学校编码：10384　　分类号_____密级_____

学号：15420081152043　　UDC_____

厦门大学

硕 士 学 位 论 文

面板数据综合评价和聚类分析及应用研究

Panel Data Clustering & Comprehensive Evaluation and Application

李建虎

指导教师姓名：周 永 强 副教授

专 业 名 称：数 量 经 济 学

论文提交日期：2011 年 5 月

论文答辩时间：2011 年 5 月

学位授予日期：2011 年 月

答辩委员会主席：_____

评 阅 人：_____

2011 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。

声明人（签名）：

年　　月　　日

摘　要

传统的多元统计分析及综合评价研究大都针对截面数据，且有关面板数据的理论多着重于计量模型的构建和参数估计，而关于如何将传统多元统计分析的方法引入到面板数据的分析中，以及如何进行面板数据的综合评价，国内外学者的研究则相对较少。

面板数据是指同时包含截面数据和时间序列数据的一种三维数据。它至少给出了三个方面的重要信息：一是各时期各样本各指标的绝对量水平；其次是指标的动态发展水平，即指标随时间变化的增量水平或增长率；第三个方面是样本各项指标发展的协调水平，即指标的变异程度或稳定性。故针对面板数据的综合评价及聚类分析应从以上三个方面展开。

本文就如何对面板数据进行综合评价及聚类分析展开系统论述，并依此展开实证分析。文章首先介绍了研究背景及意义，而后着重探讨了面板数据综合评价和聚类分析的方法和步骤，最后从综合评价及聚类分析两个方向对我国不同地区城镇居民生活水平的差异进行了详细分析，分析从绝对量水平、指标增长率及指标稳定性三个角度展开，得到相关结论，并根据结论提出了相关政策建议。

本文可能的创新之处主要体现在以下三个方面：1.本文就如何进行面板数据的综合评价给出了新的分析思路，认为应从三个角度进行，即指标的绝对量水平、指标值的增长率及指标的稳定性，并对每个角度所要进行的详细分析过程做了说明。2. 详细分析了面板数据聚类分析中距离选取应当注意的问题，并给出了面板数据聚类分析中马氏距离的定义。3.首次利用面板数据综合评价和聚类分析对我国不同地区城镇居民生活水平的差异进行分析。

关键词：面板数据；综合评价；聚类分析；生活水平

ABSTRACT

The traditional multivariate statistical analysis and comprehensive evaluation research mostly focus on cross-section data, and most of the theory of panel data focus on the construction of econometric model and parameter estimation. The research about how to introduce the traditional method of multivariate statistical analysis into the panel data analysis, and how to do the comprehensive evaluation for panel data is relatively rare.

Panel data is 3D data including cross-section data and time series data. It at least gives us three important information. First is the absolute level of samples' indicators. The second which we call incremental level or rate of growth is the dynamic development level of samples' indicators. The last which we call variation or stability of indicators is the level of coordinated development of samples' indicators. Therefore, the comprehensive evaluation and clustering analysis for panel data should be started from the above three aspects.

This article systematicly discusses the way to do the comprehensive evaluation and clustering analysis for panel data, and expandes the empirical analysis based on these. First, the article introduces the background and significance of research. Then it focuses on the methods and procedures of comprehensive evaluation and clustering analysis for panel data. At last, it analysises the differences of living standard of urban residents among regions. The analysis is done from three angels which are absolute level, rate of growth and stability of indicators. Then we make a summary of conclusions, and propose some suggestions about it.

The possible innovation of this paper is mainly reflected by the following three aspects: First, this article gives us a new idea on how to do the comprehensive evaluation for panel data, which is that our analysis should starts from three aspects: absolute level, rate of growth and stability of indicators. Then it introduces the details of analysis. Second, the article discusses the problemes that we should pay attention to when we choose a distance formula in clustering

panel data. And it gives the definition of Mahalanobis distance in clustering panel data. Third, it is the first time bringing comprehensive evaluation and clustering anlysis for panel data into the difference anlysis of living standard of urban residents among different regions.

