一带一路典型国家人类发展指数相关性探究

吴子祎、张晶、徐春辉、张恩嘉、丁文泽

摘要： 推动我国与一带一路国家之间实现经济政策协调，共同开展范围更大、水平更高、层次更深的互惠合作，是一带一路倡议的主要内容[1]。为实现这样的精准合作，必然需要对沿线各国，尤其是支点国家的发展模式进行剖析，更加明确促进政策沟通、设施联通、贸易畅通、资金融通、民心相通的努力方向，更加有的放矢。

本文首先利用 Python 软件，对“一带一路网”新闻标题进行文本分析，统计出词频热度最高的国家，分别为：白俄罗斯、俄罗斯联邦、哈萨克斯坦、斯里兰卡、土耳其、马尔代夫、柬埔寨。词频热度一定程度上反映了该国与我国之间经济政治联系的强度，重要性更强。因此，本文以上述重点国家为研究对象，使用 R 软件，基于联合国计划开发署的开放数据库，以线性回归、文本挖掘、机器学习的方法，对其综合发展水平和影响因素进行国家间横向和时间维度纵向的量化和数据可视化分析。

数据分析元数据及程序文件请访问 github：https://github.com/nickwu96/big_data_analysis_group3

1. 背景

GDP(GNP)提供了一个统一的工具，来描述不同时间和地点的经济活动。但王志平认为，GDP(GNP)根本还是围绕物的发展，但却不能说明其构成和受益分布, 也不能反映其他人们同样关心的内容，如健康、机会、安全等[2]。为更好的反映人本身的发展，联合国在 1990 年推出了“人类发展指数” (HDI)，并在此之后被广泛用于测量和比较各国的相对人类发展水平差异。作为综合性指数,HDI 将每个国家人民的健康（以出生时预期寿命度量）、教育（以预期受教育年限和平均受教育年限度量）和收入（以人均国民总收入度量）三个维度的信息统筹，综合为一个数据[3]。HDI 介于 0 和 1 之间,数值越大,代表人类发展水平越高。

根据《2018 数据更新：人类发展指数和指标》为例，总共收录了 189 个国家和地区的人类发展指数值。本文对一带一路沿线的白俄罗斯、俄罗斯联邦、哈萨克斯坦、斯里兰卡、土耳其、马尔代夫、柬埔寨等国家综合发展水平和影响因素进行国家间横向和时间维度纵向的量化和数据可视化分析，以期寻找发展中国家发展瓶颈，提出相应的政策建议。

2. 词频分析

2.1. 数据来源及方法

我们选取了“中国一带一路网”（https://www.yidaiyilu.gov.cn/）作为本次词频分析的数据来源。该网站作为官方网站，最权威地发布了与中国签订一带一路合作协议国家的动态。

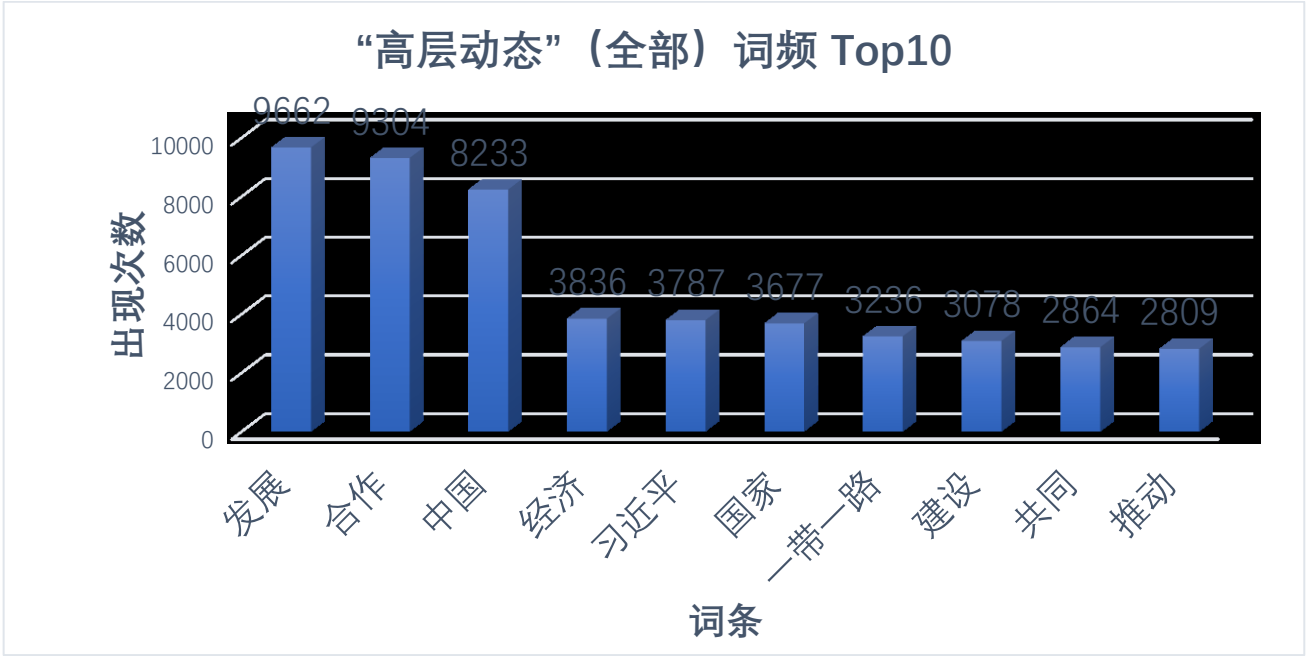
数据爬取的工具采用了 Python 3.7，安装了如下开源包：requests（用于获取网页内容）、beautifulsoup（用于解析网页内容）、pandas（用于结构化存储数据）、jieba（用于中文分词）、wordcloud（用于绘制词云）。

2.2. 数据结果及词频分析

我们选取了“中国一带一路网”上“高层动态”和“海外动态”两个版块，试通过对响应版块内新闻词频的分析，分别了解到国内高层对一带一路沿线国家的重视程度和一带一路沿线国家的出现频率。

2.2.1. “高层动态”版块

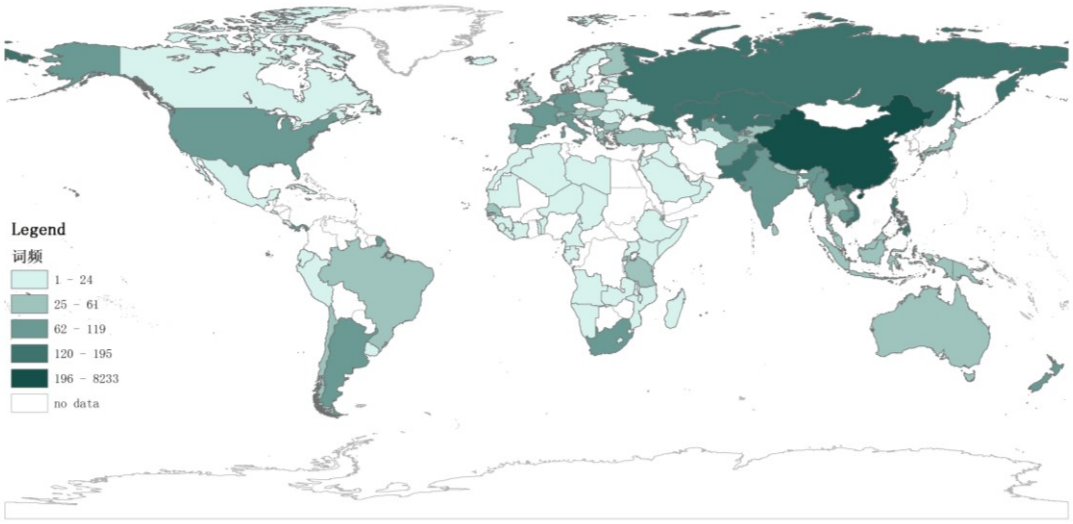
我们共统计了 35 页新闻列表，读取了 1014 条新闻，共出现了 23624 个词。词频分析前十名结果如下：



可以看到“发展”、“合作”、“一带一路”、“建设”、“共同”、“推动”等词出现的频率较高。显而易见，发展和合作是一带一路的主旋律，国内高层领导对其他国家的访问也是希望建立更广、更深入的合作伙伴关系。

