A 股股票数据分析与可视化

2022 年数据可视化课程期末报告

学号：2022103795

姓名：吕明倩

日期：2022 年 1 月 13 日

目　录

1 数据介绍

本数据集内容为 2000 年至 2021 年 A 股上市公司的常用指标，初始数据集共包含 37 个字段，其中共有 11 个名义指标以及 26 个数值指标，各字段名称以及含义解释如表 1 和表 2 所示，观察到数据集中存在大量缺失值，需要对其进行处理。

字段名称	含义	非空值数
stkcd	证券代码	49284
year	数据年份	49284
证券代码	证券代码	49284
行业代码	行业分类：细分至二级行业	49284
Industry	行业分类：制造业取两位代码，其他行业用大类	49284
Year	数据年份	49284
Loss	是否亏损：当年净利润小于 0 取 1，否则取 0	49284
Dual	两职合一：董事长与总经理是同一个人为 1，否则为 0	49284
SOE	是否国有企业：国有控股企业取值为 1，其他为 0	45078
Big4	是否四大：经由四大（普华永道、德勤、毕马威、安永）审计为 1，否则为 0	49243
Opinion	审计意见：财务报告被出具了标准审计意见取值为 1，否则为 0	49242

表 1：名义指标字段名称、含义以及非空值数

字段名称	含义	计算方法	非空值数
Size	公司规模	年总资产的自然对数	49281
Lev	资产负债率	年末总负债除以年末总资产	49281
ROA	总资产净利润率	净利润/总资产平均余额	49280
ROE	净资产收益率	净利润/股东权益平均余额	48714
ATO	总资产周转率	营业收入/平均资产总额	49280
Cashflow	现金流比率	经营活动产生的现金流量净额除以总资产	49281
REC	应收账款占比	应收账款净额与总资产的比值	48906
INV	存货占比	存货净额与总资产的比值	48309
FIXED	固定资产占比	固定资产净额与总资产比值	49278
Growth	营业收入增长率	本年营业收入/上一年营业收入-1	49029
Board	董事人数	董事会人数取自然对数	49211
Indep	独立董事比例	独立董事除以董事人数	49211
Top1	第一大股东持股比例	第一大股东持股数量/总股数	46192
Top5	前五大股东持股比例	前五股东持股数量/总股数	46192
Top10	前十大股东持股比例	前十股东持股数量/总股数	46192
Balance1	股权制衡度	第二大股东持股比例除以第一大股东持股比例	46191
Balance2	股权制衡度	第二到五位大股东持股比例的和除以第一大股东持股比例	46192
BM	账面市值比	账面价值/总市值	48676
TobinQ	托宾 Q 值	（流通股市值 + 非流通股股份数 × 每股净资产 + 负债账面值)/总资产	48368
ListAge	上市年限	ln(当年年份-上市年份 +1)	49284
FirmAge	公司成立年限	ln(当年年份-公司成立年份 +1)	49284
Dturn	月均超额换手率	当年股票月均换手率 - 去年股票月均换手率	45229
INST	机构投资者持股比例	机构投资者持股总数除以流通股本	49280
Mshare	管理层持股比例	管理层持股数据除以总股本	47633
Mfee	管理费用率	管理费用除以营业收入	48441
Occupy	大股东资金占用	其他应收款除以总资产	48686

表 2：数值指标字段名称、含义、计算方法以及非空值数

1.1 数据清洗

处理初始数据集中包含的缺失值，注意到对于 ROE、ROA、ATO 等指标，每年都有缺失值出现，而对于股权类指标 $TOP1$、$TOP5$、$TOP10$、$Balance1$、$Balance2$，缺失值集中出现在 2000 年、2001 年、2002 年以及 2020 年，对于国企判断类指标 SOE，其缺失值仅出现在 2000 年至 2004 年。

首先对于缺失值优先采用向前填充法，即若某证券本年度某字段缺失时，接受本年度公司发展情况与上一年度保持一致的假设，用上一年的数据填充，当无上一年对照数据时，采用向后填充法，用后一年的对应指标数据填充，若此时表中数据仍有缺失，则将对应的整行数据删失。

注意到对于字段 $Size$、Lev、$Cashflow$ 和 $FIXED$，采用向前填充即可处理全部缺失值，而对于其他字段仍需进行向后填充。经过数据清洗后得到的 $processed\_data$ 数据集共涉及 4016 支证券，共包含 47945 条数据，将该数据集作为后续问题分析的数据基础。

