一带一路对中国和其他沿线国家的影响及政策分析

——数据科学的视角

范皓年 邓睿哲 李润泽

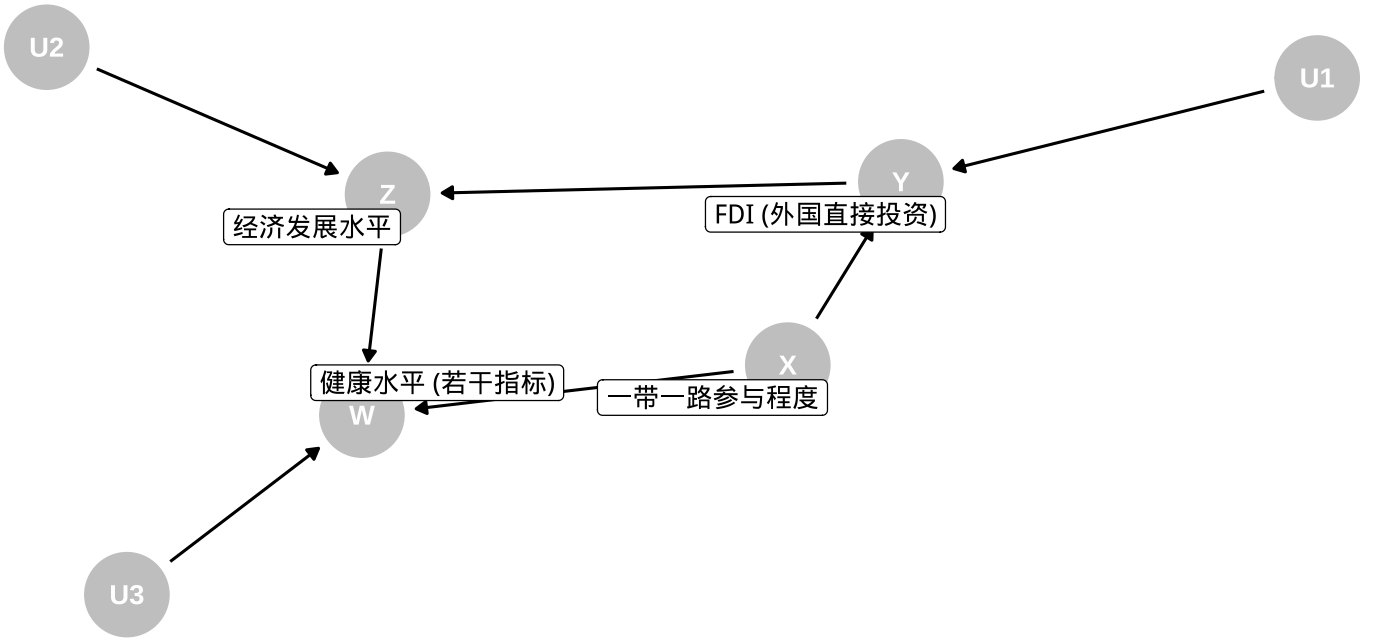
Peking University

2021-07-05

概要

主要工作

数据模型

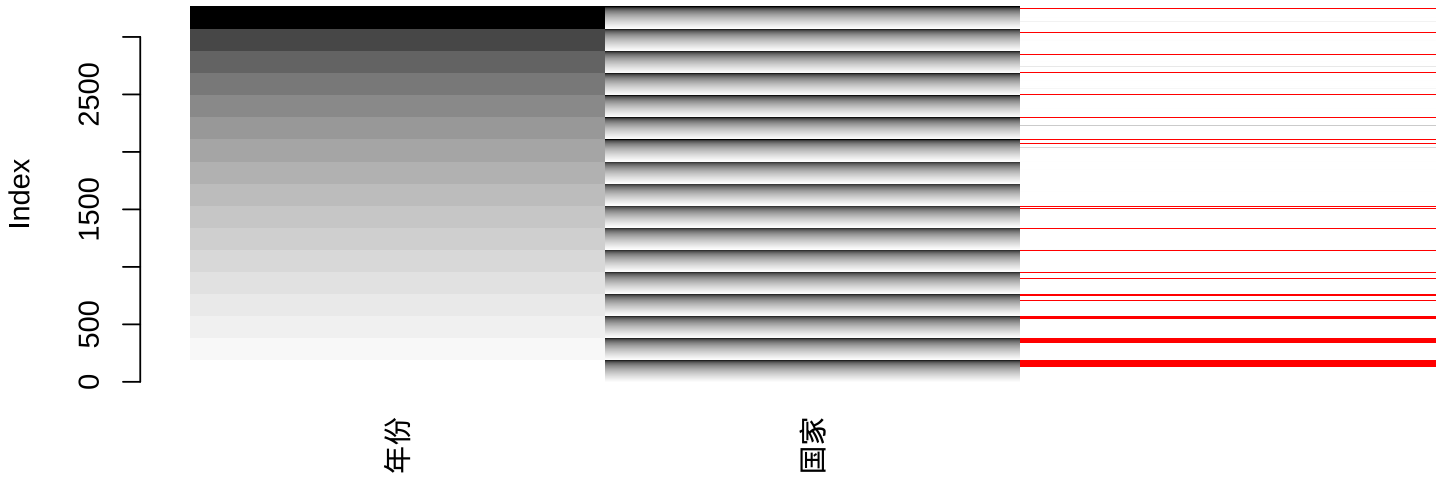


- Assumption: 不存在 $X \to Z$的线，即 $X \perp\!\!\!\perp Z|Y$.

分析技术

- 缺失数据填补
- 双重差分法
- 合成控制法

缺失数据填补



- 删除
- 填补：linear regression with bootstrap

二重差分法

二重差分法 (Difference-in-Differences, DID) 是一种经典技术. 具体来说，就是以下模型