对于词频分析的结果，我们筛选出其中出现的国家，出现次数前十的国家如下：中国 非洲 俄罗斯 哈萨克斯坦 菲律宾 巴基斯坦 新加坡 越南 印度 老挝。

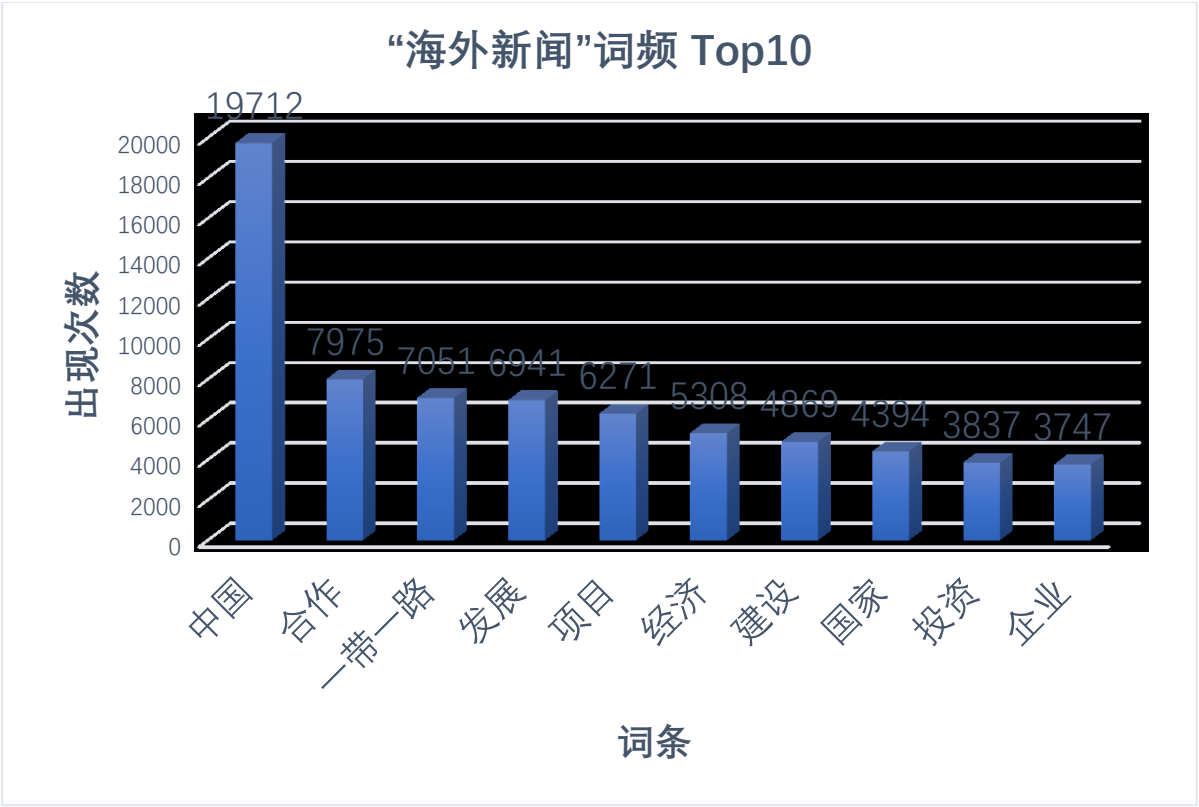
在世界地图上以颜色标记各国家出现的次数，颜色越深即出现次数越多，结果如下：



词云绘图的结果如下：

2.2.2. “海外新闻”版块

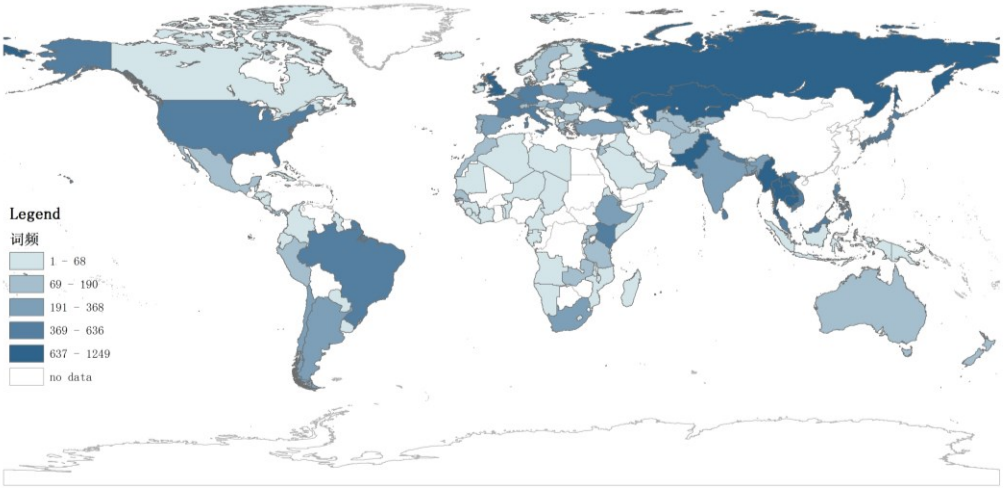
同理，我们也对“海外新闻版块”做了类似的统计。共统计了 537 页新闻列表，读取了 2677 条新闻，共出现了 53053 个词。词频分析前十名结果如下：



可以看到，与“高层动态”的结果有相似的地方：“合作”、“一带一路”、“发展”等词出现的频率较高，体现了一带一路合作发展的主旋律。但也有些许不同，如“项目”、“经济”、“建设”、“投资”、“企业”等词的频率跃居前十位，这是因为中国与一带一路沿线国家签订了大量合作协议，共商合作发展。

对于词频分析的结果，我们筛选出其中出现的国家，出现次数前十的国家如下：非洲 俄罗斯 巴基斯坦 泰国 老挝 柬埔寨 英国 哈萨克斯坦 缅甸 马来西亚

在世界地图上以颜色标记各国家出现的次数，颜色越深即出现次数越多，结果如下：



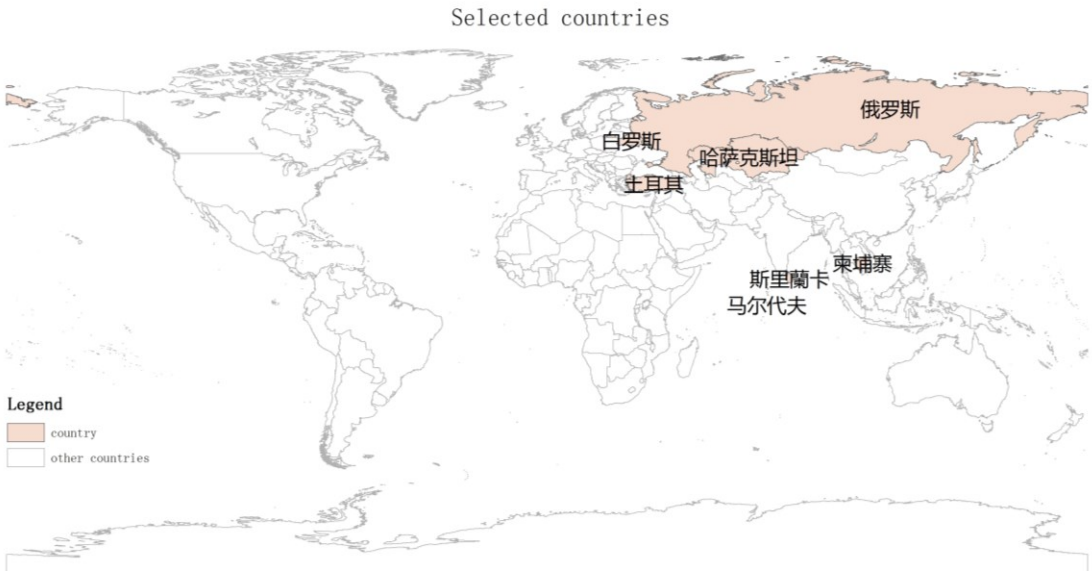
词云绘图结果如下：

2.3. 结论

考虑到加入“一带一路”计划的 131 个国家重要程度是不同的，结合各国家在一带一路网中出现的次数及地缘政治、经济水平等因素综合考虑，我们选定了 7 个国家作为我们的分析国家，将会在后续研究中重点分析这几个国家。

