【阶层研究】

理论与数据双驱动的社会分层研究

梁玉成　贾小双

【摘　要】以往社会分层研究的理论和方法可以归纳为理论和数据驱动两种范式，通过回顾和比较发现其都存在不可避免的局限性，为此提出一种结合二者优势的理论与数据双驱动的社会分层研究框架。该框架将社会阶层看作由类别和等级参数所构成的高维社会空间中聚集的子群体，基于分层理论所提出的阶层测量指标建构“社会阶层空间”，并使用机器学习算法识别出空间中的不同群体，从而进行阶层划分。使用这一框架对中国综合社会调查2017年的数据进行阶层划分，发现其既能区分出地位一致性高的、边界清晰的阶层，也能对地位不一致的、还未形成阶层的利益群体进行准确识别。此外还发现：(1)将当前中国社会划分为三个阶层是最好的划分方式，三个社会阶层在经济、声望、文化等维度上的特征分布都存在高、中、低的等级差异；(2)分层的指标并不是越多越好，对中国当前社会阶层划分和对个体阶层测量最有意义的指标是单位类型，其次是职业社会经济地位。

【关 键 词】社会分层；理论驱动；数据驱动；机器学习；社会阶层空间；阶层测量方法

【作者简介】梁玉成(1971－)，男，中山大学社会学与人类学学院教授、博士生导师；贾小双(通信作者)(1993－)，女，中山大学社会学与人类学学院博士研究生(广州 510275)。

【原文出处】《西安交通大学学报》：社会科学版，2022.1.25～37

【基金项目】国家社会科学基金项目(15ZDB172)。

社会结构是社会学的核心议题。作为社会结构最重要的维度，阶层结构的研究对于理解社会现象和社会变迁有着重要的意义，一直以来广受国内外社会学家的关注，发展出了丰富的社会分层理论，并在此基础上提出了不同的阶层测量方法。总体来看，国内外学者对阶层的理解可分为两种：一种认为阶层是等级不同的群体，只需确定一定的数量标准就可以对社会阶层进行区分，例如按照收入的高低划分为低收入群体、中等收入群体和高收入群体；另一种认为阶层是社会性质、社会属性完全不同的群体，而不仅仅是简单的上下排列的等级层次，因此需要找到阶层之间属性差异的指标来界定。传统的社会分层理论(如马克思、韦伯和涂尔干的分层理论)都体现了将阶层看作属性不同的群体这一阶层视角，即根据生产资料的占有、劳动分工等差异来界定阶层[1]。

社会学对阶层结构的测量常常将两种视角结合起来，既考虑群体的社会属性差异，也关注社会属性的等级层次，而用于分层的社会属性常常被理解为“对各类资源的占有”。李强[1]认为，社会分层的本质是资源在不同群体中的分布。因此，资源的种类和占有水平是阶层和社会地位划分的依据。他将格伦斯基提出的用于分层的七种资源[2]扩展为十种，分别是生产资料资源、财产或收入资源、市场资源、职业或就业资源、政治权力资源、文化资源、社会关系资源、主观声望资源、公民权利资源以及人力资源。这十种资源各有侧重，其不同组合可以形成不同的分层标准，而不同的分层组合所划分的阶层群体又常常相互交叉，即在一种标准下被划分为同一个阶层的群体在另一种标准下可能被分为不同的阶层群

体。基于不同的资源组合和不同的划分标准，社会学发展出了不同的阶层测量方法。但笔者发现，这些方法都存在一定的局限性：一方面，不同分层模型测量阶层地位时选用的维度(资源种类)和划分标准(资源占有水平)不同；另一方面，这些方法都面临着“分层结果无法在现实中验证”的批判。因此，本文尝试提出一种理论和数据双驱动的阶层测量，在更全面地考虑阶层测量维度的基础上，使用数据驱动的方式从现实出发进行阶层划分。

一、两种阶层测量范式

社会分层研究的首要任务是界定社会阶层，阶层测量需要研究者制定出分层的指标对社会阶层进行划分。自马克思以来，社会理论家和社会学研究者提出了大量的阶层测量理论和方法，对社会分层指标、测量方法和划分方式进行了界定和讨论，发展出了理论驱动和数据驱动两种研究范式。

(一)理论驱动的阶层测量

传统的阶层测量均属于理论驱动范式。在这一范式下，社会分层的研究者在阶层测量上存在两种不同的取向：一种是阶级分析，另一种是职业分层。前者多使用类别型(categorical)的阶级测量方法，本文称之为阶层归类法；后者多使用连续型(continuous)的阶层测量方法，本文称之为数值测量法。

阶层归类法是指研究者基于社会分层理论探索出一些有重要经济社会差异的大的阶级类别，然后将社会人群纳入这些大的类别，社会学分层理论最重要的两种阶层图示——赖特阶级分类模型和埃里克森—戈德索普层图式(EGP)都属于这一类测量方法[1,3-4]。赖特阶级分类模型是根据不同社会群体围绕物质生产资料、劳动力、组织和技术四种资产所产生的占有(控制)和剥削关系进行的阶级分类[3]；戈德索普等提出的EGP图式主要是依据职业信息进行的阶层划分，根据职业声望、职业的市场地位(职业的经济收入来源和收入水平、经济保障状况和经济提升、职业的技术能力等)、工作地位(管理权限、工作自主程度等)以及雇佣关系等特征对职业社会阶层地位进行划分[4-7]。

数值测量法是指研究者基于特定的特征计算出一个有高低等级的、连续的数值作为界定阶层地位的指标，其典型代表是职业声望量表(Occupation Prestige Scale，OPS)和社会经济地位指数(Socio-Economic Index，SEI)。职业声望量表是通过调查的方式来了解人们对国家或国际职业分类标准中的职业评价，从而计算出职业声望的评估标准[8]。目前大多数学者使用的职业声望量表是特莱曼整合60个国家与地区的85套职业声望调查数据所提出的较为稳定的、可以用于跨国比较分析的国际标准职业声望量表(Standard International Occupational Prestige Scale，SIOPS)[9-11]。社会经济地位测量则是使用每一类职业的平均教育水平和平均收入对该类型的职业声望进行回归，并基于回归方程来估计所有职业的社会经济地位指数[12-15]，目前所使用的社会经济地位指数是1992年甘泽布姆等根据国际标准职业编码(International Standard Classification of Occupations，ISCO)提出的国际标准社会经济地位量表(International Socio-Economic Index of Occupational Status，ISEI)。这一量表给出了每一个职业对应的ISCO、ISEI、SIOPS，并与十等级的EGP阶层分类相对应[16-17]。随着社会的发展，国际标准职业编码在不断更新，IESI和SIOPS也进行了相应的更新。

(二)数据驱动的阶层测量

随着大数据和计算社会科学的发展，数据驱动的阶层测量方法逐渐兴起，并在学术界和业界得到了广泛应用。与传统阶层测量方法不同，数据驱动的阶层测量主要是用于估计个体或家庭的社会经济地位(Socio-Economic Status，SES)，而不以研究整个社会的阶层结构为目的。社会经济地位是指基于个体或家庭的受教育水平、收入水平和职业水平而形成的在经济层面和社会层面相对于他人的社会位置，并且通常被划分为高、中、低三个等级[18]。传统的SES测量主要使用调查数据来获取决定SES的教育、收入、职业等传统社会分层理论所关心的阶层测量维度直接进行划分，而数据驱动的阶层测量主要依据大数据来测量个体的社会经济地位。

由于大数据难以获取经济资源、职业资源、声望资源等理论驱动分层所关注的数据，而更多地包含社交网络和生活方式等社会资本和文化资本信息，因此，基于不同社会经济地位的群体拥有不同生活

方式和社会网络的观点[19-20]。数据驱动的阶层测量主要使用手机或互联网获取的用户行为、社交网络以及环境(如居住区域)数据等,通过一定的算法对个体或家庭的社会经济地位进行预测和估计。