$$P_t^N = \mu + \frac{1}{J}\sum_{j=2}^{J+1} Y_{jt}^N$$

并用如下公式来估计.

$$\hat{P}_t^N = \frac{1}{T}\sum_{s=1}^{T}\left(Y_{1s}^N - \frac{1}{J}\sum_{j=2}^{J+1} Y_{js}^N\right) + \frac{1}{J}\sum_{j=2}^{J+1} Y_{jt}^N$$

合成控制法

合成控制法 (Synthetic Control) 是一种稍微新一点的技术. 具体来说，就是以下模型.

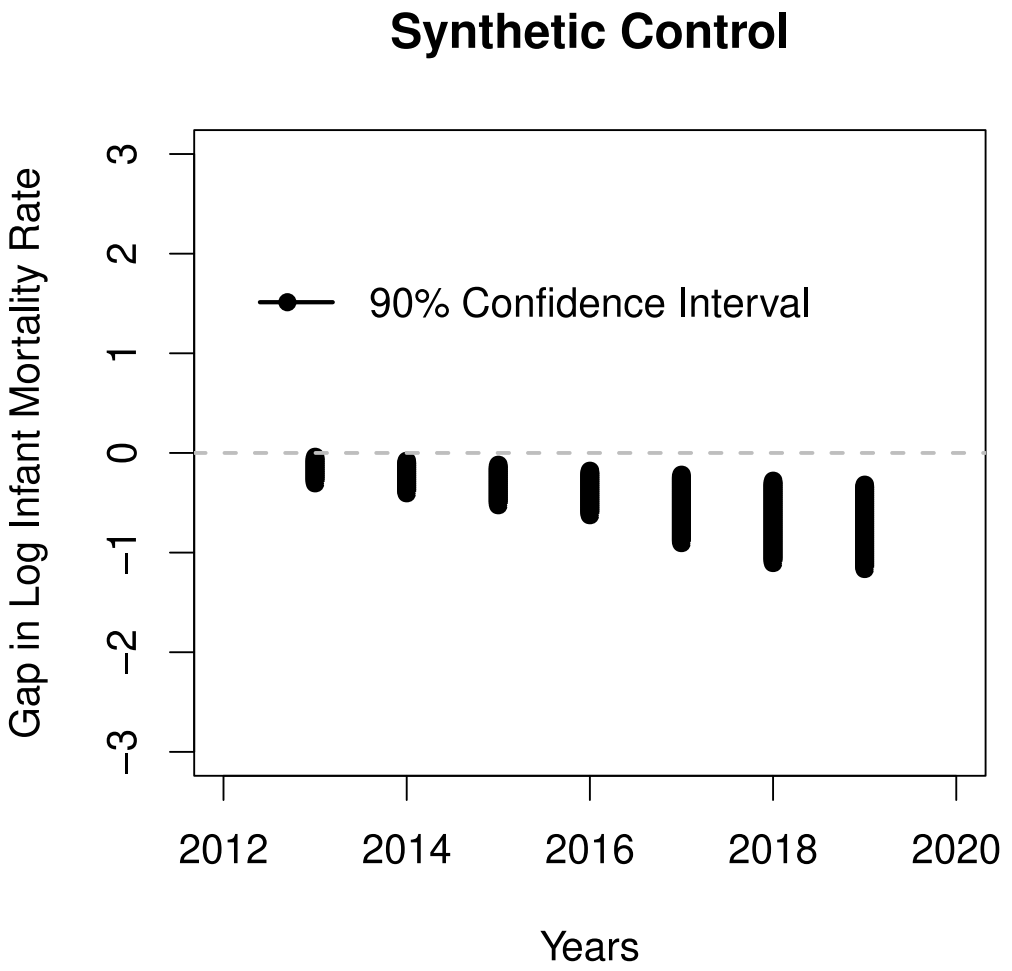
$$P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N, \text{where } w \geq 0, \ \sum_{j=2}^{J+1} w_j = 1.$$

