Study of Temporal and Spatial Generation Patterns of U.S. Power Plants

https://github.com/[redacted]/eGRID16

[redacted]

Abstract

Power plant emissions contribute a great amount Green House Gas (GHG) to the atmosphere and GHG emissions account a lot for the global mean temperature rise. Power Plant emission amount is related to its generation and in general, higher generation indicates higher fuel use and higher emission. Studying power plants generation can help us understand emission pattern. To study the temporal and spatial pattern of power plants generation can help us predict future power generation and then we can give suggestions accordingly to reduce emission. I used the Emissions & Generation Resource Integrated Database (eGRID) to do analysis. From the time series analysis and spatial analysis results, it turned out there is an installation time trend for generator annual net generation, and there are spatial distribution patterns of power plants in U.S.. From 1900s to 2016, the generation increased with fluctuation and the trend was related to policy and world fuel market. We could not predict the future electricity generation since there was no clear trend after 1987. However, from the differences of spatial distribution patterns of Power Plants in U.S., we found that using renewable energy to replace traditional energy power plants can help a lot to reduce GHG emission in the future.

Contents

List of Tables

List of Figures

<Note: set up autoreferencing for figures and tables in your document>

1 Research Question and Rationale

Public awareness of global climate change's impacts on the environment and the society has increased. Greenhouse gas (GHG) emissions account a lot for the global mean temperature rise. To achieve the goal of limiting global warming to 1.5°C, by 2030, global net human-caused emissions of carbon dioxide (CO2) would need to decrease by around 45 % from 2010 levels, and by 2050, CO2 emission would need to reach 'net zero' (IPCC, 2018). In United States, a large amount of GHG emissions come from the electric power sector. The electric power sector accounts for 38% of the total U.S. energy consumption in 2017 (US EIA, 2017). In the same year, the transportation end-use sector and other mobile combustion accounted for 1,732.02 MMT CO2Eq., which accounts 28% of the total GHG emission in the U.S., which is the second largest GHG emissions from all other sectors (the largest is the transportation sector) (US EPA, 2018). To study the temporal and spatial pattern of power plants generation can help us predict future power generation and then we can give suggestions accordingly to reduce emission.

Therefore, here I raise two research questions: (1) Is there an installation time trend for generator annual net generation? (New power generators tend to have higher or lower capacity overtime?) (2) Is there a spatial distribution pattern of Power Plants in U.S.? (Number of power plants/ Total Electricity Generation/ Equivalent CO2 emission in each state ~ States)

Here I am going to use the 2016 Emissions & Generation Resource Integrated Database (eGRID). It is a comprehensive source of data on the environmental characteristics of # almost all electric power generated in the United States. I am going to use the GEN16 tab (2016 Generators) and the PLNT 16 tab (2016 Plants). (Note: usually one power plant has several generators). The former gives information of all the generators currently in use in 2016 and the latter gives information of all the power plants in use in 2016.

2 Dataset Information

Here I am going to use the 2016 Emissions & Generation Resource Integrated Database (eGRID). It is a comprehensive source of data on the environmental characteristics of # almost all electric power generated in the United States. I am going to use the GEN16 tab (2016 Generators) and the PLNT 16 tab (2016 Plants). The dataset was retrieved at: https://www.epa.gov/energy/emissions-generation-resource-integrated-database-egrid on 2019-03-25 21:10:11 EDT. It was originally an excel file and I saved it as two separate csv files of GEN16 an PLNT16. Since the dataset has two names in the first two rows (one full name and one abbreviate one), here I only list the columns I will use later, others can check the original dataset for references.

Table 1: Dataset Information

Dataset Name	Information	Useful Column1	Useful Column2	Useful Column3	Useful Column4
GEN16	Generators Information	SEQGEN16: eGRID2016 Plant file sequence number	PSTATABB: Plant state abbreviation	GENNTAN: Generator annual net generation (MWh)	GENYRONL: Generator year on-line
PLNT16	Plants Information	SEQGEN16: eGRID2016 Plant file sequence number	PSTATABB: Plant state abbreviation	PLNGENAN: Plant annual net generation (MWh)	PLCO2EQA: Plant annual CO2 equivalent emissions (tons)

