《国际关系定量分析基础》2020 秋季

第一次小组作业 (共计 100 分)

陈道想　　黄卓尔　　潘明花　　杨霖　　赵佳鹏

截止时间：**2020 年 10 月 12 日 11：59 am**

东南亚地区 (如图 1) 是国际关系和比较政治学界关注的重点地区，本次作业将利用公开数据，对东南亚地区国家的政治、经济、社会、外交等关系进行描述。

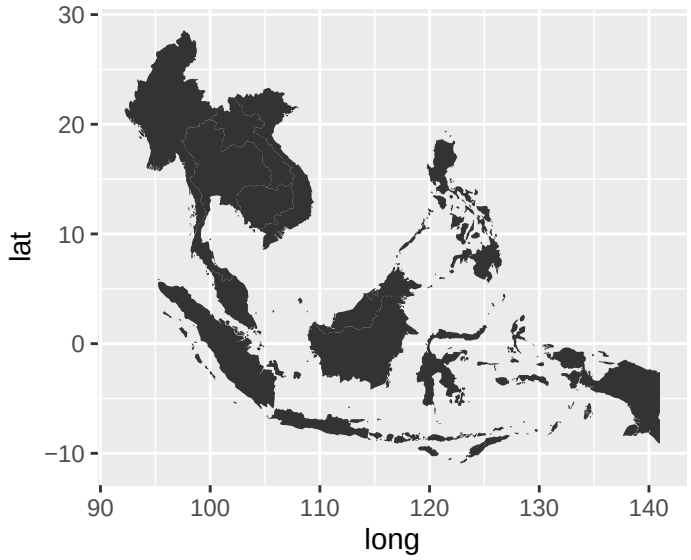

图 1: 东南亚地图

注意事项：

- 小组作业截止时间：2020 年 10 月 12 日 11：59 am
- 请直接在 R Markdown 文件中完成本次作业
- 作业在网络学堂提交，每个小组仅需提交一份
- 提交作业的文件名需以 `HW-1-Team-X.Rmd` 和 `HW-1-Team-X.pdf`(或者 `HW-1-Team-X.html`), 请将 `X` 替换为小组编号，如 `HW-1-Team-A.Rmd` 和 `HW-1-Team-A.pdf`。(若 R Markdown 出现无法 `knit` 为 `pdf` 情况，可以使用 `bookdown::html_document2:`，则会生成为 `html`)

- 请显示每道题的 R Code 于 pdf 中，注重 Code 的整洁性和可读性，可参考Google's R Style Guide
- 本次作业所需的数据和 R Packages 已经提供。本次作业需要的数据可以通过以下命令获取 (或直接 `load("terrorism.RData")`):

```
load(url("https://cc458.github.io/files/terrorism.RData"))
load(url("https://cc458.github.io/files/conflict.RData"))
```

其中，

- `terrorism.RData` 包括三个数据集: `region_map_shp`, `gtd_region`, `wdi`;
- `conflict.RData` 包括四个数据集 `polity`, `acled_cnty`, `ideal_point_wide`, `ucdp_cnty`

R 基础问题（共 15 分）

1.(**10 分**) 请用 `knitr::kable` 创建一个 8*3 的表格，总结这 7 个数据框。表格除第一行为 header 外，其余每一行表示一个数据框；除第一列为数据框的名称外，其余两列分别为每一个数据框的变量 (variables) 和观测量（observations）数。（提示：可以先创建一个包含这些信息的新数据框，然后再使用 `kable` 创建表格；也可以利用 R Markdown 手动创建）。

```
library(knitr)

# 获取数据维度
num <- matrix(nrow = 7, ncol = 2)
num[1,] <- dim(region_map_shp)
num[2,] <- dim(gtd_region)
num[3,] <- dim(wdi)
num[4,] <- dim(polity)
num[5,] <- dim(acled_cnty)
num[6,] <- dim(ideal_point_wide)
num[7,] <- dim(ucdp_cnty)

# 创建一个数据新的数据集
df <- data.frame(dataset = c("region_map_shp","gtd_region","wdi",
                             "polity","acled_cnty","ideal_point_wide",
                             "ucdp_cnty"),
                 variable_num = c(num[,1]),
                 observations_num = c(num[,2]))
```

表 1: 数据集信息

dataset	variable_num	observations_num
region_map_shp	17769	18
gtd_region	9585	136
wdi	319	10
polity	241	4
acled_cnty	107	4
ideal_point_wide	1664	5
ucdp_cnty	17	5

```
# 利用 kable 产生一个新的表格
kable(df, caption = " 数据集信息")
```

2.(**5 分**)stargazer 是政治学常用的产生统计表格的软件包，请利用 stargazer 提供一个关于 wdi 数据的描述性统计表格。

```
# 加载软件包
library(stargazer)

# 使用 stargazer 命令产生新的描述性表格
stargazer(wdi, header=FALSE, type='latex', title = " 描述性统计结果",
                      digit.separator = "")
```

表 2: 描述性统计结果

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
year	319	2004.000	8.380	1990	1997	2011	2018
gdppc	296	7568.071	13061.030	95.188	715.872	6037.739	66188.780
gdpgrowth	304	3.707	4.478	−37.002	2.232	5.885	14.173
fdi	292	5.078	5.349	−2.757	1.861	6.110	28.598
gender	134	0.639	0.381	0.000	0.333	1.000	1.000
milexp	283	2.310	1.558	0.190	1.265	3.083	8.675
poverty	82	14.833	17.348	0.000	0.875	22.775	66.700
pop	319	50243727.000	63772880.000	258721	4677895.0	69610521	267663435

数据可视化问题（共 85 分）

3.(**10 分**) 数据 region_map_shp 是一个包含空间信息的数据集，其中变量 terratck 包含了东南亚 1991-2006 年之间所遭遇的恐怖主义袭击数量总和。同时数据集 gtd_region 记录了东南亚 1991-2006 年每一次恐怖袭击的经纬度地理位置（变量 longitude 和 latitude)。请利用 ggplot2 这一软件包产生如下地图 (图 2)，描述各国在此期间的恐怖主义数量分布以及被袭击的地点，并简要描述你对关于东南亚恐怖袭击活动地理分布的观察。

```
ggplot() +
# 设置各国地图形状
  geom_polygon(data = region_map_shp,
               aes(x = long,
                   y = lat,
                   group = group,
                   fill = terratck),
               size = 0.25) +
  coord_fixed() +
# 设置各国地图颜色填充规则
  scale_fill_manual(values = c("#009E73", "#F0E442", "#0072B2","#D55E00", "#CC79A7"),
                    name = "Total number of attacks",
                    na.value = "gray",
                    labels = c("1 - 10","10 - 100","100 - 500","500 - 1000",
                               ">2000", "NA")) +
# 将各次袭击地点加至地图上并设置颜色填充规则
  geom_point(aes(x = gtd_region$longitude,
                 y = gtd_region$latitude,
                 color = "black"),
             alpha=.5, size = 0.5, na.rm= FALSE) +
  scale_color_manual(name = "",
                     labels = c("Attack locations"),
                     values = "black")  +
# 增加横纵坐标
  labs(y = "latitude", x = "longitude")
```