于是就有估计 $\hat{P}_t^N = \sum_{j=2}^{J+1} \hat{w}_j Y_{jt}^N$.

而 w的估计

$$\begin{aligned} \hat{w} =& \arg\min_w \sum_{i=1}^{T} \left(Y_{1t}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N \right)^2 \\ & \text{subject to } w \geq 0, \ \sum_{j=2}^{J+1} w_j = 1. \end{aligned}$$

- P值，置信区间的计算 (Chernozhukov et al., 2021)



程序技术

Non-standard evaluation, NSE

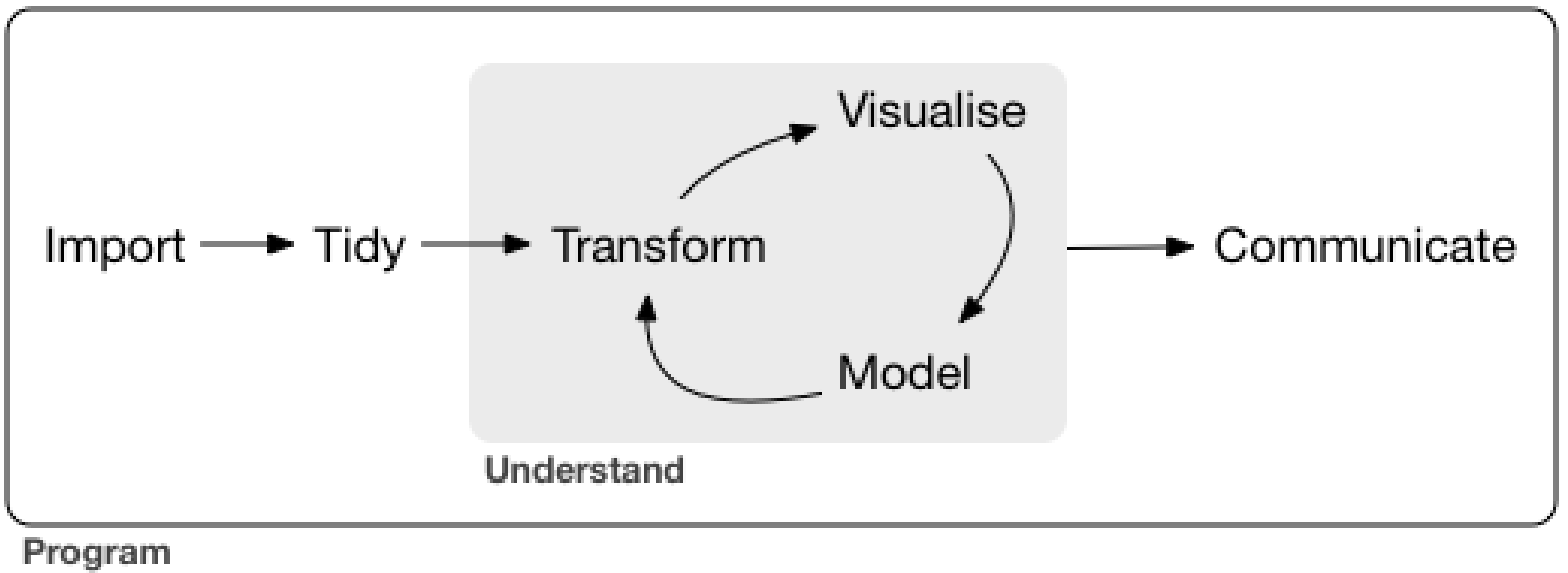
```
### Use lazy evaluation to replicate a func
repli <- function(fun) {
  ex <- substitute(fun)

  for (i in seq_along(country_list)) {
    # ...
    eval(ex, envir = globalenv())
  }
}

repli(placebo_specification_test())
```