3 Exploratory Data Analysis and Wrangling

```
#Load dataset
GEN16 <- read.csv("./Data/Raw/egrid_GEN16.csv")
PLNT16 <- read.csv("./Data/Raw/egrid_PLNT16.csv")

#data wrangling
#This dataset has two names, one is the full name, and the first row is the abbreviati
#change column names to the abbr.
names(GEN16) <- lapply(GEN16[1, ], as.character)
names(PLNT16) <- lapply(PLNT16[1, ], as.character)

#filter out the first row and use abbr. as column names
GEN16 <- GEN16[2:26184,]
PLNT16 <- PLNT16[2:9710,]



#numeric data has comma, convert factor to numberic
class (GEN16$GENNTAN)
```

```
## [1] "factor"
```

```
GEN16$GENNTAN <-as.numeric(gsub(",", "", GEN16$GENNTAN))
PLNT16$PLNGENAN <-as.numeric(gsub(",", "", PLNT16$PLNGENAN))
PLNT16$PLCO2EQA <-as.numeric(gsub(",", "", PLNT16$PLCO2EQA))

#Year data as.Date
class (GEN16$GENYRONL)
```

```
## [1] "factor"
```

```
GEN16$GENYRONL<-as.Date(GEN16$GENYRONL,format = "%Y")

#filter data by the sequence of time
GEN16 = GEN16[order(GEN16[,'GENYRONL']),]

#GEN16 - sum the totoal generation by year
GEN16sel <- GEN16 %>%
  select(SEQGEN16, PSTATABB, GENNTAN, GENYRONL) %>%
  filter(!is.na(GENNTAN)) %>%
  filter(GENNTAN>0) %>%
  group_by(GENYRONL)%>%
  summarise(GENSUM = sum(GENNTAN))

#PLNT16 - sum the totoal generation/CO2/plant numbers by state
PLNT16sel <- PLNT16 %>%
```

```r
  select(SEQPLT16, PSTATABB, PLPRMFL, PLNGENAN, PLCO2EQA) %>%
    filter(!is.na(PLNGENAN)&!is.na(PLCO2EQA)) %>%
  filter(PLNGENAN>0)%>%
  group_by(PSTATABB)%>%
  summarise(PLNTGEN = sum(PLNGENAN),
            ECO2 = sum(PLCO2EQA),
            Count=n())

#summary code for GEN16
colnames(GEN16sel)
```

```
## [1] "GENYRONL" "GENSUM"
```

```r
class(GEN16sel$GENSUM)
```

```
## [1] "numeric"
```

```r
class(GEN16sel$GENNTAN)
```

```
## Warning: Unknown or uninitialised column: 'GENNTAN'.
```

```
## [1] "NULL"
```

```r
summary(GEN16sel)
```

```
##     GENYRONL                GENSUM
##  Min.   :1891-04-15   Min.   :      876
##  1st Qu.:1925-10-14   1st Qu.:   794214
##  Median :1956-04-15   Median : 10767508
##  Mean   :1956-03-21   Mean   : 33200403
##  3rd Qu.:1986-10-14   3rd Qu.: 60662462
##  Max.   :2017-04-15   Max.   :213550203
```

```r
dim(GEN16sel)
```

```
## [1] 123   2
```

```r
head(GEN16sel)
```

```
## # A tibble: 6 x 2
##   GENYRONL   GENSUM
##   <date>      <dbl>
## 1 1891-04-15  24330
## 2 1893-04-15   1512
## 3 1896-04-15  21453
## 4 1898-04-15  25514
## 5 1899-04-15    876
## 6 1900-04-15  20899
```

```r
#summary code for PLNT16
colnames(PLNT16sel)
```

```
## [1] "PSTATABB" "PLNTGEN"  "ECO2"     "Count"
```

```r
class(PLNT16sel$PLNTGEN)
```

```
## [1] "numeric"
```

```r
class(PLNT16sel$ECO2)
```

```
## [1] "numeric"
```

```r
summary(PLNT16sel)
```

```
##     PSTATABB     PLNTGEN                ECO2                Count       
##  AK     : 1   Min.   :    76474   Min.   :    18470   Min.   :   2.0  
##  AL     : 1   1st Qu.: 34625868   1st Qu.: 11970062   1st Qu.:  61.0  
##  AR     : 1   Median : 60445059   Median : 31234830   Median :  96.0  
##  AZ     : 1   Mean   : 80006416   Mean   : 40119484   Mean   : 146.5  
##  CA     : 1   3rd Qu.:107747773   3rd Qu.: 54001623   3rd Qu.: 148.0  
##  CO     : 1   Max.   :453941341   Max.   :239363582   Max.   :1163.0  
##  (Other):45
```

```r
dim(PLNT16sel)
```

```
## [1] 51  4
```

```r
head(PLNT16sel)
```

```
## # A tibble: 6 x 4
##   PSTATABB   PLNTGEN     ECO2 Count
##   <fct>        <dbl>    <dbl> <int>
## 1 AK         6339538  2944890   128
## 2 AL       142863565 65536193    69
## 3 AR        60445059 33927595    52
## 4 AZ       108734651 50946063   112
## 5 CA       197956373 44797489  1163
## 6 CO        54679959 40213412   147
```

```r
#save new datasets
#write.csv(GEN16sel, file = "./Data/Processed/GEN16sel_Processed.csv",row.names=FALSE)
#write.csv(PLNT16sel, file = "./Data/Processed/PLNT16sel_Processed.csv",row.names=FALS
#data prep for shiny app
#PLNT16orisel <- PLNT16 %>%
#  select(SEQPLT16, PSTATABB, PLPRMFL, PLNGENAN, PLCO2EQA) %>%
#    filter(!is.na(PLNGENAN)&!is.na(PLCO2EQA)) %>%
#  filter(PLNGENAN>0)
#write.csv(PLNT16orisel, file = "./Data/Processed/PLNT16ori_Processed.csv",row.names=F
```