1.2 指标相关性分析

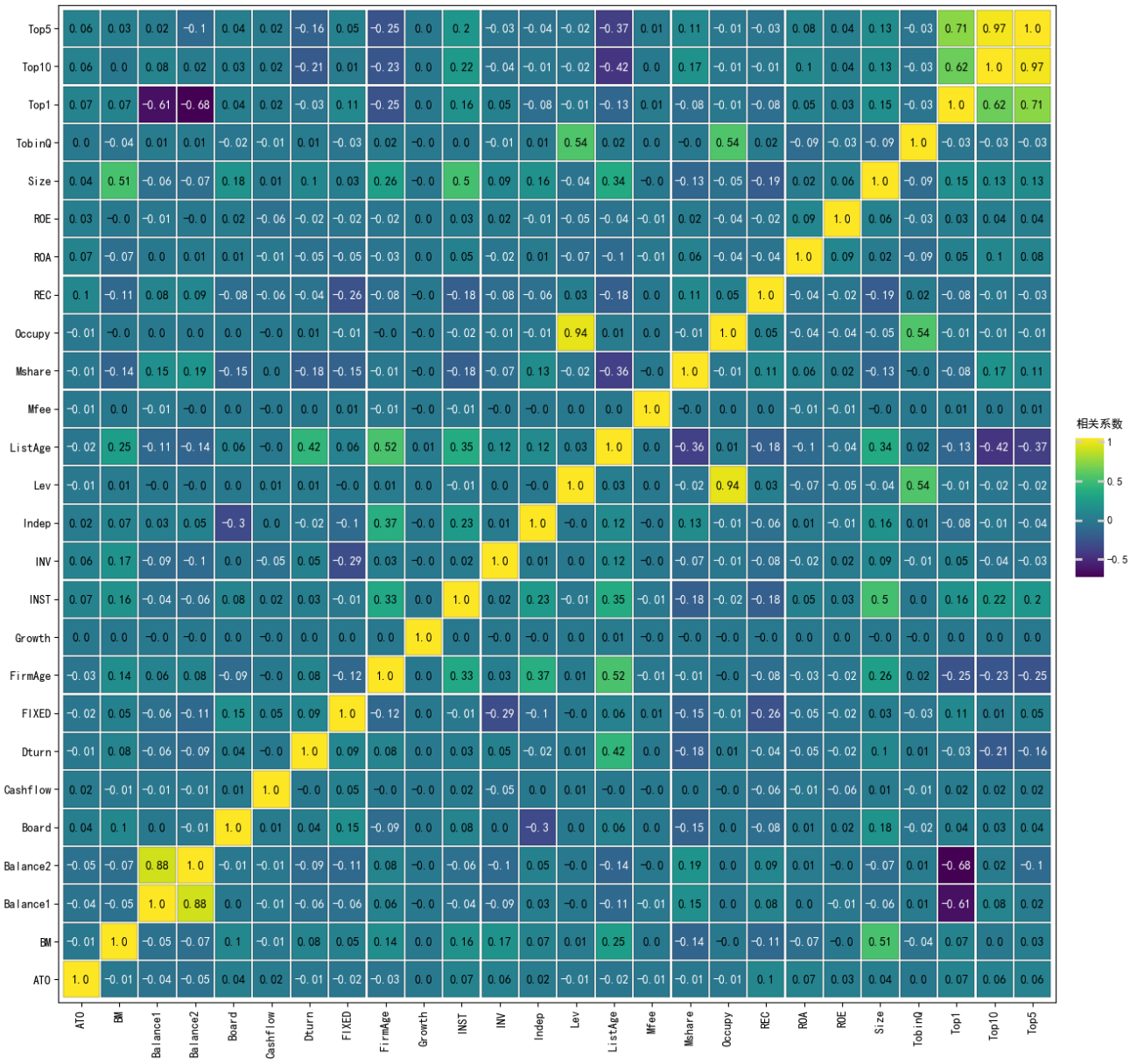

图 3：各指标之间的相关性

计算全部数值型指标的相关系数，发现大多数指标之间的相关性非常低，可以近似认为大多数指标是不相关的。注意到除了资产负债率与大股东资金占用的相关系数高达 0.94，具备超强的正相关性外，与股东持股比例有关的指标 $Top1$、$Top5$ 和 $Top10$ 之间也具有强的正相关性，此外两个股权制衡度的正相关系数也达到了 0.88，同时股权制衡度也与 $Top1$ 之间存在较强的负相关性，相关系数超过-0.61，这意味着第一大股东的持股比例越高，对该证券的股权制衡度越低，将股权集中在一个人手中不利于实现股权制衡，个人承担的风险性也越大。

1.3 问题陈述

通过对数据的清洗整理与初步分析来发掘感兴趣的问题。

首先注意到数据集中存在 5 个 0-1 分类指标，其中指标 SOE 标记了每支证券每年是否属于国有控股企业，当该证券本年度属于国有控股企业时，SOE 取值为 1，否则取值为 0。基于此可以实现对数据集的分类，判断每支证券是否属于国有控股企业，观察每支证券在观测时间段内是否发生过企业类型的转变，统计各年份企业类型的转变情况来分析不同年份之间的表现差异，还可以研究国企与非国企在股东持股比例以及股权制衡等问题上的异同点。

其次，每支股票均存在行业类型的划分，一级行业类型可以分为以下 19 类，行业代码以及对应的行业类型统计如表 4 所示：

代码	行业	代码	行业
A	农、林、牧、渔业	K	房地产业
B	采矿业	L	租赁和商务服务业
C	制造业	M	科学研究和技术服务业
D	电力、热力、燃气及水生产和供应业	N	水利、环境和公共设施管理业
E	建筑业	O	居民服务、修理和其他服务业
F	批发和零售业	P	教育
G	交通运输、仓储和邮政业	Q	卫生和社会工作
H	住宿和餐饮业	R	文化、体育和娱乐业
I	信息传输、软件和信息技术服务业	S	综合行业
J	金融业		

表 4：行业分类

因此也可以根据一级行业类型对数据集进行划分，观察不同行业的盈利表现、经营状况以及各行业财务指标随时间的变化发展趋势，来达到挖掘行业特点的目的。

2 国企控股证券与非国企控股证券

2.1 企业类型的分类与转变

首先统计所有证券的企业控股类型。对于数据集中包含的 4016 支证券而言，他们所属的企业类型可以分为三大类：第一类和第二类分别是在整个时间跨度内始终属于国企控股和始终属于非国企控股的证券，分别占比 22.3% 和 64.9%，而第三类则是在数据集时间范围内发生过企业类型转换的证券，总占比为 12.8%。

因此对于本数据集而言，非国企控股的证券数目要远远多于其他类型，而始终被国企控股的证券数目大约仅为非国企控股证券数目的三分之一，这反映出大部分企业仍然属于非国企控股类型。

此外，对于第三类发生过类型转变的证券，还可以进行更详细的划分。关注每支证券的企业属性随时间的变化情况，又可以细分为三类，分别是从非国企控股转为国企控股的证券、从国企控股转为非国企控股的证券以及发生过多次企业类型转变的证券，发生前两类转变的证券数目大致相同，在发生过企业类型转变的证券中约有 17.2% 发生过两次或两次以上的企业类型转变。