- 只有在真正用到fun的时候才会对其进行求值，其中fun的返回值并不必须良定

具体细节

- The Workflow
- 数据集说明
- 数据清洗
- 数据分析
- 数据可视化

The Workflow



此图取自 *R for Data Science*，以CC BY-NC-ND 3.0 US发布.

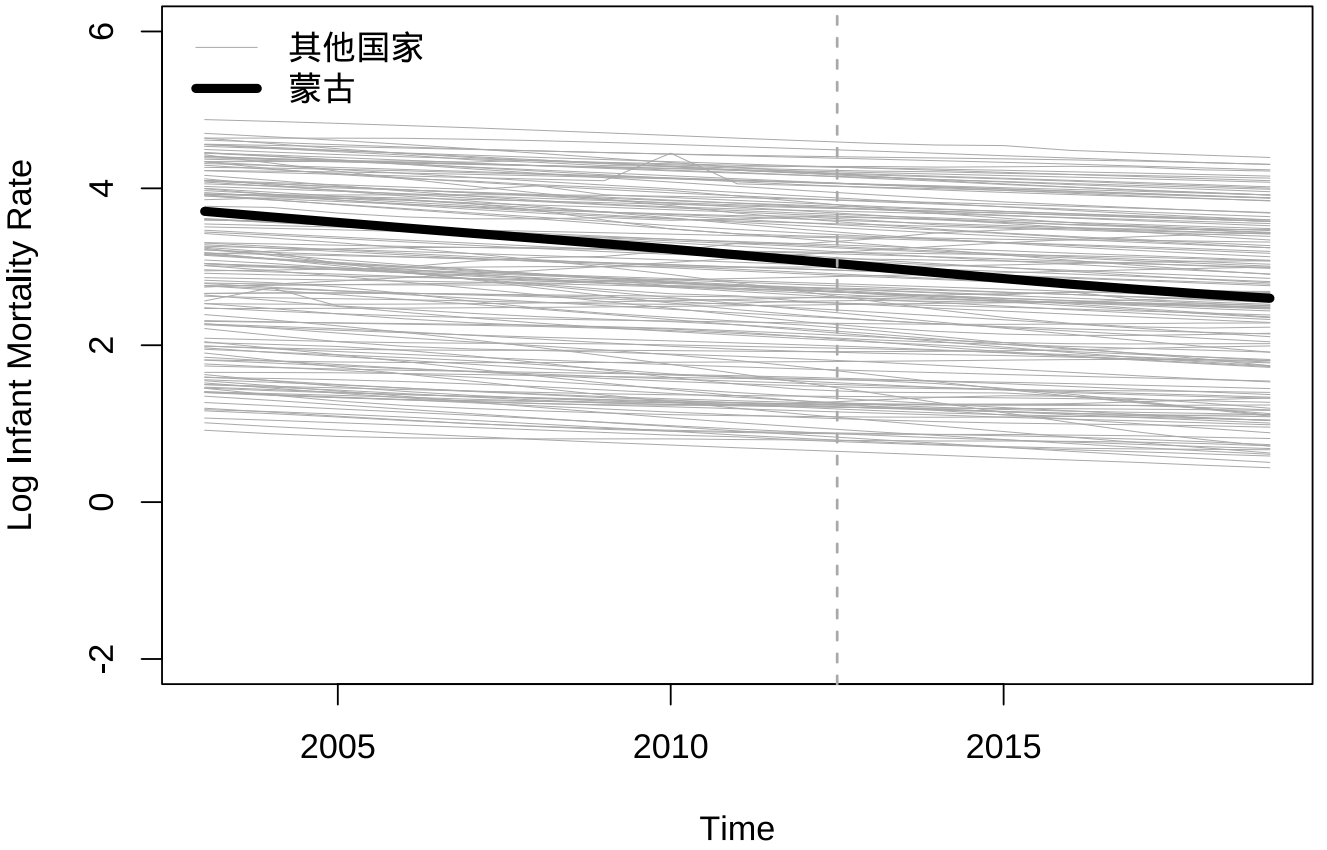
数据集说明

国际贸易数据（`/data/investment/FDI_untidy.csv`）下载自CEIC数据库. 我们引用CEIC全球数据库中“实际利用的外国资本：按地区分类”和“对外直接投资：国别”两个数据集，逐条下载了中国在利用其他国家资本量（月度数据），和中国对其他国家的直接投资（月度数据）. CEIC数据库覆盖的时间范围为1985年12月至2021年4月，每条数据均以月为统计单位.

其他用到的数据集来自世界健康数据集，包括：

- under5MortalityRate.csv 数据集记录了1962-2019年不同国家5岁以下儿童死亡率（每千人的死亡人数）.
- infantMortalityRate.csv 数据集记录了1962-2019年不同国家出生婴儿死亡率（每千人的死亡人数）.

画个图



数据清洗

- 什么样子的数据结构才算是整理好?
- 每行应该代表一个观察(observation)，每列应该代表一个变量(variable).
- 利用tidyverse，结合**正则表达式**，进行数据清洗. 其部分步骤如下页所示.

数据清洗

```
simplified_df <- raw_df %>%
  filter(X1 %>% str_detect("^\\d"))
```

```
fliped_df <- simplified_df %>%
  pivot_longer(c(-时间), names_to = "observation", values_to = "val")
```

```
str %>%
  str_replace(pattern = "(.*):(总计|一带一路)", replacement = "1/\\2/\\2
  str_replace(pattern = "::", replacement = ":") %>%
  str_replace(pattern = "(.*):(.*洲):*(.*)", replacement = "1/\\2/\\3")
```

```
df <- fliped_df %>%
  mutate(observation = observation %>% stdize()) %>%