在使用生活方式特征预测阶层地位的研究中,研究者从多个生活方式的多个维度预测个体的SES,如活动轨迹、电话沟通模式、消费模式、社交媒体上讨论的话题以及使用的语言和社交媒体上的表现等。其方法一般是将手机或社交媒体上记录的海量个体行为数据转化为结构化的数据(即个案—变量式的数据),用以刻画个体生活方式的特征,然后根据这些特征来预测个体的SES等级、收入或职业类别[21-25]。在使用社会网络特征进行预测时,研究者通常依据社会网络分析(Social Network Analysis,SNA)中整体网分析的各项网络结构指标(如中心性、度分布等)[26],而非个体及其朋友的社会人口属性来预测个体的SES。

根据生活方式或社会网络特征预测个体或家庭的SES等级以后,还需要结合直接测量SES的相关数据——如用户居住小区的房价、普查或社会调查发布的地区社会经济水平、用户的职业类别等作为用户的"真实"SES,来验证基于行为和网络预测的准确性。由此可见,数据驱动的阶层测量实际上把阶层测量作为一个分类任务去完成,研究者基于个体的行为或社会网络特征,采用机器学习的方法对用户进行分类,并通过特征筛选、优化算法等方式来提高分类的准确性。在实际操作中,支持向量机(Support Vector Machine,SVM)、随机森林(Random Forest)等有监督机器学习分类方法和词聚类与词嵌入(Word Cluster and Embedding)、K均值聚类(K-means Clustering)等无监督聚类方法常被用于划分用户的SES等级。

(三)两种阶层测量范式的比较

如表1所示,通过对理论与数据驱动的阶层测量方法及其理论依据进行梳理,发现理论驱动的阶层测量更关注经济资源(如生产资料的占有、收入与财富等)、与经济资源直接相关的资源(如职业类别、市场资源、劳动关系等权力资源以及受教育程度、技能水平等人力资本)以及声望资源,而并未将(除人力资本外的)文化资本和社会网络资源纳入社会分层的维度;而数据驱动的阶层测量正好相反,只考虑文化资本和社会网络资源,实际上这种分异的产生是由于数据和方法的局限性。

在理论驱动的阶层测量发展之时,研究者只能使用调查数据进行研究,调查数据中更多地包含教育、收入、职业等核心变量,而较少包含生活方式数据;在分析方法上,由于人脑的思考维度是有限的,理论驱动的分层模型只能考虑有限维度的社会属性,无法处理高维的特征,加之传统的实证分析大多采用线性模型,由于存在地位不一致的可能,阶层地位并不一定是各种资源的线性组合。因此,研究者只能选用更重要的维度对阶层进行测量。由于每种理论驱动的方法都只考虑特定维度的资源,在阶层

表1　　理论与数据驱动的阶层测量方法比较

阶层测量方法			划分依据	操作化
理论驱动	阶层归类	赖特	三种控制权:对资本、物质资料和劳动的控制权 四种控制级别:全部、部分、微量和没有控制	测量四种资产:生产资料、劳动力、组织和技术资产
		戈德索普模型与EGP图式	职业信息(市场地位)	职业的收入水平,经济保障状况、晋升机会、技术水平
			雇佣关系(工作地位)	雇佣关系、职业在组织中所处的权威和控制系统的位置、工作自主性
	数值测量	职业声望	职业声望高低	调查对各职业的评价 国际标准职业声望量表(SIOPS)
		社会经济地位	社会经济地位指数(SEI)	邓肯社会经济地位指数模型(收入、声望、教育程度) 国际标准社会经济地位指数(ISEI)
数据驱动	预测 基于特征分类		生活方式、社交网络等特征机器学习模型的分类	活动轨迹、通话模式、社交媒体使用特征、社会网络结构等

划分的方式上也存在差异，因此不同流派的分层研究者对究竟应该如何进行划分争论不休[27-29]。此外，不同国家或地区、不同时期的社会发展情况存在差异[30]，研究者基于不同的数据测量出的阶层结构能在多大程度上反映社会现实也难以验证[31]。

对于数据驱动的阶层测量而言，手机和互联网产生的大数据主要是对个体使用痕迹的记录，通过这些记录很容易得出个体的移动轨迹、通话模式以及社交媒体上的信息。因此，基于大数据的阶层测量只能根据文化和社会网络等信息来推测。但由于大数据很难获取教育程度、收入、职业等隐私信息，通常用社区房价、地区SEL等作为替代，因此对SES的预测结果难以验证。此外，若特征维度较高，过于复杂的黑箱算法也使得分层结果难以解释。

实际上，很多大型社会综合调查的数据包含行为、态度、生活方式等文化资本和社会网络的数据，只是因方法的限制使得研究者未能将其纳入阶层测量中；而机器学习方法和技术不仅可以用于大数据的分析，同样可以用于调查数据的分析。为克服纯理论和纯数据驱动的阶层测量方法的不足，本文尝试将两种方法的优势结合起来，提出一种理论与数据双驱动的阶层测量方法。

二、理论与数据双驱动的阶层测量方法——基本框架设计

阶层测量的目的是分析社会的阶层结构，从而分析结构形成的原因及其影响。因此，研究者所得出的阶层结构必须符合社会现实。然而，有学者对我国分层研究的四种模式进行分析后提出了尖锐的批评，认为“关于中国分层的几种不同模型只不过是几种不同的关于当前中国社会分层状况的概念或分类游戏而已，并且四种模型经过一番操作能够实现相互转化”，并认为“关于当代中国社会分层状况的几种模式，其是非对错本质上与‘事实’①无关，因而也不可能通过将它们与‘事实’对比，看谁更符合‘事实’(或能获得更多‘事实’支持)的方法来对它们的是非对错加以判断。它们之间的差异，实质上只是几种关于社会分层之话语系统之间的差异”[31-32]。这一观点启发了笔者，即在进行阶层划分时应该从社会事实出发进行阶层结构测量，避免从理论上对阶层进行定性的分类。但如何根据社会事实来划分阶层呢？前文提到，阶层是社会属性和等级不同的群体，是对不同资源占有水平不同的群体，那么阶层划分就是根据社会成员的属性和等级将社会成员划分为不同的群体，而如何选择用于区分阶层的属性和等级，则需要借助分层理论的帮助。基于这一观点，本文建构了理论和数据双驱动的阶层测量框架。

(一)社会结构、布劳空间与社会阶层

布劳在《不平等与异质性》中建构了其宏观社会结构理论，认为社会结构可以用类别参数和等级参数来描述。类别参数是指将人口平行地划分为界限明确的若干个亚群体的特征，包括性别、种族、宗教、国籍、居住地、语言、职业、婚姻状况等；等级参数是将人口按高低秩序划分为若干层次的特征，包括教育、收入、财富、权力等。布劳认为，社会结构的分化一般有异质性和不平等两种形式，异质性是水平分化，指人口在由类别参数所表示的各群体之间的分布；不平等是垂直分化，指由等级参数所表示的地位分布。此外，他还用相交性表示社会结构中几条轴线的人口分布共变情况。类别参数和等级参数构成了多维空间，而人口在这一多维空间中的分布则构成社会结构[33-34]。这一“多维空间”被命名为布劳空间，所有社会人口特征都是布劳空间的潜在坐标轴[35-36]。

社会阶层是社会结构最核心的维度，因此可以认为，社会阶层是由类别参数和等级参数共同决定的。如上文所述，不同社会阶层既是异质性的群体，也是在等级秩序的阶梯中占有不同位置的群体。因此，参照社会结构的定义，可以将社会阶层看作人口在由类别参数和等级参数所构成的高维社会空间中的分布所形成的次级群体，那么阶层划分就是去识别这些群体。基于这一思想，本文建构了理论与数据双驱动的阶层测量框架：第一步，建构社会阶层空间，即基于分层理论提出的对阶层划分有意义的资源(阶层测量的维度)，将其操作化为可测量的变量作为社会空间的维度，建构出社会空间；第二步，使用无监督聚类的方法识别在高维社会空间中形成的次级群体，从而进行阶层划分。