Keywords: panel data；comprehensive evaluation；clustering analysis；living standard

目　录

Contents

1 绪论

1.1 研究背景及意义

传统的多元统计分析及综合评价研究大都针对截面数据，且其理论与方法已经较为成熟。从二十世纪七十年代末以来，面板数据（Panel Data）的研究便异军突起，时至今日，其理论方法已日渐成熟，涌现出了大量相关的理论和经验分析文章，形成了现代计量经济学的一个相对独立的分支[1]。绝大多数有关面板数据的理论[2]，或者是着重于计量模型的构建，建立诸如单方程、联立方程、变截距、变系数、线性、非线性等模型，或者是着重于模型参数估计方法的研究。关于如何将传统多元统计分析的方法引入到面板数据的分析中，以及如何进行面板数据的综合评价，以达到充分利用面板数据所含丰富信息的目的，国内外学者的研究则相对较少（详见下一节）。

面板数据是指同时包含截面数据和时间序列数据的一种三维数据，既具有空间维度特征又具有时间维度特征。从横截面上看，它是由若干个体在某一时刻构成的截面观测值，从纵剖面上看则是一个时间序列。面板数据可以用三下标变量表示，例如 x_{ikt}， $i=1,2,\cdots,N$， $k=1,2,\cdots,P$， $t=1,2,\cdots,T$。N表示面板数据中含有的样本数；P表示指标变量总数；T表示时间序列的最大长度。因此，依据这三个维度，面板数据至少给出了三个方面的重要信息[3]：一是各时期各样本各指标的绝对量水平；其次是指标的动态发展水平，即指标随时间变化的增量水平或增长率；第三个方面是样本各项指标发展的协调水平，即指标的变异程度或稳定性。显然，如果针对面板数据的统计分析及综合评价没有同时考虑到上述三个方面，就没有充分挖掘其所含的丰富信息。

传统的聚类分析方法虽然已经比较丰富和成熟，但它的分析对象一般是固定时期的不同个体截面数据，不能充分利用不同时期积累的众多数据，故往往

[1] Hsiao C., Analysis of Panel Data [M].北京:北京大学出版社,2005.21-92.
[2] Hsiao T. P. and Chih Y. Y. Comparison of Linear and Nonlinear Models for Panel Data Forecasting: Debt Pol icy in Taiwan [J]. Review of Pacific Basin Financial Markets and Policies, 2005,(3): 525-541.
[3]李因果，何晓群.面板数据聚类方法及应用[J].统计研究,2010,(9):74-79.

不能满足人们分析问题的需要。基于单一的固定时期的聚类分析往往忽视了指标的动态发展趋势及其发展状态，无法预测其未来发展轨迹及样本所属类别。例如:在企业竞争力聚类分析中，竞争力是随着时间动态变化的，仅仅固定在某一年度的截面数据分析就显得有些片面，如果根据一个较长时期积累的面板数据进行聚类分析则显得较为合理。如何针对面板数据进行聚类分析，找出有效的研究方法就是一个非常有意义的问题。

聚类分析需要处理两个核心问题:一是用什么统计量来描述样本之间的相似程度；二是采用何种聚类方法确定类与类之间的相似程度。从一组复杂数据产生一个相当简单的类结构，必然要求进行“相关性”或“相似性”度量。在将经典的聚类分析方法应用于面板数据时，易于想到的方法是:(1)对样本按年度一一进行聚类，但这样处理显然会容易造成各年度分类结果的不一致；(2)采取退化方法，对指标在时间序列维度上求均值，然后进行聚类，这样的聚类结果显然抹杀了样本指标的发展速率及其动态趋势；(3)取各个年度指标的平均发展速率进行聚类，这种聚类结果显示了指标的动态性，却抹杀了指标的绝对量水平值。显然，上述三种聚类思路都存在缺陷，对面板数据的聚类:一方面要考虑样本指标间的绝对距离，另一方面必须考虑其时间序列的动态发展特征。很难想象一个呈正向增长指标的个体应该会和一个呈杂乱甚至负向增长指标的个体聚为一类。因此，对面板数据的聚类相似性测度指标的设计，必须考虑绝对量、动态发展趋势及指标稳定性三方面的信息。

如果设计出了恰当的统计量来度量面板数据在绝对量、动态发展趋势及指标稳定性三方面的信息，那么就可以从三个角度对样本进行更为合理的聚类。从绝对量角度看，我们可以按绝对量相似程度把样本分类，绝对量结构较为相似的一类。从指标动态发展趋势角度看，我们可以按指标增长率相似程度对样本进行归类，增长率结构较为相似归为一类。从指标稳定性角度看，我们可以根据指标稳定性相似程度对样本进行归类，稳定性结构类似的归为一类。我们亦可以对三个角度的度量赋予权重，综合考虑样本间的相似度，并进行分类。例如，通过对表现地区经济发展水平的面板数据的分析，我们将相似的地区聚为一类，根据分类，对比分析不同样本聚为一类的原因及不同类的特征，进而进行深入分析，期望找到适应不同地区的政策指导。