  separate(col = "observation", into = c("type", "地区", "国家"), sep =
  spread(key = "type", value = "val")
```

数据分析

数据建模和分析是传统上受重视的技术. 其主要内容已经详述，这里不再赘述.

- 本项目评估了：对测试进行安慰剂检验 (placebo test) 的情况（`sens.csv`），测试的P值（`p.noeff.csv`），结果的置信区间（`ci.csv`）
- 并以`pdf`文件，绘制出了按国家分类的置信区间的情况

数据可视化

可视化工具

项目将世界经济及其相关的数据，展示在世界地图上，考虑Python语言相对于JavaScript具有更好的数据处理能力，我们使用基于Apache Echarts的Pyecharts.

我们主要做了如下几个可视化工作：

- 将2003到2019年的中国对外直接投资总额表示在地图上
- 将世界健康数据集中预期寿命和5岁以下死亡率分性别表示在图中

我们从图中可以定性地看出中国外企对于一带一路沿线国家的投入，以及相应国家的经济水平、生活水平的优化.

文件结构

```
visualization
├── README.md
├── data
│   ├── FDI_filled_m.csv
│   ├── FDI_useful.csv
│   ├── LE.csv
│   ├── UFMR_m.csv
│   ├── country_ce.json
│   ├── syno_dict.json
│   └── world_country.json
├── mytool.ipynb
├── raw_plot
│   └── ...
├── out
│   ├── 五岁以下死亡率.html
│   ├── 外商直接投资情况-filled.html
│   ├── 外商直接投资情况.html
│   └── 预期寿命.html
├── FDI.py
└── world_health.ipynb
```

可视化相关的脚本以及输出结果全部储存在./visualization中.

其中./visualization/data/是可视化所用到的数据，不仅包括我们绘图所需的数据，包括对外直接投资FDI*.csv、健康相关数据LE*.csv和UFMR*.csv等，还包括中英对照表country_ce.json、以及国家名的同义对照表syno_dict.json等工具数据.

raw_plot/目录是用R生成的原始数据变化情况，其中一带一路参与国家以加粗线绘制.

mytool.ipynb为工具和测试用notebook，用于生成工具json和进行原型开发测试.