(二)社会阶层空间的建构与分割

建构社会阶层空间需要先描绘出社会空间的

“轴线”，即定义用于阶层划分的维度。李强总结了过往分层理论所使用的阶层划分的10种资源：生产资料资源、财产或收入资源、市场资源、职业或就业资源、政治权力资源、文化资源、社会关系资源、主观声望资源、公民权利资源以及人力资源。但这一分类过于细致，导致这10种资源并非互斥，如文化资源包含了人力资本，职业或就业资源中也包含收入、生产资料和市场资源等信息，在操作化时较难进行测量。陆学艺[27]根据我国特色，将阶层划分要素综合为5个：职业或劳动分工、经济资源、组织资源(也称权力资源)、文化(技术)资源和单位地位或制度分割，但这种归类也忽视了社会网络资源、除人力资本外的文化资本以及声望资源和公民权利。在对二者进行综合的基础上，本文将用于社会分层的要素归纳为7类，分别是：(1)经济资源，主要指收入状况，包括个人收入与家庭收入；(2)职业与声望，整合了组织资源(有无管理权限)、职业资源(职业类型、工作状况)和职业声望；(3)单位地位或制度分割，包括户口、单位类型、体制以及党员身份等；(4)社会资本；(5)民权资源；(6)人力资本；(7)文化资本，主要包括人力资本以外的其他文化资本，如生活方式、消费结构等。结合获取数据的情况将上述要素操作化为可测量的变量，即社会空间的坐标轴。

构建好社会空间的下一步是进行阶层划分。由于本文没有理论预设，并不清楚人口在这个高维空间中是如何分布的，因此并不知道社会可以划分为多少个阶层以及每个阶层拥有什么样的特征。为此，本文采用数据驱动的方式，使用无监督(unsupervised)的机器学习聚类(clustering)算法来帮助识别人口在这个空间中的分布状况，寻找高维空间中聚集在一起的一个个“团体”来进行阶层划分。

聚类算法的目标是将样本划分为若干个不相交的子集，每个子集叫作一个“簇”(cluster)，每个簇对应这个子集一些潜在的特质，如高教育程度、高收入等。聚类算法事先并不清楚这些特质的存在，而是通过学习数据的分布结构找到内在性质和规律而自动形成的簇。聚类算法的聚类逻辑是“物以类聚”，即将拥有相似特征的样本划分到同一个簇，而不同簇的样本之间尽可能不同，即簇内相似度(intra-cluster similarity)高而簇间相似度(inter-cluster similarity)低。因而，“相似度”或称“距离”是聚类算法簇划分的重要依据。常见的相似度或距离测量方式有欧式距离(Euclidean distance)、曼哈顿距离(Manhattan distance)、余弦相似性、图中连边概率等。不同的聚类算法采取不同的相似度或距离计算方式，当前常见的聚类算法可以分为5类：划分式的聚类(如K-means聚类算法及其变种)、层次聚类、基于密度的聚类、基于网格的聚类、基于图的聚类(如谱聚类)和基于模型的聚类(如采用最大期望算法的高斯聚类)。在实际应用中，选择哪种聚类算法取决于数据特征和算法的性能表现。而在运行完聚类算法对样本进行簇划分之后，还需要选取适当的性能度量指标对聚类的效果进行评估，以分析聚类算法是否实现了簇内相似度最高而簇间相似度最低的目标。值得一提的是，无监督的聚类算法需要研究者自己定义簇的个数，因此在实际研究中需要通过不断调试模型参数来找到最佳的聚类簇数[37-38]。

在提出理论与数据双驱动阶层测量方法的基本框架后，如何对7个分层要素进行操作化以建构社会空间、选取何种聚类算法以及如何设定模型参数还需要研究者根据具体数据所包含的信息和模型的表现来决定。为此，本文使用中国社会综合调查(CGSS)2017年数据②来建构我国的社会阶层空间，并通过聚类算法来对我国的社会阶层进行划分。

三、我国的社会分层——理论与数据双驱动阶层测量方法的应用

CGSS 2017共收集了12582个样本，根据上述社会分层的7大要素，笔者在数据中选出相关变量对每个要素进行操作化，操作化过程见表2。其中，CGSS 2017的职业编码采用ISCO-08编码，为获得职业声望和职业社会经济地位，本文使用R语言中的ISCO08 ConveRsions程序来生成SIOPS-08和ISEI-08；社会资本的测量参考边燕杰[39]测量城市居民社会资本的方法；网络异质性的测量根据受访者所认识的人中有几个列出的职业类别：网顶为受访者的社会网络中的最高声望，平均网络质量为受访者网络中的平均声望；阅读习惯包括月均读书本数、电子书本数，日均看报纸/杂志数以及日均手机阅读新闻咨询小

表2 分层要素的操作化

分层要素	操作化
经济资源	个人月收入对数、家庭月收入对数
职业与声望	职业编码、声望、职业地位、当前工作状况、工作管理权限、单位类型
单位地位或制度分割	户口、体制、党员身份、城乡
社会资本	网络异质性、网顶(网络中的最高声望)、平均网络质量(平均声望)
民权资源	医疗保险、养老保险
人力资本	受教育年限、听英语能力、说英语能力
文化资本	阅读习惯、生活方式、家庭支出结构(消费、投资)

时数;生活方式来源于问卷A部分生活方式模块中对媒体的使用情况、闲暇时间的活动、在空闲时间做什么事情三个量表,本文将量表进行重新编码,转换成虚拟变量[③]。

因聚类模型不允许数据存在缺失值,但有些样本在职业类型等关键变量上的答案缺失且无法填补,因此本文删除了关键变量缺失的样本,最后得到9726个样本。为检验清理后样本是否会导致关键变量与原样本在分布上的差异,选取收入、受教育程度两个常用于测量社会阶层的重要指标进行检验。从分布形态上看,清理后样本的收入、教育年限[④]和原样本分布形态较为一致,如图1所示。同时对原样本和清理后样本进行了独立样本T检验,结果显示二者在收入和教育程度的分布没有显著差异[⑤]。综上,可以认为删除职业等关键信息缺失的样本并不会导致清理后样本重要指标分布与原样本之间的偏差,本文对样本的清理没有损害原样本的代表性。

(一)社会经济地位分层与地位不一致

为验证在社会阶层空间中通过无监督聚类算法所划分的簇是否能够作为社会阶层,本文先在低维度的社会空间进行探索,以便分析每个簇的阶层特征,并与传统阶层测量方法进行对比。具体而言,以数值型阶层测量方法——社会经济地位指数模型为基准,首先使用社会经济地位理论中考虑的4个关键变量:收入、职业类别、职业声望、受教育年限进行阶层划分,并将结果与国际社会经济地位指数(ISEI)进行对比。

在模型选择上,首先使用K均值聚类算法、高斯混合聚类算法和凝聚层次聚类算法对样本进行聚类。这些聚类算法需要事先设定聚类的簇数,为便于后期对每一类别的特征描述,本文将聚类簇数控制在10类以下,因此模型的簇参数(n-cluster)被设定为3~10共8种选择。此外,凝聚层次聚类算法可以选用不同的相似性(距离)测量方式和凝聚(合并)的规则,本文对三种相似性(距离)测量方式(欧氏距离、曼哈顿距离和余弦相似性)和所有的凝聚规则(计算簇间邻近性的规则,包括单链、全链、组平均和ward方法)都进行了尝试,从中选择聚类效果最好的模型进入下一步分析。聚类的效果使用CH得分(Calinski Harabasz Score)来衡量,得分越高表示簇内相似性越

