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*

目录

1 SUMMARY **2**

1.1 概要 2

2 前言 **2**

2.1 环境 2

3 主要结果 **2**

3.1 数据模型 3

3.2 分析技术 3

3.3 程序技术 5

4 具体流程 **6**

4.1 The Workflow 6

4.2 Import 6

4.3 Tidy 6

4.4 Understand 8

4.5 Communicate 9

5 总结 **12**

5.1 建议 12

5.2 展望 12

*名拼音序.

参考文献 **12**

1 SUMMARY

1.1 概要

2 前言

2.1 环境

2.1.1 R info

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.2 LTS
##
## Locale:
##   LC_CTYPE=zh_CN.UTF-8       LC_NUMERIC=C
##   LC_TIME=zh_CN.UTF-8        LC_COLLATE=zh_CN.UTF-8
##   LC_MONETARY=zh_CN.UTF-8    LC_MESSAGES=zh_CN.UTF-8
##   LC_PAPER=zh_CN.UTF-8       LC_NAME=C
##   LC_ADDRESS=C               LC_TELEPHONE=C
##   LC_MEASUREMENT=zh_CN.UTF-8 LC_IDENTIFICATION=C
##
## Package version:
##   dplyr_1.0.6      ggdag_0.2.3      ggplot2_3.3.3    lubridate_1.7.10
##   mice_3.13.0      purrr_0.3.4      readr_1.4.0      showtext_0.9-2
##   stringr_1.4.0    tidyr_1.1.3      tidyverse_1.3.1  VIM_6.1.0
```

2.1.2 python info

本项目的 python 部分使用 python 3.8.8 生成，部分包版本号如下：

conda 4.10.1; **pyecharts** 1.9.0; **numpy** 1.20.1; **json5** 0.9.5; **pandas** 1.2.4; **jupyterlab** 3.0.14

3 主要结果

3.1 数据模型

我们的数据模型如图所示：

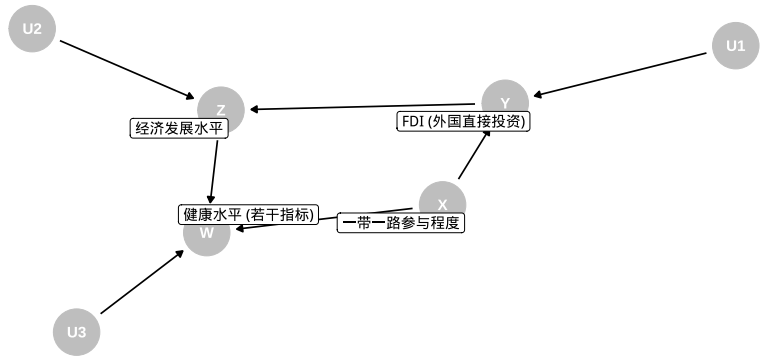

图 1: 数据模型示意图

此图在 R 语言中，用 **ggdag**[1] 生成. 是有向无环图 (Directed acyclic graph, DAG)，边代表因果作用[2].

在该模型中，我们假定

3.2 分析技术

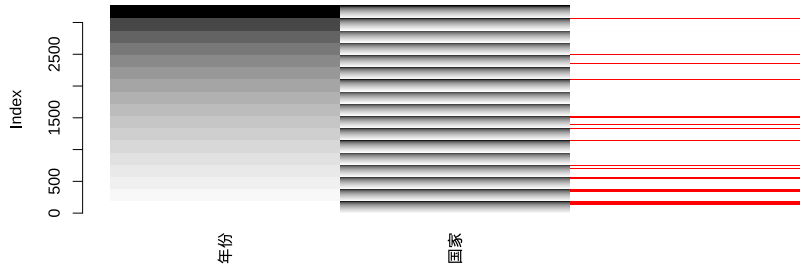

图 2: 缺失数据示意图

首先注意到数据集中存在许多缺失数据[3].

我们利用两种方法分析.

3.2.1 二重差分法

二重差分法 (Difference-in-Differences) 是一种经典技术. 它运用以下模型[4].

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

并用如下公式来估计.

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

3.2.2 合成控制法

合成控制法 (Synthetic Control) 是另一种经典技术. 它运用以下模型[5].

$$P_t^N = \sum_{j=2}^{J+1} w_j Y_{jt}^N, \text{ where } w \geq 0 \text{ and } \sum_{j=2}^{J+1} w_j = 1 \ . \ (3)$$

