Assignment 8: Time Series Analysis

Sarah Ko

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on time series analysis.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A08_TimeSeries.pdf") prior to submission.

The completed exercise is due on Tuesday, 19 March, 2019 before class begins.

Brainstorm a project topic

1. Spend 15 minutes brainstorming ideas for a project topic, and look for a dataset if you are choosing your own rather than using a class dataset. Remember your topic choices are due by the end of March, and you should post your choice ASAP to the forum on Sakai.

Question: Did you do this?

> ANSWER: Yes I did! I chose my own dataset and have described my project goals in the Sakai forum.

Set up your session

2. Set up your session. Upload the EPA air quality raw dataset for PM2.5 in 2018, and the processed NTL-LTER dataset for nutrients in Peter and Paul lakes. Build a ggplot theme and set it as your default theme. Make sure date variables are set to a date format.

```
# load packages
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2.
```

```
## v ggplot2 3.1.0     v readr   1.3.1
## v tibble  2.1.1     v purrr   0.3.2
## v tidyr   0.8.3     v stringr 1.4.0
## v ggplot2 3.1.0     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.3
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

```
## -- Conflicts ------------------------------------------------------------------- tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 3.5.2
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.5.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(multcompView)
```

```
## Warning: package 'multcompView' was built under R version 3.5.2
```

```
library(lme4)
```

```
## Warning: package 'lme4' was built under R version 3.5.3
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following object is masked from 'package:tidyr':
##
##     expand
```

```
##
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:nlme':
##
##     lmList
```

```
library(trend)
```

```
## Warning: package 'trend' was built under R version 3.5.2
```

```
# get working directory
getwd()
```

```
## [1] "C:/Users/Sarah/Documents/Duke/Year 2/Spring 2019/Data Analytics/Environmental_Data_Analytics"
```

```
# set wd to filepath of Environmental_Data_Analytics to use relative filepath
# load EPA air quality raw dataset for PM2.5 in 2018
Air_PM25_2018 <- read.csv("./Data/Raw/EPAair_PM25_NC2018_raw.csv")

# load NTL-LTER for nutrients in Peter and Paul Lakes
NTL_Nutrients_PeterPaul <-
  read.csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaulGathered_Processed.csv")

# build ggplot theme, set as default theme
SKotheme <- theme_gray(base_size = 15) +
theme(axis.text = element_text(color = "black"),
legend.position = "right",
plot.title = element_text(hjust = 0.5))
theme_set(SKotheme)

# check date variables
class(Air_PM25_2018$Date)
```

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

```
class(NTL_Nutrients_PeterPaul$sampledate)
```

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

```r
# change date variables to date must define original format of data
Air_PM25_2018$Date <- as.Date(Air_PM25_2018$Date, format = "%m/%d/%y")
NTL_Nutrients_PeterPaul$sampledate <- as.Date(NTL_Nutrients_PeterPaul$sampledate, format = "%Y-%m-%d")