国家	收入水平
俄罗斯	中高等收入国家
白罗斯	中高等收入国家
柬埔寨	中低等收入国家
土耳其	中高等收入国家
哈萨克斯坦	中高等收入国家
斯里兰卡	中低等收入国家
马尔代夫	中高等收入国家

在地图上标示出选定的国家如下：



3. 方法论和数据

3.1. 方法论

本研究的整体思路为（图 3-1）：首先选取因变量与自变量指标，然后通过采集并整理数据，调整数据结构及数据清洗等步骤，确定变量数据。然后通过描述性分析、相关矩阵、VIF 检验等预实验确定最终指标。最后通过回归分析：按区域回归及按年份回归两种形式，找出具有显著意义的指标。最后通过机器学习验证，从而对结果进行分析和讨论。

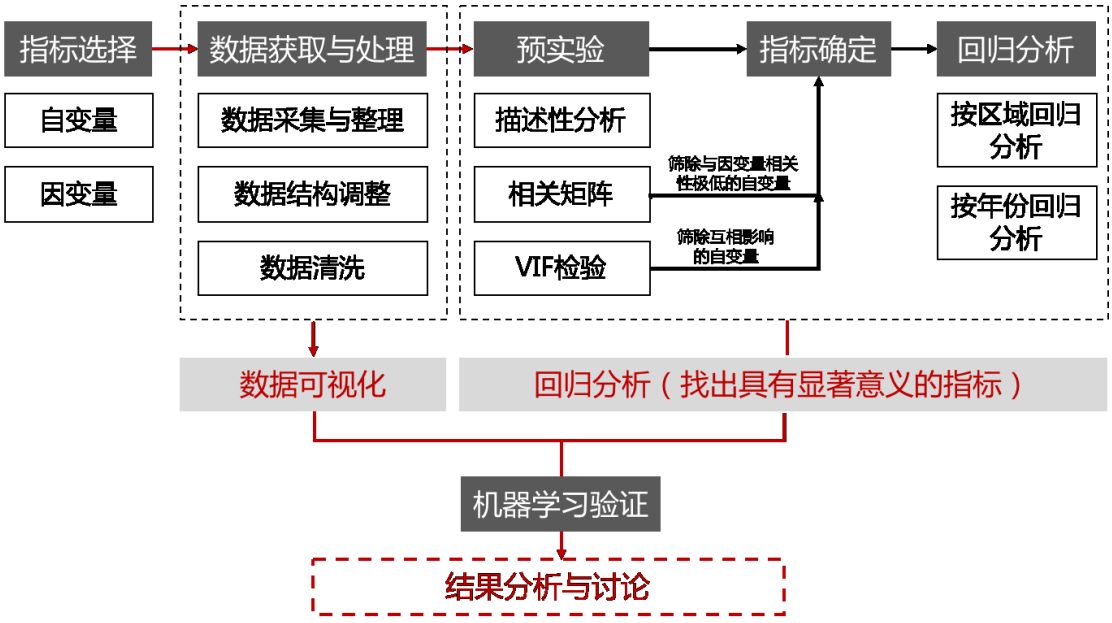


图 3-1 研究思路

3.2. 指标选取

3.2.1. 因变量选取

本研究选取的因变量为由联合国开发计划署发布的人类发展指数，其作用是指导在发展中国家援助战略的制定。数据范围包括 1990~2017 全世界绝大多数国家或地区的数据。涉及内容包括涉及内容：健康寿命、教育获得及收入水平（图 3-2）。

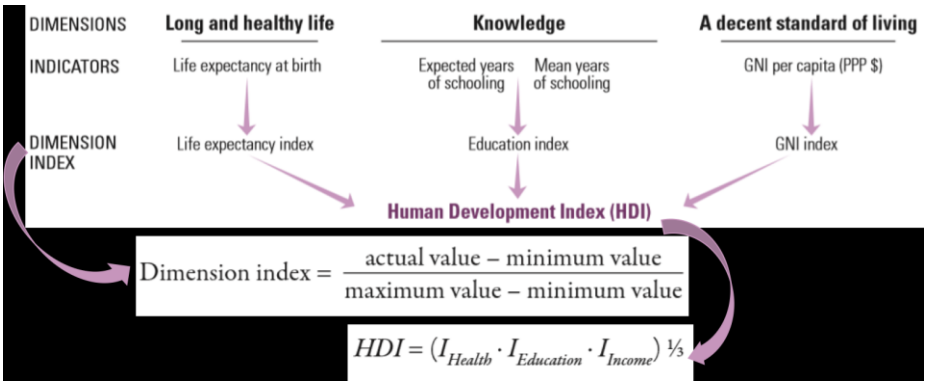


图 3-2 人类发展指数内涵

由于人类发展指数涉及三方面的内容，为更好地研究人类发展指数的影响因素，因此本研究还选取三个子项作为因变量进行深入研究分析，分别为教育指数、收入指数和寿命指数。

3.2.2. 自变量选取

在不与因变量重叠的基础上选取本研究的自变量。数据来源于世界银行数据库 World Bank Database。初步选取 14 个指标作为研究对象，其基本情况如表 3-1:

指标名称	数据年份	可获取性	备注
利差	1980-2017	大部分有	贷款利率减去存款利率
存款利率	1980-2017	大部分有	
股票交易总额	1975-2017	大部分有	占 GDP 的比例
商品出口	1960-2017	基本全有	现价美元
商品进口	1960-2017	基本全有	现价美元
国际旅游收入	1995-2017	基本全有	占总出口百分比
国际旅游支出	1995-2017	基本全有	占总进口百分比
高科技出口	1988-2018	大部分有	现价美元
男性劳动力参与率	1990-2018	基本全有	占 15 岁以上男性人口百分比
识字率	1980-2016	一部分有	青年男性占 15-24 岁男性人口的百分比

城市化率	1960-2017	基本全有	
抚养比	1960-2017	基本全有	
铁路总公里数	1996-2017	大部分有	
通电率	1990-2016	大部分有	占人口的百分比

表 3 自变量的信息

3.3. 数据描述

对人类发展指数、教育指数、收入指数及寿命指数进行空间可视化及平均值的时间变化分析。分别得到以下结果：

3.3.1. 人类发展指数：

整体来看，人类发展指数的在空间上分布不均，大致可以分为三个层次：北美、欧洲、大洋洲属于第一层次，整体人类发展指数较高；亚洲和南美属于第二层次，整体人类发展水平居中；非洲属于第三层次，人类发展水平较为落后。

从时间维度看，整体人类发展指数不断提高，七国平均人类发展指数逐渐超过世界平均水平。中国人类发展指数与七国平均水平趋近，并赶超世界平均水平。中国 2013 年提出“一带一路”战略，时间点位于中国赶超平均水平之后，反映出中国国力强盛后才提出更加关注发展的国家级战略。

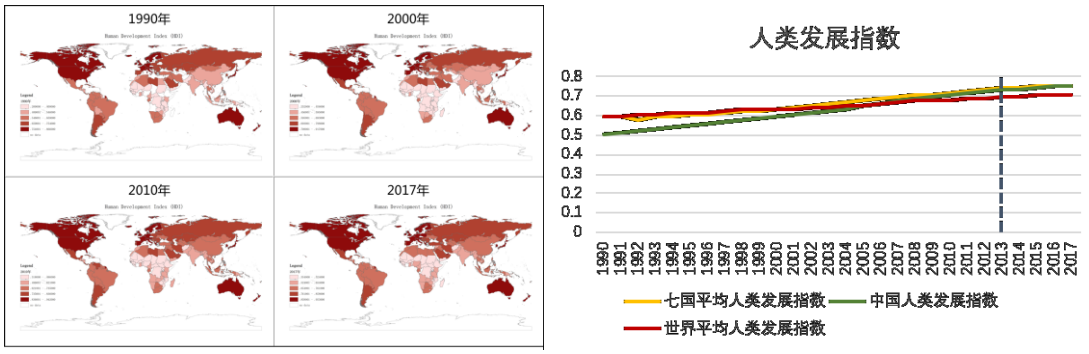


图 3-3 人类发展指数空间分布（左）及时间变化趋势（右）

3.3.2. 教育指数：

整体来看，教育指数与人类发展指数相似，在空间上体现出三个层次。但亚洲整体发展速度较快，不断向第一层次水平靠近。非洲局部地区的教育指数发展也相对较快，撒哈拉以南非洲整体水平高于北非。