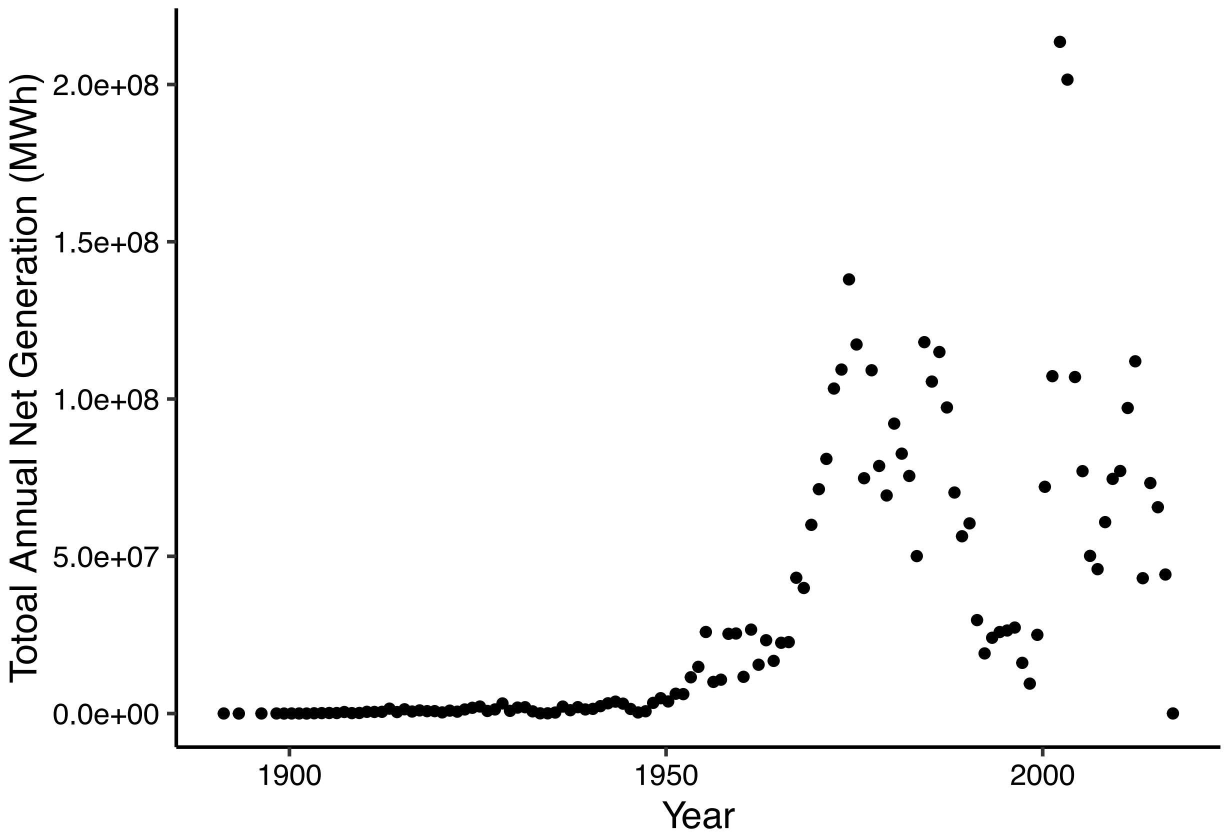

Figure 1: Overview of Generator online year

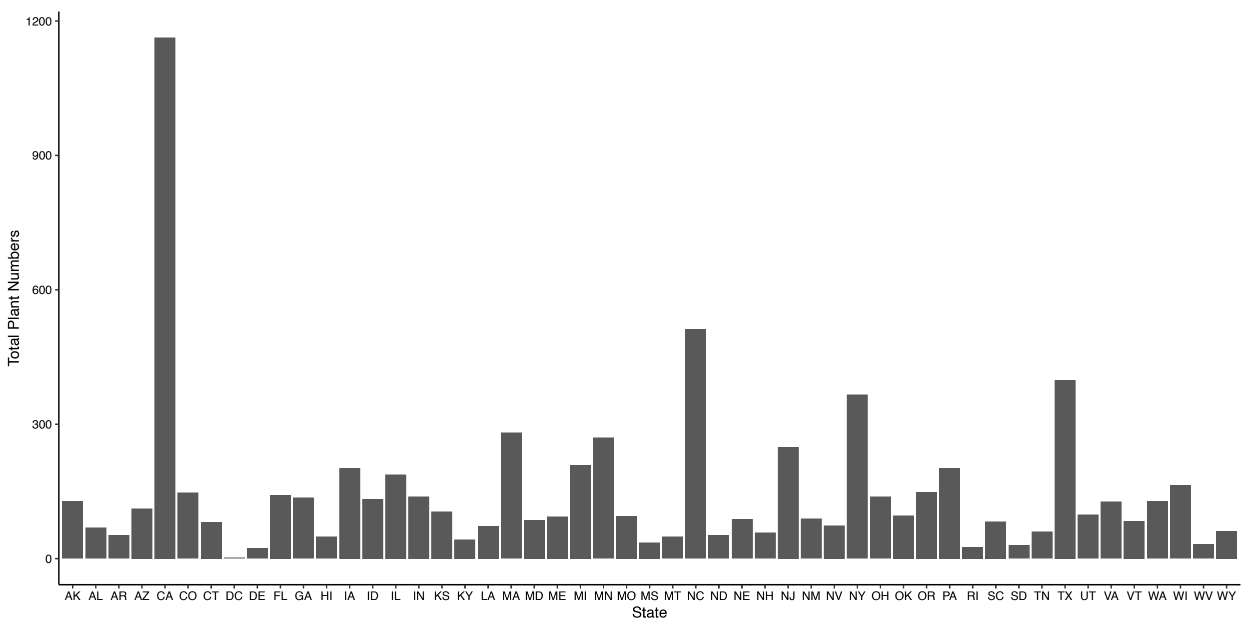

Figure 2: Plant Numbers

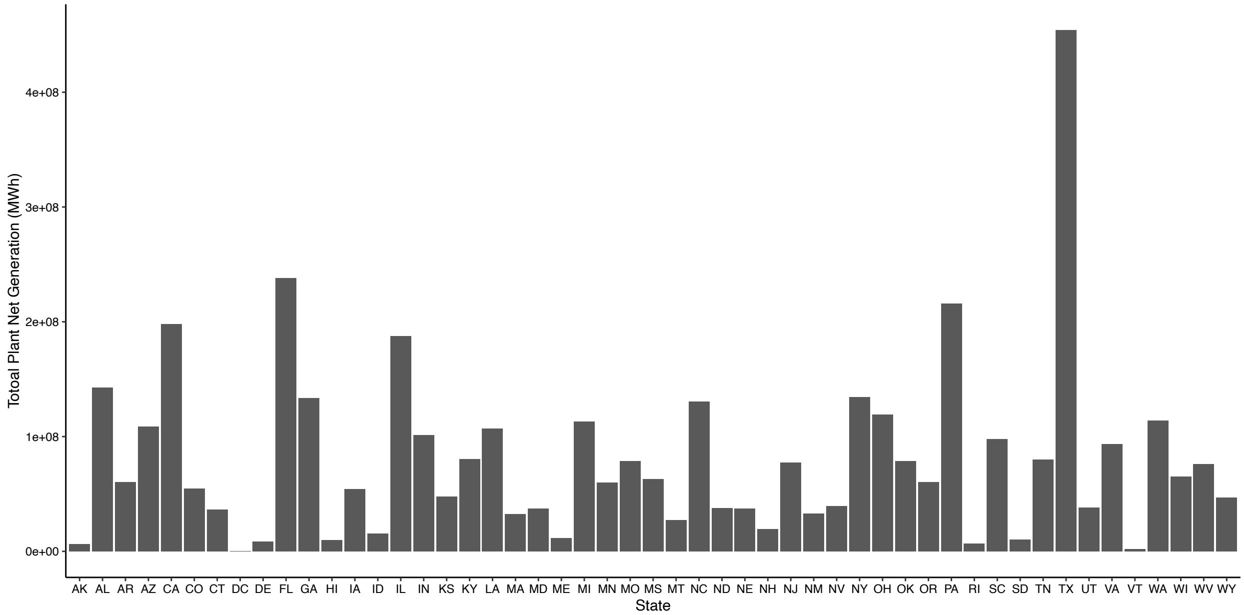

Figure 3: Plant Generation

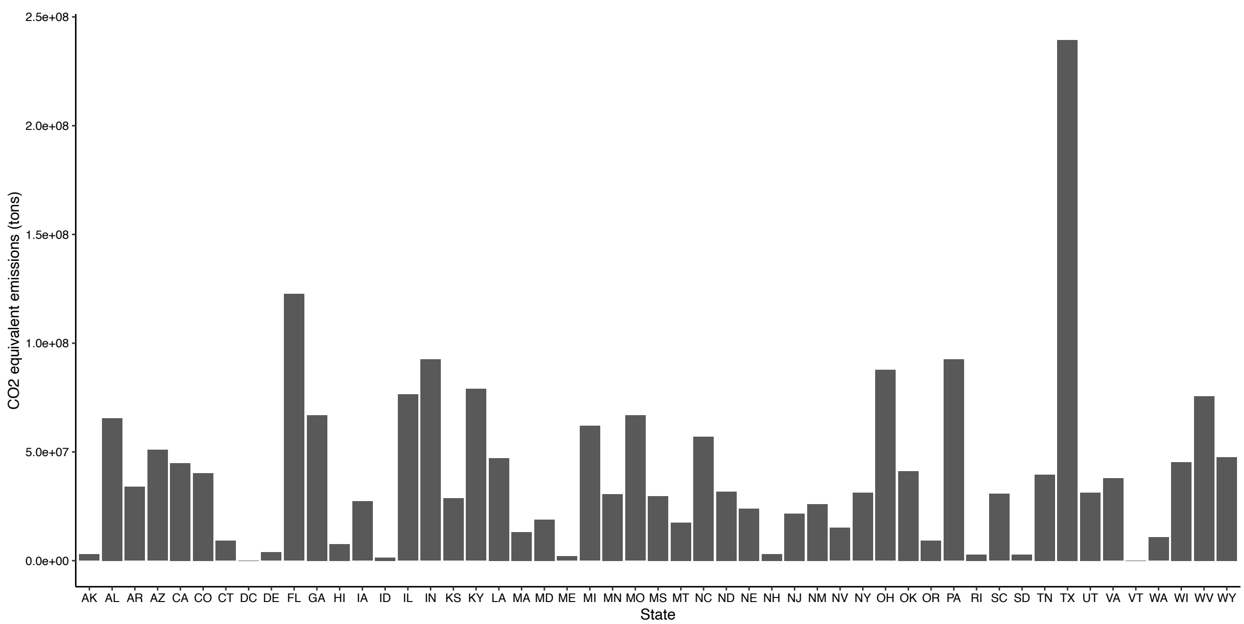

Figure 4: Plant Equivalent CO2 Emissions

4 Analysis

```
#Q1 Time series analysis on GEN16
# Use GLM to see if there is a significant time trend
GENTest.fixed <- gls(data = GEN16sel,
                     GENSUM ~ GENYRONL,
                      method = "REML")
summary(GENTest.fixed) # significnat trend t=10.23, p<0.005.
```

```
## Generalized least squares fit by REML
##   Model: GENSUM ~ GENYRONL
##   Data: GEN16sel
##        AIC      BIC    logLik
##   4562.782 4571.169 -2278.391
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 44750764 3129296.8 14.30058       0
## GENYRONL        2295     224.3 10.22992       0
##
##  Correlation:
##          (Intr)
## GENYRONL 0.361
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.6061370 -0.5704619 -0.1403653  0.3843417  4.3789847
##
## Residual standard error: 32367813
## Degrees of freedom: 123 total; 121 residual
```

According to GLM, time has a significant effect on the total annual net generation of generators ($t=10.23$, $p<0.05$).

```
# Run a Mann-Kendall test
mk.test(GEN16sel$GENSUM)
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM
## z = 11.311, n = 123, p-value < 2.2e-16
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 5.175000e+03 2.092503e+05 6.897241e-01
```

```r
# there is a trend over time according to this test (p<0.001), Z is positive, positive

# Test for change point
pettitt.test(GEN16sel$GENSUM) #changing point at 58 - Year 1952