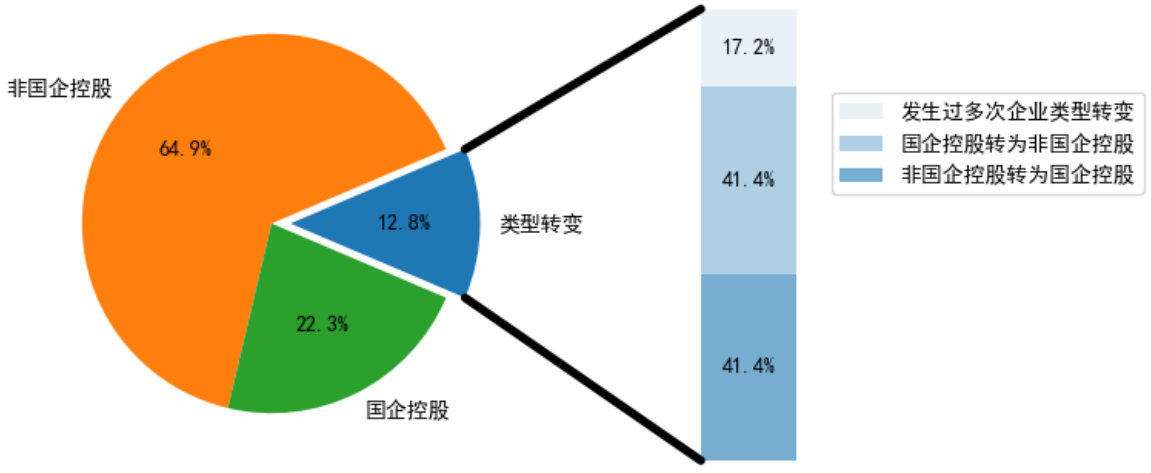

图 5：国企控股与非国企控股的证券数量统计

进一步探究发生企业类型转变的时间结点是否具有某种规律，逐年分别统计每年转变为国企控股和转变为非国企控股的证券数目，不难发现在 2006 年和 2007 年附近，有大量证券转变为非国企控股，而在 2019 年至 2021 年则出现大量证券转变为国企控股的现象。这意味着近三年的经济大环境，使得不少企业成功发展为国企企业，这一点与当下的企业发展趋势相吻合。

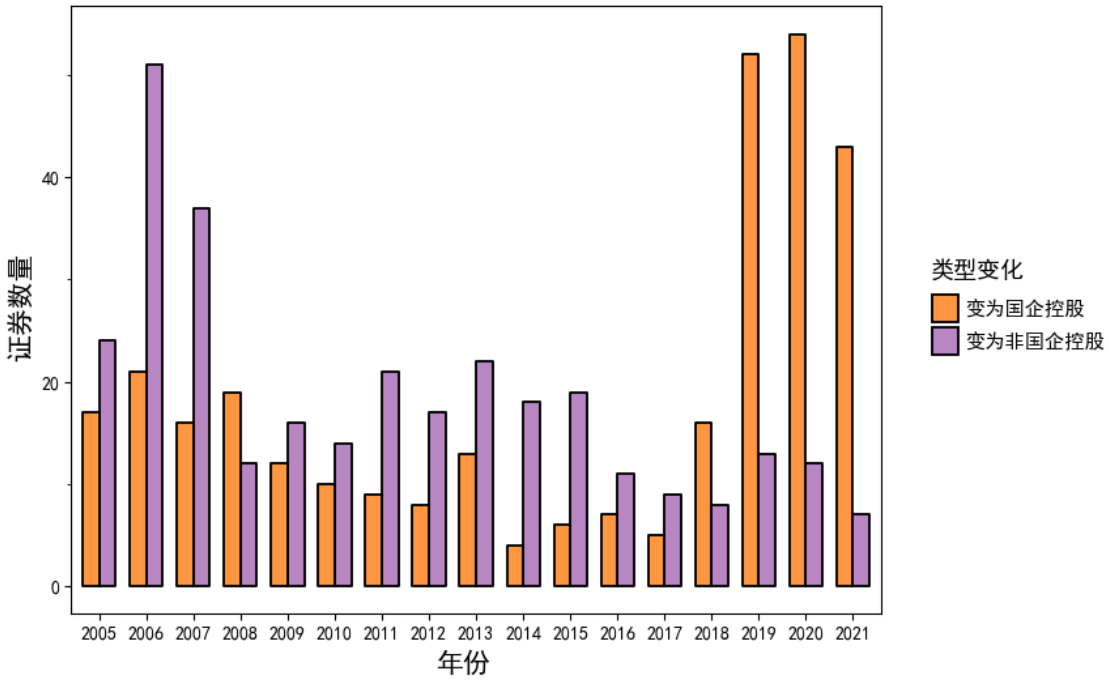

图 6：每年企业类型转变数量统计柱形图

2.2 各企业类型的变化差异

虽然近三年有大量企业转型为国有企业，但从总体上看，近几年非国企控股证券总数显著多于国企控股证券总数。图 7 展示了数据集中每年是否属于国企控股的证券数目，二者均随时间呈现出增长趋势，但与国企控股证券数目相比，非国企控股证券数目的增长明显更快，在 2000 年国企控股证券暂时多于非国企控股证券，但伴随着全国经济的高速发展，非国企控股证券数目在 2010 年实现首次超越。

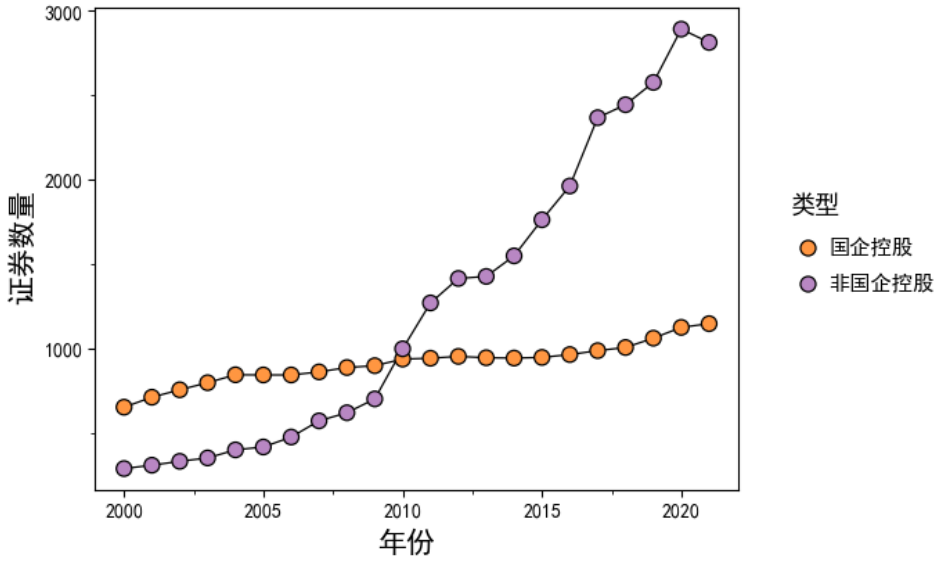

图 7：各企业类型证券数目的时间变化图

为了探究国企控股证券与非国企控股证券在股权性质上的差异性，绘制第一大股东持股比例、前五大股东持股比例、前十大股东持股比例以及股权制衡度的散点图和核密度曲线。不难发现结论具有一致性，对于国企控股证券，三个股东持股比例指标的峰度明显更低，这意味着国企控股企业的股权更为分散，而对于非国企控股企业其股权更集中地掌握在少数人手中。对于股权制衡度指标，国企控股证券反而具有更高的峰值，第二到五位大股东持股比例与第一大股东持股比例的比值更高，也反映出国企控股企业的第一大股东持股比例较低的情况。