# confirm date variables
class(Air_PM25_2018$Date)
```

```
## [1] "Date"
```

```r
class(NTL_Nutrients_PeterPaul$sampledate)
```

```
## [1] "Date"
```

Run a hierarchical (mixed-effects) model

Research question: Do PM2.5 concentrations have a significant trend in 2018?

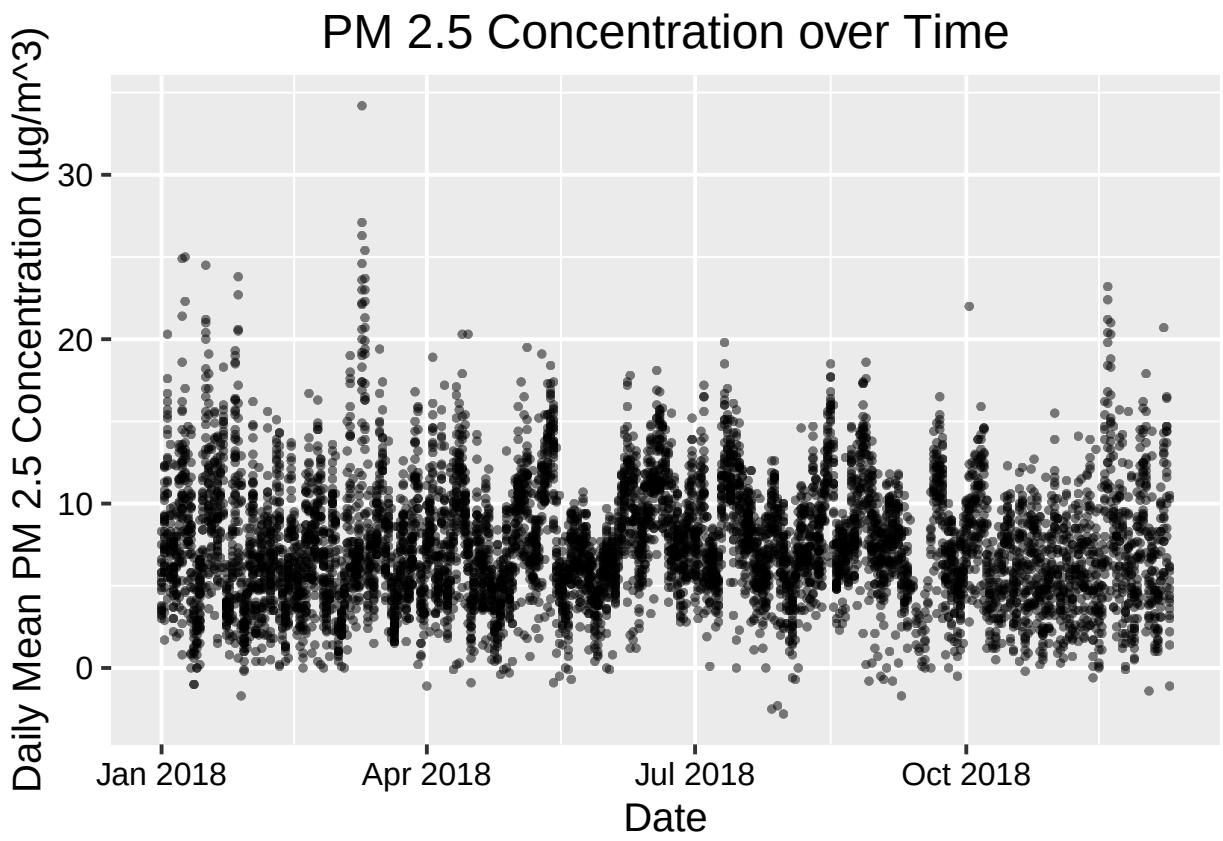
3. Run a repeated measures ANOVA, with PM2.5 concentrations as the response, Date as a fixed effect, and Site.Name as a random effect. This will allow us to extrapolate PM2.5 concentrations across North Carolina.

3a. Illustrate PM2.5 concentrations by date. Do not split aesthetics by site.

```r
# 3) run repeated measures ANOVA
PM25_repeat_measures <- lme(data = Air_PM25_2018,
                     Daily.Mean.PM2.5.Concentration ~ Date,
                     random = ~1|Site.Name)
PM25_repeat_measures
```

```
## Linear mixed-effects model fit by REML
##   Data: Air_PM25_2018
##   Log-restricted-likelihood: -20297.38
##   Fixed: Daily.Mean.PM2.5.Concentration ~ Date
## (Intercept)        Date
## 20.14183588 -0.00074241
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:    1.841425 3.457061
##
## Number of Observations: 7611
## Number of Groups: 24
```

```r
# 3a) graph PM2.5 by date
PM25_date <- ggplot(data = Air_PM25_2018, aes(x = Date, y = Daily.Mean.PM2.5.Concentration)) +
  geom_point(alpha=0.5, stroke=0) +
  ggtitle("PM 2.5 Concentration over Time") + # add main title
  xlab("Date") +
  ylab("Daily Mean PM 2.5 Concentration (\U003BCg/m^3)")
print(PM25_date)
```



3b. Insert the following line of code into your R chunk. This will eliminate duplicate measurements on single dates for each site. PM2.5 = PM2.5[order(PM2.5[,'Date'],-PM2.5[,'Site.ID']),] PM2.5 = PM2.5[!duplicated(PM2.5$Date),]

3c. Determine the temporal autocorrelation in your model.

3d. Run a mixed effects model.