从时间维度看，整体教育指数不断提高。七国平均教育指数逐渐超过世界平均水平。中国教育指数趋近世界平均水平，但距七国平均水平仍有差距。中国 2013 年提出“一带一路”战略后第二年，中国教育指数超过世界平均水平。

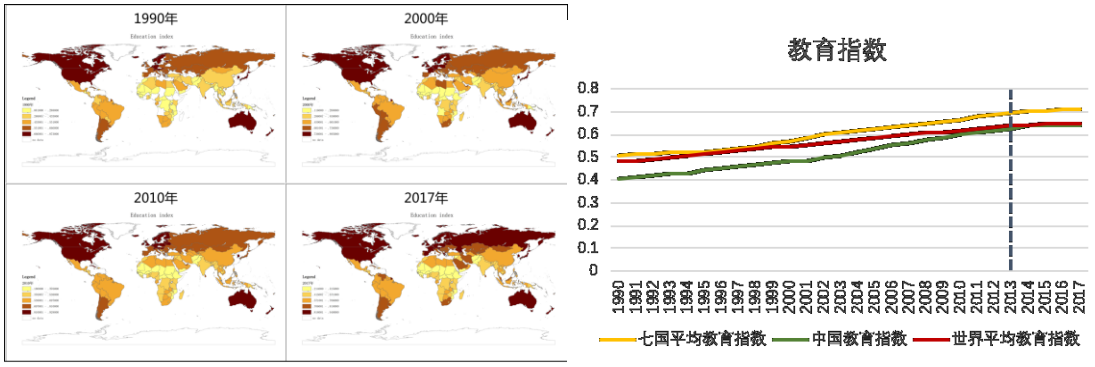


图 3-4 教育指数空间分布（左）及时间变化趋势（右）

3.3.3. 收入指数：

整体来看，收入指数的空间分布与教育指数相似。比较明显的是自 1990 年，中国每十年上升一个层次，目前位于中等偏上水平。北非部分国家的收入指数相较于其他国家而言呈现波动状态。

从时间维度看，整体收入指数也不断提高。七国平均收入指数逐渐超过世界平均水平。中国收入指数近十年内赶超世界平均水平，与七国平均水平趋近。自 2013 至今，中国收入指数仍持续上升，并在近几年（2016 年、2017 年）微超过七国平均水平。

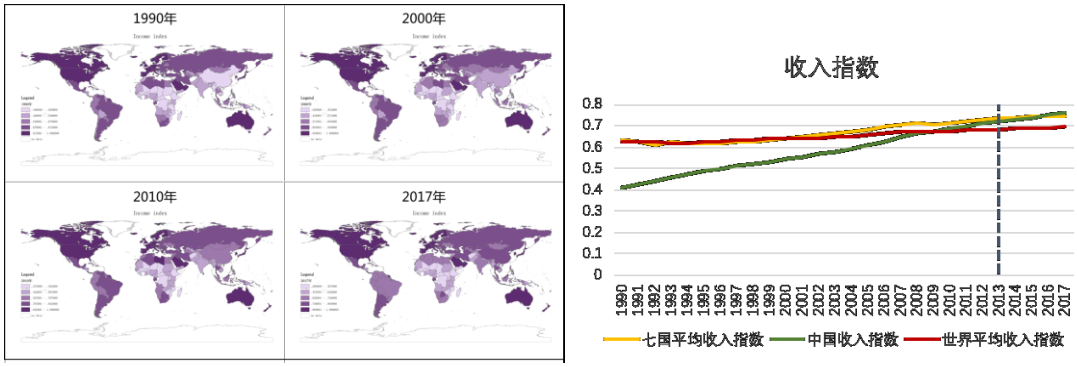


图 3-5 收入指数空间分布（左）及时间变化趋势（右）

3.3.4. 寿命指数：

整体来看，寿命指数的空间分布与收入指数相似，体现出寿命与收入之间的相关关系。东亚地区的寿命指数明显高于西亚地区，北非局部地区的寿命指数相对较高。中国寿命指数较之周边国家尤其是除日本韩国以外的国家而言，一直都较高，与教育指数和收入指数的相关性不强。

从时间维度看，七国平均寿命指数接近世界平均水平。中国寿命指数一直超过世界平均水平及七国平均水平。

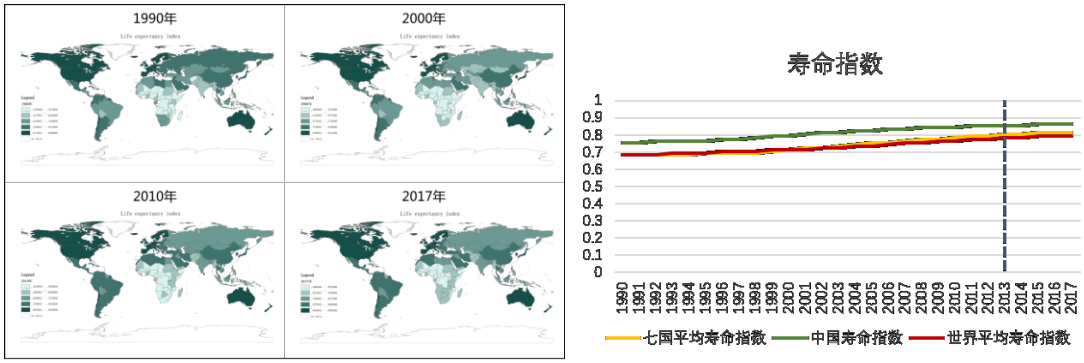


图 3-6 寿命指数空间分布（左）及时间变化趋势（右）

整体来讲，从空间分布角度，四项指标尤其是三项人类发展指数的分项指标的空间分布较为相似，呈现出三个层次的空间分布特征。北美、欧洲、大洋洲属于第一层次，整体人类发展指数较高；亚洲和南美属于第二层次，整体人类发展水平居中；非洲属于第三层次，人类发展水平较为落后。从时间发展趋势来看，四项指标都呈现出逐年增长的趋势。2007 及 2008 两年是中国赶超世界平均水平的重要时期，然而距七国平均水平仍有一定差距，距世界顶尖水平差距较大，因此中国更加重视发展和合作。2013 年，中国提出“一带一路”战略，高举和平发展的旗帜，积极发展与沿线国家的经济合作伙伴关系，推动中国及沿线国家的发展。由此可见，“一带一路”战略是国家发展到一定水平之后为更加持续地发展和交流提出的重要战略，是中国不断向世界领先水平靠近的重要手段，更是带动区域发展，尤其是相对落后的非洲等地区发展的重要合作倡议。

4. 回归分析

4.1. 时间维度分析：

时间维度的分析是指，将 2002 年到 2017 年的每一个年份作为一个分析单元，七个国家作为每个分析单元的样本进行分析，分析随着时间的变化，各自变量指标对因变量的影响程度，其中缺失值用平均值填充。

country	eduIndex	expIndex	incIndex	HDI	Deposit interest rate	Stocks traded, total value	Merchandise exports	Merchandise imports	International tourism, receipts	International tourism, expenditures	High-technology exports	Employment in industry, male	urbanization	Dependency Ratio	railways	electricity
Belarus	0.84	0.82	0.77	0.81	7.00	13.00	2.92E+10	3.42E+10	3.10	2.97	6.01E+08	42.03	78.13	46.07	5459	100
Russia	0.83	0.79	0.83	0.82	5.86	9.22	3.54E+11	2.38E+11	3.64	10.88	9.17E+09	37.48	74.29	46.59	85545	100
Kazakhstan	0.81	0.77	0.82	0.80	7.00	0.48	4.85E+10	2.96E+10	3.56	4.39	1.77E+09	30.12	57.34	53.66	16040.3	100
Sri Lanka	0.75	0.85	0.71	0.77	9.00	0.87	1.14E+10	2.1E+10	26.58	9.48	6.90E+07	29.08	18.38	51.70	29000	96
Turkey	0.69	0.86	0.83	0.79	15.29	44.31	1.57E+11	2.34E+11	15.17	2.08	3.05E+09	31.57	74.64	49.50	10207	100
Maldives	0.56	0.89	0.74	0.72	3.57	13.00	3.18E+08	2.36E+09	87.72	10.38	1.56E+07	17.57	39.38	37.99	29000	100
Cambodia	0.49	0.76	0.53	0.58	1.53	13.00	1.21E+10	1.55E+10	25.41	5.04	1.56E+07	28.55	22.98	55.50	650	50

