DOI:10.13546/j.cnki.tjyjc.2014.07.022

面板数据多指标聚类和变系数模型的方法与实证

李　峥[1],刘云霞[1,2]

(1.厦门大学 经济学院统计系;2.福建省统计科学重点实验室,福建 厦门 361005)

摘　要:文章通过将面板数据的多指标聚类分析和变系数模型相结合,分析了我国53个国家级开发区2003~2010年的经济运行情况。在聚类结果的基础上,通过建立面板模型总结了各类开发区经济运行的特点,找出了影响各类开发区经济运行的主要因素。结果表明这种结合是合理并且有效的。

关键词:面板数据多指标聚类;面板数据模型;国家级经济技术开发区

中图分类号:F222.3　　**文献标识码**:A　　**文章编号**:1002-6487(2014)07-0011-04

0　引言

自1984年我国设立首批国家级经济技术开发区以来,发展到现在形成了包含经济特区、经济技术开发区、高新技术产业开发区等各类特殊经济区域在内的多元开发区体系。开发区已成为我国经济的重要载体,然而改革后各个开发区发展情况参差不齐,造成开发区发展水平不一致的原因是值得探讨的问题。由于2003~2010年开发区的经济运行指标是面板数据,本文采用面板数据聚类分析对开发区进行分类,再进一步用变系数回归模型分析各类开发区的特点。

1　方法介绍

1.1　多指标面板数据聚类方法

面板数据聚类分析分为单指标和多指标两种形式。单指标面板数据聚类方法简单易行,所以现在大多数研究讨论的是单指标面板数据聚类方法。然而现实生活中大多数经济现象呈现出更为复杂的形态,单一指标通常难以充分反映研究对象的特征,因而,本文采用多指标面板数据聚类分析方法。假设总体共有N个样品,每个样品有p个指标值,时间跨度为T,$x_{ij}(t)$表示第i个样品的第j个指标在t时刻的指标值。其中,$i\in[1,N]$;$j\in[1,p]$;$t\in[1,T]$。

聚类分析要考虑两个基本问题,第一是确定如何度量样品之间的相似性,第二是确定使用何种系统聚类方法。首先,借鉴欧式距离来测度样品之间的相似程度,但由于此时涉及三个维度,一般的"欧式距离"不再适用,故使用一种包含三个维度的"欧式时空距离",其表达式为:

$$d_{rk}=\left\{\sum_{t=1}^{T}\sum_{j=1}^{P}\left[x_{rj}(t)-x_{kj}(t)\right]^2\right\}^{1/2} \tag{1}$$

其中,d_{rk}表示样品r与样品k之间的"欧式时空距离"。两两样品间的"欧式时空距离"组成一个对角线元素为0的对称矩阵,如式2所示。

$$\begin{bmatrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & 0 & & \\ \cdots & \cdots & \cdots & \ddots & \\ d_{N1} & d_{N2} & \cdots & d_{NN-1} & 0 \end{bmatrix} \tag{2}$$

其次,需确定把类与类聚集成一个新类所依照的准则,此处使用离差平方和法(Wald法)。多指标面板数据的离差平方和与截面数据的离差平方和函数有所不同,记第g类样品间的离差平方和为S_g,构造S_g的函数为:

$$S_g=\sum_{t=1}^{T}\sum_{j=1}^{P}\sum_{i\in i^g}\left[x_{ij}(t)-\bar{x}_j^g(t)\right]^2 \tag{3}$$

其中,i^g表示第g类中所有样品序号的集合;$\bar{x}_j^g(t)$表示第g类所有样品的第j个指标在t时间的平均值。

Ward法指出,当第r类与第k类合并时,增加的离差平方和可作为度量这两类之间距离的指标,将增加的离差平方和最小的两类优先合并。第r类和第k类合并时,考察它们之间的距离函数为(假设k类有两个样品p和q,当k类有多个样品时,可认为是最后一次合并时的两类,并可递归下去):

$$D_{rk}=\Delta S_{rk}=\frac{n_r+n_p}{n_r+n_k}S_{rp}+\frac{n_r+n_q}{n_r+n_k}S_{rq}-\frac{n_r}{n_r+n_k}S_{pq} \tag{4}$$

其中,n_r表示第r类样品的数目,S_{rp}表示第r类与第p类合并成新类的离差平方和,其它符号具有类似的意义。

1.2　面板数据变系数模型

常用的面板数据模型有混合回归模型、固定效应模型

基金项目:中央高校基本科研业务费专项资金资助项目(0140ZK1008)