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  GEN16sel$GENSUM
## U* = 3670, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              58
```

```r
#GEN16sel[58,]

# Run separate Mann-Kendall for each change point
mk.test(GEN16sel$GENSUM[1:57])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[1:57]
## z = 6.8907, n = 57, p-value = 5.551e-12
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 1.002000e+03 2.110267e+04 6.278195e-01
```

```r
# there is a trend over time according to this test (p<0.001), Z is positive, positive
mk.test(GEN16sel$GENSUM[58:123])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[58:123]
## z = 2.8224, n = 66, p-value = 0.004767
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 5.110000e+02 3.265167e+04 2.382284e-01
```

```r
# there is a trend over time according to this test (p<0.05), Z is positive, positive

# Is there a second change point?
```

```r
pettitt.test(GEN16sel$GENSUM[58:123]) #there is! 17 - Year 1969
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  GEN16sel$GENSUM[58:123]
## U* = 681, p-value = 0.0001447
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                               17
```

```r
#GEN16sel[75,]

# Run separate Mann-Kendall for each change point
mk.test(GEN16sel$GENSUM[58:74])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[58:74]
## z = 2.5951, n = 17, p-value = 0.009455
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##           S        varS         tau
##  64.0000000 589.3333333   0.4705882
```

```r
# there is a trend over time according to this test (p<0.05), Z is positive, positive
mk.test(GEN16sel$GENSUM[75:123])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[75:123]
## z = -2.0084, n = 49, p-value = 0.0446
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S           varS           tau
##  -234.0000000 13458.6666667    -0.1989796
```

```r
# there is a trend over time according to this test (p<0.05), Z is negative, negative

# Is there a third change point?
pettitt.test(GEN16sel$GENSUM[75:123]) #there is! 1987
```

```
##
##  Pettitt's test for single change-point detection
```

```
##
## data:  GEN16sel$GENSUM[75:123]
## U* = 322, p-value = 0.01123
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                               19
```

```r
# GEN16sel[94,]
mk.test(GEN16sel$GENSUM[75:93])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[75:93]
## z = 0.69971, n = 19, p-value = 0.4841
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##          S       varS        tau
##  21.000000 817.000000   0.122807
```

```r
# no trend!
mk.test(GEN16sel$GENSUM[94:123])
```

```
##
##  Mann-Kendall trend test
##
## data:  GEN16sel$GENSUM[94:123]
## z = 1.1775, n = 30, p-value = 0.239
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##           S        varS         tau
##   67.000000 3141.666667    0.154023
```

```r
# no trend!
# no trend before and after the changing point
```

According to the Mann-Kendall test, there is a trend over time for total annual net generation of generators ($p<0.001$), Z =11.311, which indicates a positive trend over time. According to pettitt test. There are three changing points: Year 1952, Year 1969 and Year 1987 ($p<0.05$).