为了识别权重 w，我们需要一些假设. 我们假定结构冲击项 u_t 在同一时段内互不相关，即:

$$E\left(u_i Y_{jt}^N\right) = 0 \text{ for } 2 \leq j \leq J+1.$$

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

而 w 的估计

$$\widehat{w} = \arg\min_{w} \sum_{i=1}^{T}\left(Y_{1t}^N - \sum_{j=2}^{J+1} w_j Y_{jt}^N\right)^2 \text{ s.t.w} \geq 0 \text{and} \sum_{\mathrm{j=2}}^{\mathrm{J+1}} \mathrm{w_j} = 1.$$

此后，我们利用 Chernozhukov et al.[6] 的方法分析其置信区间.

3.2.3 缺失数据填补

我们的数据集，正和许多类似的真实世界数据集一样，存在着许多缺失数据 `NA`. 缺失数据的处理方式不外乎删除或填补.

- 对于我们的 investment 数据集，其缺失普遍存在，故我们采用填补的方法.
- 对于本次大赛提供的世界健康数据集，其缺失更有规律，即对于任意一个国家，缺失一个时间点的数据意味着缺失此前所有数据. 恰当地选择时间范围，再删去个别几个缺失较大的国家[1]，就在可以避免填补的同时保留大部分数据. 因此，我们选择删去.

[1] 这些国家往往那时才成立，例如从母国中分裂出.

对于前者，我们调用 R 包 **mice**[7]，采用 linear regression with bootstrap 的方法进行缺失数据填补. 首先，对数据进行 bootstrap 重抽样，进行线性回归插值，然后计算均值得到最后结果. 其中，多元线性回归的步骤采用 Schafer 的算法[8]，其步骤如下：

1. Calculate the cross-product matrix $S = X'_{obs}X_{obs}$
2. Calculate $V = (S + \text{diag}(S)\kappa)^{-1}$, with some small ridge parameter κ.
3. Calculate regression weights $\hat{\beta} = VX'_{\text{obs}}\, y_{\text{obs}}$.
4. Draw a random variable $\dot{g} \sim \chi^2_\nu$ with $\nu = n_1 - q$.
5. Calculate $\dot{\sigma}^2 = \left(y_{obs} - X_{obs}\hat{\beta}\right)' \left(y_{obs} - X_{obs}\hat{\beta}\right)/\dot{g}$.
6. Draw q independent $N(0,1)$ variates in vector $\dot{z}_1$.
7. Calculate $V^{1/2}$ by Cholesky decomposition.
8. Calculate $\dot{\beta} = \hat{\beta} + \dot{\sigma}\dot{z}_1 V^{1/2}$.
9. Draw n_0 independent $N(0,1)$ variates in vector $\dot{z}_2$.
10. Calculate the n_0 values $y_{imp} = X_{mis}\dot{\beta} + \dot{z}_2\dot{\sigma}$.

3.3 程序技术

3.3.1 Non-standard evaluation in R

本项目使用了一种在 R 中非常重要的技术，即 Non-standard evaluation，又称 lazy evaluation.[9]

具体来说，让我们看一段重要代码：

