	III 区
江 西	3	III 区	新 疆	1	IV 区

通过表 3 各省份的综合风险得分，对各省的风险水平进行排名，若某个省的风险综合得分越高，那么该省棉花生产的风险水平就越高，排名就越靠前。可以看到，新疆的综合得分远高于其他省份而高居第一位；北京和上海这两个城市职能和规模都相似的大型城市的棉花种植风险水平非常相近，地理位置相近的省份一般也具有相似的风险水平。根据各省份农业生产和经济发展等方面的状况，能对当地政府、保险行业以及巨灾债券的投资者们有所建议和启示。

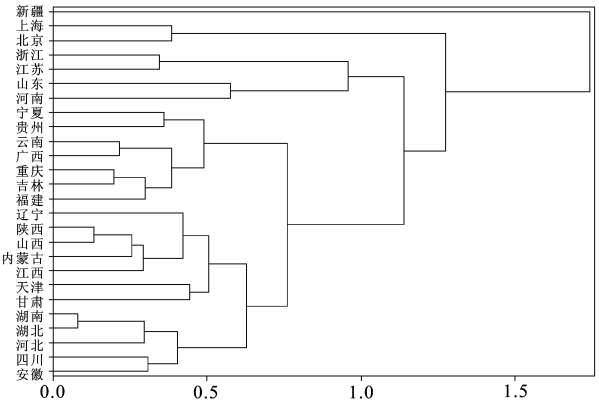

图 2　面板数据聚类树状图

从图 2 结果可看出，聚类分析法得出的聚类结果与因子分析法得出的结果在个别省份上稍有不

同：因子分析法给出了较为明确的综合风险得分和排名；聚类分析法只是对相近风险水平的省份进行分类，能更直观地看出不同省份所处的风险区域。

分析谱系图可以将各省份分为四个风险区域（见图3）[①]：北京、上海为第Ⅰ区域；江苏、浙江、山东、河南为第Ⅱ区域；新疆为第Ⅳ区域；余下的各省份为第Ⅲ区域。再结合因子分析得出的各省份综合因子得分和排名，便完成了对中国各省份棉花生产的风险区划，确定了相应的风险水平和风险等级。

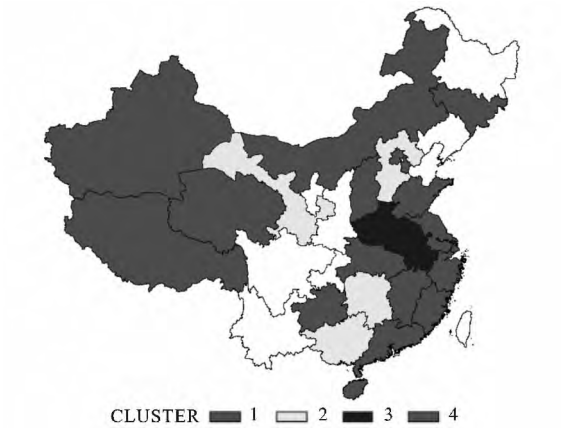

图3　面板数据聚类地图

五、结论与政策建议

（一）结论

通过对中国各个省市进行风险区划和农业巨灾保险费率厘定，可以发现经济较为发达的省市拥有较强的防灾抗灾减灾能力，所面临的风险也越低，经济发展对于降低农业巨灾风险有着非常重要的作用，因为经济发展能够为农民群众提供更高的收入水平，同时政府也能将更多的资源用于农业基础设施的建设，从而能够减少自然灾害对农业生产的影响和减少地区所遭受的损失；同时，风险水平较高的省市具有较高的纯保费，这与所预期的“拥有较高风险的地区应承担较高的保险费率”相一致，基本达到了本文最初的期望。

面板数据聚类的难点在于如何提取面板数据的时序特征、截面特征或总体特征。本文应用ad-AP分别计算每个个体的最佳聚类中心，组成高质量的新数据集X^*，将面板数据聚类问题转化为数据集X^*的聚类问题，本文将新数据集X^*作为截面数据进行聚类分析。

实例分析表明该聚类方法的结果符合逻辑，说明了聚类结果的有效性：ad-AP得到的最佳聚类中心为个体的样品，能够代表个体不同发展阶段，具有很好的解释性；聚类结果体现出个体不同发展阶段与其他个体的关系，具有很好的实用性。因此，对具有明显动态特征的面板数据的聚类分析，本文的方法为之提供了一种新的分析途径。

（二）政策建议

中国农业保险开展实践较短，数据积累不足，数据质量不高；风险区划工作严重滞后，各地的农业风险管控能力较弱；精算研究有待加强，专业人才缺乏；行业资源未能有效整合，分类费率系统机制没有真正形成；农业保险存在保险制度不完善、保险业务不断萎缩、农民群众缺乏风险意识、道德风险和逆向选择问题较为严重等问题，这些问题本质上可以通过精确的风险区划来规避风险和规范行为。

作为对照，美国的农业保险发展可提供一定的借鉴经验：

1. 以立法促进农业保险数据库建设。美国1938年出台了《联邦农作物保险法》，确定了农业保险精算的地位，美国农业风险数据的积累、风险评估和产品研发工作交给联邦农作物保险公司（FCIC）完成。1996年，美国农业部成立了风险管理局（RMA），建立起了更为详尽的农作物风险数据库，对各地每种农作物保险的赔付率和费用率进行跟踪分析，专门承担起美国农业保险政策及法规的制定、监管、再保险和综合服务等职能，标志着美国农业保险的定价进入了更加科学化和系统化的发展阶段。以下将从数据基础、保险责任、经营费用、费率厘定、费率调整等方面，对美国农业保险定价的主要政策进行对此分析。

