EPS贸易数据分析

胡华平; 向树林

2021-02-18

Table of Contents

# 1. 讨论要点

## 1.1 目前进展

本次讨论

1). 要达到的**目标**：完整地得到**分析数据集**（dataset for analysis），为实证分析做好准备！

定义 1: (分析数据Dataset for Analysis) 当明确好理论模型（实证模型）后，我们往往需要对原始数据进行清洗、整理、计算、变换等数据操作，从而得到一份马上可以进行实证建模的**数据集**，我们称之为**分析数据集**（dataset for analysis）。

2). 已经完成的**工作**：

- 工作进展1：进一步清洗数据  
- 工作进展2：得到分析数据表

## 1.2 文献讨论

暂无

## 1.3 面临问题

讨论提出：

1. 两位数和四位数产品代码的抉择问题。因为跟我们的研究目标有关系。
2. 一带一路政策的冲击，没有说清楚。

## 1.4 下一步工作

清单如下：

1. Rmd使用技巧，进行总结。
2. git操作demo要进一步完善。
3. 文章写作，实证分析。分样本比较（一带一路国家 VS 非一带一路国家）

# 2. 工作进展1：进一步清洗数据

读取已经合并好的数据，并进一步清洗，具体包括：

* 去掉原数据中的**千分位分隔符**（逗号,）。
* 正确变换列类型，例如列value（贸易额）的类型应该为number（数值型）。
* 变换数值单位。例如列value（贸易额）的单位由**美元**变换为**万美元**。
* 筛除不必要的数据行。例如，删除所有**贸易量**行（用不到），以及**贸易额**的无数据行（因为没有产生贸易活动）。
* 加入**国家**的编码等信息。需要用到tbl\_nation的数据表，先处理好匹配需要的key，然后使用left\_jion()函数进行匹配。
* 删除重复行，例如产品0909就存在重复情形（原因可能来自于最开始的数据抓取环节）

# 读取数据  
tbl\_hs2012 <- read\_rds("../data/eps/tbl-hs2012.rds")  
  
tbl\_nation <- read\_rds("../data/eps/list-nation-hs2012.rds") %>%  
 rename("country" = "nation", "country\_code"="code",  
 "country\_id"="index")  
#str(tbl\_nation)  
  
# 转换数据形式  
tbl\_hsclear <- tbl\_hs2012 %>%  
 mutate(value = str\_replace\_all(value, ",", "")) %>%  
 mutate(value = as.numeric(value),  
 value = 0.0001\*value) %>%  
 filter(str\_detect(var, "美元"), !is.na(value)) %>%  
 left\_join(., tbl\_nation, by = "country")%>%  
 arrange(code\_prod, year,month, country\_id) %>%  
 distinct()  
  
#head(tbl\_hsclear)  
#unique(tbl\_hsclear$year)  
#str(tbl\_hsclear)

# 3. 工作进展2：得到分析数据表

## 3.1 计算得到贸易总额及其月度变动数据

### 3.1.1 理论表达

表示**分产品每个月**的加总贸易额，表示滞后1期变量。

表示**分产品**下的**月度贸易变动额**。

### 3.1.2 代码操作

**具体思路**如下：

1. 我们注意到原数据中实际上已经有**分产品**下的**月度贸易总额**。数据。也即country变量下包含有总值类别，因此可以直接使用filter()函数过滤得到**分产品**下的**月度贸易总额**。
2. **分产品**下的**月度贸易变动额**可以先计算得到滞后变量，然后再差分得到贸易变动额（$\Delta X}= X\_{l0} - X\_{l1}$）。具体使用group\_by() + dplyr::lag()函数操作得到滞后变量。

以下为具体R代码操作：

xc <- tbl\_hsclear %>%  
 arrange(code\_prod, year, month) %>%  
 filter(country\_id==1) %>%  
 #unite(col = "ym", year, month, sep = "-", remove = FALSE) %>%  
 select(code\_prod,product, year, month, value) %>%  
 group\_by(product) %>%  
 mutate(x\_l0 = value,   
 x\_l1 = dplyr::lag(value,n=1)) %>%  
 ungroup() %>%  
 mutate(delta\_x = x\_l0 - x\_l1)