```r
### Using lazy evaluation to replicate a func along country list
repli <- function(fun) {
  ex <- substitute(fun)

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

其中的 `substitute`

4 具体流程

4.1 The Workflow

根据 *R for Data Science*[10]，

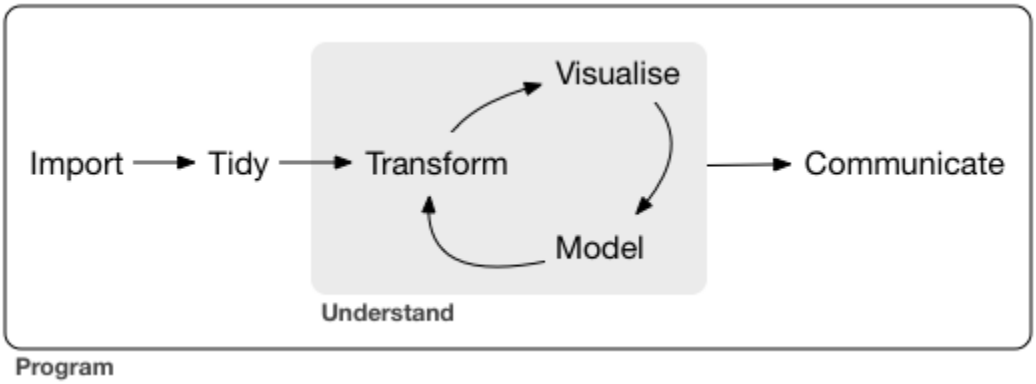

图 3: The Data Science Workflow[2]

4.2 Import

国际贸易数据（`/data/investment`）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，并经过清洗得到可供分析的数据集。

其中，实际利用的外国资本分为直接投资和其他投资方式。直接投资包括中外合资，合资开发和独资企业，外资参股公司以及共同开发；其他投资方式包括补偿贸易和加工组装。数据来源为国家统计局和中华人民共和国商务部。

对外直接投资指中国国内投资者以现金、实物、无形资产等方式在国外及港澳台地区设立、购买国（境）外企业，并以控制该企业的经营管理权为核心的经济活动。数据统计来源为中华人民共和国商务部。

// TODO - **R. Li**

4.3 Tidy

tidy data[11]

```
raw_df <- read_csv("./data/investment/FDI_untidy.csv")


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

[2]This picture is from *R for Data Science*, released under CC BY-NC-ND 3.0 US.

```
    rename(时间 = X1)

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

  stdize <- function(str) {
    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")
  }

  sep_df <- fliped_df %>%
    mutate(observation = observation %>% stdize()) %>%
    separate(col = "observation", into = c("type", " 地区", " 国家"), sep = "/")

  df <- sep_df %>% spread(key = "type", value = "val")
}

raw_df %>%
  process() %>%
  write_csv("./data/investment/FDI_tidy.csv")

cont <- raw_df %>%
  filter(X1 == " 状态") %>%
  as_vector() %>%
  .[. == " 继续"] %>%
  names()
raw_df %>%
  select(X1, all_of(cont)) %>%
  process() %>%
  write_csv("./data/investment/FDI_tidy_cont.csv")
```

```
raw_df <- read_csv(
  file = "./data/investment/FDI_tidy_cont.csv",
  col_types = cols(
    时间 = col_date(format = "%m/%Y")
  ),
  guess_max = 50000
```

```
)

df0 <- raw_df %>%
  filter(!is.na(国家))

# 对外直接投资: 非金融类: 累计 为一带一路数据所特有
OBOR_col <- " 对外直接投资: 非金融类: 累计"

df <- df0 %>%
  filter(国家 != " 一带一路" & 国家 != " 总计") %>%
  select(-all_of(OBOR_col))

df <- df %>%
  filter(month(时间) == 12) %>%
  mutate(年份 = as.integer(year(时间)), .keep = "unused", .before = 1) %>%
  filter(年份 >= 2002)

df <- df %>%
  select(names(df) %>% str_subset(pattern = " 投资 (和其他)*$", negate = TRUE)) %>%
  filter(!is.na(`对外直接投资: 截至累计`))

df %>% write_csv(file = "./data/investment/FDI_useful.csv")

df1 <- df0 %>%
  filter(国家 == " 一带一路" & !is.na(.[OBOR_col])) %>%
  select(时间, all_of(OBOR_col)) %>%
  mutate(
    年份 = as.integer(year(时间)),
    月份 = as.integer(month(时间)),
    .keep = "unused", .before = 1) %>%
  arrange(年份, 月份)

df1 %>% write_csv(file = "./data/investment/FDI_OBOR.csv")
```

4.4 Understand

```
fdi <- read_csv(
  file = "./data/investment/FDI_useful.csv",
```

```
  col_types = cols(
    年份 = col_double(),
    国家 = col_factor()
  )
) %>% unite(col = 国家, 地区, 国家)

country_name <- fdi[[" 国家"]] %>% unique()