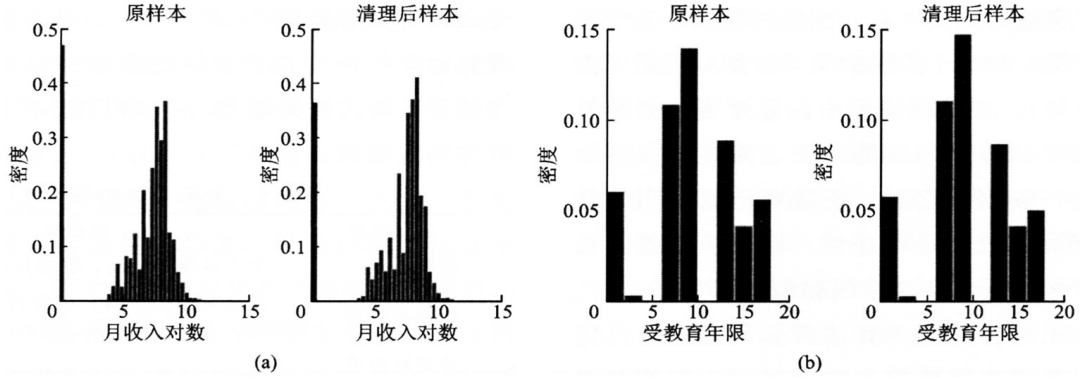

图1 清理后样本与原样本的收入与教育程度分布对比

大而簇间相似性越小，即聚类效果越好。图2是三种聚类模型在不同参数设置下的聚类效果展示。

由图2可知，总体而言K-means算法的聚类效果最好，其中簇数为7的K-means聚类模型与其他模型相比能够最大限度地将相似的人划分在同一个子群体，并将不同的子群体分开。因此本文选用簇数为7的K-means聚类模型来进一步分析不同子群体在各个维度上的特征。由于聚类模型所产生的类别编号没有实际意义，为更直观地观察每个子群体在不同属性上的差异，需要对簇标签进行重新排序。为此，本文选用社会经济地位得分作为排序的标准，以计算每一簇社会经济地位得分的均值，并按照从小到大的顺序对7个簇进行排序，按照顺序对簇标签进行重新编码，然后考察这7个子群体在收入、社会经济地位得分、职业声望得分和教育程度4个维度上的差异，从而评估模型是否实现了阶层划分。

总体而言，该模型从收入、教育、声望和职业所构成的社会阶层空间中识别出了属性和等级不同的7个子群体。由表3可看出，7个阶层的规模差异较大，其中第6阶层的规模最小，仅占总人口3.61%，而第5阶层人数最多，占总人口23.29%。

表3　　7个阶层的人数分布

阶层	人数	占比/%
1	1374	14.13
2	1586	16.31
3	894	9.19
4	1544	15.87
5	2265	23.29
6	351	3.61
7	1712	17.60
总计	9726	100.00

如图3所示，在特征分布上，7个子群体的社会经济地位水平(ISEI得分的分布)和社会声望存在较大的阶梯式差异，根据社会声望分层和社会经济地

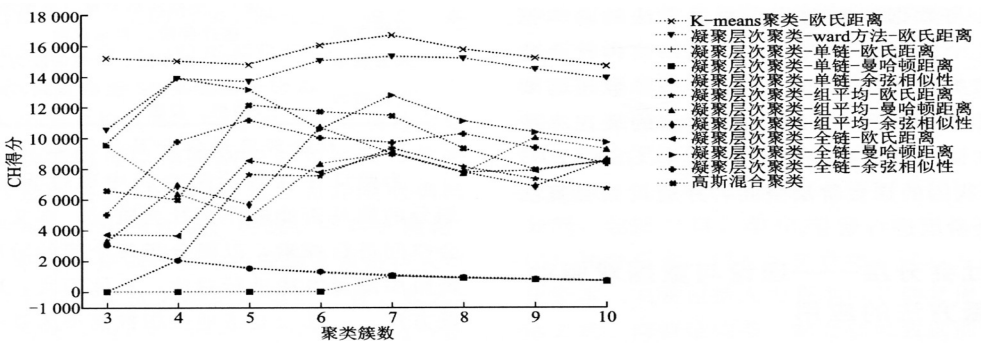

图2　不同模型的聚类效果

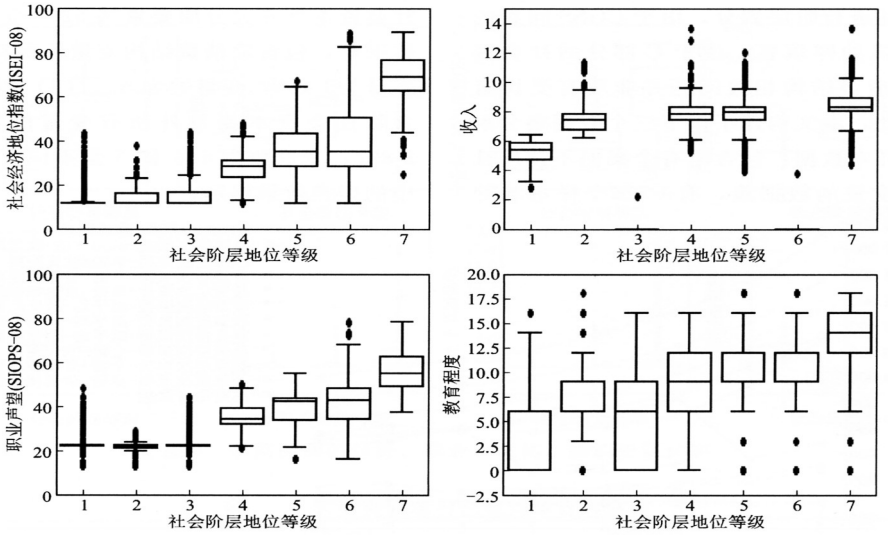

图3　7个阶层的社会经济地位特征

位分层理论可以认为这7个子群体即可作为7个不同的阶层，但是7个阶层的收入水平和教育程度并不完全与阶层等级相符合，尤其是第3和第6阶层的收入水平与其阶层地位完全不匹配，这两个阶层中大多数人的收入为0。

产生这种结果是因为CGSS 2017询问的是受访者去年的收入，而有些受访者2016年处于无工作状态，笔者在处理职业数据时，将目前无工作但是曾经有工作的受访者曾经的职业作为其职业编码，主要是考虑到在现实中职业地位在时间上更具有稳定性，曾经处于高职业地位的个体，其社会经济地位很难因目前收入的减少而产生大幅波动。

如图4所示，当具体分析收入为0的第3阶层和第6阶层的职业类型时，由于职业编码大致是按照职业等级从高到低的顺序进行编码，ISCO的数值越小，说明职业的等级越高，通过对比两个阶层的职业类型分布，无须查看职业编码表便可发现，同样是收入均值和中位数为0的两个群体，第6阶层的职业类型相较于第3阶层的职业类型而言等级更高。也就是说，本文的模型自动识别出了无收入人群中的职业社会经济地位和职业声望不同的两个社会阶层。而在职业社会地位和职业声望相似的第2、3阶层中，模型又通过收入和教育信息识别出了同样处于较低职业地位的两个不同的社会阶层。