### 3.1.3 过程解释

下面表1简单展示了得到的数据表结果：

表 1: （其中5类产品）分产品的月度贸易额及变动（单位：万美元）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| index | code\_prod | product | year | month | x\_l0 | x\_l1 | delta\_x |
| 344 | 0201 | 0201 - 鲜、冷牛肉 | 2016 | 09 | 577 | 456 | 121 |
| 345 | 0201 | 0201 - 鲜、冷牛肉 | 2016 | 10 | 537 | 577 | -40 |
| 346 | 0201 | 0201 - 鲜、冷牛肉 | 2016 | 11 | 557 | 537 | 19 |
| 347 | 0201 | 0201 - 鲜、冷牛肉 | 2016 | 12 | 726 | 557 | 169 |
| 4323 | 1001 | 1001 - 小麦及混合麦 | 2016 | 09 | 10214 | 8731 | 1483 |
| 4324 | 1001 | 1001 - 小麦及混合麦 | 2016 | 10 | 6029 | 10214 | -4185 |
| 4325 | 1001 | 1001 - 小麦及混合麦 | 2016 | 11 | 628 | 6029 | -5401 |
| 4326 | 1001 | 1001 - 小麦及混合麦 | 2016 | 12 | 4743 | 628 | 4115 |
| 4509 | 1005 | 1005 - 玉米 | 2016 | 09 | 516 | 693 | -177 |
| 4510 | 1005 | 1005 - 玉米 | 2016 | 10 | 530 | 516 | 14 |
| 4511 | 1005 | 1005 - 玉米 | 2016 | 11 | 1152 | 530 | 622 |
| 4512 | 1005 | 1005 - 玉米 | 2016 | 12 | 3873 | 1152 | 2721 |
| 4569 | 1006 | 1006 - 稻谷、大米 | 2016 | 09 | 10111 | 8702 | 1409 |
| 4570 | 1006 | 1006 - 稻谷、大米 | 2016 | 10 | 9517 | 10111 | -594 |
| 4571 | 1006 | 1006 - 稻谷、大米 | 2016 | 11 | 14275 | 9517 | 4759 |
| 4572 | 1006 | 1006 - 稻谷、大米 | 2016 | 12 | 19960 | 14275 | 5684 |
| 5256 | 1201 | 1201 - 大豆，不论是否破碎 | 2016 | 09 | 307852 | 329275 | -21423 |
| 5257 | 1201 | 1201 - 大豆，不论是否破碎 | 2016 | 10 | 224872 | 307852 | -82980 |
| 5258 | 1201 | 1201 - 大豆，不论是否破碎 | 2016 | 11 | 338230 | 224872 | 113358 |
| 5259 | 1201 | 1201 - 大豆，不论是否破碎 | 2016 | 12 | 392724 | 338230 | 54494 |

## 3.2 计算得到贸易变动分解数据表

### 3.2.1 理论表达

其中：

* 表示由于新国家加入（Newcoming）贸易（与上月相比），引发的贸易额变动:
* 表示由于原国家退出（Exit）贸易（与上月相比），引发的贸易额变动：
* 表示由于**继续留存**国家**贸易量**增加（Growth）（与上月相比），引发的贸易额变动：
* 表示由于**继续留存**国家**贸易量**缩减（Shrink）（与上月相比），引发的贸易额变动：

至此，我们还可以分别计算上述4种情形下的：贸易变动()、国家名单()、国家数量()、贸易密度()，具体变量定义为：

1. **新进入**情形(Newcoming)：贸易变动()、国家名单、国家数量，贸易密度。
2. **退出**情形(Exit)：贸易变动()、国家名单、国家数量，贸易密度。
3. 继续留存下业务量**增长**情形(Growth)：贸易变动()、国家名单、国家数量，贸易密度。
4. 继续留存下业务量**缩减**情形(Shrink)：贸易变动()、国家名单、国家数量，贸易密度。

### 3.2.2 代码操作

**基本思路**：

* 实现新的数据形式变换（按产品分年月，“折叠”行数据）。利用group\_by() + nest()函数，构造出list-column data.frame。
* 构造**当期**的数据块和**滞后期**的数据块。具体要用到lag()函数。

要注意对初始项的处理，可以是“保持”、“NA”。我们选择了“保持”。

* 比较**当期**数据块和**滞后期**数据块的差异，分别计算得到**进入或退出**的国家名单、国家数量、贸易密度、贸易额变动。

具体要用到平行计算包purrr::map2()函数，同时还需要编写专门的函数compare.tbls()函数来得到自己想要的上述几个变量。

具体的R代码如下：

注意：因为数据集比较大，而且设计到多个计算，尽管purrr::map2()已经进行了平行化运算，但是还是会耗费一定计算时间，大约1-2分钟。因此，**强烈建议**一次计算，然后保留结果。