fdi_na <- fdi %>%
  tidyr::complete(年份, 国家) %>%
  rename(对外直接投资 = `对外直接投资: 截至累计`)

fdi_lg <- fdi_na %>%
  mutate(lg = log(对外直接投资), .keep = "unused")

fill_a_country <- function(.dt, .cn) {
  res <- .dt %>%
    filter(国家 == .cn) %>%
    mice(method = "norm.boot", m = 1, maxit = 3) %>%
    complete()
  if (any(is.na(res$lg))) {
    non_na <- !(res$lg %>% is.na())
    res$lg <- res$lg[non_na][1]
  }
  return(res)
}

fdi_filled <- country_name %>% map(~fill_a_country(fdi_lg, .x))

result <- fdi_filled %>%
  reduce(rbind) %>%
  mutate(对外直接投资 = exp(lg), .keep = "unused") %>%
  separate(col = 国家, into = c(" 地区", " 国家"), sep = "_")

result %>% write_csv("./data/investment/FDI_filled.csv")
```

4.5 Communicate

本节说明项目中所用到的可视化相关工具、组件、流程。

4.5.1 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)[12] 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。

4.5.2 文件结构

可视化相关的脚本以及输出结果全部储存在./visualization 中。

```
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
├── obor_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
```

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。

`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` 为利用随机森林算法填充部分缺失数据之后的 FDI 图像。

4.5.3 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 pyecharts 图的属性
from pyecharts.charts import Timeline, Map              # 选取 pyecharts 基本类型
from pyecharts.globals import ThemeType                 # 选取 pyecharts 主题
import numpy as np                                      # python 数值计算工具
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 图结构
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对照表

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
for year in range(2003, 2019+1):                        # 循环添加不同年份的数据到 timeline 图中
    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) ",# 设定图层名
             [[ce_dict[row[' 国家']], row[3]]           # 读入数据，使用 dataframe 方法进行筛选
                 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)) # 在地图中不显示对应国家的数值
        .set_global_opts(
            title_opts=opts.TitleOpts(title=f"{year}年外商直接投资情况"), # 设定当前页的标题
            visualmap_opts=opts.VisualMapOpts(
                max_=df[df.iloc[:, 0] == year].iloc[:, 3].max()),    # 重设图例范围
            toolbox_opts=opts.ToolboxOpts(),                         # 打开工具箱组件，便于后续使用鼠标调节


        )
    )
    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
```

5 总结

5.1 建议

5.2 展望

参考文献

[1] BARRETT M. ggdag: Analyze and Create Elegant Directed Acyclic Graphs[M]. 2021.

[2] PEARL J, GLYMOUR M, JEWELL N P. Causal inference in statistics: a primer[M]. Wiley, 2019.

[3] KOWARIK A, TEMPL M. Imputation with the R Package VIM[J]. Journal of Statistical Software, 2016, 74(7): 1–16.

[4] DOUDCHENKO N, IMBENS G W. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis[R]. Working Paper Series, 22791, National Bureau of Economic Research, 2016.

[5] ABADIE A, GARDEAZABAL J. The Economic Costs of Conflict: A Case Study of the Basque Country[J]. The American Economic Review, American Economic Association, 2003, 93(1): 113–132.

[6] CHERNOZHUKOV V, WÜTHRICH K, ZHU Y. An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls[J]. Journal of the American Statistical Association, Taylor & Francis, 2021, 0(ja): 1–44.

[7] VAN BUUREN S, GROOTHUIS-OUDSHOORN K. mice: Multivariate Imputation by Chained Equations in R[J]. Journal of Statistical Software, 2011, 45(3): 1–67.

[8] SCHAFER J L. Analysis of incomplete multivariate data[M]. Chapman & Hall/CRC, 1997.

[9] WICKHAM H. Advanced R[M]. CRC Press, 2019.

[10] WICKHAM H, GROLEMUND G. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data[M]. 第 1 版. Paperback; O'Reilly Media, 2017.

[11] WICKHAM H. Tidy data[J]. The Journal of Statistical Software, 2014, 59(10).

[12] LI D, MEI H, SHEN Y, 等. ECharts: A declarative framework for rapid construction of web-based visualization[J]. Visual Informatics, 2018, 2(2): 136–146.