图 4-1 原始数据-以 2017 年为例

4.1.1. 相关性分析

首先对原始数据进行相关性分析，以此探究各指标间的相关程度，为后面回归自变量的选取提供支撑。使用 R 语言进行回归分析并绘制相关系数矩阵图。得到的结果如下图所示。

可以看出，商品出口量与商品进口量的相关系数为 0.90，商品出口量与高科技出口量相关系数为 0.99，因此，可保留商品进口量，剔除商品出口量和高科技产品出口量。银行存贷款利率差和股票贸易值相关系数为 0.62，且这两个指标在这七个国家中的数据缺失值较多，由此剔除银行存贷款利率差和股票贸易值。铁路总公里数数据缺失较大，也可剔除掉。因此，最终只保留商品进口量、男性劳动力占比、城市化率和抚养比这四个指标完成下面的回归。

2017	Deposit.interest.rate	Stocks.traded..total.value	Merchandise.exports	Merchandise.imports	International.tourism..receipts	International.tourism..expenditures	High.technology.exports.	Employment.in.industry..male.	urbanization	Dependency.Ratio	railways	electricity
Deposit.interest.rate	1.00	0.62	0.25	0.57	-0.34	-0.42	0.16	0.26	0.44	0.09	-0.08	0.54
Stocks.traded..total.value	0.62	1.00	0.23	0.60	0.01	-0.51	0.11	0.04	0.43	-0.13	-0.25	0.05
Merchandise.exports	0.25	0.23	1.00	0.90	-0.43	0.24	0.99	0.45	0.59	-0.08	0.80	0.28
Merchandise.imports	0.57	0.60	0.90	1.00	-0.41	-0.04	0.84	0.40	0.65	-0.03	0.53	0.30
International.tourism..receipts	-0.34	0.01	-0.43	-0.41	1.00	0.48	-0.43	-0.88	-0.49	-0.64	-0.05	-0.03
International.tourism..expenditures	-0.42	-0.51	0.24	-0.04	0.48	1.00	0.29	-0.38	-0.38	-0.44	0.76	0.14
High.technology.exports.	0.16	0.11	0.99	0.84	-0.43	0.29	1.00	0.44	0.58	-0.08	0.84	0.30
Employment.in.industry..male.	0.26	0.04	0.45	0.40	-0.88	-0.38	0.44	1.00	0.64	0.28	0.13	0.15
urbanization	0.44	0.43	0.59	0.65	-0.49	-0.38	0.58	0.64	1.00	-0.26	0.22	0.57
Dependency.Ratio	0.09	-0.13	-0.08	-0.03	-0.64	-0.44	-0.08	0.28	-0.26	1.00	-0.32	-0.54
railways	-0.08	-0.25	0.80	0.53	-0.05	0.76	0.84	0.13	0.22	-0.32	1.00	0.37
electricity	0.54	0.05	0.28	0.30	-0.03	0.14	0.30	0.15	0.57	-0.54	0.37	1.00

图 4-2 相关性系数矩阵

4.1.2. 选取指标的含义

商品进口量，单位为现价美元，数据来源于世界贸易组织，商品进口具体指收到的来自世界其他地方的以现价美元计算的商品的离岸价值。

抚养比是被抚养人口（15 岁以下或 64 岁以上人口）与劳动年龄人口（15-64 岁人口）之比。数据体现为每百名劳动年龄人口中被抚养人口所占的比例。

城市化率为城市人口占总人口比例，是衡量一个国家城市化发展水平的指标。

男性劳动力参与率是指年龄在 15 岁及 15 岁以上的人口中从事经济活动的人口比率。所有在特定阶段为货物和服务的生产提供劳力的人员。具体来源于国际劳工组织的劳动力市场主要指标数据库。

4.1.3. 回归分析

1）随时间变化各指标对人类发展水平的影响

（1）对人类发展指数的影响

商品进口量对人类发展水平影响很小，回归系数量级为 10^(-13)，可以忽略不计；人类发展指数与抚养比、城镇化率呈负相关，与男性劳动力参与率呈正相关。这说明抚养比越大、城市化水平越高，人类发展水平越低，与认知一致。抚养比越大，说明国家的人口结构问题越大，中青年的对青少年和老年人口的抚养压力越大，导致人类发展水平较低；城市化水平越高，人类发展水平越低，原因可能是当前这几个国家发展到逆城市化阶段，

随着城市化进程的推进，反而使其教育、健康寿命和平均收入水平下降。而男性劳动力参与率越高，人类发展指数越高。

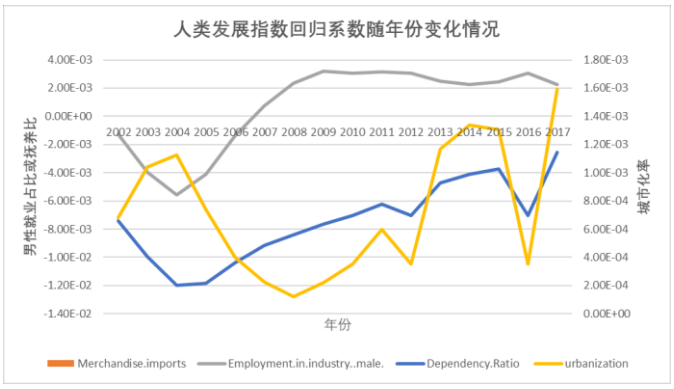


图 4-3 随时间变化各因素对人类发展指数的影响

（2）对教育、健康水平和收入水平的影响

从系数大小上看，商品进口量（橙色）对人类发展水平影响很小，回归系数量级为 10^(-13)，可以忽略不计；城镇化率（黄色）的影响较小，系数量级为 10^(-3)；男性劳动力参与率（灰色）和抚养比（蓝色）的系数量级为 10^(-2)，影响较大。