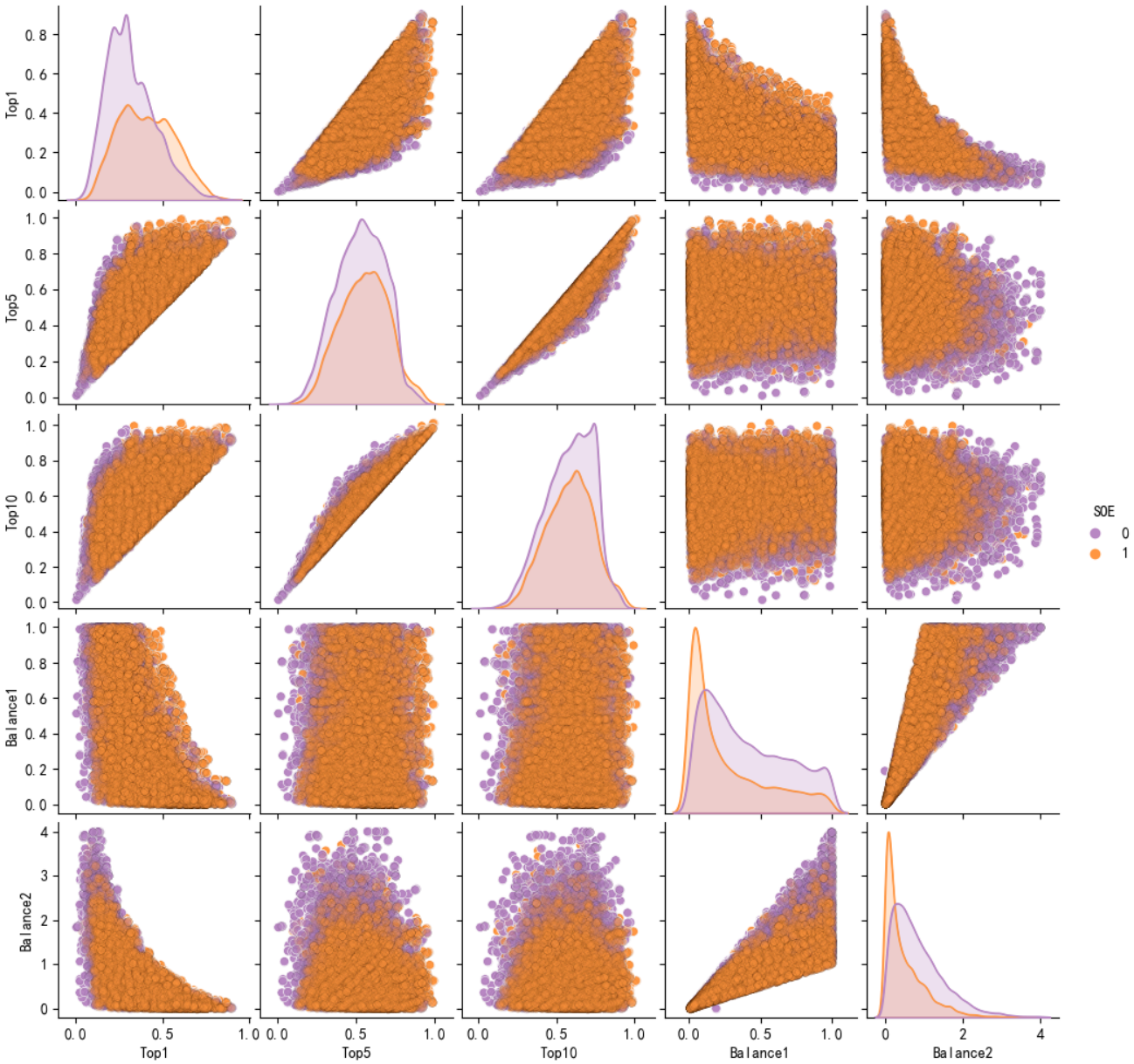

图 8：各企业类型的股东持股比例和股权制衡对比图

3 不同行业数据对比分析

将证券按照表 4 中的行业类型进行分类，由于 C 类制造业中包含的证券数目过多，所以将属于制造业的证券取其行业代码的前两位，划分为 4 小类，其余行业用大类进行分类，探究不同行业的财政表现、发展速度以及时间变化趋势。

3.1 财务指标差异

注意到数据集中包含大量的数值型财政指标，大多数均可以反应出公司的经营发展状况。首先关注不同行业的公司规模差异，发现采矿业（行业 B）和金融业（行业 J）的公司规模的平均水平普遍高于其他行业，而金属、机械、仪器仪表等方面的制造业（行业 C4）以及住宿餐饮业（行业 H）的公司规模相对较低。对于不同行业内部之间的公司规模差异方面，发现教育行业（行业 P）的公司规模差异化最小，同属于金融业（行业 J）的公司之间的规模差异在所有行业中是最大的。