首先，编写定制函数compare.tbls()：

#dt2 <- tbl\_compare$dt\_l0[1][[1]]  
#dt1 <- tbl\_compare$dt\_l1[1][[1]] %>%  
 #mutate(country=c(LETTERS[1:5]))  
  
# build function for map2   
## it can compare difference with two data.frame, also bypass some calculate  
compare.tbls <- function(dt1, dt2){  
 # new coming country  
 c\_new <- setdiff(dt2$country, dt1$country)  
 n\_new <- length(c\_new)  
 v\_new <- dt2 %>%  
 filter(country %in% c\_new) %>%  
 summarise(value = sum(value, na.rm = T)) %>%  
 unlist()  
 d\_new <- ifelse(n\_new!=0, v\_new/n\_new, NA)  
   
 # exit country  
 c\_exit <- setdiff(dt1$country,dt2$country )  
 n\_exit <- length(c\_exit)  
 v\_exit <- dt1 %>%  
 filter(country %in% c\_exit) %>%  
 summarise(value = sum(value, na.rm = T)) %>%  
 unlist()  
 d\_exit <- ifelse(n\_exit!=0, v\_exit/n\_exit, NA)  
   
 # comon country  
 common\_c <- intersect(dt2$country, dt1$country)  
 dt2\_com <- dt2 %>%  
 filter(country %in% common\_c) %>%  
 add\_column(period = "T2", .before = "country")  
 dt1\_com <- dt1 %>%  
 filter(country %in% common\_c)%>%  
 add\_column(period = "T1", .before = "country")  
 ### case if no common country  
 dt\_null <- data.frame(matrix(rep(NA,6),nrow = 1 ))   
 names(dt\_null) <- c('country','country\_id','T1','T2','status','change')  
 #paste0(paste0("'",names(dt\_com) ,"'"), collapse = ",")  
 if (length(common\_c)==0) {  
 dt\_com <- dt\_null %>%  
 filter(!is.na(country))  
 } else{  
 dt\_com <- bind\_rows(dt2\_com, dt1\_com) %>%  
 spread(key = period, value = value) %>%  
 mutate(status = ifelse(T2<T1, "S", "G"),  
 change = T2-T1)  
 }  
   
 c\_com <- unlist(dt\_com$country)  
 n\_com <- length(c\_com)   
   
 # for grow country  
 dt2\_grow <- dt\_com %>%  
 filter(status=="G")  
 c\_grow <- unlist(dt2\_grow$country)  
 n\_grow <- length(c\_grow)  
 v\_grow <- dt2\_grow %>%  
 summarise(value = sum(change, na.rm = T)) %>%  
 unlist()  
 d\_grow <- ifelse(n\_grow!=0, v\_grow/n\_grow, NA)  
   
 # for shrink country  
 dt2\_shrink <- dt\_com %>%  
 filter(status=="S")  
 c\_shrink <- unlist(dt2\_shrink$country)  
 n\_shrink <- length(c\_shrink)  
 v\_shrink <- dt2\_shrink %>%  
 summarise(value = sum(change, na.rm = T)) %>%  
 unlist()  
 d\_shrink <- ifelse(n\_shrink!=0, v\_shrink/n\_shrink, NA)  
   
 # all results  
 out<- data.frame(c\_new = I(list(c\_new)), n\_new = n\_new,   
 v\_new =v\_new, d\_new = d\_new,  
 c\_exit = I(list(c\_exit)), n\_exit = n\_exit,   
 v\_exit =v\_exit, d\_exit = d\_exit,  
 c\_com = I(list(c\_com)), n\_com = n\_com,  
 c\_grow = I(list(c\_grow)), n\_grow = n\_grow,   
 v\_grow =v\_grow, d\_grow = d\_grow,  
 c\_shrink = I(list(c\_shrink)), n\_shrink = n\_shrink,   
 v\_shrink =v\_shrink, d\_shrink = d\_shrink)  
 return(out)  
}