FDI.py为对外直接投资可视化脚本，出于易用性，其中render()函数中给出的文件名，在得到成品文件后稍后手动更改为中文.

world_health.ipynb为世界卫生健康相关数据可视化脚本，前两个cell分别用于绘制世界国家预期寿命和5岁以下死亡率，第三个cell尝试将不同的性别绘制在同一张图中，但是由于timeline和gender两个尺度只能分开调整，所以在时间纵向对比时并不方便，我们将结果绘制为三个图构成的Page Echarts图.

./visualization/out/是可视化的文件，成品文件名已经更改，相对清楚. 注意其中外商直接投资情况-filled.html为利用算法填充部分缺失数据之后的FDI图像.

流程

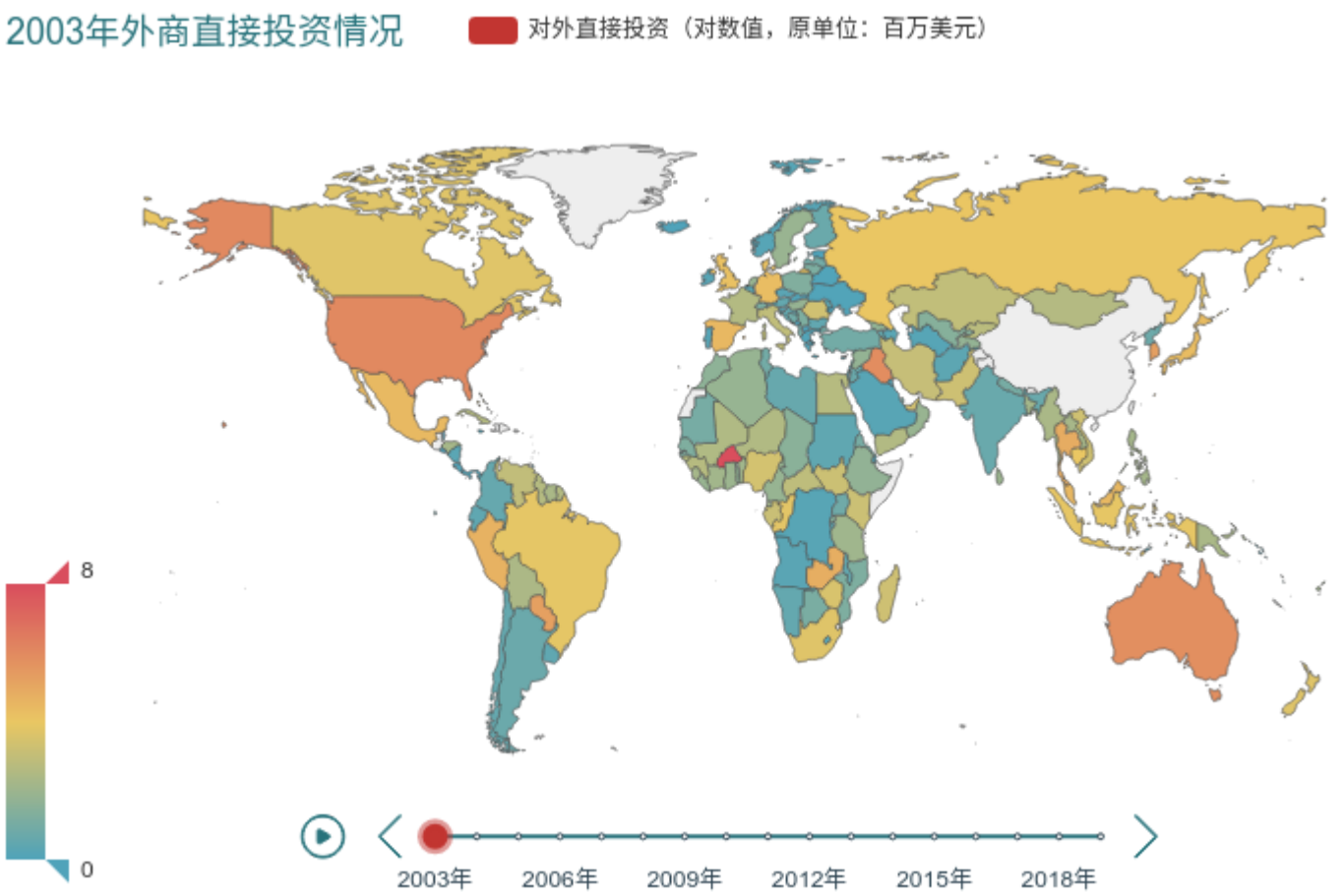
以FDI（对外直接投资）为例，我们讲述项目中使用的pyecharts可视化方法，相对其他几个可视化工作，其中使用了对数化、相对复杂，故说明后其余同理.

```
import pandas as pd                                          # 数据分析组件
import json                                                  # 用于导入工具json
from pyecharts import options as opts                        # 用于调整pyechart
from pyecharts.charts import Timeline, Map                   # 选取pyecharts基
from pyecharts.globals import ThemeType                      # 选取pyecharts主
import numpy as np                                           # python数值计算工
tl = Timeline(init_opts=opts.InitOpts(
    theme=ThemeType.INFOGRAPHIC,
    bg_color='white',
    page_title='外商直接投资情况'
))                                                           # 生成timeline图结
with open("./data/country_ce.json", 'r', encoding='utf-8') as f:
    ce_dict = json.load(f)                                   # 导入国家名称中英文

df = pd.read_csv('./FDI_filled_m.csv')                       # 生成dataframe
df.iloc[:, 3] = df.iloc[:, 3].apply(np.log1p)                # 将数值列对数化
```