我们观察发现，东南亚地区恐怖袭击活动地理分布不均。具体而言，呈现出以下几个主要的特点：

- 菲律宾是发生恐怖袭击活动数量最多的国家；
- 文莱、新加坡、东帝汶、越南发生恐怖袭击活动的数量较少；

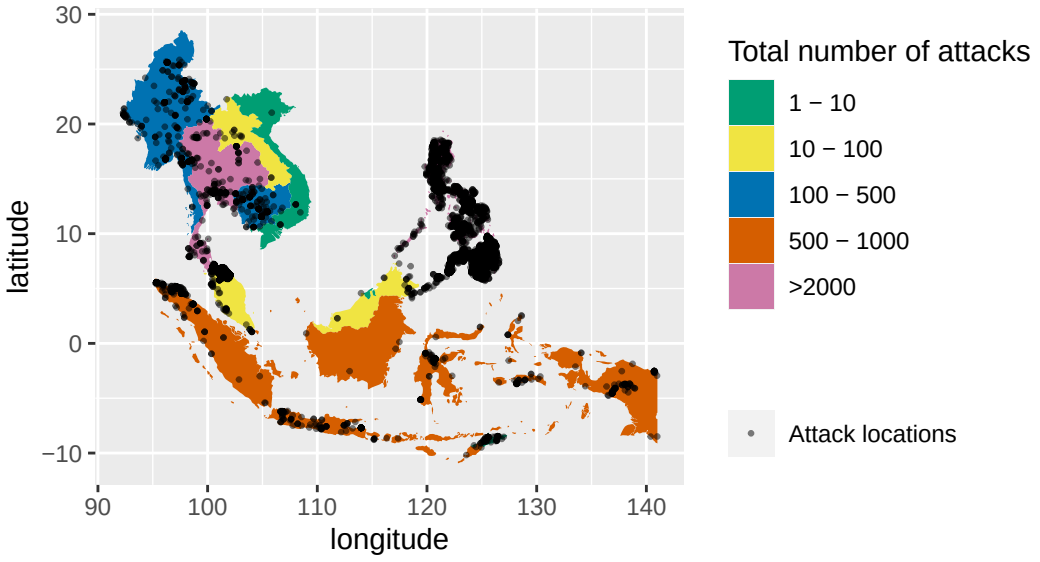

图 2: 东南亚恐怖袭击数量分布

- 在缅甸、泰国、柬埔寨等国，恐怖袭击活动地理分布较其它国家均匀；
- 在马来西亚，恐怖袭击活动大部分集中于西马来西亚北部；而在印度尼西亚，恐怖袭击活动大部分集中于苏门答腊岛西北部、爪哇岛一带、苏拉威西岛中部与塞兰岛一带。

4.(**10 分**) 数据集 `wdi` 中的变量 `gdpgrowth` 记录了东南亚各国在 1990-2018 年之间的国民生产总值增长率。请利用 `ggplot2` 绘制各国在此时间段内的国民生产总值增长率随时间而变化的折线图（图 3）。并据此图简要谈谈 1997-1998 年亚洲金融危机对东南亚国家经济增长的冲击（提示：可以使用 `facet_wrap` 分别绘制, 也可以绘制到同一张图）

```
# 初始化作图数据集与横纵坐标变量
ggplot(wdi,aes(x = year,y = gdpgrowth)) +
  geom_point(colour = "pink", size = 2) +
  geom_line(colour = "black") +
# 增加 1997 与 1998 年灰色竖线以方便观察
  geom_vline(xintercept = 1997, color = "grey") +
  geom_vline(xintercept = 1998, color = "grey") +
  labs(x = "Year", y = "GDP per capita growth (annual %)",
       caption ="Source: World Bank Group 2020") +
  scale_x_continuous(breaks = seq(1990, 2018, 10)) +
# 分别绘制各国折线图并排列齐整，设定纵坐标不统一
  facet_wrap(vars(country),
             nrow = 4,
             ncol = 3,
```

```
            scales = "free") +
  theme(strip.text.x = element_text(size = 10, color='white',angle=0),
        strip.background = element_rect(fill = "#525252", color='#525252'))
```