```
# 3b) eliminate duplicate measurements on single dates for each site
PM2.5 <- Air_PM25_2018

PM2.5 = PM2.5[order(PM2.5[,'Date'],-PM2.5[,'Site.ID']),]
PM2.5 = PM2.5[!duplicated(PM2.5$Date),]

# 3c) determine temporal autocorrelation
ACF(PM25_repeat_measures)
```

```
##    lag          ACF
## 1    0  1.000000000
## 2    1  0.473017989
## 3    2  0.143093030
## 4    3  0.060500838
## 5    4  0.061574447
## 6    5  0.087756109
## 7    6  0.061116723
## 8    7  0.007595491
## 9    8  0.025491472
## 10   9  0.057872193
```

```
## 11  10  0.095911195
## 12  11  0.086519308
## 13  12  0.041507759
## 14  13  0.041091743
## 15  14  0.008663124
## 16  15 -0.012810524
## 17  16 -0.016388970
## 18  17 -0.023436707
## 19  18  0.020967717
## 20  19  0.032373855
## 21  20 -0.046770645
## 22  21 -0.086974675
## 23  22 -0.045009633
## 24  23  0.014507171
## 25  24  0.046279402
## 26  25  0.021031653
## 27  26 -0.017185250
## 28  27  0.008158717
```

```
# take the 2nd value (the innermost group level) to define the degree of autocorrelation = 0.473017989

# 3d) run mixed effects model
PM25_mixed_effects <- lme(data = PM2.5,
                    Daily.Mean.PM2.5.Concentration ~ Date,
                    random = ~1|Site.Name, #specify autocorrelation structure of order 1
                    method = "REML") #define method as restricted maximum likelihood
summary(PM25_mixed_effects)
```

```
## Linear mixed-effects model fit by REML
##  Data: PM2.5
##        AIC      BIC    logLik
##   1865.215 1880.543 -928.6076
##
## Random effects:
##  Formula: ~1 | Site.Name
##         (Intercept) Residual
## StdDev:    1.650184 3.559209
##
## Fixed effects: Daily.Mean.PM2.5.Concentration ~ Date
##                Value Std.Error  DF   t-value p-value
## (Intercept) 90.46502  34.57133 339  2.616764  0.0093
## Date        -0.00473   0.00195 339 -2.425102  0.0158
##  Correlation:
##      (Intr)
## Date -0.999
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -2.38072443 -0.63365107 -0.09616694  0.61426094  3.42056220
##
## Number of Observations: 343
## Number of Groups: 3
```

Is there a significant increasing or decreasing trend in PM2.5 concentrations in 2018?

ANSWER: There is a significant decreasing trend in PM2.5 concentrations in 2018. The model

equation is: concentration of PM2.5 (ug/m3) = 90.46502 - 0.00473*date. The p values corresponding to the intercept and the date variable are significant, with p = 0.0093 and p = 0.0158 respectively.

3e. Run a fixed effects model with Date as the only explanatory variable. Then test whether the mixed effects model is a better fit than the fixed effect model.

```
# 3e) run fixed effect model with Date as the only explanatory variable
PM25_date_only <- gls(data = PM2.5,
                      Daily.Mean.PM2.5.Concentration ~ Date,
                      method = "REML") # no random effect, no autocorrelation structure
summary(PM25_date_only)
```

```
## Generalized least squares fit by REML
##   Model: Daily.Mean.PM2.5.Concentration ~ Date
##   Data: PM2.5
##        AIC      BIC    logLik
##   1865.202 1876.698 -929.6011
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 98.57796  34.60285  2.848840  0.0047
## Date        -0.00513   0.00195 -2.624999  0.0091
##
##  Correlation:
##      (Intr)
## Date -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.3531000 -0.6348100 -0.1153454  0.6383004  3.4063068
##
## Residual standard error: 3.584321
## Degrees of freedom: 343 total; 341 residual
```