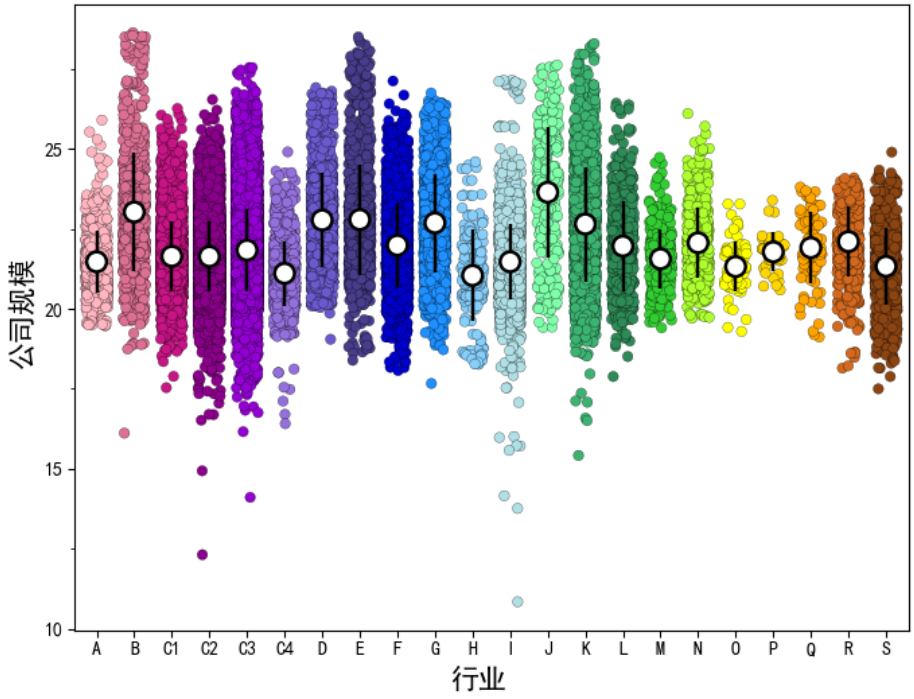

图 9：不同行业公司规模差异

其次关注各类可以反应行业经营状况和行业特点的关键指标，部分指标在不同行业之间的表现情况呈现出一致性的特点，在这里选取存货占比和固定资产占比两个行业差异化明显的指标做出详细说明。

存货占比表示存货净额与总资产的比值，反映出该行业的存货储备水平，也反映出该

行业对供需要求的调节能力。

通过图 10（左）可以直观发现，房地产行业（行业 K）的存货占比水平远远高于其他行业，这意味着对于房地产行业，其房源储备是相当充足的，房源供给稳定，甚至有可能出现供大于需的情况，这是因为阻碍房产消费的并不是房源数目的短缺，而是房价过高带来的消费能力不足。

此外对于电热燃水生产供应业（行业 D）、交通运输仓储和邮政业（行业 G）、住宿餐饮业（行业 H）、教育业（行业 P）以及卫生社会工作（行业 Q）而言，其存货占比非常小，在总资产中几乎不存在存货占比，这一点与上述行业的行业特点相符，例如对于电热燃水生产供应业，存储成本大且存储更新的周期短，因此并不需要储备大量的电热燃水，及时稳定产出保证使用者不短缺即可达到行业目的。

固定资产占比表示固定资产净额与总资产的比值，与流动资产相反，固定资产反应出行业非可实时交换的资产水平，主要来自于设备、土地等固定资产的折旧价值。

通过图 10（右）不难发现电热燃水生产供应业（行业 D）以及交通运输、仓储和邮政业（行业 G）的固定资产占比水平显著高于其他行业，结合实际情况，电热燃水生产供应业有大量的生产专业设备以及全国范围内的供应管道，交通运输、仓储和邮政业有大量的海陆空运输交通工具，这些都属于固定资产范畴，而对于金融业（行业 J）和房地产业（行业 K）并不依赖大量的固定资产，金融业中资产主要集中在金融产品上，房地产行业的资产则主要集中在商品房源上。

除此之外还注意到对于居民服务、修理和其它服务业（行业 O）的固定资产占比内部差异化极大，虽然大部分企业的固定资产占比不足 35%，但存在少数企业的固定资产占比超过 75%，究其原因推测是部分维修行业可能存在大型专业设备依赖。

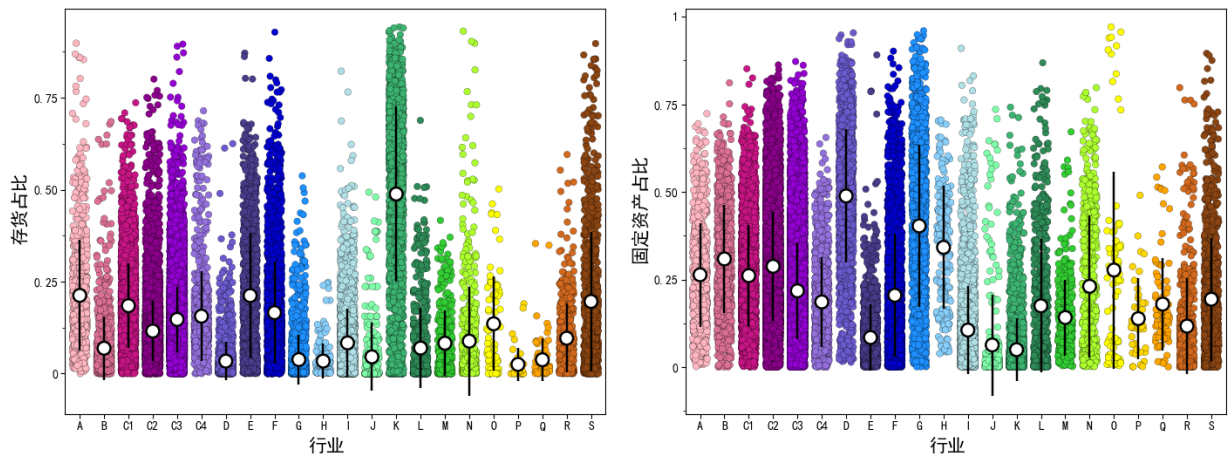

图 10：不同行业在存货占比（左）和固定资产占比（右）上的差异

3.2 发展速度差异

为了探究不同行业的企业发展速度之间的差异性，企业从成立到上市所花费的时间周期可以作为一个重要的参考指标，注意到数据集中的字段 $ListAge$ 和 $FirmAge$ 分别蕴含了公司的上市年份和成立年份信息，经过简单的数据处理即可获得 4016 支证券对应企业的上市所用年限。

首先忽略证券的行业属性，统计每支证券从成立到上市所花费的时间信息，汇总如图 11 所示，有 308 家企业不到一年就完成上市目标，证券代码为 603700 的宁水集团从成立到上市花费的年限最长，高达 60 年，该公司主要从事机械水表和智能水表的研发、生产、销售，属于典型的细分行业龙头。此外上市年限为 11 年的公司数目最多，有 368 家企业，从总体上看，所有公司的平均上市年限为 9.5 年左右。