以上这种收入与职业的社会经济地位和声望不匹配现象在社会分层理论中被称为“地位不一致”。当使用多个维度进行社会分层时，阶层群体在不同维度上的等级排序可能存在差异，当这种差异过大时，即可以认为产生了地位不一致。地位一致和不一致的程度可以用“地位结晶化”的概念来衡量。高地位结晶化(地位一致)指运用各种分层标准得到的结果都是一致的；低地位结晶化(地位不一致)指运用各种分层标准得到的结果都是不一致的[40]。根据个人在n个地位测量维度和m个等级排序体系下所取得的地位排序组合状况，人们的地位一致性程度又可进一步划分，有学者根据三个维度和三个等级将其划分为地位一致者、中等地位不一致者、绝对地位不一致者和两个地位差四种不同的类型[41]。吉登斯认为，地位一致性程度是判断群体已经形成了阶层还是只是利益群体的关键，如果某个群体各个维度的地位水平高度相关，那么该群体就可以称作一个边界相对清晰的、定型化的阶层；如果某个群体各个维度的地位水平相关程度不明显，那么该群体还不能称作一个相对封闭的、定型化的阶层，只能说该群体在某个维度上成为一个利益群体但没有形成阶层[42-43]。按照上述观点，在以上7个阶层中，第1、2、4、5、7阶层地位一致性程度较高，是边界相对清晰、定型化的阶层；而第3、6阶层的地位一致性相对较低，其阶层的边界相对不够清晰，这恰好体现出本文提出的分层模型的优势。因为如果按照职业划分，第3阶层的成员可能被归到第2阶层，而第6阶层可能被分到第5和第7阶层；而若按照经济进行分层，第3、6阶层会被划分到同一个阶层中，而本文的分层模型既识别出了定型化的阶层，也识别出了这两个特殊的阶层边界不够清晰的群体。

(二)高维社会阶层空间的阶层划分

在使用简洁模型验证了理论与数据双驱动模型的分层效力之后，笔者根据分层理论所涵盖的7大要

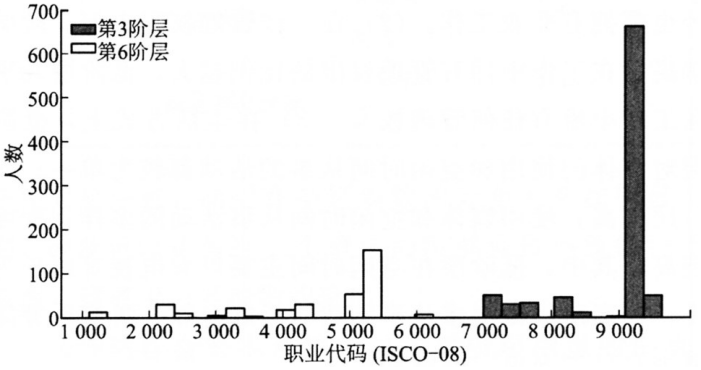

图4 第3和第6阶层的职业分布

素将样本映射到高维空间进行阶层划分。由于CGSS问卷的B、C、D部分是随机抽样填答，位于C部分的社会资本和位于D部分的消费结构相关问题并非所有受访者都进行了回答，因此，本文构建了以下三个数据集分别进行分析。(1)全样本数据：包含所有个案但不使用社会资本和消费结构变量的数据集，有9726个样本和除社会资本外6大分层要素共52个变量。(2)社会资本数据集：包含除消费结构变量外所有变量的数据集，样本量为3430，变量数为55。(3)消费结构数据集：包含除社会资本变量外所有变量的数据集，样本量为2897，变量数为64。图5是不同聚类模型在高维空间中的聚类效果展示。

首先使用全样本数据集建构社会阶层空间来进行阶层划分。为避免各变量的量纲不同对计算聚类所造成的偏差，在对数据进行零均值(Z-score)标准化[⑥]后，使用与简洁模型相同的算法和参数设置对7大分层要素所构成的52维空间中的样本进行聚类。结果显示，在这一空间中，使用K-means聚类算法将样本聚集成3个子群体的CH得分最高，聚类效果最好。因此，本文采纳最佳模型的结果将群体划分为三个阶层，并按照三个子群体的平均社会经济地位得分高低进行排序，以此顺序将其定义为低、中、高三个阶层。表4为三个阶层的人数分布情况，其中，中等阶层规模最大，占总人口的42.22%；高阶层的规模相对较小，占总人口的24.86%。

表4　三个阶层的人数分布

阶层	人数	百分比/%
低阶层	3202	32.92
中等阶层	4106	42.22
高阶层	2418	24.86
总计	9726	100.00

然后考察不同阶层在各维度上的特征，从而评估模型的社会分层效果。如图6所示，模型所划分出

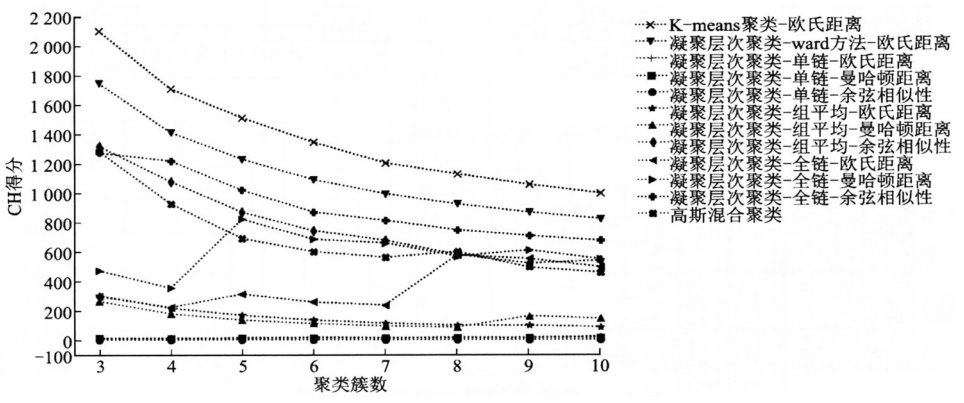

图5　不同聚类模型在高维空间中的聚类效果

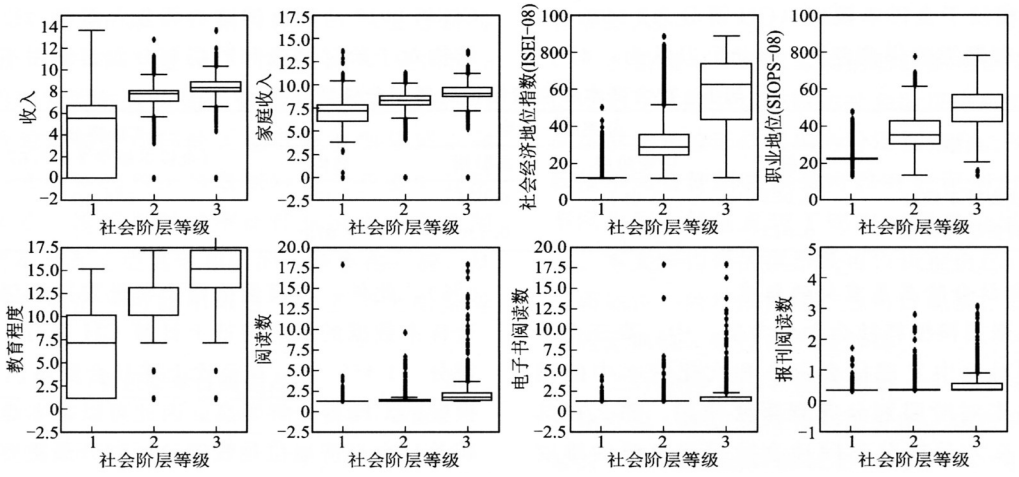

图6　不同阶层的经济、职业、文化、技能资源分布

的低、中、高三个阶层在经济收入水平、职业社会经济地位、职业声望、受教育程度和阅读习惯等经济资源、职业资源、人力资本和文化资本四大阶层要素8个维度上的等级次序完全一致。