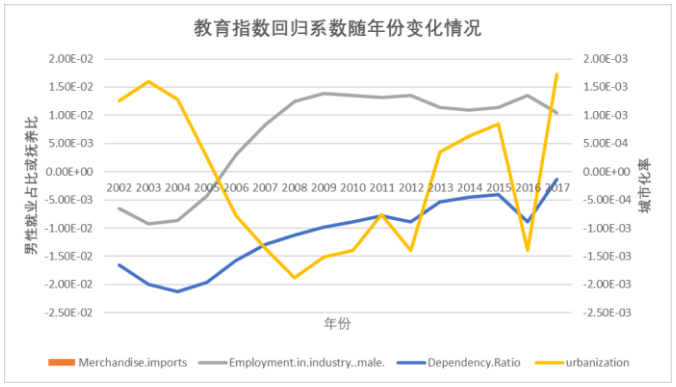


图 4-4 各指标对教育水平的影响

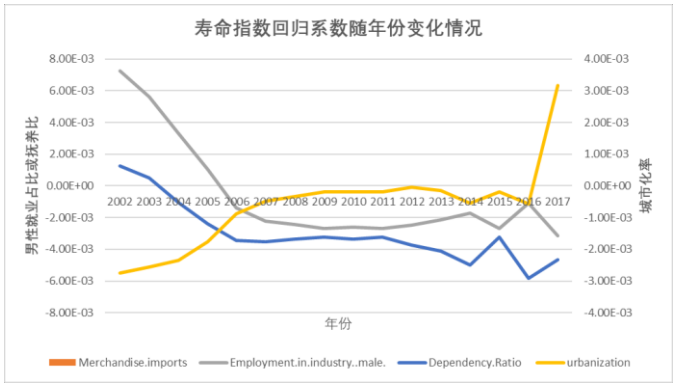


图 4-5 各指标对寿命健康水平的影响

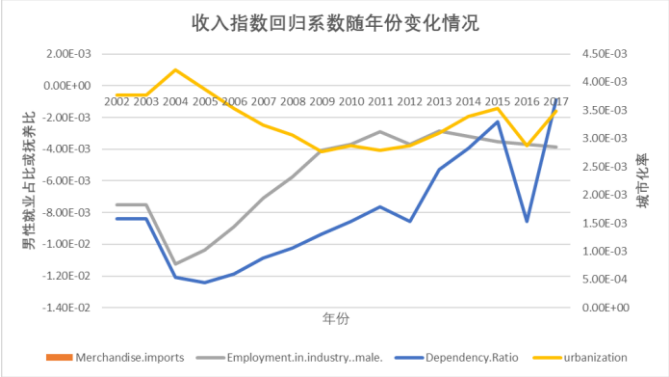


图 4-6 各指标对收入健康水平的影响

（3）无显著性结果的原因分析

分析当前出现与认知相悖或者无明显结果的原因有以下几点：

一是因变量选取的问题。人类发展指数本身存在局限性，首先该指标由教育指标、寿命健康指标和收入指标三个简单乘积再开根号得到，没有考虑在发展过程中的其他因素，也没有衡量教育、健康和收入对人类发展水平的影响程度。三个分指标由五个简单指标通过极差归一化的简单方式计算得到，没有构建科学合理的指标体系，对人类发展水平的评价考虑不够全面，可能无法真实反映一个国家的人类发展水平。二是自变量指标选取。受数据可获得性限制、选取合理性等因素影响，指标不具有代表性。在选取指标的过程中，首先需要避开跟教育、健康寿命、收入相关的指标。于是考虑国家贸易往来、交通、人口结构等方面，根据指标数据的完整性、排除指标的共线性后，最终剩余的四个指标可能不够合理，导致没有规律。三是选取的国家数量太少，只选取了七个国家，有四个变量，相对来说拟合的样本量不够多。四是国家发展水平存在巨大差异，按照世界银行的分类，选取的七个国家分布于亚洲、欧洲，它们的收入水平不同，有中高等、中低等收入国家，各自发展水平和发展阶段不同，将它们作为一类进行分析时，可能不具有代表性。

国家	收入水平	地区
俄罗斯	中高等收入国家	欧洲与中亚地区（不包括高收入）
白罗斯	中高等收入国家	欧洲与中亚地区（不包括高收入）
柬埔寨	中低等收入国家	东亚与太平洋地区（不包括高收入）
土耳其	中高等收入国家	欧洲与中亚地区（不包括高收入）
哈萨克斯坦	中高等收入国家	欧洲与中亚地区（不包括高收入）
斯里兰卡	中低等收入国家	南亚

马尔代夫	中高等收入国家	南亚

表 4 世界银行对各国发展情况的分类

4.2. 国家维度分析：

国家维度的分析是指，将每个国家 2000 年至 2016 年数据整体作为一个分析单元，对白俄罗斯、俄罗斯、柬埔寨等七个国家影响 HDI 因素进行回归分析各自变量指标对因变量的影响程度，并比较之间的差异。

白俄罗斯	国家名	HDI指数	受教育指数	预期寿命指数	经济发展指数	商品出口（现价美元）	商品进口（现价美元）	国际旅游，收入（占总出口的百分比）	国际旅游，支出（占总进口的百分比）	工业男性就业人员（占男性就业的百分比）	城市化率	抚养比	通电率（占人口的
白俄罗斯	2000	0.683	0.668	0.729	0.654	7326000000	8546000000	2.667499078	3.243640757	40.54999924	69.973	47.11470693	100
白俄罗斯	2001	0.689	0.677	0.73	0.662	7451000000	8286000000	3.47239953	5.739999291	40.59600067	70.458	46.46152271	100
白俄罗斯	2002	0.696	0.687	0.733	0.671	8020900000	9092300000	3.428558146	5.805499251	40.68700027	70.939	45.81885264	100
白俄罗斯	2003	0.705	0.696	0.736	0.683	9945600000	11558000000	3.115379314	3.767096657	40.81600189	71.415	45.14530088	100
白俄罗斯	2004	0.715	0.706	0.739	0.7	13773700000	15490800000	2.417055599	3.039575256	40.97299957	71.887	44.39199122	100
白俄罗斯	2005	0.725	0.716	0.744	0.715	15979000000	16708000000	1.973050187	3.000924704	41.14199829	72.354	43.54112718	100
白俄罗斯	2006	0.741	0.745	0.749	0.729	19734000000	22351000000	1.859123847	2.928968874	41.34799957	72.816	42.89809448	100
白俄罗斯	2007	0.757	0.774	0.756	0.741	24275300000	28693100000	1.783892088	2.440241058	41.50699997	73.273	42.11445227	100
白俄罗斯	2008	0.774	0.804	0.764	0.755	32570800000	39381300000	1.61525403	2.104259877	41.74499893	73.726	41.34668533	100
白俄罗斯	2009	0.784	0.828	0.772	0.754	21304200000	28509000000	2.315890795	2.513511797	41.80699921	74.172	40.77717088	100
白俄罗斯	2010	0.792	0.829	0.781	0.766	25283500000	34884400000	2.269492896	2.032509279	41.94900131	74.672	40.51082392	100
白俄罗斯	2011	0.798	0.833	0.789	0.774	41418700000	45759000000	1.605170928	1.529829473	41.75999832	75.188	40.78428785	100
白俄罗斯	2012	0.803	0.836	0.796	0.776	46059900000	49404400000	1.900339932	1.93264291	41.87799835	75.697	41.32039577	100
白俄罗斯	2013	0.804	0.835	0.803	0.775	37203100000	43022700000	2.624522943	2.87364994	42.15800095	76.199	42.07561949	100
白俄罗斯	2014	0.807	0.835	0.808	0.778	36126500000	40613700000	2.840456413	2.986860554	42.2120018	76.693	42.93996287	100
白俄罗斯	2015	0.805	0.836	0.812	0.769	26660400000	30291500000	3.088621798	3.049200082	41.9280014	77.181	43.82544795	100
白俄罗斯	2016	0.805	0.838	0.814	0.764	23537300000	27609900000	3.404963428	2.974074482	43.45999908	77.661	44.96682397	100

图 4-7 原始数据-以白俄罗斯为例

4.2.1. 相关性分析

利用 SPSS 软件按国家，以 HDI 为因变量，商品出口、商品进口、国际旅游收入、国际旅游支出、工业男性就业人员、城市化率、抚养比、通电率等数据进行多重共线性检验，排除存在明显多重共线性的自变量，对剩余自变量与 HDI 进行回归分析。以白俄罗斯为例：

相关性

		商品出口（现价美元）	商品进口（现价美元）	国际旅游，收入（占总出口的百分比）	国际旅游，支出（占总进口的百分比）	抚养比	通电率（占人口的百分比）
商品出口（现价美元）	Pearson 相关性	1	.983**	.789**,*	-.719**	-.770**	.c
	显著性（双侧）		.000	.000	.001	.000	.
	N	17	17	17	17	17	17
商品进口（现价美元）	Pearson 相关性	.983**	1	.702**,*	-.756**	-.833**	.c
	显著性（双侧）	.000		.002	.000	.000	.
	N	17	17	17	17	17	17
国际旅游，收入（占总出口的百分比）	Pearson 相关性	.789**,*	.702**,*	1	-.531**,*	-.480**	.c
	显著性（双侧）	.000	.002	.855	.028	.051	.
	N	17	17	17	17	17	17
国际旅游，支出（占总进口的百分比）	Pearson 相关性	-.719**	-.756**	-.531**,*	1	.783**,*	.c
	显著性（双侧）	.001	.000	.028		.000	.
	N	17	17	17	17	17	17
抚养比	Pearson 相关性	-.770**	-.833**	-.480**	.783**,*	1	.c
	显著性（双侧）	.000	.000	.051	.000	.022	.
	N	17	17	17	17	17	17
通电率（占人口的百分比）	Pearson 相关性	.c	.c	.c	.c	.c	.c
	显著性（双侧）	.	.	.	.	.	
	N	17	17	17	17	17	17

**. 在 .01 水平（双侧）上显著相关。

*. 在 0.05 水平（双侧）上显著相关。

c. 因为至少有一个变量为常量，所以无法进行计算。

系数[a]

模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	(常量)	-.020	.124		-.163	.874		
	商品出口（现价美元）	-7.659E-013	.000	-.200	-1.114	.294	.013	75.800
	商品进口（现价美元）	1.437E-012	.000	.413	1.988	.078	.010	101.332
	国际旅游，收入（占总出口的百分比）	.012	.002	.638	6.823	.000	.049	20.528
	国际旅游，支出（占总进口的百分比）	-.002	.002	-.040	-.694	.505	.126	7.938
	抚养比	-.005	.002	-.211	-3.046	.014	.089	11.228

a. 因变量: HDI指数

对相关性较高的变量重点进行 VIF 检验，去除多重共线性的影响。

系数[a]