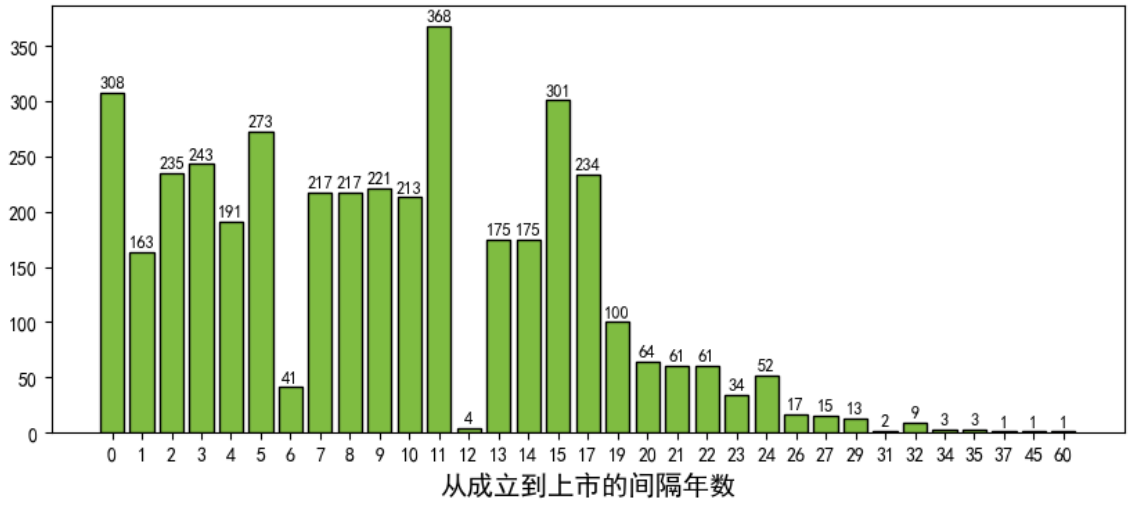

图 11：各企业上市所用年限统计图

考虑不同行业属性，进一步讨论不同行业的企业在上市年限上的表现差异。注意到有部分企业的行业类型并非是一成不变，在企业发展过程中出现过转型现象，对于这些企业我们暂时不做考量，仅统计行业类型未发生过转变的 3256 家企业的上市年限，结果汇总在图 12 中。

如果以上市年限作为分类指标，每类中的证券行业大多以制造业（行业 C）类居多，房地产业（行业 K）的发展速度很快，大约在 5 年内即可完成上市，而对于文化、体育和娱乐业（行业 R）上市年限主要集中在 5 到 15 年的区间内，信息传输、软件和信息技术服务业（行业 I）大部分需要经过 10 年左右的发展才能完成上市，除此之外批发和零售业（行业 F）的发展速度业基本高于所有行业的平均水平。

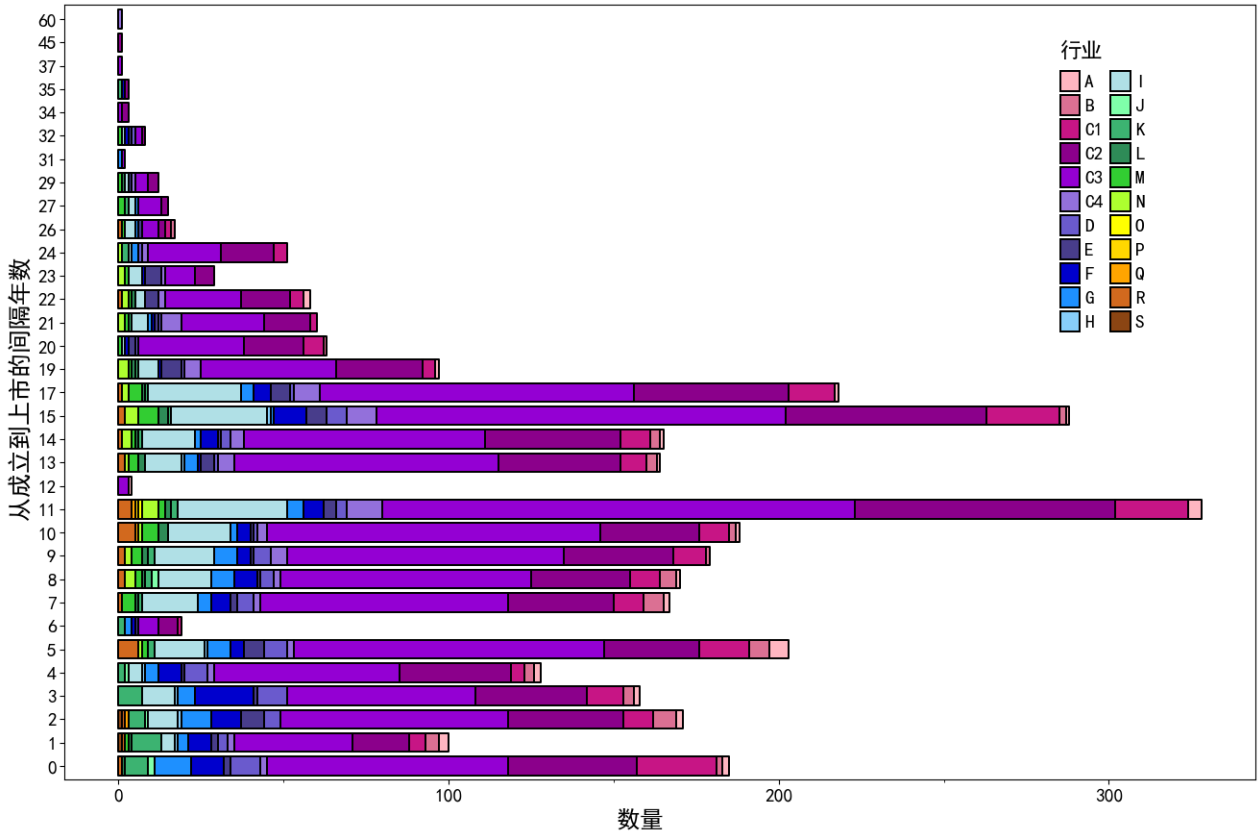

图 12：各行业上市年限的分类统计图

3.3 时间变化趋势

上述分析主要是依托于数据表中的截面数据，在这里关注不同行业的财务指标随时间的变化情况，首先进行企业总资产净利润率 ROA 的时间序列分析，绘制处理完异常值的不同行业总资产净利润率的时间序列图。