表5描述了三个阶层在制度与民权资源上的差异。首先，高、低两个阶层的成员呈现出明显的城乡二元分割，低阶层中91%的成员为农业户口，74%居住在农村地区；而高阶层中80%以上成员为非农业户口，且仅有8%的成员当前居住在农村地区。其次，在党员身份和工作单位体制方面，各阶层党员和体制内人员的比例随着阶层等级的升高而上升。最后，在养老和医疗保险方面，高阶层参与基本养老保险、商业医疗和养老保险的比例更高。此外，三个阶层在职业资源和文化资源其他方面也存在较大差别：(1)在工作经历及当前工作状况方面，低阶层主要由从未工作过和曾经务农现在无工作的人口组成，而高阶层和中等阶层的大部分人当前正在从事非农工作，对于中高阶层当前无工作的人，曾经也都拥有非农工作。(2)在工作管理权限方面，阶层越高，在工作中拥有管理权限的比例越大，低阶层几乎在工作中没有任何管理权⑦。(3)在生活方式上，低阶层对媒体的使用和空闲时间从事的活动都较为单一；而阶层越高，使用媒体和空闲时间从事活动的多样化程度越高。其中，低阶层在空闲时间主要以看电视等娱乐为主，而高阶层则更多从事阅读、锻炼身体、听音乐等能够为自己“充电”的活动⑧。

由此可见，模型所划分的三个阶层既在等级参数分布上存在着低、中、高的差异，且在8个维度上的等级次序完全一致，同时在类别参数的分布上存在明显差别，因此可以认为这三个群体的阶层地位一致性程度较高，形成了边界清晰的阶层。

为评估社会资本和消费结构对阶层测量的影响，本文继续加入社会资本特征和消费结构特征进行阶层划分，并用“类别不一致率”作为测量社会资本特征与消费结构对阶层划分的指标。类别不一致率的计算方式是：以上述全样本数据的阶层划分结果为基准阶层类别C，类别不一致率即为使用其他特征(数据集)进行阶层划分之后的类别C_i与基准类别C不一致成员的比例⑨。

分别使用社会资本数据集和消费结构数据集来建构阶层社会空间，为与基准类别进行对比，同样使用簇数为3的K-means聚类模型进行阶层划分，且按

表5　　三大阶层制度与民权资源占有情况

变量			低阶层		中等阶层		高阶层	
			人数	比例	人数	比例	人数	比例
			3202	0.33	4106	0.42	2418	0.25
单位体制或制度分割	户口	农业	2926	0.91	1986	0.48	417	0.17
		非农	276	0.09	2120	0.52	2001	0.83
	居住地区	农村	2358	0.74	1096	0.27	183	0.08
		城市	844	0.26	3010	0.73	2235	0.92
	党员身份	非中共党员	3105	0.97	3828	0.93	1649	0.68
		中共党员	97	0.03	278	0.07	769	0.32
	单位体制	体制外	3155	0.99	2698	0.66	1084	0.45
		体制内	47	0.01	1408	0.34	1334	0.55
民权资源	基本医疗保险	未参加	259	0.08	336	0.08	135	0.06
		参加	2943	0.92	3770	0.92	2283	0.94
	基本养老保险	未参加	980	0.31	1201	0.29	411	0.17
		单价	2222	0.69	2905	0.71	2007	0.83
	商业医保	未购买	3126	0.98	3732	0.91	1774	0.73
		购买	76	0.02	374	0.09	644	0.27
	商业养老保险	未购买	3139	0.98	3840	0.94	2003	0.83
		购买	63	0.02	266	0.06	415	0.17

照ISEI对模型所得出的簇标签进行排序，得到低、中、高三个阶层，如表6所示。总体而言，两个模型的分类不一致率较低，对样本的阶层划分均与全样本模型的阶层划分相差不大，加入社会资本和消费结构特征后，分别仅有5.9%和7.5%的成员阶层类别发生了变化。从模型的表现上来看，加入社会资本和消费结构变量后，模型的CH得分相较于全样本模型(CH得分为2098.67)大幅降低，模型聚类效果变差[10]。因此可以认为，对于CGSS 2017所调查的这一部分人而言，社会资本和消费结构对于阶层测量和阶层划分而言作用不大[11]。

(三)区分社会阶层最重要的维度

上文通过聚类模型将社会划分为低、中、高三个阶层，并在模型对比中发现社会资本和文化资本中的消费结构特征并未对阶层划分起到重要作用。那么在其他特征中，什么才是区分不同社会阶层最重要的维度呢？由于聚类模型的类别划分原理综合了所有特征属性来计算样本之间的相似性，是一个“黑箱操作”，故无法得知人们哪些特征上的相似或差异在决定被划分到哪个群体时的作用更大。为了找出社会分层最重要的维度，本文将这个问题转化为机器学习分类模型的特征选择问题，即从个体的阶层类别来反推识别阶层类别最重要的特征。具体而言，以全样本聚类模型对个体的阶层归类作为个体的真实阶层类别，然后基于全样本数据中的52个特征训练出能够准确识别每个个体阶层类别的决策树模型，最后比较每个特征对于模型分类的重要性，重要性最高的特征便是判断个体阶层类别的最重要特征。

在训练决策树时，首先将全部样本按照1:1的比例随机划分成训练集和测试集两个部分，用训练集训练模型，用测试集评估模型预测效果。在不经过任何调参的情况下，使用不同的初始状态运行1000次，模型均可达到90%左右的准确率[12]，因此可以认为该简洁模型对个体阶层的识别能力可以达到分析要求[13]。

通过对全样本阶层聚类模型中所使用的52个特征(变量)的重要性进行分析，发现绝大部分(96%)特征对判断个体阶层的重要性都不足0.1。笔者选取了对预测阶层类别的重要程度大于0.01的特征(变量)在图7中进行展示，预测个体阶层类别最重要的是“单位类型”，其次是“职业社会经济地位得分”。也就是说，当前在我国阶层测量中最重要的因素是单位类型和职业社会经济地位得分。笔者通过仅使用以上两个特征对全样本数据进行簇数为3的K-means聚类分析来对这一发现进行进一步验证，结果发现，仅使用两个特征的聚类模型的平均类别不一

表6 社会资本和消费结构阶层划分模型分类

模型	分类不一致率				CH得分(全样本模型2098.67)
	低阶层	中等阶层	高阶层	总体	
社会资本模型	0.054	0.097	0.026	0.059	676.60
消费结构模型	0.013	0.083	0.130	0.075	414.67

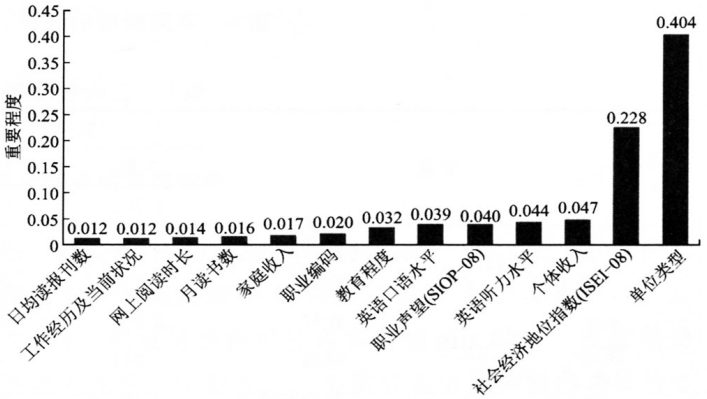

图7 阶层测量特征的重要性

致性为27.2%，也就是说，仅使用单位类型和职业社会经济地位便能够实现对70%以上的人口群体进行准确的阶层划分。并且，模型对低阶层的识别更好(低阶层的类别不一致性低至6%)。此外，该模型按两个特征聚类模型的阶层分类在全样本数据的所有特征上计算CH得分，所得出的CH得分(1458.51)甚至比上述社会资本模型和社会网络模型的CH得分都要高。因此可以基本确认单位类型和职业社会经济地位是我国阶层划分最重要的维度。

那么，各阶层的单位类型到底存在何种差异呢？如表7所示，三个阶层成员的单位类型均有各自的特征：低阶层主要由务农和无工作的人口构成，高阶层则有50%以上的成员主要来自国家机关、事业单位、国企、集体企业等体制内的工作单位，且有部分(30%左右)成员来自民营企业、私企和外企；而中等阶层的单位类型则以个体工商户和自由职业者(34.61%)以及民营、私企(25.6%)为主，也有一部分成员(35.7%)来自体制内工作单位。