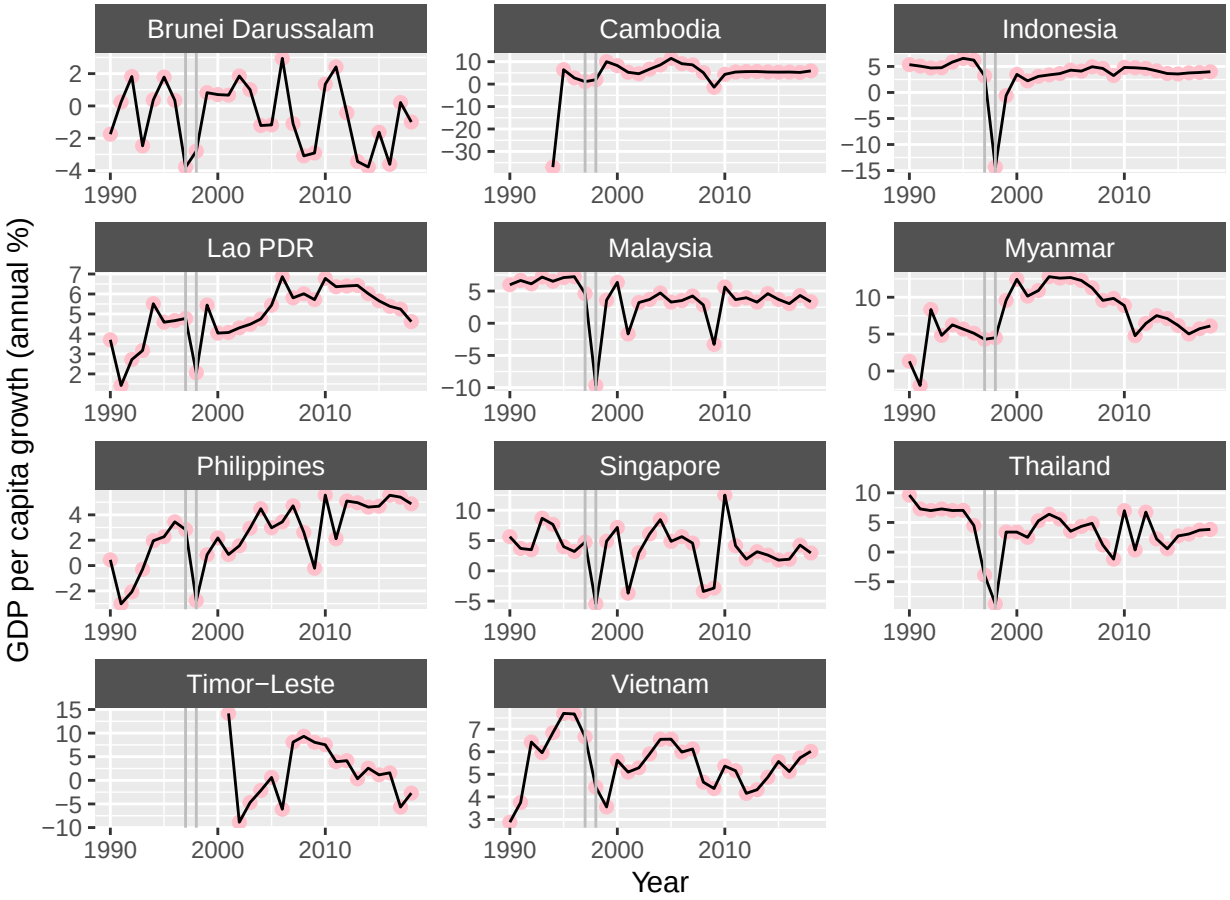

图 3: 东南亚各国国民生产总值增长率变化（1990-2018）

我们观察发现，1997-1998 年亚洲金融危机对东南亚国家经济增长的冲击是不一的。具体而言，呈现出以下几个主要的特点。

- 柬埔寨、缅甸两国的 GDP 增长率在危机前后没有明显变化，说明危机对这两个国家的 GDP 增长基本没有影响。
- 文莱、印尼、老挝、马来西亚、菲律宾、新加坡、泰国、越南等国的 GDP 增长率在危机到来后出现了不同程度的下降：
  - 印尼、马来西亚、新加坡等国的经济增速跌幅均超过 10%；
  - 文莱、菲律宾、老挝、越南等国的经济增速跌幅较小，且后两者在危机中仍然维持了经济正增长；

  - 危机对上述各国经济增速的负面影响维持了 1 到 2 年左右不等。
- 东帝汶在亚洲金融危机期间仍然被印尼所占领，但金融危机对印尼的冲击间接促成了东帝汶在 1999 年通过全民公决独立。

5.(**10 分**) 数据集 `wdi` 中 `gdppc` 表示人均国民生产总值 (GDP per capita，以 2018 年美元为单位)，变量 `milexp` 表示军费开支占国民生产总值的比值。请利用 `ggplot2` 描述 `gdppc` 与 `milexp` 之间的关系，并讨论你是否发现什么规律。

```
# 初始化作图数据集与横纵坐标变量
ggplot(wdi, aes(x = gdppc,y = milexp)) +
    geom_point() +
# 在整个横坐标范围上进行回归
    geom_smooth(colour = "yellow") +
    geom_smooth(method = "lm",
                colour = "springgreen4") +
# 考虑到数据以 11500 美元为界出现明显的差异，进行分段回归
    geom_vline(xintercept = 11500, color = "grey") +
    geom_smooth(data = subset(wdi, gdppc >= 11500),
                colour = "royalblue") +
    geom_smooth(data = subset(wdi, gdppc < 11500),
                colour = "tomato") +
    labs(x = "GDP per capita",
         y = "Ratio of military expenditure to GDP (%)",
         caption = "Source: World Bank Group 2020")
```

```
# 初始化作图数据集与横纵坐标变量
ggplot(wdi, aes(x = gdppc,y = milexp)) +
    geom_point() +
    geom_smooth(aes(color = country)) +
    labs(x = "GDP per capita",
         y = "Ratio of military expenditure to GDP (%)",
         caption = "Source: World Bank Group 2020") +
# 分别绘制各国折线图并排列齐整，设定横纵坐标不统一
    facet_wrap(vars(country),
         nrow = 4,
         ncol = 3,
         scales = "free") +
    theme(strip.text.x = element_text(size = 12, color='white',angle=0),
```

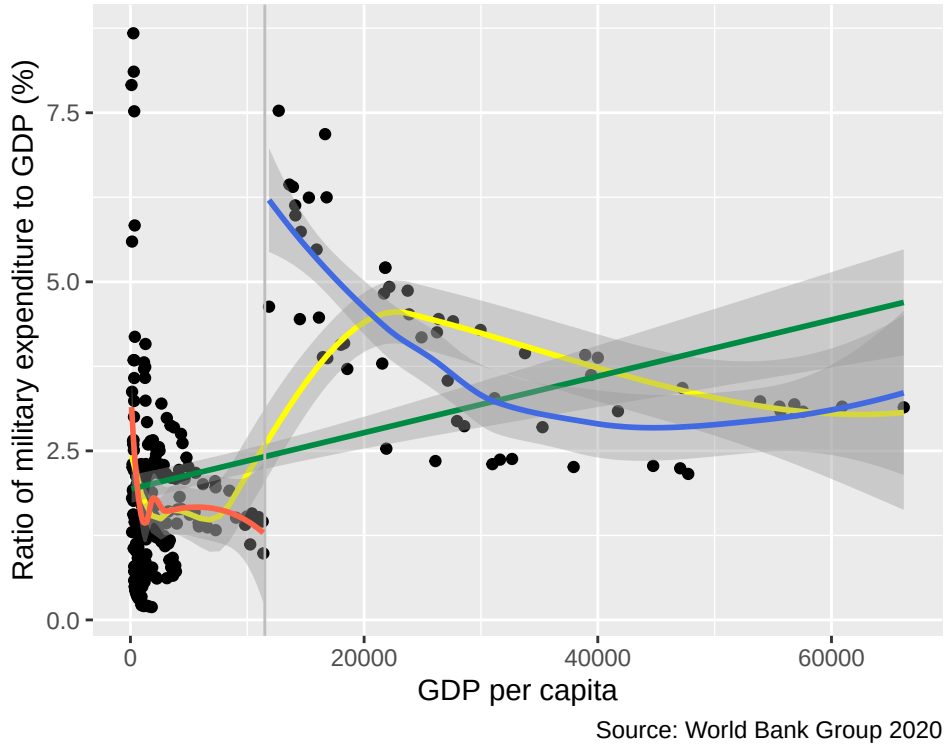

图 4: 人均 GDP 与军费开支占 GDP 比值的关系（东南亚地区）

```
        legend.position = "none",
        strip.background = element_rect(fill = "#525252", color='#525252'))
```

我们注意到，在所定的时间范围内，文莱与新加坡的人均 GDP 均高于 11500 美元，而其它国家的人均 GDP 均小于 11500 美元。为了分别体现它们的规律，我们以 11500 美元为界在图 4 中进行了分段回归。同时，我们也在整个横坐标范围上进行了回归。

另外，我们在图 5 中描绘了东南亚各国 `gdppc` 与 `milexp` 的关系。观察发现，`gdppc` 与 `milexp` 之间存在如下关系和规律。

- 由图 5 可见，对于大部分国家而言，`gdppc` 与 `milexp` 呈负相关关系，即军费开支占 GDP 的比值随着人均 GDP 的增长而降低。
- 由图 4 可见，若将东南亚国家作为一个整体看待，随着人均 GDP 的增加，军费开支占 GDP 的比值呈现先降低，后升高，再降低的趋势。整体上看，二者呈正相关关系。
  - 当人均 GDP 低于 7500 美元左右时，二者大体呈负相关关系；
  - 当人均 GDP 高于 22500 美元左右时，二者亦大体呈负相关关系；
  - 当人均 GDP 高于 7500 美元左右但低于 22500 美元左右时，二者大体呈正相关关系。这一部分的正相关关系对整体正相关关系贡献很大，可由下面一点得到解释。
- 由图 4 可见，若以 11500 美元的人均 GDP 为界将东南亚国家分为两类看待，则每类国家内