As is shown in the figure (Figure 5), before the first changing point 1952, the annual net generation of generators in U.S. increased very slowly each year and after 1952, the speed of generation change became faster. Annual net generation kept growing until 1969, this is the second change point. After 1969, there was a negative trend of annual net generation. The third changing point is Year 1987, and there was no clear pattern in 1969-1987 or after 1987.

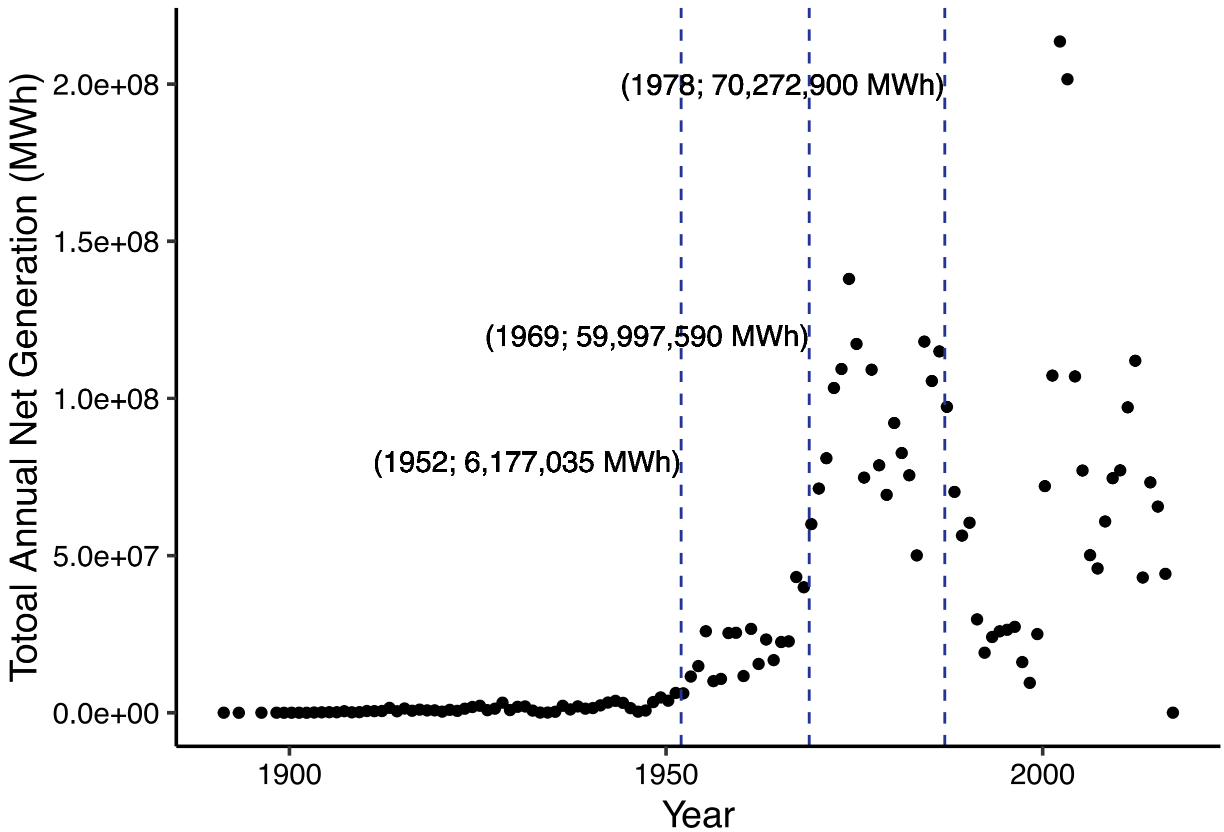

Figure 5: Generators Time Series Analysis

```
#Q2 spatial distribution analysis on PLNT16
#data wrangling
#add fips number to PLNT16sel data
library(usmap)
state_map <- us_map(regions = "states")
PLNT_State<- merge(state_map, PLNT16sel, by.x = "abbr", by.y = "PSTATABB")%>%
  select(abbr, fips, full, PLNTGEN, ECO2, Count)
PLNT_State = PLNT_State[!duplicated(PLNT_State$abbr),]

#join PLNT16 data to counties map data
states_sf<- st_read('./Data/RAW/States.shp')
```

```
## Reading layer `States' from data source `C:\Users\[redacted]\Desktop\eGRID16\Data\Raw\
## Simple feature collection with 52 features and 1 field
## geometry type:  MULTIPOLYGON
## dimension:      XY
## bbox:           xmin: -179.1743 ymin: 17.91377 xmax: 179.7739 ymax: 71.35256
## epsg (SRID):    4269
## proj4string:    +proj=longlat +datum=NAD83 +no_defs
```

```
st_crs(states_sf)
```

```
## Coordinate Reference System:
##   EPSG: 4269
##   proj4string: "+proj=longlat +datum=NAD83 +no_defs"
```

```
PLNT_State_merge <- merge(states_sf, PLNT_State, by.x = "STATEFP", by.y = "fips")
#mapview(PLNT_State_merge)
```