需要说明的是，虽然单位类型是划分我国社会阶层最重要的维度，且三个阶层的单位类型构成的确存在差异，但这并不意味着可以依据单位类型直接对社会阶层进行划分，这也是本文构建的模型和传统分层模型的最大差别，即传统的分层方法是以变量为中心，即可以根据一些有重要经济社会差异的类别变量(如单位类型)对模型进行定类划分，或者根据可以反映社会经济等级的连续变量(如ISEI)进行“划线切割”，但单位类型相同的人可能会被划分到不同的阶层，职业社会经济地位得分相同的人也可能被划分到不同的阶层。而本文构建的理论与数据双驱动模型的分层是以人群为中心，根据不同的特征计算出人与人之间的距离，并将类别和等级属性相似的人聚集在一起，从而保证阶层内部成员尽可能相似且不同阶层之间差异的最大化。

四、结论与讨论

如何测量和划分社会阶层是社会分层研究者长期争论的焦点。我国社会学研究者对如何分层做出了许多尝试，提出了丰富的阶层测量方法和分层体系，得出了对我国社会阶层结构的不同看法，但这些研究都面临一个问题——阶层测量和阶层划分的真实性和有效性难以在现实中得到验证。在大数据时代，尽管一些研究者做出了基于社会现实(大量的、真实的数据)测量阶层的尝试，但由于其数据的限制导致其测量指标可能并不是区分阶层的关键。本文在回顾社会分层理论和研究中的经典分层理论、方法模型和具有代表性的分层研究后，将当前的社会分层方法归纳为理论和数据驱动的两种阶层测量范式，通过对比两种范式下的分层方法，发现二者各自存在弊端。为此，尝试提出将二者结合起来的理论与数据双驱动的阶层测量框架：理论驱动在于根据过往分层理论中提出的对阶层划分有意义的资源(要素)整合了7种分层要素及其操作化方法，基于布

表7 各阶层单位类型占比 %

单位类型	低阶层	中等阶层	高阶层
党政机关、人民团体、军队	0.16	1.95	8.85
国有/集体事业单位	0.41	9.11	26.05
国有企业	0.22	17.73	17.33
集体企业	0.69	5.50	2.94
村居委会等自治组织	0.25	1.41	1.74
民营、私营企业	4.15	25.60	26.84
外资、合资企业	0.12	1.14	3.35
民办非企业、社团等社会组织	0.12	0.51	1.24
个体工商户和自由职业者	10.27	34.61	10.38
其他	1.66	2.19	1.24
务农	44.60	0.17	0.04
无工作	37.35	0.07	0
合计	100	100	100

劳的宏观社会结构理论来构建分层的社会阶层空间;数据驱动在于使用无监督聚类方法,完全由机器决定社会应该分为几个阶层,以及每个阶层包含哪些人。

在数据与理论双驱动的阶层测量框架下,本文使用CGSS 2017数据对中国的社会阶层进行划分。首先使用简洁模型来验证所提出的分层框架和方法的有效性,发现使用无监督聚类模型可以有效识别出社会空间中的不同阶层,并且发现了现实中存在阶层地位不一致现象。本文所构建的模型既可以识别出已经形成阶层边界的高地位一致性的阶层,也可以识别出阶层边界尚不清晰的低地位一致性的利益群体。

之后建构了包含经济资源、职业资源、人力资本、文化资源、单位地位和制度分割、民权资源6大分层要素共52个维度的高维社会阶层空间,并使用聚类模型进行阶层划分,结果显示,高维空间中的人口可被划分为三个子群体,通过比较三个子群体在收入、声望、职业社会经济地位、人力资本和文化资本上的差异,发现这三个边界清晰的群体代表着我国社会低、中、高三个阶层,且这三个阶层具有高地位一致性。接着使用社会资本模型和消费结构模型对人口进行分层,通过对比这两个模型与全样本模型的分类一致性,发现加入社会资本和消费结构的相关变量并不会引起分层结果的改变,并且在考虑更多特征后,模型的性能反而下降。也就是说,社会资本和消费结构特征对社会分层的作用不大。

那么社会分层最重要的指标究竟为何呢?本文进一步使用机器学习的决策树模型分析了每个指标(特征)对于测量(预测)个体阶层等级的重要性。结果发现,在我国,单位类型是社会分层最重要的指标,职业社会经济地位次之,而其他特征对估计个体社会阶层的重要性微乎其微。进一步使用仅含有单位类型和社会经济地位得分的模型进行阶层划分,结果巩固了这一结论:当仅考虑单位类型和职业社会经济地位水平时,模型对70%人口的阶层划分与考虑52个指标时并无差异。

本研究还存在一些需要改进之处。首先,在分层指标的操作化上,由于数据的局限性,对社会资本的测量较为简单,只考虑了个体的网络规模和网络所蕴含资源的最高可达性和异质性,还需要收集更多社会网络结构和整体网的数据,将个体网络结构和个体在整体网络结构中所处的位置纳入社会资本的测量中。其次,本文所得出的"社会资本和消费结构特征对于阶层划分意义不大"的结论是基于CGSS 2017的数据得出的结果,但由于CGSS 2017在询问社会网络相关议题时,只是随机选择了1/3的受访者进行填答,因此样本量较全样本而言有较大损失。虽然进行了多种验证发现这一结论具有稳健性,但若条件允许,在同一个样本上进行比较更为严谨。最后,本研究只是基于CGSS 2017数据,得出结论的稳健性还需要进一步使用其他数据进行验证。

此外,本研究仅仅是对社会分层方法上的探索,并基于这一方法对我国的社会阶层划分做出尝试。今后可以努力的方向还有很多,例如使用其他国家的数据,利用这一方法进行国际社会分层的比较等。同时,阶层是一种社会结构的维度,当前对社会阶层的划分主要采取的是地位结构观这一理论视角,即把阶层视为属性和等级不同的群体,但社会结构还有另外一种理论视角——网络结构观,在这一视角下,对群体的划分一般采用社团分割的办法——基于人与人之间实际存在的交往关系所形成的群体分化来进行阶层划分,这种方法仍然值得探索。社会分层包含两个层面,一是测量和划分阶层,本文已进行了探索;二是理解阶层结构是如何形成的以及如何随着社会的发展而产生变化的,这也是笔者下一步努力的方向。

注释:

①因为陆学艺的阶层划分是根据十大阶层在经济资源、组织资源、文化资源上的差异来划分的,谢立中认为,若要证明这一分层符合现实,也应该证明十大阶层经济资源、组织资源、文化资源上的差异,但李春玲却分析了十大阶层在收入、声望、社会经济地位指数、消费等方面的差异,所以这里的"事实"标了引号。

②在比较了CLDS、CGSS等全国大型综合调查历年数据后发现,CGSS 2017数据能够更加全面地涵盖上述社会分层的7个要素,而其他年份的CGSS数据以及CLDS的数据存在关

键模块的缺失，故选用CGSS 2017数据。

③A28、A31题答案中的1～2编码为0，3～5编码为1；A30题答案中的1～3编码为1，4～5编码为0。

④本文将受教育程度处理成了受教育年限(连续变量)。

⑤篇幅所限，T检验结果未列出，如有需要可向笔者索取。

⑥首先将类别变量按照一定顺序重新编码成定序变量，然后使用零均值标准化的方法将转化后的类别变量和连续变量标准化为Z-score值，其计算方式为Z-score=(原始值-均值)/标准差。

⑦篇幅所限，三个阶层工作经历及当前工作状况分析结果未列出，感兴趣的读者可向笔者索要。

⑧篇幅所限，三个阶层休闲方式差异分析结果未列出，感兴趣的读者可向笔者索要。

⑨为消除样本量变化所带来的差异，本文也测试了以去掉这两个样本中的社会资本和消费结构特征的数据所得出的分层结果作为基准，并与社会资本模型和消费结构模型的结果进行对比，其类别一致性均比较高。

⑩CH得分的差异也可能是因为模型的样本量不同所导致。

⑪这一结论还需要更多数据验证，因CGSS 2017对社会资本的测量比较简略，且回答了社会资本和消费结果问题的样本量较少。

⑫1000个模型准确率均值为0.91，标准差为0.005，最大值为0.92，最小值为0.89。

⑬由于本文关注的是特征的重要性而不是模型的预测准确率，模型能够识别绝大多数样本的阶层，说明当前的特征分析机制已经满足了模型预测的需求。

参考文献：

[1]李强.社会分层十讲[M].2版.北京：社会科学文献出版社，2011.