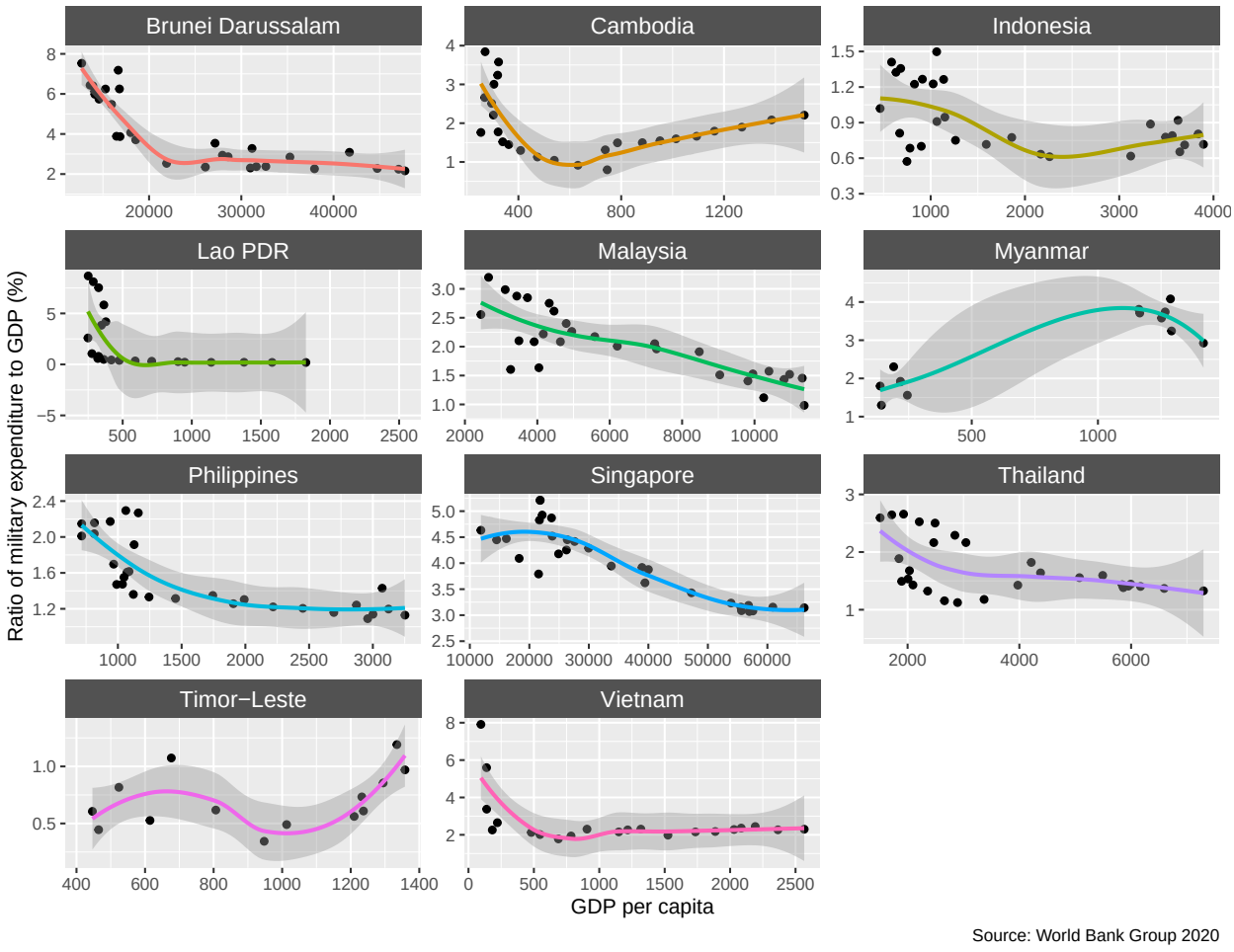

图 5: 人均 GDP 与军费开支占 GDP 比值的关系（东南亚各国）

部的 gdppc 与 milexp 呈负相关关系。这可能暗示着 11500 美元左右的人均 GDP 是区分拥有不同 gdppc 与 milexp 水平国家的界限。

6.(**10 分**) 数据集 wdi 中 fdi 表示当年外国直接投资（净流入）占国民生产总值的比例。请 ggplot2 及其 facet_wrap 命令，描述东南亚各国在 1990-2018 年之间外国直接投资的变化情况，并据此简要讨论你观察到何种模式和规律。(提示：参考第 4 题)

```
# 初始化作图数据集与横纵坐标变量
ggplot(wdi,aes(x = year,y = fdi)) +
  geom_point(colour = "pink", size = 2) +
  geom_line(colour = "black") +
  labs(x = "Year", y = "Ratio of FDI (net inflow) to GDP (%)",
       caption ="Source: World Bank Group 2020") +
  scale_x_continuous(breaks = seq(1990, 2018, 10)) +
# 分别绘制各国折线图并排列齐整，设定横纵坐标不统一
  facet_wrap(vars(country),
             nrow = 4,
             ncol = 3,
             scales = "free") +
  theme(strip.text.x = element_text(size = 10, color='white',angle=0),
        strip.background = element_rect(fill = "#525252", color='#525252'))
```

我们观察发现，东南亚各国 FDI 的变化模式不一。我们将各国大致分为如下几类。

- 高水平增长：新加坡
  - 新加坡 FDI 占 GDP 比例多年维持在 10% 以上，且在图示时间段内整体呈现增长趋势。
- 出现尖峰后走低并稳定：文莱、马来西亚、泰国
  - 这三国在图示时间段内均出现了一个高峰值，且在高峰值出现后 FDI 占 GDP 比例大致呈现走低的趋势并趋于稳定。
- 出现多个尖峰：柬埔寨、老挝、菲律宾、越南
  - 这四国在图示时间段内均出现了多个高峰值，但高峰值之后 FDI 占 GDP 比例的变化趋势不一。
- 出现低于 0 的负值：印度尼西亚
  - 1998-2001 年，印度尼西亚的 FDI 占 GDP 比例出现了连续负值，说明 FDI 产生了净流出。这可能是由 1997-1998 年亚洲金融危机所导致的。
- 东帝汶与缅甸的观测数据较其他国家少，因而暂时难以描述它们 FDI 的变化模式。不过图上可见，这两国在图示时间段内均出现了一个高峰值。