目前综合评价的研究大多都针对截面数据，且研究方法繁多[4]，但是针对面板数据的综合评价研究则比较匮乏（详见下一节）。实际上，面板数据的综合评价也应从绝对量水平、指标增长率及指标稳定性三方面进行分析，这样才能充分利用其所提供的信息对样本展开全方位的评价。

针对每一个方面，我们还可以做出多种分析。比如说针对绝对量，我们可以比较同一样本在不同时间的绝对量水平变化，对同一时间不同样本的绝对量水平进行比较，比较不同样本在不同时间的的绝对量水平，以及从整体时间范围来看样本的绝对量水平差异。指标增长率及指标稳定性也可以展开类似的分析。而这些都是截面数据无法做到的。

1.2 文献综述

1.2.1 面板数据综合评价研究现状

传统综合评价主要研究截面数据，对面板数据综合评价的研究成果比较稀少。

Liping Yu 等（2009）开发出利用面板数据进行学术期刊评价指标选取的方法，发现此法的有效性，并指出不同学科的期刊不应该用同一标准进行评估，简化的指标，可以降低成本，提高效率以及期刊评价的准确性。耿修林（2006）基于统计距离建立了针对面板数据的综合评价方法，并将此应用于企业绩效水平的横向（时间）和纵向（样本间）对比分析。王培等（2009）应用因子分析法对多指标面板数据进行了分析，并利用用综合评分法对各地区的工业企业生产效率进行了分类，得出应用因子分析的结果与现实基本相符。董锋等（2009）利用改进的因子分析方法对“江苏省科技成果转化专项资金资助企业”的面板数据进行分析，根据改进公式得出 40 家企业的公共因子得分和综合总得分。

可见，现有的面板数据综合评价都主要针对指标的绝对量，几乎没有考虑到指标的增长率及指标的稳定性，这无疑遗漏了面板数据的重要信息，使评价结果显得片面。

1.2.2 面板数据聚类分析研究现状

如何对面板数据进行有效的聚类分析，国内外学者的研究还相对较少，

[4]胡永宏，贺思辉.综合评价方法[M].北京:科学出版社,2000.

比较有代表性的如下：

Bonzo D. C.和 Hermosilla A. Y.（2002）等统计学家开创性的将多元统计方法引入到面板数据的分析中来。Bonzo D. C.运用概率连接函数代替传统的距离函数来定义聚类标准，把聚类过程当做一种优化问题，利用随机启发式技术优化目标函数，并运用自适应模拟退火方法(ASA)对面板数据进行了聚类分析。Ren J.(2009)基于费舍尔次序集群理论，通过 Frobenius 准则重建了 Ward 函数，提出了一种多变量面板数据序聚类方法。Guangli Nie 等(2010)借鉴欧氏距离定义了用于面板数据聚类的的距离，此距离可以对时间进行加权，并利用银行信用卡数据进行了实证，检验其聚类效果。

国内，朱建平(2007)最早展开对面板数据聚类分析的研究，可谓开启了此领域研究的先河，为以后的研究奠定了基础。他系统阐述了面板数据的统计描述方法，并定义了用于单指标面板数据聚类分析的相似性指标，用它对我国不同区域的城镇居民人均收入及人均消费支出的差异进行了实证分析。郑兵云(2008)分析了面板数据的数据格式和数字特征，基于欧氏距离重构了多指标面板数据的距离函数，但在时间维度上取均值，将问题过度简单化，退化为截面数据的聚类思路，存在明显的信息损失缺陷和假定缺陷。汪凯(2009)基于单指标面板数据的聚类分析对安徽省农村居民收入和消费的地区差异进行分析。刘兵(2010)提出可以根据样品中各个指标的时序数据的趋势特征来考虑是否应该进行压缩或如何进行压缩，分别就水平趋势、非水平趋势及线性趋势的数据系列给出了其聚类统计量及聚类方法。但它要求同个面板数据中时序趋势属于同一类，故应用范围受到限制。

上述文章中所提供的聚类分析算法没有详细考察面板数据的动态分类统计特征。

李因果、何晓群(2010)的研究为面板数据聚类分析的研究指出了一个新的方向。他综合考虑了面板数据的“绝对指标”、“增量指标”及其“时序波动”特征，在重构面板数据相似性测度的距离函数和Ward聚类算法的基础上，提出了面板数据聚类方法，并以财政金融面板数据为例，对中国14个沿海开放城市进行了聚类分析。