Since our research scale is U.S., it also includes Alaska and this made it very difficult to map it using ggplot (AL will be far away from other states and the whole map will be very small to show). Therefore, here I use the plot_usmap function instead. As the maps show (Figure 6, 7, 8), there are spatial distribution patterns. For total plant numbers, California ranks first but in general, there are more power plants in east coast states than in west coast. Texas is another exception in the southern part that has relatively larger numbers of power plants. For plant annual net generation, in general, east coast states also have relatively higher net generation than west coast states, and Texas ranks first. Accordingly, Texas also has highest annual CO2 equivalent emissions, and east coast states have higher emissions than west coast states.

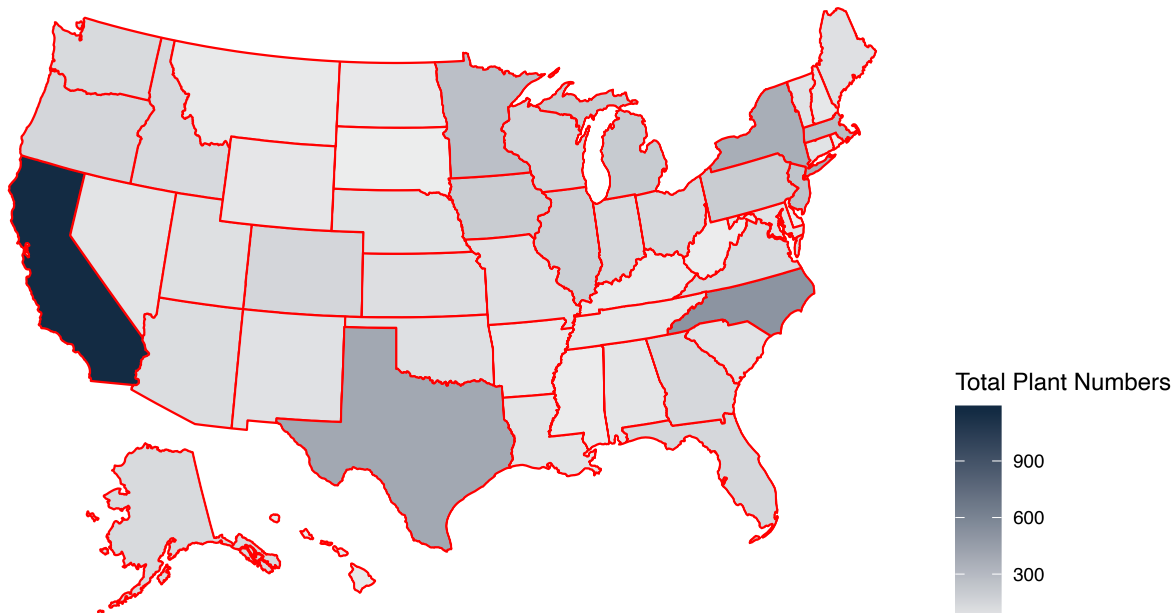

Figure 6: Map of Plant Numbers

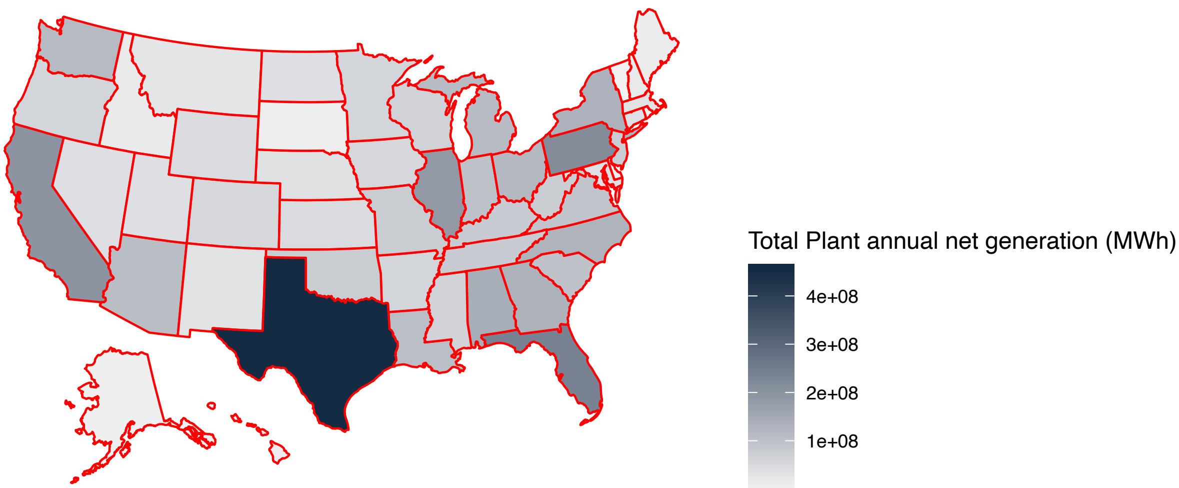

Figure 7: Map of Plant Generations

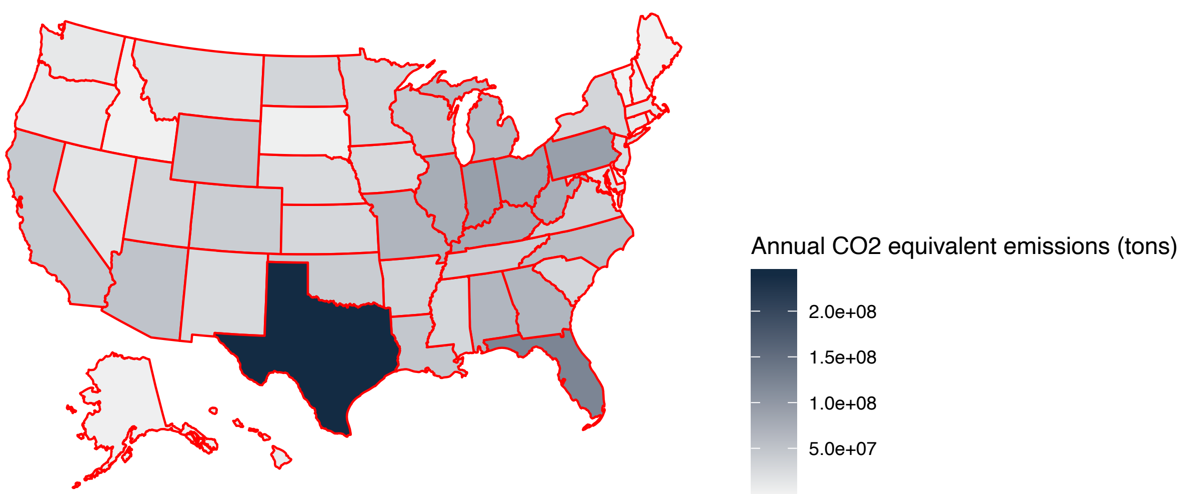

Figure 8: Map of Equivalent CO2 Emissions

5 Summary and Conclusions

There are also three changing points of the whole period: 1952, 1969, 1987. After 1952, there is a huge increase in power generation each year, this might be related to the technology thrive at the end of 1940s. The first hydraulic fracturing treatment was pumped on a gas well operated by Pan American Petroleum Corp in the Hugoton field in 1947 and it was the beginning of period of rapid electric industry growth (Oilscams.org, Access: 2019-4-14). However, at the end of 1960s, energy crisis caused the reduction of electricity generation and this energy crisis peaked at 1973 (Energy Crisis (1970s), Access: 2019-4-14). This accord with our second changing point here at 1969 and after 1969, there was a negative trend of annual net generation.