7.(**10 分**) 请利用可视化方法，简要描述并讨论数据集 wdi 中 fdi, gdppc, gdpgrowth 和 milexp

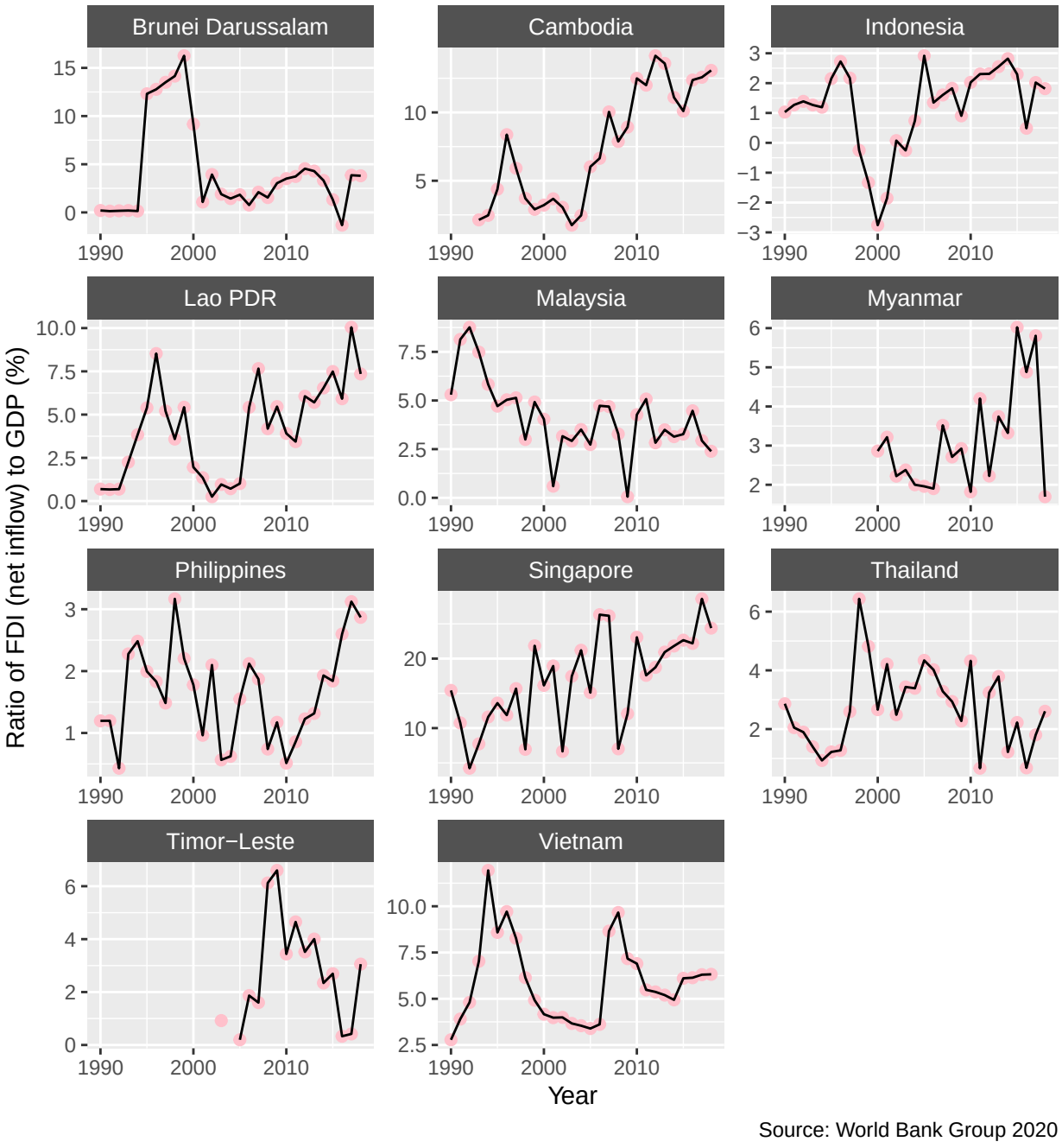

图 6: 东南亚各国 FDI（净流入）占 GDP 比例的变化（1990-2018）

这四个变量的相关关系。

```
library(GGally)
# 选定需要进行相关性分析的变量并作简图
# 相关系数小数位数可调整
ggcorr(wdi[, c("fdi", "gdppc", "gdpgrowth", "milexp")],
       palette = "RdBu", label = TRUE, name = "correlation", label_round = 3)
```

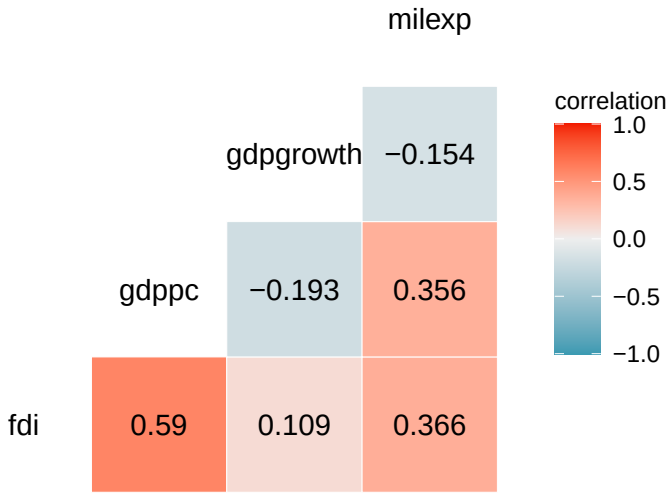

图 7: 数据集 wdi 中 fdi, gdppc, gdpgrowth 和 milexp 四个变量的相关关系简图

```
# 选定需要进行相关性分析的变量并作详图
ggpairs(wdi[, c("fdi", "gdppc", "gdpgrowth", "milexp")])
```

我们观察发现，`fdi`, `gdppc`, `gdpgrowth` 和 `milexp` 这四个变量的相关关系如下。

- `gdpgrowth` 和 `gdppc`，`gdpgrowth` 和 `milexp`，`gdpgrowth` 和 `fdi` 三对变量相关系数分别为-0.193，-0.154 与 0.109，可以认为它们不相关。
- 剩余的各对变量均呈正相关关系。`fdi` 和 `gdppc`，`fdi` 和 `milexp`，`gdppc` 和 `milexp` 之间的相关系数分别为 0.590，0.366 与 0.356。
  - FDI 一般能够拉动经济增长。
  - `gdppc` 和 `milexp` 呈正相关关系已由图 4 说明。