作者简介:李　峥(1987-),女,福建人,硕士研究生,研究方向:数据挖掘。
刘云霞(1978-),女,山西人,助理教授,研究方向:数据分析。

和随机效应模型三种。这里只介绍本文采用的模型。

个体固定效应模型可定义为：

$$y_{it}=\alpha_i+X_{it}'\beta+\varepsilon_{it},i=1,2,\cdots,N;t=1,2,\cdots T \quad (5)$$

其中y_{it}为被解释变量，α_i表示对于i个个体有i个不同截距项，X_{it}为$k\times1$阶回归变量列向量，β为$k\times1$阶回归系数列向量，ε_{it}为误差项。其特点是α_i是随机变量，且变化与X_{it}有关。

个体随机效应模型与个体固定效应模型形式类似，不同之处在于它的α_i是随机变量，且变化与X_{it}无关。

固定效应模型和随机效应模型又都有变截距和变系数模型。变截距模型是指回归的斜率系数都相同而截距不同；变系数模型则考虑了斜率系数的变异性。由于现实生活中不断变化的经济结构或社会经济背景有时会导致反映经济结构的参数随时间或横截面个体不同而变化，采用不变系数模型无法很好地刻画经济结构的这种变化，所以有必要考虑变系数模型。

变系数固定效应面板数据模型的一般形式如下：

$$y_{it}=\alpha_i+x_{it}b_i+u_{it}(i=1,\cdots,N;t=1,\cdots,T) \quad (6)$$

其中，y_{it}为被解释变量，$x_{it}=(x_{1it},x_{2it},\cdots,x_{kit})$为$1\times k$维解释变量，$N$为截面个数，$T$为每个截面观测期总数，$k$为解释变量个数，$\alpha_i$为模型截距项，$b_i=(b_{1i},b_{2i},\cdots,b_{ki})'$为对应于$x_{it}$的系数向量，$u_{it}$为随机误差项。

在建立变系数模型前，需检验面板数据模型的基本形式，即应该建立固定效应模型还是随机效应模型，通常使用Hausman检验，其原理为：构建的统计量是组内估计量和广义估计量之差。原假设与备择假设是（仅考虑个体效应模型）：

H_0：个体效应与回归变量无关（个体随机效应回归模型）

H_1：个体效应与回归变量相关（个体固定效应回归模型）

检验统计量为：

$$H=\frac{(\dot{\theta}-\tilde{\theta})^2}{\dot{S}^2-\tilde{S}^2}\sim\chi^2(1) \quad (7)$$

基于上述两种方法的结合，作者认为可以先对所研究面板数据进行聚类分析，把样本分为几个不同的类别，再用面板数据变系数模型来研究各类的特点。

2 实证分析

本文以49个国家级经济技术开发区和5个享受国家级经济技术开发区政策的其他类开发区为样本，其中拉萨开发区由于数据不全，略去对其研究，因此共53个样本。选取2003~2010年各个开发区的7项产出类指标，包括地区生产总值(x_1)、工业总产值(x_2)、工业增加值(x_3)、税收收入(x_4)、出口总额(x_5)、进口总额(x_6)和实际利用外资金额(x_7)。数据来源于《中国开发区年鉴》、中国商务部网站和中国开发区网站。

2.1 面板数据多指标聚类分析

2.1.1 聚类结果

对2003~2010年53个开发区的面板数据进行观测，发现各个开发区的各项指标在不同年份的大小关系变化较为复杂，无法直接对它们的经济运行情况进行判断和区分。因此有必要用聚类分析来研究各个开发区的相似性和差异性。将原始数据标准化，根据上文描述的方法进行面板数据聚类分析。聚类结果显示，可以把开发区分为四个层次，分类结果如下。

第一类：天津，广州，昆山，苏州工业园区，这一类的经济运行情况最好。此类开发区位于东部地区，起步较早，所依托的城市工业基础较为雄厚，因而无论在工业发展、对外贸易，还是招商引资方面都在所有开发区中遥遥领先，说明其经济综合实力强，可谓中国开发区建设的典范。第二类：大连，青岛，上海金桥出口加工区，这一类的经济运行情况次好。此类开发区也位于东部地区，毗邻深水良港，具有广阔的腹地，对外经济活跃，相比于第一类，其各项发展指标虽稍微逊色，但是发展迅速，在基础设施、服务设施、生活设施、园区环境等方面都在努力赶追。