[2]格伦斯基.社会分层[M].2版.王俊，译.北京：华夏出版社，2005.

[3]赖特.阶级[M].刘磊，吕梁山，译.北京：高等教育出版社，2006.

[4]GOLDTHORPE J H, HOPE K. The Social Grading of Occupations: A New Approach and Scale[M]. Oxford: Clarendon Press, 1974.

[5]GOLDTHORPE J H, LLEWELLYN C. Class mobility in modern Britain: Three theses examined[J]. Sociology, 1977, 11(2): 257-287.

[6]GOLDTHORPE J H, PAYNE C. On the class mobility of women: Results from different approaches to the analysis of recent British data[J]. Sociology, 1986, 20(4): 531-555.

[7]侯利明，秦广强.中国EGP阶层分类的操作化过程：以中国综合社会调查(CGSS)数据为例[J].社会学评论，2019(2)：16-26.

[8]VALDES D M, DEAN D G. The North-Hatt occupational prestige scale: A modest replication[J]. The American Journal of Economics and Sociology, 1965, 24(3): 257-260.

[9]BIAN Y. Chinese occupational prestige[J]. International Sociology, 1996, 11(2): 161-186.

[10]TREIMAN D J. A standard occupational prestige scale for use with historical data[J]. Journal of Interdisciplinary History, 1976, 7(2): 283-304.

[11]TREIMAN D J. Occupational Prestige in Comparative Perspective[M]. New York: Academic Press, 1977.

[12]DUNCAN O D. A Socioeconomic Index for All Occupations[M]. New York: Free Press, 1961.

[13]BLAU P M, DUNCAN O D. The American Occupational Structure[M]. New York: Free Press, 1978.

[14]GILLIAN S, DAVID L F. A revised socioeconomic index of occupational status[J]. Social Science Research, 1981, 10(4): 364-395.

[15]李春玲.当代中国社会的声望分层：职业声望与社会经济地位指数测量[J].社会学研究，2005(2)：74-102.

[16]GANZEBOOM H B G, De GRAAF P M, TREIMAN D J. A standard international socio-economic index of occupational status[J]. Social Science Research, 1992, 21(1): 1-56.

[17]GANZEBOOM H B G, TREIMAN D J. Internationally comparable measures of occupational status for the 1988 international standard classification of occupations[J]. Social Science Research, 1996, 25(3): 201-239.

[18]BRADLEY R H, CORWYN R F. Socioeconomic status and child development[J]. Annual Review of Psychology, 2002, 53(1): 371-399.

[19]凡勃仑.有闲阶级论[M].蔡受百，译.北京：商务印书馆，1964.

[20]布尔迪厄.区分：判断力的社会批判[M].刘晖，译.北京：商务印书馆，2015.

[21]DING S, HUANG H, ZHAO T, et al. Estimating socioeconomic status via temporal-spatial mobility analysis: A case study of smart card data[C]. International Conference on Computer Communication and Networks(ICCCN), 2019: 1-9.

[22]XU Y, BELYI A, BOJIC I, et al. Human mobility and socioeconomic status: Analysis of Singapore and Boston[J]. Computers, Environment and Urban Systems, 2018(72): 51-67.

[23]BLUMENSTOCK J, CADAMURO G, et al. Predicting poverty and wealth from mobile phone metadata[J]. Science, 2015, 350(6264): 1073-1076.

[24]PREOTIUC-PIETRO D, VOLKOVA S, LAMPOS V, et al. Studying user income through language, behaviour and affect in social media[J]. Plos One, 2015, 10(9): e0138717.

[25]LAMPOS V, ALETRAS N, GEYTI J K, et al. Inferring the socioeconomic status of social media users based on behaviour and language[C]. Chain: European Conference on Information Retrieval, 2016: 689-695.

[26]EAGL N, MACY M, CLAXTON R. Network diversity and economic development[J]. Science, 2010, 328(5981): 1029-1031.

[27]陆学艺.当代中国社会阶层研究报告[M].北京:社会科学文献出版社,2002.

[28]孙立平.断裂:20世纪90年代以来的中国社会[M].北京:社会科学文献出版社,2003.

[29]孙立平.失衡:断裂社会的运作逻辑[M].北京:社会科学文献出版社,2004.

[30]李春玲.中国社会分层与流动研究70年[J].社会学研究,2019(6):27-40.

[31]谢立中.当代中国的阶级或阶层结构:两种不同话语系统的"真实性"辨析[J].山东社会科学,2016(3):61-74.

[32]李春玲.断裂与碎片:当代中国社会阶层分化实证分析[M].北京:社会科学文献出版社,2005.

[33]布劳.不平等和异质性[M].王春光,谢圣赞,译.北京:中国社会科学出版社,1991.

[34]BLAU P M. A macrosociological theory of social structure[J]. American Journal of Sociology, 1977, 83(1): 26-54.

[35]MCPHERSON J M, RANGER-MOORE J R. Evolution on a dancing landscape: Organizations and networks in dynamic Blau Space[J]. Social Forces, 1991, 70(1): 19-42.

[36]SURHONE L M, TIMPLEDON M T, MARSEKEN S F. Blau Space[M]. Montana: Betascript Publishing, 2010.

[37]周志华.机器学习[M].北京:清华大学出版社,2016.

[38]SAXENA A, PRASAD M, GUPTA A, et al. A review of clustering techniques and developments[J]. Neurocomputing, 2017 (267): 664-681.

[39]边燕杰.城市居民社会资本的来源及作用:网络观点与调查发现[J].中国社会科学,2004(3):136-146.

[40]伦斯基.权力与特权:社会分层的理论[M].关信平,陈宗显,谢晋宇,译.杭州:浙江人民出版社,1988.

[41]JACKSON E F. Status consistency and symptoms of stress[J]. American Sociological Review, 1962, 27(4): 469-480.

[42]吉登斯.社会的构成[M].李康,李猛,译.北京:生活·读书·新知三联书店,1998.

[43]董运生.地位不一致与阶层结构变迁[D].长春:吉林大学,2006.

On the Social Stratification Research Driven by Theory and Data

Liang Yucheng　Jia Xiaoshuang

Abstract: The theories and methods of previous social stratification research can be summarized into two paradigms: theory-driven research and data-driven research. Comparison and review of these two paradigms revealed that both of them have inevitable limitations. Thus, a framework that combines the advantages of these two paradigms is proposed, which is called theory-and-data-driven social stratification research framework. To define social classes, this framework regards social class as subgroups gathered in a high-dimensional social space composed of nominal and graduated parameters, and constructs the space of social class using dimensions which are selected from the criteria proposed by the social stratification theories, and uses machine learning algorithms to identify the subgroups in this space. When using this framework to identify the social classes of samples from CGSS 2017, it's found that this framework can not only distinguish social classes that have high consistencies and clear boundaries, but also accurately identify subgroups with inconsistent status which have not yet formed a social class. In addition, it is found that: (1) Chinese society can be stratified into three social classes which can be graded into high, medium or low level on economy, prestige, culture and other feature dimensions;(2) Using as many dimensions as possible is not necessary, the most useful criterion for social stratification is Danwei, followed by ISEI.

Key words: social stratification; theory-driven; data-driven; machine learning; space of social class; measurement of social class