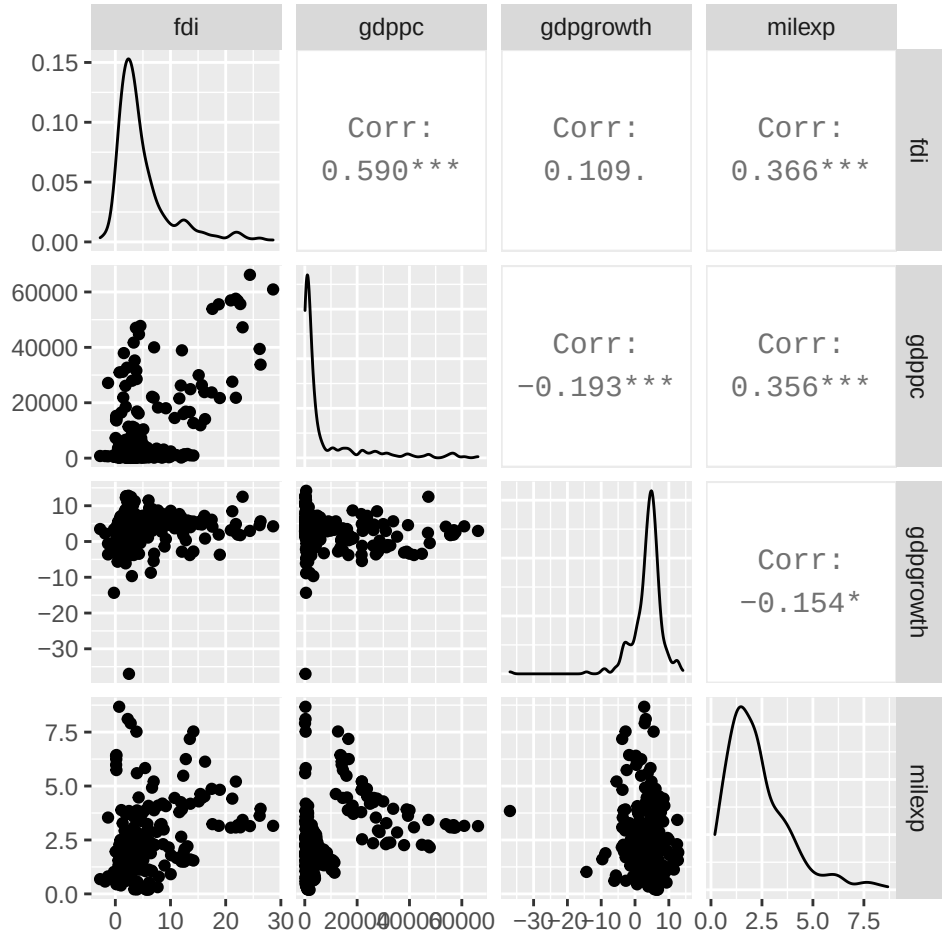

图 8: 数据集 wdi 中 fdi, gdppc, gdpgrowth 和 milexp 四个变量的相关关系详图

注：各对变量相关系数的显著性参见图 8 。

8.(**10 分**) 数据集 `polity` 是国际关系中最常见的用来测量国家整体类型的数据。其中的变量 `polity2` 的值域为 [-10, +10]，即 “最不民主”(-10) 到 “最民主”(+10)。请利用 `ggplot` 简要描述这一变量的分布情况。

```
# 可以使用 facet_wrap 函数分国家进行展示
# 但这可能导致规律与模式不够明显
# 我们选用条形图对东南亚地区的民主指数进行展示
ggplot(polity, mapping = aes(x = factor(polity2))) +
  geom_bar() +
  scale_x_discrete() +
  ylab("Count") +
  xlab("Polity2 Index")
```

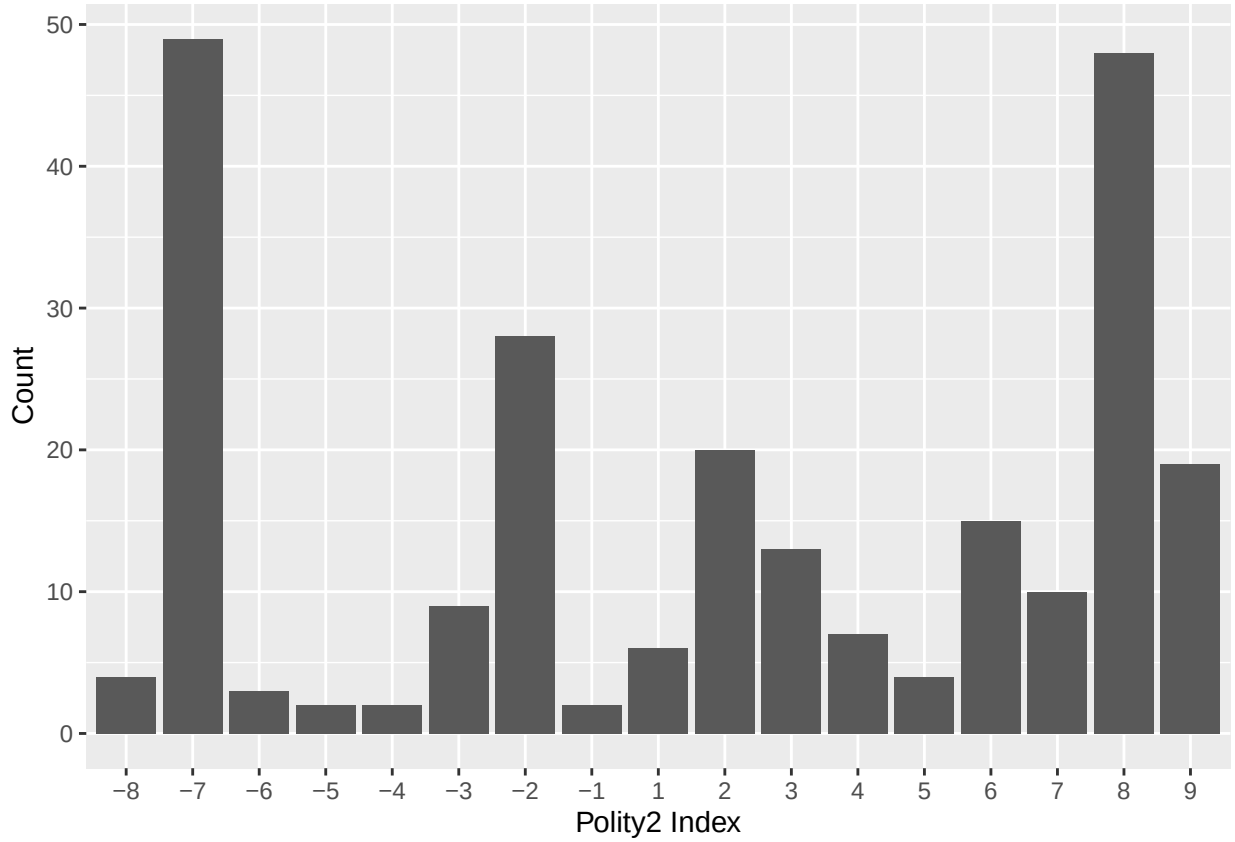

图 9: 民主指数 polity2 的变量分布图

9.(**10 分**) 数据集 `acled_cnty` 记录了 2010-2019 年东南亚国家经历的 “抗议”（protest）和 “骚乱”（riots）数量。利用 `ggplot2` 绘制各国经历的抗议和骚乱变化情况，并比较抗议和骚乱在各国内部

的差异情况。（提示：利用 ggplot 中的 linetype 和 facet_wrap 命令）

```
# 本段代码定义了 interger_break 函数
# 目的是使老挝一图的纵坐标不出现非整数刻度
library("scales")
integer_breaks <- function(n = 5, ...) {
  breaker <- pretty_breaks(n, ...)
  function(x) {
     breaks <- breaker(x)
     breaks[breaks == floor(breaks)]
  }
}
# 初始化作图数据集、横纵坐标变量与线条类型
ggplot(data = acled_cnty,
       aes(x = year,
           y = events,
           group = event_type,
           linetype = event_type)) +
  geom_point(aes(shape = event_type)) +
  geom_line(aes(color = event_type))  +
  labs(x = "Year", y = "Social unrest",
       caption="Source: ACLED 2020") +
  scale_x_continuous(breaks = seq(1991, 2019, 5)) +
  scale_y_continuous(breaks = integer_breaks()) +
# 使用 expand = c(0,1)，可使老挝一图纵坐标出现 0、1、2、3，更为美观
# 但这样会导致其它国家的图中有个别点溢出
# 下面分别绘制各国折线图并排列齐整，设定纵坐标不统一
  facet_wrap(vars(country),
             nrow = 3,
             ncol = 3,
             scales = "free_y") +
  theme(strip.text.x = element_text(size = 12, color='white',angle=0),
        legend.position = "bottom",
        legend.title = element_blank(),
        strip.background = element_rect(fill = "#525252", color='#525252'))
```