第三类：杭州，南京，宁波，上海漕河泾，烟台，北京，长春，广州南沙，沈阳，武汉，这一类的经济运行情况较差。此类有两个中部地区开发区，其余均是东部地区开发区，它们大多位于我国省会城市或直辖市，政治、经济、交通、教育等资源条件和配套设施齐全，有一定的区位指向。与前两类开发区相比经济综合实力稍弱，但发展潜力较大。

第四类：贵阳，南宁，石河子，兰州，东山，西宁，乌鲁木齐，成都，郑州，昆明，湛江，太原，银川，上海虹桥，海南洋浦，惠州大亚湾，福清融侨，秦皇岛，连云港，萧山，呼和浩特，温州，福州，威海，营口，长沙，南昌，宁波大榭，上海闵行，重庆北部新区，芜湖，哈尔滨，厦门海沧，西安，合肥，南通，这一类的经济运行情况最差。大部分中部地区和全部西部地区都在此类中。此类开发区经济运行虽差，但经营成本低，且有国家政策扶持，蕴藏巨大的发展潜力。

2.1.2 聚类结果的有效性验证

以2010年为例，对四类开发区的7项指标求平均值，结果如表1所示。

表1 53个开发区面板聚类分析结果和各类2010年各指标平均值

类别	包含的开发区	2010年七项指标平均值						
		x_1 亿元	x_2 亿元	x_3 亿元	x_4 亿元	x_5 亿美元	x_6 亿美元	x_7 亿美元
一	天津，广州，昆山，苏州工业园区	1436.34	4365.74	1037.39	242.80	277.61	248.22	19.02
二	大连，青岛，上海金桥出口加工区	930.85	2708.85	618.20	202.24	64.53	82.88	15.88
三	杭州，南京，宁波，上海漕河泾，烟台，北京，长春，广州南沙，沈阳，武汉	570.98	1749.27	343.29	115.06	66.98	61.81	5.11
四	剩余的36个开发区	218.73	567.98	155.65	36.95	11.00	10.09	1.86

通过对比可发现，7项指标值基本都是呈现从第一类到第四类依次递减的趋势，证明分类是合理的。同时，从

图1(2010年四类开发区综合经济指标对比)和图2(2010年四类开发区对外贸易指标对比)中也可以更直观地看到各类开发区的各项指标值差异。由此,验证了聚类分析结果的有效性。

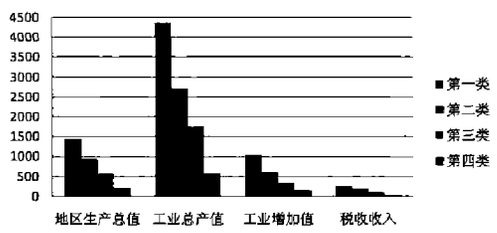

图1 2010年四类开发区综合经济指标(单位:亿元)

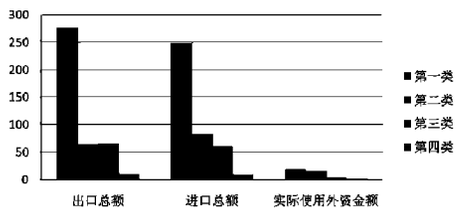

图2 2010年四类开发区对外贸易指标(单位:亿美元)

2.2 面板数据模型

为了进一步探究聚类后各类开发区经济运行的主要驱动因素,对各类开发区面板数据分别拟合回归模型。由于工业总产值和工业增加值高度相关,在此剔除工业总产值这一变量,以GDP为因变量,其它5个指标为自变量。影响地区经济运行的除不可观测的宏观因素外,个体在5个自变量上的不同表现也可能导致估计结果的差异。因此本文采用变系数面板数据模型来对四类面板数据进行估计,并使用混合最小二乘估计方法。在估计过程中,采用逐步剔除不显著变量的方法,筛选出显著的变量。前三类的Hausman检验结果在5%的显著性水平下都拒绝原假设,必须建立个体固定效应模型,第四类由于建立变系数模型效果欠佳,而Hausman检验结果在5%显著性水平下不能拒绝原假设,因此直接建立随机效应模型。