观察图 13 中的数据变化情况，注意到采矿业（行业 B）在 2002 年到 2008 年出现总资产净利润率逐年增长的现象，并且该行业 ROA 水平显著高于行业平均水平。住宿和餐饮业（行业 H）在 2008 年出现短暂高峰，推测可能是受到北京奥运会的影响。卫生和社会工作（行业 Q）的总资产净利润率从 2007 年开始逐年攀升，并在 2010 年至 2014 年始终维持较高水平，从侧面反映出该时间段内卫生和社会工作行业发展势头良好。

反观金融业（行业 J）在 2007 年由于遭受金融危机的影响，总资产净利润率出现负增长。此外从 2019 年开始，近几年的居民服务、修理和其他服务业（行业 O）受到重创，ROA 显著低于往年水平，并在 2020 年出现严重的负指数现象，推测这种现象可能是由于新冠疫情所导致的，隔离封城等防疫措施并不利于服务行业的发展。

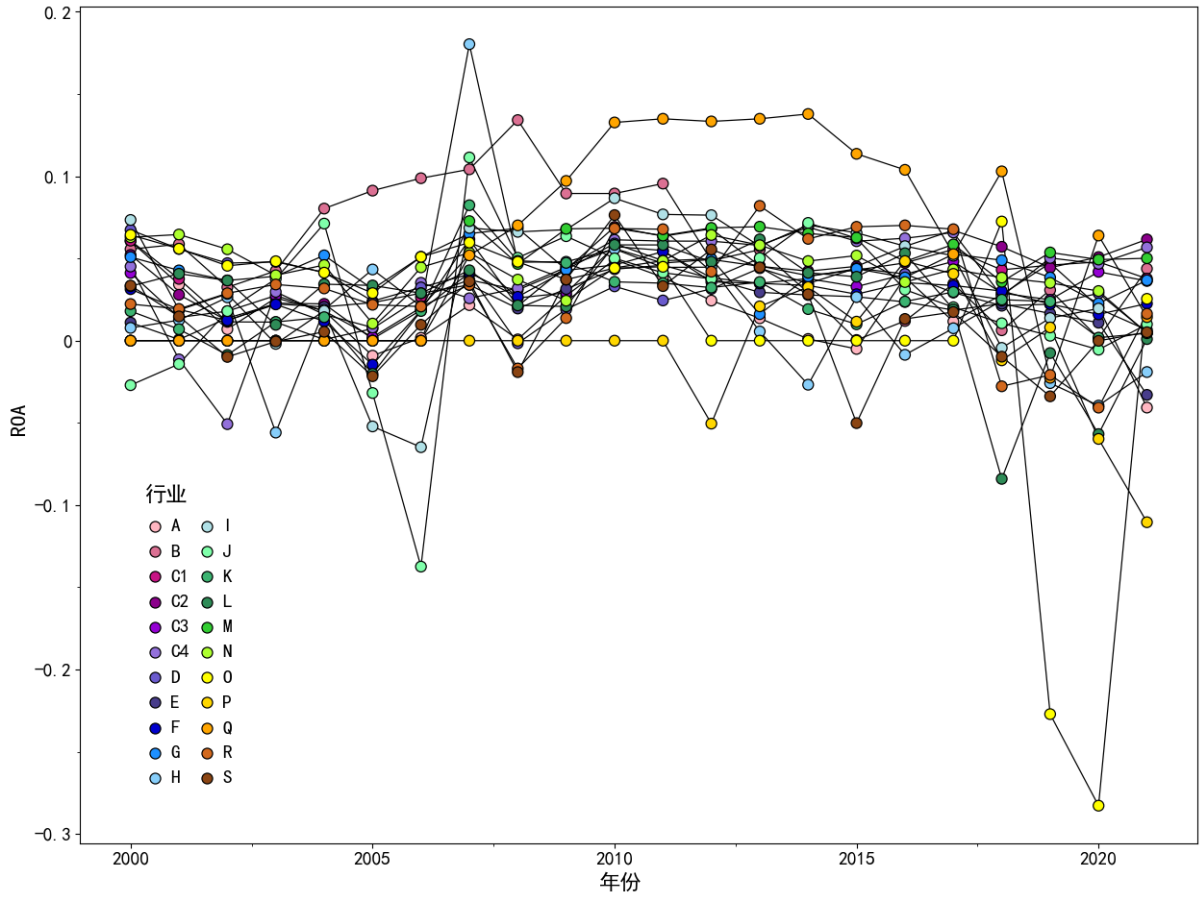

图 13：各行业总资产净利润率时间序列图

其次，现金流比率（$Cashflow$）也是反应公司经营状况的重要指标，该比率用于衡量企业经营活动所产生的现金流量可以抵偿流动负债的程度，比率越高，说明企业的财务弹性越好。不同行业由于其经营性质的不同，经营活动产生的现金净流量的差别较大，导致该比率存在行业差异，因此通过分析该比率，可判别企业财务状况是否良好，公司运行是否健康。

通过绘制不同行业现金流比率的瀑布图，来观察各个行业现金流比率在不同年份的表现情况。可以发现大多数行业的现金流比率均值维持在 0.1 左右，并且随年份的数值变化并不大，常年保持稳定水平。值得注意的是，金融业（行业 J）的现金流比率的波动率较大，在 2007 年金融危机时该行业的现金流比率均值高达 0.4，之后的几年也出现极速下降和迅速上升的情况，先后出现多次峰值，这意味着金融行业的财务弹性波动相对较大。同样地，对于卫生和社会工作行业（行业 Q）也曾在 2006 年附近出现现金流比率猛增的现象，但之后现金流比率逐渐下降至正常水平。