我们观察发现，2010-2019 年东南亚国家经历的抗议和骚乱数量呈现出以下几个主要的特点。

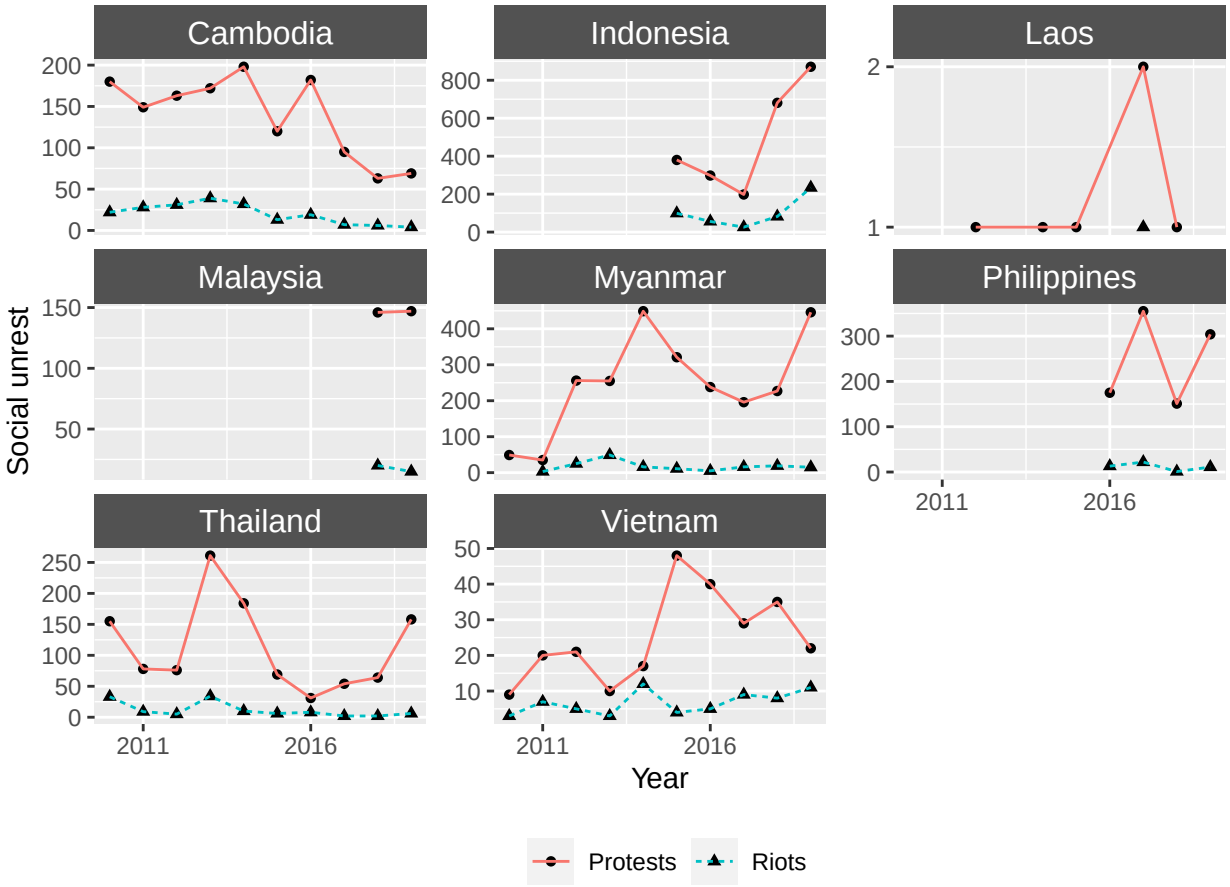

图 10: 东南亚各国国内抗议与骚乱数量（2010-2019）

- 各国内部抗议的数量均高于同年骚乱的数量；
- 各年间抗议的数量可能会发生较大的波动，但骚乱相对于抗议而言，其数量变化起伏不大；
- 各国内部抗议与骚乱的变化趋势有一定相似性；
- 特别地，老挝国内经历的抗议和骚乱数量极低，远远小于图上所列的其它国家。

10.(**5 分**) 数据集 `ucdp_cnty` 记录了东南亚国家 1946-2019 年之间经历的三种冲突（`type_of_conflict2`: 国内冲突、国际冲突以及国际化的国内冲突）的数量 (`conflict`)。请利用 `ggplot` 描述各国分别经历的这三类冲突分布情况。注意其中的缺失值（`NA`），并简要说明为何有缺失值。

```
ggplot(ucdp_cnty,
       aes(x = countyname,
           y = conflict,
           fill = factor(type_of_conflict2))) +
  geom_bar(position="dodge", stat="identity") +
# 使横坐标各国名称不重叠
  scale_x_discrete(labels = function(x){sub("\\s","\n",x)}) +
  scale_fill_manual(values=c("yellow", "red", "blue")) +
  labs(x = "", y = "Count", caption="Source: UCDP 2019") +
  theme_bw() +
  theme(strip.text.x = element_text(size = 12, color='white',angle=0),
        legend.position = "bottom",
        legend.title = element_blank(),
        strip.background = element_rect(fill = "#525252", color='#525252'))
```

查阅 UCDP 提供的数据文档可以推测，缺失值存在的原因在于 1946-2019 年之间越南存在南北越分裂时期。这一阶段南北越之间的冲突被认为是“国际冲突”，但若将其标注为 `Vietnam` 而非 `NA` 则有失妥当。

11.(**10 分**) 数据集 `ideal_point_wide` 记录了东南亚国家在 1973-2018 年之间在联合国大会中投票是否同意中国、印度、美国和俄罗斯（苏联）的情况。利用 `ggplot` 分布绘制东南亚国家在此期间的立场变化。

```
ggplot(ideal_point_wide,
       aes(x = year, y = agreement),
       color = factor(type)) +
  geom_line(aes(color = factor(type)))  +
  labs(x = "Year", y = "UNGA voting agreement",
       caption ="Source: Bailey, Strezhnev, and Voeten (2017, JCR)") +
  scale_x_continuous(breaks = seq(1973, 2018, 10)) +
```