对第一类开发区建立固定效应变系数模型,结果如表2所示。

表2 第一类面板数据回归结果

地区	C	x_3	x_6
昆山	9.6171(0.7577)	1.6002(0.0000)	-1.5308(0.0002)
苏州工业园区	-6.7905(0.8276)	1.6002(0.0000)	-0.0223(0.9039)
天津	14.5206(0.7513)	1.6002(0.0000)	-1.8528(0.0028)
广州	-97.5443(0.0014)	1.6002(0.0000)	-0.8572(0.0670)
$R^2=0.9949, \bar{R}^2=0.9931$		$F=557.9457, p-value=0.0000$	

注:括号内为各拟合值的p值。

从表2可看出,影响地区增加值的主要因素是工业增加值和进口值,且工业增加值对地区增加值的影响是固定的(即对拟合的所有开发区其系数是固定的,下同),系数是显著的,而进口值的影响则是变系数的,即不同开发区有不同的系数。除苏州工业园区之外,其它三个地区的进口值在10%的显著性水平下都是显著的,且都为负。进口值一般能够从侧面反映一个地方的消费能力,一个进口发达的地方往往其经济发展也好。这一类开发区处于东部沿海,是少数拥有码头的国家级开发区,在稳固工业的基础上,表现出了强劲的对外贸易能力,成为国内开发区中的先行者,以此大力带动地区经济的快速发展。

对第二类开发区建立固定效应变系数模型,结果如表3所示。

表3 第二类面板数据回归结果

地区	C	x_3	x_5	x_4
大连	28.1769(0.1883)	1.1298(0.0000)	1.1123(0.0000)	1.6922(0.0059)
青岛	40.3560(0.0246)	1.1298(0.0000)	1.1123(0.0000)	0.9069(0.0067)
上海金桥出口加工区	-103.3686(0.0046)	1.1298(0.0000)	1.1123(0.0000)	0.2884(0.0499)
$R^2=0.9972, \bar{R}^2=0.9960$		$F=819.4509, p-value=0.0000$		

注:括号内为各拟合值的p值。

从表3可看出,影响地区增加值的主要因素是工业增加值、出口值和税收收入。工业增加值和出口值对地区增加值的影响是固定的,税收收入是变系数的,且在5%的显著性水平下都是显著的。一方面,工业仍然是这些开发区经济发展的基石,另一方面,它们都有丰富的港口资源和颇具规模的出口加工区,使开发区作为出口贸易主要渠道的作用得以充分体现。而税收收入通常能很好地反映地区的经济发展状况,这类开发区制定合理的税收政策,并且由于产业结构的调整,这些地区旅游业、商业、服务业等第三产业的比重上升,其税收弹性较大,因而税收收入对GDP的贡献作用变得显著。可以发现,在前两类开发区中,没有中部或西部地区的开发区,说明东部开发区具有明显的区位优势,工业基础雄厚,交通便利,科技实力较强,投资环境优越,外贸也较发达。

对第三类开发区建立固定效应变系数模型,结果如表4所示。

表4 第三类面板数据回归结果

地区	C	x_3	x_4
烟台	22.6647(0.4212)	0.5946(0.0000)	4.2419(0.0000)
宁波	-23.7663(0.5559)	0.5946(0.0000)	4.2913(0.0000)
上海漕河泾	-11.8175(0.6797)	0.5946(0.0000)	6.5012(0.0000)
杭州	29.0141(0.3050)	0.5946(0.0000)	2.1706(0.0008)
北京	-2.8087 (0.9122)	0.5946(0.0000)	2.3091(0.0000)
南京	9.6977(0.7155)	0.5946(0.0000)	2.4273(0.0065)
沈阳	-108.5892(0.0001)	0.5946(0.0000)	7.6230(0.0000)
武汉	64.4205(0.0043)	0.5946(0.0000)	1.0263(0.0027)
长春	45.1831(0.0900)	0.5946(0.0000)	5.1572(0.0000)
广州南沙	60.8272 (0.0026)	0.5946(0.0000)	1.1107(0.0000)
$R^2=0.9730, \bar{R}^2=0.9639$		$F=106.4458, p-value=0.0000$	

注:括号内为各拟合值的p值。

从表4可看出,影响地区增加值的主要因素是工业增加值和税收收入。工业增加值对地区增加值的影响是固定的,税收收入是变系数的,且在1%的显著性水平下都是显著的。与第二类类似,它们也是通过税收方面的政策和产业结构的调整来影响经济发展。同时,为了与第二类作比较,把出口值作为自变量纳入模型,发现拟合效果不显著。值得一提的是,这类中的长春和武汉为中部地区的开发区。长春和武汉开发区充分发挥它们在中部地区崛起过程中的引擎作用,表现出良好的发展态势。