其次，其余的数据**变换**、**计算**和**导出**过程如下：

# nest dataset and then compare difference  
require(purrr)  
rm\_list <- c("var", "ym", "country\_code")  
tbl\_compare <- tbl\_hsclear %>%  
 select(-all\_of(rm\_list)) %>%  
 filter(country\_id!=1) %>%  
 group\_by(code\_prod, product, year, month) %>%  
 nest() %>%  
 ungroup() %>%  
 rename("dt\_l0"="data") %>%  
 group\_by(code\_prod) %>%  
 mutate(dt\_l1 = lag(dt\_l0,  
 default = list(first(dt\_l0)))) %>%  
 # purrr with custom function compare.tbls  
 mutate(compare = map2(.x = dt\_l1, .y = dt\_l0, .f = compare.tbls))  
  
write\_rds(tbl\_compare, here("data/eps/tbl-compare-newcoming-exit-01-raw.rds"))   
  
# unnest data table  
tbl\_ne <- tbl\_compare %>%  
 select(-dt\_l0, -dt\_l1) %>%  
 unnest(cols = compare)  
  
write\_rds(tbl\_ne, here("data/eps/tbl-compare-newcoming-exit-02-unnest.rds"))

### 3.2.3 过程解释

#### （1）最开始的数据集

首先，下表2给出了最开始清洗好的数据集tbl\_hsclear，为了演示方便，我们选择**HS编码**为1005的农产品（玉米），年月范围为2012/01-2020/03

表 2: 步骤1 2012/01-2020/03期间玉米产品月度贸易数据（单位：万美元）

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| code\_prod | product | year | month | country | country\_id | value |
| 1005 | 1005 - 玉米 | 2012 | 01 | 印度 | 12 | 0.76 |
| 1005 | 1005 - 玉米 | 2012 | 01 | 老挝 | 20 | 19.68 |
| 1005 | 1005 - 玉米 | 2012 | 01 | 德国 | 108 | 170.29 |
| 1005 | 1005 - 玉米 | 2012 | 01 | 智利 | 161 | 0.04 |
| 1005 | 1005 - 玉米 | 2012 | 01 | 秘鲁 | 182 | 7.45 |
| 1005 | 1005 - 玉米 | 2012 | 01 | 美国 | 197 | 24861.96 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 印度 | 12 | 27.00 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 老挝 | 20 | 120.50 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 德国 | 108 | 162.65 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 法国 | 109 | 0.65 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 奥地利 | 119 | 0.13 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 阿根廷 | 153 | 1.29 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 巴西 | 159 | 0.39 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 秘鲁 | 182 | 4.21 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 美国 | 197 | 17108.77 |
| 1005 | 1005 - 玉米 | 2012 | 02 | 澳大利亚 | 200 | 5.40 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 缅甸 | 7 | 5.13 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 老挝 | 20 | 42.04 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 德国 | 108 | 96.21 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 法国 | 109 | 11.45 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 俄罗斯联邦 | 142 | 2.10 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 阿根廷 | 153 | 1.22 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 智利 | 161 | 0.07 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 圭亚那 | 173 | 0.16 |
| 1005 | 1005 - 玉米 | 2012 | 03 | 美国 | 197 | 15196.05 |

#### （2）分年月按产品“折叠”后的数据集

简要信息如下：

tbl\_compare <- read\_rds(here("data/eps/tbl-compare-newcoming-exit-01-raw.rds"))  
  
#dplyr::glimpse(tbl\_compare)