In the late 1970s, the government published more regulations on the electricity generation industry such as the National Energy Act in 1978 to exert more control on the electricity industry. In 1990s, the government published Congress passes Bush's Energy Policy Act (EPACT) to deregulate the electricity industry (Ballotpedia, Access: 2019-4-14). The back and force between strict and loose policies led to the fluctuations of the generator generations, and this might be a cause why there is a changing point of 1987 according to the pettitt test but there was no trend before and after this year according to the Mann-Kendall test.

From the maps, we can see that there was spatial distribution pattern for power plants generation in 2016. In general, eastern states had higher power plant numbers, electricity net generation, and equivalent CO2 emissions than western states. California ranked first for total power plants numbers but its annual net generation and equivalent CO2 emission were both much less than Texas. The main reason for this was the type of power plants. In Texas, there are more traditional power plants using coal and natural gas as fuel, but in California, there are more renewable energy power plants. Compared to traditional power plants (600 MW, 0.75), renewable power plants have lower nameplate capacity and capacity factor (1000MW, 0.1). Therefore, even though California had more power plants, it had lower electricity generation. Also, compared to traditional power plants, renewable power plants use clean energy and barely have any emissions, so California's equivalent CO2 emissions were also much lower than Texas.

To sum up, there is an installation time trend for generator annual net generation. From 1900s to 2016, the generation increased with fluctuation and the trend was related to policy and world fuel market. It was very difficult to predict the future electricity generation since there was no clear trend after 1987. However, from the differences of spatial distribution patterns of Power Plants in U.S., we can give suggestions of using renewable energy to replace traditional energy power plants. This can help a lot to reduce GHG emission in the future.