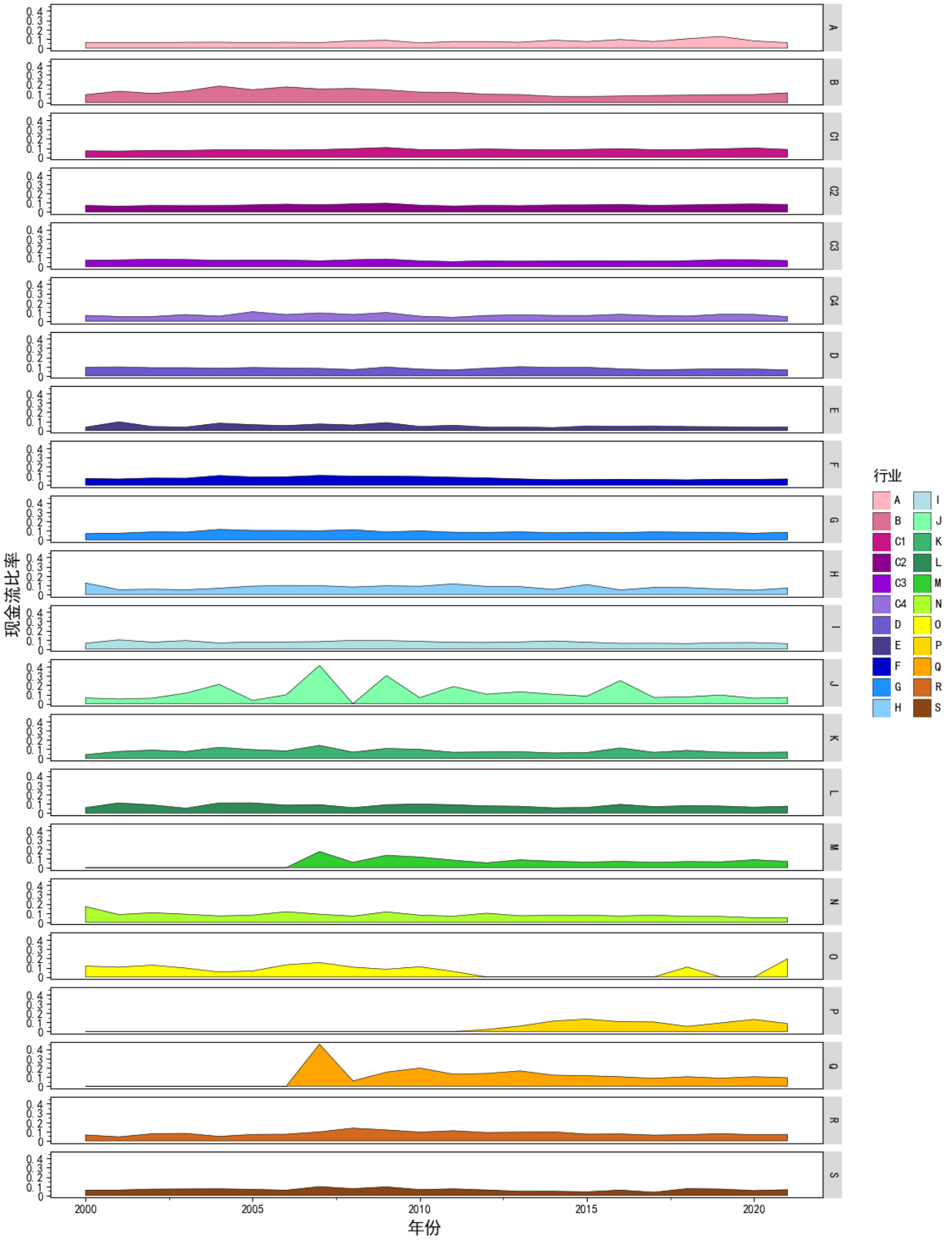

图 14：各行业现金流比率瀑布图

4 分析结果汇总

通过对 2000 年至 2021 年部分 A 股股票进行数据分析与可视化，主要从两个角度入手，解决了以下问题。

第一，关注国企控股企业与非国企控股企业的表现差异。按照观测时段内的企业属性将证券进行分类，划分为国企控股证券、非国企控股证券以及发生过类型转变的证券三大类，对于第三类又可以细分为由非国企控股转为国企控股、由国企控股转为非国企控股以及发生过多次企业类型转变三类。

研究发现，非国企控股证券数目约为国企控股证券数目的三倍，同时有 12.8% 的企业发生过企业类型转变，其中转变为非国企控股企业主要发生在 2006 年和 2007 年，而 2019 年至 2021 年则集中出现大量企业转变为国企控股企业。此外注意到随着年份的增加，国企控股企业和非国企控股企业的数目均呈现增长态势，且国企控股企业数目增长速度明显更缓慢，起初非国企控股企业数目少于国企控股企业数目，但在 2010 年实现反超，并一路高歌猛进。

关于两类企业在股权性质上的差异，可以通过股东持股比例和股权制衡度反应出来。对于国企控股企业，其大股东持股比例明显更低，股权制衡度峰值更高，这意味着国企控股企业的股权结构更为分散，大股东的绝对话语权水平更低。

第二，由于每支证券均含有行业属性，可以根据行业类型进行划分，观察不同行业在财政经营发展方面的差异性。

首先关注公司规模，发现采矿业和金融业的平均公司规模高于其他行业，并且金融业的行业内部规模差异化最大，而金属、机械、仪器仪表等方面的制造业以及住宿餐饮业的公司规模相对较小。关于不同行业资金结构方面的差异，主要可以通过存货占比和固定资产占比反应出来，房地产行业的存货占比水平远远高于其他行业，而电热燃水生产供应业以及交通运输仓储和邮政业的存货占比普遍偏低，但固定资产占比显著高于平均水平，这意味着上述两类行业的发展依赖于大量的固定资产，但对于具有灵活属性的金融业，其固定资产占比就很小。

发掘企业从成立到上市所花费的年限属性，来反应企业的发展速度。从总体上看，所有公司的平均上市年限为 9.5 年左右，考虑不同行业的上市年限差异，房地产行业的发展速度最快，多数在 5 年内即可完成上市，文化、体育和娱乐业的上市年限主要集中在 5 到 15 年的区间内，信息传输、软件和信息技术服务业大部分需要经过 10 年左右的发展才能

完成上市。

以时间为线，纵向关注不同行业随时间的发展变化。对于总资产净利润为代表的财务指标，采矿业和卫生社会工作行业分别在 2002 年和 2007 年出现连续五年的总资产净利润提升现象，住宿餐饮业则是在 2008 年出现短暂高峰期。反观金融行业，由于受到 2007 年金融危机的影响，总资产净利润出现显著的负增长现象，并且居民服务、修理和其他服务业从 2019 年开始总资产净利润低于-20%，但在 2021 年得到恢复，推测大概率是新冠疫情的影响所致。

此外，通过不同行业随时间变化的现金流比率来反应不同行业的经营状况差异。可以发现大多数行业的现金流比率均值维持在 0.1 左右，而金融业则具有相当显著的财务弹性现象，该行业的现金流比率的波动率非常大，在 2007 年金融危机时迅速上升至 0.4，并在之后也多次出现短暂峰值。