分年月按产品“折叠”后的数据集，详细的数据集形态见图1

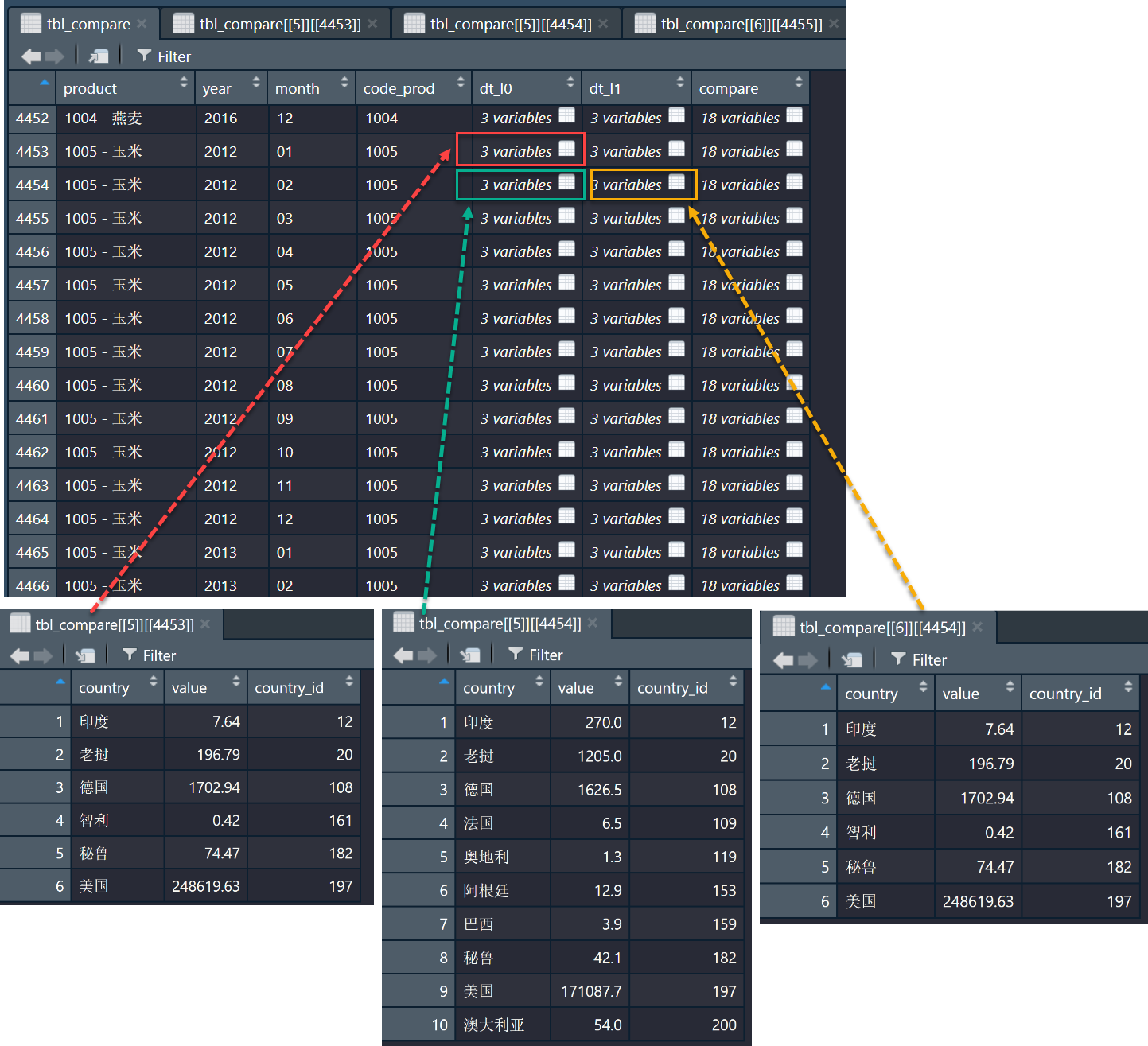


图 1: 步骤2 数据“折叠”及形态关系

#### （3）进行月度比较得到的变动分解数据集

首先，我们需要：

* 读取数据文件.rds
* 去掉分组属性ungroup()
* 将0全部替换为NA。具体需要用到dplyr::na\_if()函数

tbl\_ne <- read\_rds(here("data/eps/tbl-compare-newcoming-exit-02-unnest.rds")) %>%  
 ungroup() %>%  
 mutate\_if(is.numeric, dplyr::na\_if, y=0)

下表3中，我们将看到对年月贸易变动的计算和比较数据集示例：

表 3: 步骤3 贸易变动计算

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| product | year | month | code\_prod | n\_new | v\_new | d\_new | n\_exit | v\_exit | d\_exit | n\_com | n\_grow | v\_grow | d\_grow | n\_shrink | v\_shrink | d\_shrink |
| 1005 - 玉米 | 2012 | 01 | 1005 |  |  |  |  |  |  | 6 | 6 |  |  |  |  |  |
| 1005 - 玉米 | 2012 | 02 | 1005 | 5 | 79 | 16 | 1 | 0.42 | 0.42 | 5 | 2 | 1271 | 635 | 3 | -77641 | -25880 |
| 1005 - 玉米 | 2012 | 03 | 1005 | 4 | 75 | 19 | 5 | 371.34 | 74.27 | 5 | 1 | 108 | 108 | 4 | -20577 | -5144 |

## 3.3 整合为完整的**分析数据集**

会上操作。