对第四类开发区建立随机效应模型,结果如表5所示。

表5　　第四类面板数据回归结果

地区	C	x_3	地区	C	x_3
贵阳	12.3015	1.2978(0.0000)	连云港	1.4818	1.2978(0.0000)
石河子	3.8217	1.2978(0.0000)	萧山	-9.0287	1.2978(0.0000)
南宁	9.9338	1.2978(0.0000)	温州	9.5792	1.2978(0.0000)
兰州	12.5456	1.2978(0.0000)	呼和浩特	-12.5208	1.2978(0.0000)
东山	5.0003	1.2978(0.0000)	福州	18.6077	1.2978(0.0000)
西宁	4.6413	1.2978(0.0000)	威海	6.0119	1.2978(0.0000)
乌鲁木齐	4.8582	1.2978(0.0000)	营口	52.4966	1.2978(0.0000)
成都	46.0699	1.2978(0.0000)	长沙	26.2820	1.2978(0.0000)
郑州	33.2472	1.2978(0.0000)	南昌	12.5314	1.2978(0.0000)
昆明	10.3104	1.2978(0.0000)	宁波大榭	40.5045	1.2978(0.0000)
湛江	3.7492	1.2978(0.0000)	上海闵行	-32.0793	1.2978(0.0000)
太原	-1.6484	1.2978(0.0000)	重庆北部新区	15.2394	1.2978(0.0000)
银川	9.6949	1.2978(0.0000)	芜湖	-26.1357	1.2978(0.0000)
上海虹桥	70.1679	1.2978(0.0000)	哈尔滨	3.4294	1.2978(0.0000)
海南洋浦	11.7020	1.2978(0.0000)	厦门海沧	10.5846	1.2978(0.0000)
惠州大亚湾	12.7461	1.2978(0.0000)	西安	15.7058	1.2978(0.0000)
福清	-16.6278	1.2978(0.0000)	合肥	2.0080	1.2978(0.0000)
秦皇岛	25.6608	1.2978(0.0000)	南通	6.3473	1.2978(0.0000)
$R^2=0.9730$、$\bar{R}^2=0.9639$　$F=106.4458$、$p-value=0.0000$					

注：括号内为各拟合值的p值。

从表5可看出，影响地区增加值的主要因素只有工业增加值。这些地区起步较晚，发展较慢，所依托的资源环境条件也较为欠缺，虽然在某些传统工业领域和工业发展模式上有一定优势，但这也使其产业结构难以优化升级，经济发展存在瓶颈。而离海较远造成这些地区的对外贸易步履维艰，物力资源不够丰富，人力资源稀缺更是导致其科技创新思维落后。特别是西部地区，深居内陆，人口稀少，一直是我国经济欠发达的地区。这些不仅导致其工业产值、税收等的滞后发展，还使经济开发区作为对外开放港口的功能难以发挥，虽然工业增加值是地区增加值增长的动力，但与其它类开发区相比仍然颇有差距。

总的来说，无论是哪一类开发区，工业都是其经济发展的基础，2010年，53个开发区工业增加值平均达到地区增加值的67%，有些开发区甚至达到90%以上。在稳定工业的基础上各类开发区有不同特点，对地区增加值起主要作用的因素各不相同。通过对比，可以清晰地看到各类开发区的优势和劣势，找出经济发展的差距，有利于各个开发区取长补短，调整发展战略，经济发展较好的地区要继续保持优势，为发展较差地区树立榜样，而经济发展较差的地区则要汲取经验，积极转变发展方式，争取更为广阔的发展前景。

3　结论

运用多元统计的方法分析面板数据是当今学术领域的热点研究之一，本文把多指标面板数据聚类分析和变系数模型相结合，不仅从动态的角度对研究对象进行了区分，而且进一步深入探讨了各类研究对象的区别。通过开发区经济运行的实证分析，表明该方法能很好地解决面板数据的聚类问题，分类效果好，并且有效区分了各类别中影响经济运行的主要因素。

参考文献：

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

[2]孙遇春，徐吉祥，张建同，孙启承. 国家级经济开发区发展水平的比较与评估[J].统计与决策，2010,(14).

[3]朱立龙、张建同、孙遇春.我国国家级经济技术开发区综合指标评价研究[J].科学管理研究，2008,(8).

[4]任娟.指标面板数据聚类方法及其应用[J].统计与决策，2012,(4).

（责任编辑/亦　民）