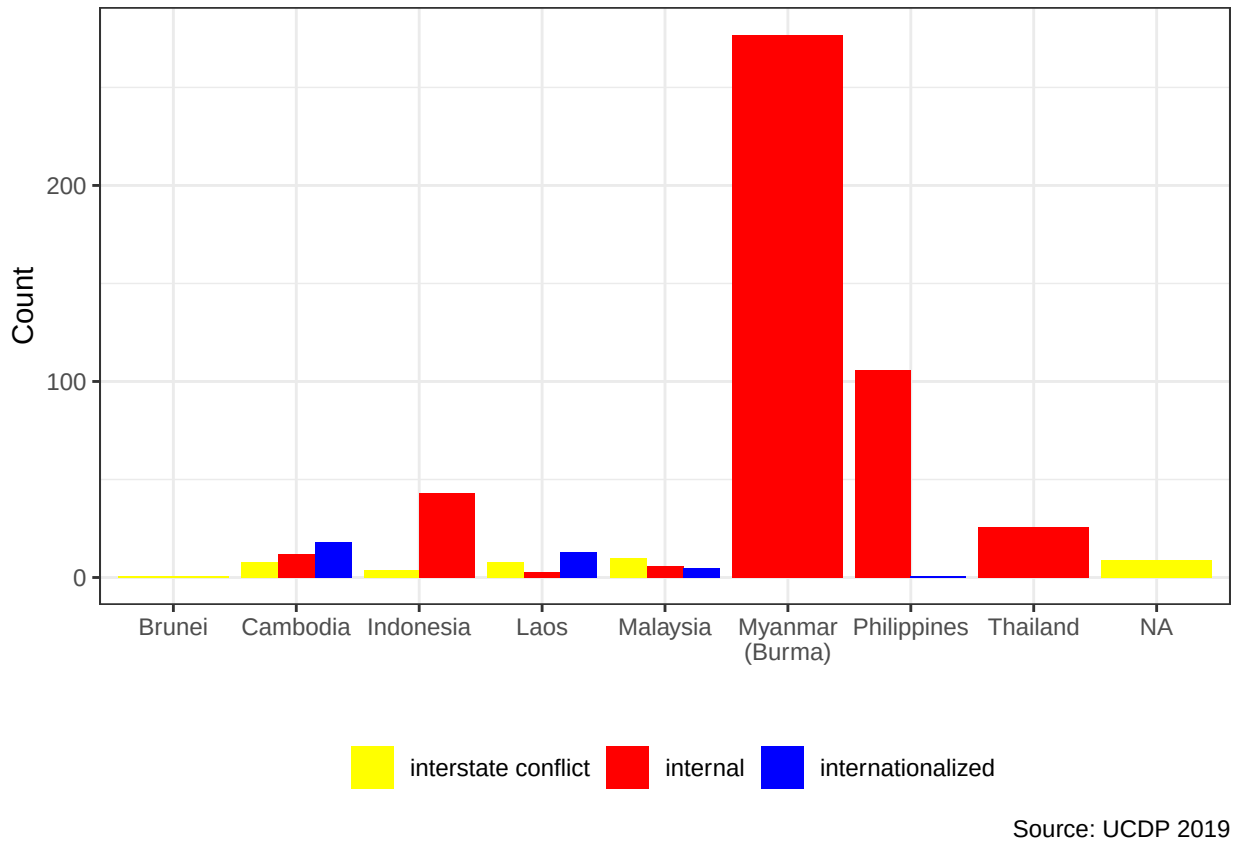

图 11: 东南亚各国冲突数量（1946-2019）

```
  scale_y_continuous(breaks = seq(0, 1, 0.5)) +
  facet_wrap(vars(countryname),
             nrow = 4,
             ncol = 3,
             scales = "fixed") +
  theme(strip.text.x = element_text(size = 12, color='white',angle=0),
        legend.position = "bottom",
        legend.title = element_blank(),
        strip.background = element_rect(fill = "#525252", color='#525252')) +
  scale_colour_manual(values = c("blue","green","gray", "red"))
```

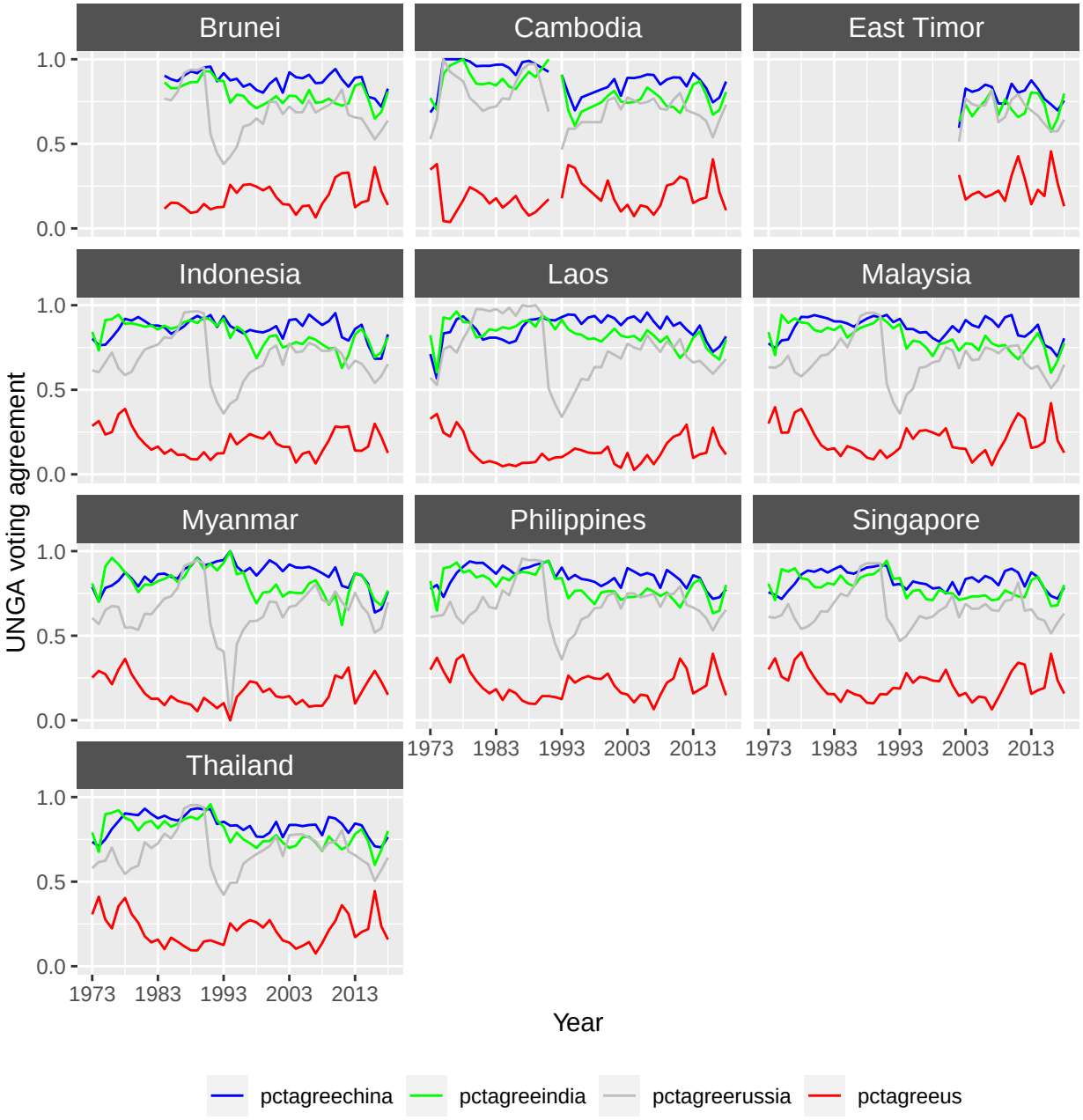

图 12: 东南亚各国在联合国大会投票同意中印美俄的情况（1973-2018）