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

*名拼音序.

1 前言

1.1 概要

1.2 环境

1.2.1 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
```

1.2.2 python info

// TODO

2 主要结果

2.1 数据模型

我们的数据模型如图所示：

此图[1] 是有向无环图 (Directed acyclic graph, DAG)，边代表因果作用[2].

2.2 分析

我们利用 (Chernozhukov et al., 2021)[3] 的方法进行分析.

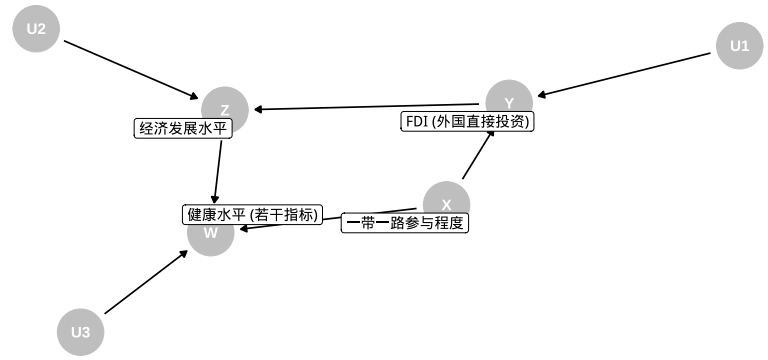

图 1: 数据模型示意图

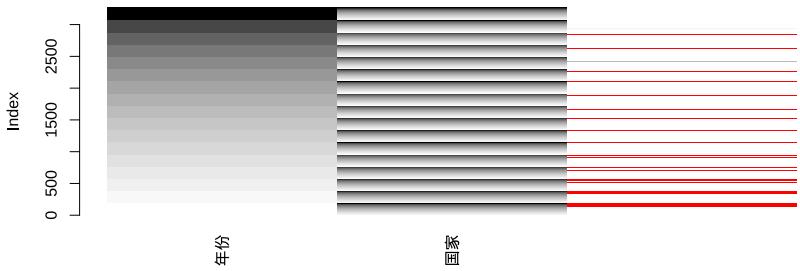

图 2: 缺失数据示意图

首先注意到数据集中存在许多缺失数据[4]，使用 linear regression with bootstrap 进行缺失数据填补.[5]

2.3 程序

2.3.1 Non-standard evaluation in R

3 具体流程

3.1 The Workflow

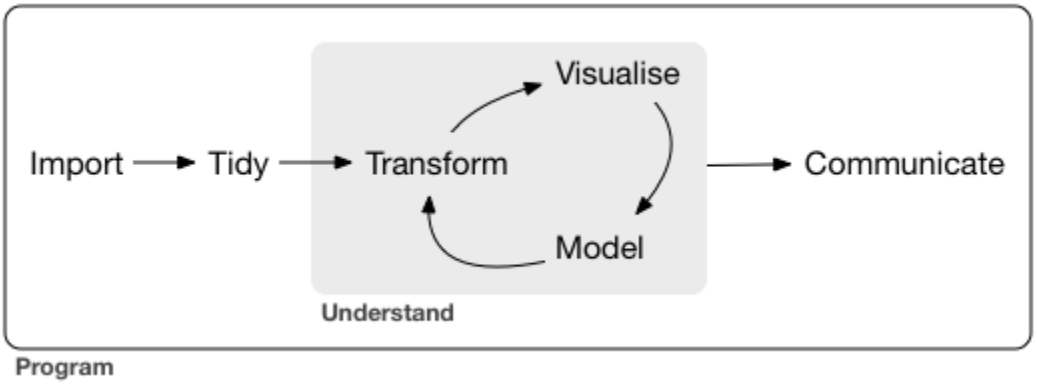

图 3: The Data Science Workflow[1]

3.2 Import

// 需要数据集的完整描述和获取方式

// TODO - **R. Li**

3.3 Tidy

tidy data[7]

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

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

[1]This picture is from *R for Data Science*[6], released under CC BY-NC-ND 3.0 US.

```
  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")
```

3.4 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")
```

3.5 Communicate

本节说明项目中所用到的可视化相关工具、组件、流程。

3.5.1 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)[8] 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。

3.5.2 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。

3.5.3 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
```

4 总结

4.1 建议

4.2 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] 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.

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

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

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

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

[8] 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.