2. 建立基于风险区划的分类费率系统。美国农业保险定价确保费率厘定的公平性。在农业保险的费率厘定中，美国农业风险管理局使用了各地农业生产的历史平均损失记录、农民的特征偏好、作物种类、防灾措施等费率因子，使分类费率系统因地而异、因人而异。中国也应该引入给予风险区划的分类费率系统，根据农民的实际风险状况实施风险差别费率，避免和缓解低风险人群对高风险人群的“交

① 本文分析范围仅限于中国内地31个省市和自治区，不包括香港、澳门和台湾地区。本文所绘制的数据统计地图只是示意图，并非行政区划图，且仅限于大陆地区和海南岛，不包括台湾地区，也不包括海南省三沙地区，没有给出南海九段线和东海海域边界。

叉补贴”现象。

3. 建立分类费率系统之上的信度调整机制。美国农业风险管理局引入了信度模型，对长期偏高或偏低的费率水平进行信度调整。例如 2011 年 11 月，风险管理局对玉米和大豆的保险费率进行了适当的下调；中国在商业车险中引入了完备的信度调整机制，其他险种限于数据的信度不够，没有实现信度调整。随着农业险数据的积累，终将过度到信度定价阶段，但可能造成两个阶段使用的费率因子交叉，出现过度奖惩效应，则需要结合分类费率模型和信度定价建立统一分析框架，构建广义线性混合模型及其扩展模型[13]。

4. 政府补贴的同时公平兼顾效率。为了激励商业保险公司参与农业保险，美国政府同中国政府一样提供保费补贴。2011 年前，美国政府补贴的平均比例约为 19%，而 2011 年后这一比例下降为 11% 左右；中国政府补贴往往没有考虑区划，政府补贴可能会使保费不公平，从而造成逆选择。

综上所述，可见中国的农业风险分析还需要一个相当的过程。

参考文献：

[1] 张峭，王克. 我国农业自然灾害风险评估与区划[J]. 中国农业资源与区划，2011(3).

[2] 庹国柱，丁少群. 农作物保险风险分区和费率分区问题的探讨[J]. 中国农村经济，1994(8).

[3] 周延，郭建林. 农业巨灾保险风险区划及费率厘定研究[J]. 江西财经大学学报，2011(6).

[4] 于洋. 农作物产量保险区域化差别费率厘定的可行性——基于非参数核密度估计实证[J]. 统计与信息论坛，2013(10).

[5] Bonzo D C，Hennoeilla A Y. Clustering Panel Data via Perturbed Adaptive Simulated Annealing and Genetic Algorithms [J]. Advances in Complex Systems，2002(4).

[6] Nie G L，Chen Y B，Zhang L. Credit Card Customer Analysis Based on Panel Data Clustering [J]. Procedia Computer Science，2012().

[7] 任娟，陈圻. 基于形状特征的多指标面板数据聚类方法及其应用[J]. 统计与信息论坛，2011(10).

[8] 杨娟，谢远涛. 基于密度的面板数据聚类分析[J]. 统计与信息论坛，2014(2).

[9] Frey B J，Dueck D. Clustering by Passing Messages between Data Points[J]. Science，2007(315).

[10] Bodenhofer U，Kothmeier A，Hochreiter S. AP Cluster：An R Package for Affinity Propagation Clustering [J]. Bioinformatics，2011 (27).

[11] 王开军，张军英，李丹，等. 自适应仿射传播聚类[J]. 自动化学报，2007，33(12).

[12] Kaufman L，Rousseeuw P J. Finding Groups in Data：An Introduction to Cluster Analysis [M]. New York：John Wiley & Sons，1990.

[13]谢远涛，李政宵. 基于联合定价模型的奖惩因子的扩展与比较[J]. 统计与信息论坛，2015(6).

Agriculture Risk Regionalization Analysis Based on Panel Data Clustering with Affinity Propagation

XIE Yuan-tao[1]，YANG Juan[2]，LIU Hao-yu[3]

(1. School of Insurance and Economics，University of International Business and Economics，Beijing 100029，China；
2. Institute of Comprehensive Development，Chinese Academy of Science and Technology for Development，
Beijing 100038，China；3. Deloitte China，Beijing 100738，China)

Abstract：Variables for individuals are developed with dynamic characteristics in many panel data sets when we deal with agriculture insurance pricing. In papers for panel data clustering，the similarity coefficients are computed by the numerical character，distribution character，and fluctuant character，but the clustering results cannot reflect the dynamic characteristics. This paper proposes the method to apply adaptive affinity propagation clustering (ad-AP)，which is improved from affinity propagation clustering，to optimize panel data set，and compute the best exemplars of each individual which constitute a new data set. Then panel data clustering analysis is transform into the new dataset clustering analysis. Experimental results on china agriculture insurance show the validity，practicability and interpretability of the design for panel data with dynamic characteristics.

Key words：panel data clustering；affinity propagation (AP)；adaptive affinity propagation (ad-AP)；cluster center

(责任编辑：郭诗梦)