流程

```
for year in range(2003, 2019+1):                                    # 循环添加不同年份的
    map = (
        Map()                                                       # 生成一个年份的地图
        .add(df.columns.tolist()[-1]+"（对数值，原单位：百万美元）",# 设定图
             [[ce_dict[row['国家']], row[3]]                        # 读入数据，使用data
                 for _, row in df[df.iloc[:, 0] == year].iterrows()],
             maptype="world",                                       # 设定为世界地图
             is_map_symbol_show=False,                              # 不描点
             )
        .set_series_opts(label_opts=opts.LabelOpts(is_show=False)) # 右
        .set_global_opts(
            title_opts=opts.TitleOpts(title=f"{year}年外商直接投资情况"),
            visualmap_opts=opts.VisualMapOpts(
                max_=df[df.iloc[:, 0] == year].iloc[:, 3].max()),    #
            toolbox_opts=opts.ToolboxOpts(),                                #

        )
    )
    tl.add(map, f"{year}年") # 将当前图层加入timeline结构中
tl.render("./out/vis.html") # 生成临时文件
```

可交互图形 by ECharts



总结

分析结果

```
ci_csv %>% filter(`max(ci.sc)`<0) %>% group_by(国家) %>% nest() %>% .[[
```

```
## [1] "哈萨克斯坦" "斯洛文尼亚" "拉脱维亚"    "黑山"
```

```
ci_csv %>% filter(`min(ci.sc)`>0) %>% group_by(国家) %>% nest() %>% .[[
```

```
## [1] "新加坡"    "格鲁吉亚" "白俄罗斯"
```

```
ci_csv %>% .[["median(ci.sc)"]] %>% mean()
```

```
## [1] -0.03983766
```

- 6年间，婴儿死亡率对数 -0.04（均值 -0.18）

分析结果

```
cii_csv %>% filter(`max(ci.sc)`<0) %>% group_by(国家) %>% nest() %>% .[
```

```
## [1] "缅甸"       "文莱"       "巴林"       "也门共和国" "叙利亚"
## [6] "阿塞拜疆"   "斯洛文尼亚" "匈牙利"
```

```
cii_csv %>% filter(`min(ci.sc)`>0) %>% group_by(国家) %>% nest() %>% .[
```

```
## [1] "印度尼西亚"
```

```
cii_csv %>% .[["median(ci.sc)"]] %>% mean()
```

```
## [1] -0.4027635
```

- 6年间，投资额对数 -0.40
- 在相同的经济发展水平下，参与一带一路能够相比预期增加沿线国家的国民健康水平，或者狭义来说，**降低婴儿死亡率**.

不足和展望

关于本幻灯片

R Markdown syntax,
Powered by **xaringan** and remark.js

THANKS