模型		非标准化系数		标准系数	t	Sig.	共线性统计量	
		B	标准 误差	试用版			容差	VIF
1	(常量)	-.537	.224		-2.396	.032		
	商品出口（现价美元）	.035	.005	.553	6.862	.000	.563	1.776
	抚养比	-.004	.002	-.197	-2.071	.059	.405	2.468

a. 因变量: HDI指数

对通过多重共线性检验后的变量进行回归分析。

SUMMARY OUTPUT

回归统计	
Multiple R	0.883237
R Square	0.780107
Adjusted R Square	0.748694
标准误差	0.023418
观测值	17

方差分析

	df	SS	MS	F	ignificance F
回归分析	2	0.027237	0.013619	24.83372	2.49E-05
残差	14	0.007677	0.000548		
总计	16	0.034914			

	Coefficients	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	0.843655	0.203932	4.136943	0.001007	0.406264	1.281046	0.406264	1.281046
X Variable 1	2.89E-12	7.52E-13	3.843298	0.001791	1.28E-12	4.5E-12	1.28E-12	4.5E-12
X Variable 2	-0.00354	0.004389	-0.80649	0.433445	-0.01295	0.005874	-0.01295	0.005874

对商品出口和抚养比作为自变量与 HDI 进行回归，显示抚养比 P 值超过 0.05，不显著。商品出口 P 值为 0.0018，显著，系数为近似于零的数字。对俄罗斯、柬埔寨等其他国家分析结果类似。分析有以下主要原因：1、样本量有限。因研究对象中多为发展中及不发达国家，统计的基础数据缺失严重，本次回归分析只选取了 2000 年以后的数据；2、自变量选取需优化。因变量为 DHI 指数，由预期寿命、预期受教育年限、经济发展情况组成，选取的自变量应更为具有代表性。

5. 机器学习

5.1. 方法与实验结果

注：构建模型之后使用 tensorflow 的插件 tensorboard 查看模型。

5.1.1. 共享参数的残差网络：

由于数据量相对比较小，因此在这次的网络架构中，采用相对浅层的经典 resnet18 作为初始结构，在其上 fork weights share 的模块（先做了 resnet18 之后再 share wieghts，具体在后面有详细的记录）。可以看到，总共有 4 个 stage，g2-g5。 每个 stage 有两个 res block，b1-b2。每个 block 的有两个 3×3Conv。在每个 stage 的第一个 block 的第一个 conv 做 downsample。

Loss 由两部分组成，一个 regression loss named ‘xent-loss’， 另一个是 L2reg，用于控制模型复杂度，在后面，我们重点关注 regression loss。

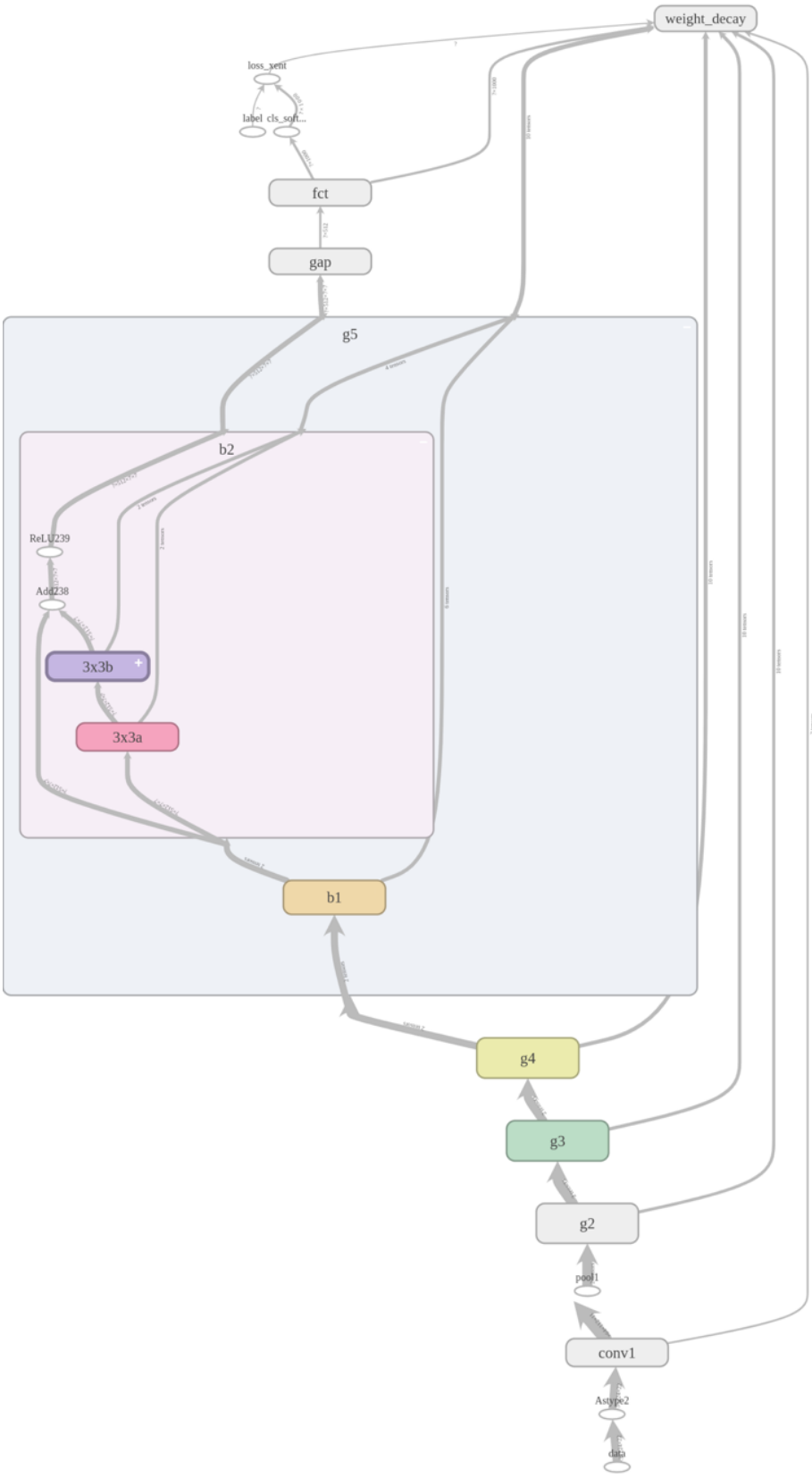


图 4. Resnet18 的网络架构

图 5. 网络中 G4b2 的结构概览，红线标出的部分表示权重已经共享

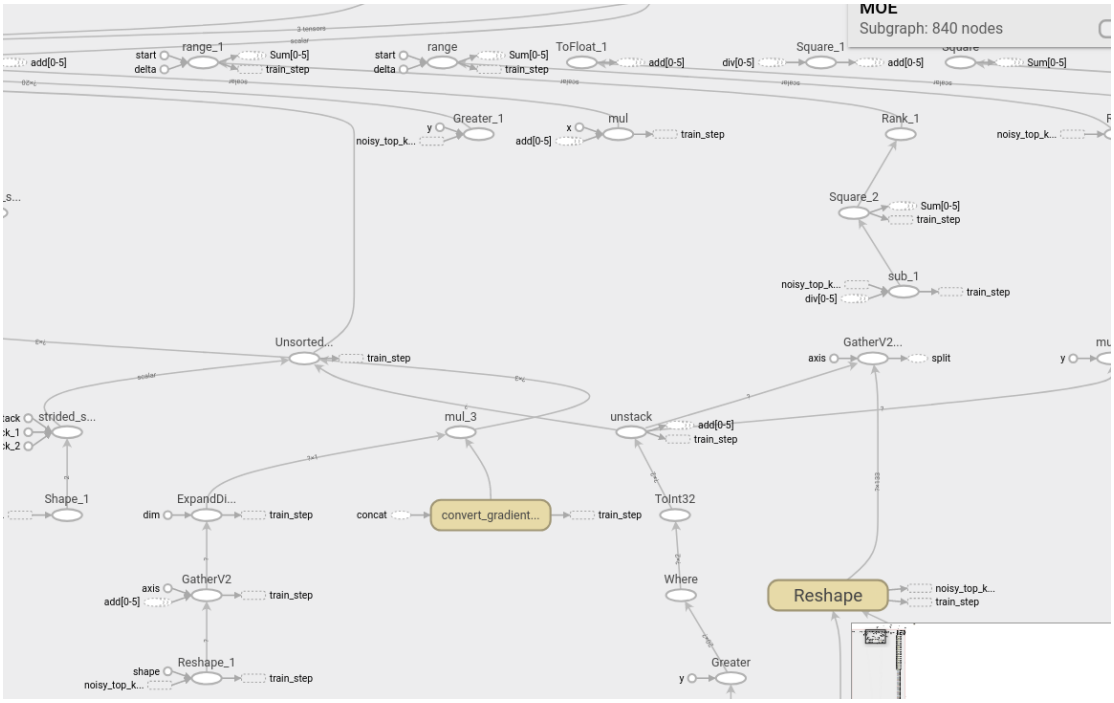


图 6.局部架构展示

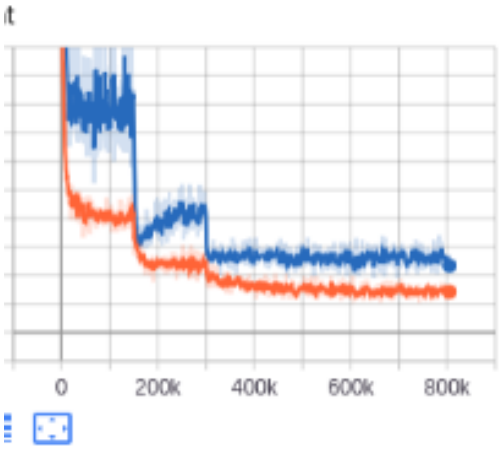
Baseline 的训练：

随机抽取大约十分之一的数据作为独立测试集，剩下的样本作为训练集使用。在训练时，采用十折交叉验证来对学习率，扩增步长等参数进行调优和选择。每一折训练 70 轮。训练使用的机器为清华大学生命学院高性能计算 CPU 集群中的一个核心。共使用 9 组参数组合。选择一个 cross validation 过程，记录梯度和 loss，将十折平均后作图，观察训练过程中的变化。

注：cross Validation 只用在选择 baseline 的 hyper paras 的时候，共训练 90 折，6300 个 epoch，选出 model performance 最大的那一组 hypers 进行后面所有模型的训练。

在后面模型的训练中，每一个实验都训练 120 个 epoch，使用全部训练集， 之后将这 120 个 models 使用独立测试集测试一遍，报道 error 最低的那一个作为这个实验的结果。

loss 的变化如下图：



在独立测试集上进行测试，得到的数据为，HDI L2 error：0.319。

这个 L2 的水平有些太高了，回过头看 training set 的 L2 error， 发现只有 0.047 左右，差距说明了数据量难以支撑 Resnet 18 网络的表征量，因此加强力的 regularization， 引入 share_resnet，具体介绍在前面。引入参数共享后，training set 的 L2 error 仍然只有 0.043 左右，但是在独立测试集上，HDI L2 error 降低到了：0.286 左右，虽然 overfitting 还是很大，但是比刚才好多了。

b．在以相关性为 keypoint 的传统数据分析方法中（感谢我们组的吴子祎，张晶，徐春辉以及张恩嘉在这方面辛苦的研究与探索），我们分析选出了和 HDI 关系最为密切的指标：受教育指数，预期寿命指数，经济发展指数，商品进口，工业男性就业人员（占男性就业的百分比），城市化率，使用传统的特征工程方法，挨个去掉上面的 feature，在独立测试集上观察 L2 error 的变化。

结果如下表；

Feature	L2
All	0.286
Remove EducationIndex	0.291
Remove ExpectedAgeIndex	0.283
Remove EconomicIndex	0.289
Remove Merchandise.imports	0.285
Remove Employment.in.industry..male.	0.287
Remove Urbanization	0.287

5.1.2. 长短期记忆网络

由于我们研究的对象是随着时间变化的，也就是说，某一年的状况不是凭空产生的，而是受前几年的影响，此外，它也不是孤零零的“就这样了”，它的状态也是会影响到后

面几年的状态。之前我们用了卷积的结构来“整合不同年份的情况，将年份作为一个独立的 feature 输入，虽然深层的卷积会将 receptive field 扩展到整个 feature 的所有维度，并且试图寻找高维空间中的某种拓扑相关性，但是对于带有时序信息 feature，RNN 还是会有更好的表现，这里我们选择 LSTM 进行训练。得到的结果类似。

Feature	L2
All	0.136
Remove EducationIndex	0.139
Remove ExpectedAgeIndex	0.136
Remove EconomicIndex	0.137
Remove Merchandise.imports	0.136
Remove Employment.in.industry..male.	0.136
Remove Urbanization	0.141

5.2. 讨论与结论

虽然上面的网络存在 overfitting 问题，但是从得到的实验结果也还是能在某种程度上说明问题。

首先在 share res 的基础上做的结果，去除之后 L2 error 减小的指标有：ExpectedAgeIndex，Merchandise.imports 两项，其余均为增加。在 LSTM 上的结果，在当前精度下，L2 减小的没有，和 L2 持平的有：ExpectedAgeIndex，Merchandise.imports，以及 Employment.in.industry..male，其余 L2 增加。

在 share res 的实验中，去除 Employment.in.industry..male 之后 L2 error 增加很少，可认为是偶发或者由于随机过程的影响，所以将其也归类为去除之后 L2

减小的类别。

由于 LSTM 的 overfitting 进一步减小，所以在合理的范围内，去除某个维度的 feature 但是对模型的回归预测性能不产生影响，我们还是可以认为该 feature 是含有冗余或者“逆信息”的，干扰了对最终结果 HDI 的拟合。同时，在 LSTM 和 share res 的实验中，EducationIndex，EconomicIndex 和 Urbanization 同时出现了去除该 feature 之后 L2 error 的升高，说明在某种程度下，HDI 的回归预测是需要这些指标的，这些指标带有其他指标所没有的互补信息，也就是说，HDI 和这些信息高度相关，且缺了这些信息，可能导致 HDI 的预测不完全。

结合我们前面的传统分析方法得到的结论，我们认为：

在三大指标中（教育，经济以及预期寿命），具有目前不可替代作用的是教育和经济，预期寿命包含的信息可以被其他指标用某些简单函数表达出来。在我们筛选得到的四个关键小指标中，具有不可替代作用的是城镇化率，其他指标均可被拟合或者直接包含在大指标中。

结合之前的传统分析方法，具体为：

商品进口量对教育、经济发展水平影响很小，可以忽略不计。城镇化率对教育、经济发展水平影响的影响较小。男性劳动力参与率和抚养比的影响稍微大一些，但是总体来说也很小。

商品进口量对健康寿命、人类发展水平影响很小，可以忽略不计。健康寿命水平与城镇化率相关，男性劳动力参与率和抚养比相关。总体人类发展指数与城镇化率等多个指标相关，但是其中某些指标带有冗余信息，也就是说其信息可被其他指标所表示。

6. 展望与总结

本次小组报告利用《大数据分析 A》课程中学到的知识，使用 Python 语言，对“一带一路网”新闻内容进行数据爬取、数据清洗、数据整理、文本分析、统出词频热度最高的国家。本文综合词频热度和专家访谈结果，重点选取 7 个典型国家为研究对象。使用 R 软件，基于联合国计划开发署和世界银行的开放数据库，进行线性回归、文本挖掘、数据可视化，并配套机器学习验证。

但是由于自变量选取过少、研究国家太少等原因，本次报告结果并不显著。在后续的研究中，我们会继续增加自变量，做更加深入的研究。

感谢本组全部成员的专业的态度和辛勤的付出，谢谢你们！

参考文献

[1] 国家发展改革委、外交部、商务部，《推动共建丝绸之路经济带和 21 世纪海上丝绸之路的愿景与行动》，2015 年 3 月。

[2] 王志平，《“人类发展指数”(HDI):含义、方法及改进》，《上海行政学院学报》，2007 年 03 期。

[3] 联合国开发计划署,《2018 数据更新：人类发展指数和指标》,2015 年 3 月。

[4] Habibzadeh, M., Jannesari, M., Rezaei, Z., Baharvand, H. & Totonchi, M. in 10th International Conference on Machine Vision (ICMV). (2018).