```
anova(PM25_mixed_effects, PM25_date_only)
```

```
##                    Model df      AIC      BIC    logLik   Test  L.Ratio
## PM25_mixed_effects     1  4 1865.215 1880.543 -928.6076
## PM25_date_only         2  3 1865.202 1876.698 -929.6011 1 vs 2 1.986919
##                    p-value
## PM25_mixed_effects
## PM25_date_only      0.1587
```

Which model is better?

ANSWER: The mixed effects model has an AIC value of 1865.215. The fixed effect model with only date as the explanatory variable has a slightly lower AIC value of 1865.202, indicating it to be the better model. However, the p value is 0.1587, which indicates that the models do not have a significantly different fit.

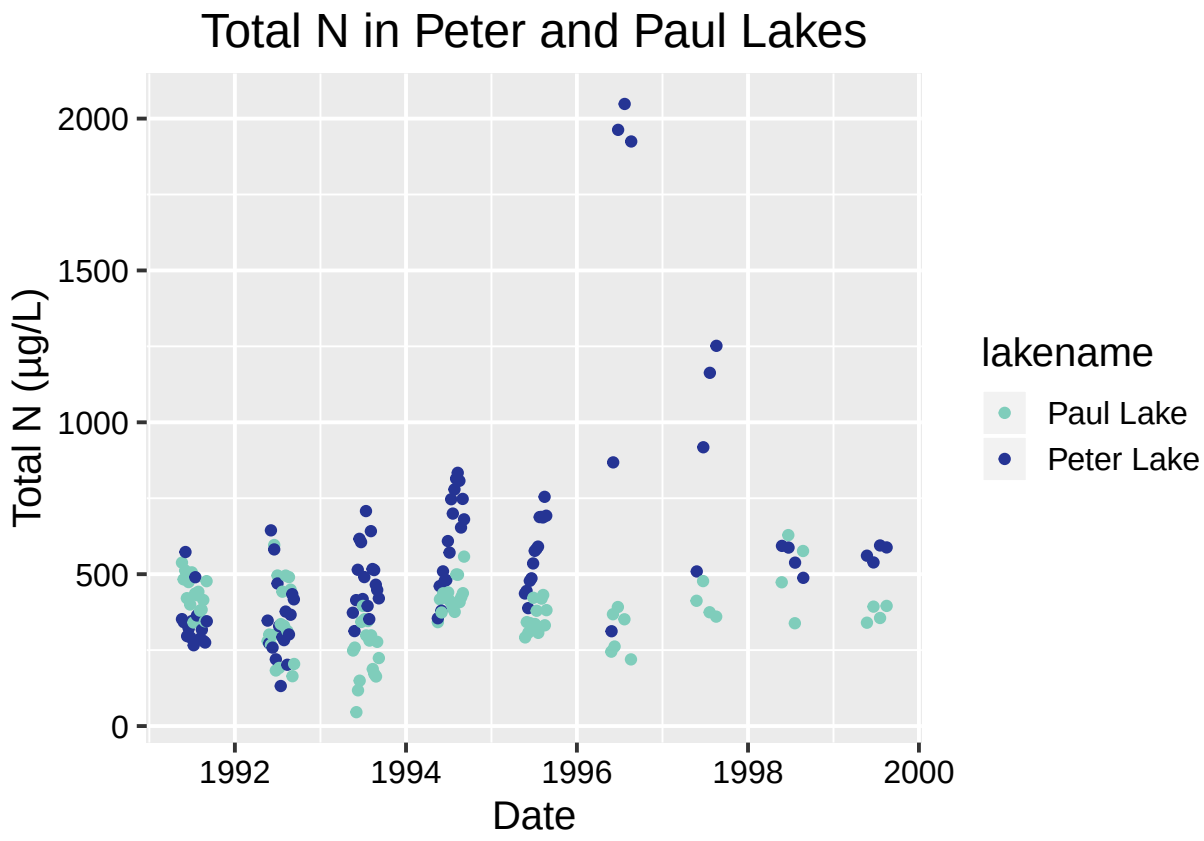
Run a Mann-Kendall test

Research question: Is there a trend in total N surface concentrations in Peter and Paul lakes?

4. Duplicate the Mann-Kendall test we ran for total P in class, this time with total N for both lakes. Make sure to run a test for changepoints in the datasets (and run a second one if a second change point is likely).

```r
# Wrangle dataset
NTL_PeterPaul_TotalN <-
  NTL_Nutrients_PeterPaul %>%
  select(-daynum, -year4) %>%
  filter(nutrient == "tn_ug") %>%
  filter(depth == 0)

# initial visualization of the data
Total_N_plot <- ggplot(NTL_PeterPaul_TotalN, aes(x = sampledate, y = concentration,
                                                 color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494")) +
  ggtitle("Total N in Peter and Paul Lakes") + # add main title
  xlab("Date") +
  ylab("Total N (\U003BCg/L)")
print(Total_N_plot)
```




```r
Peter.N.surface <- filter(NTL_PeterPaul_TotalN, lakename == "Peter Lake")
dim(Peter.N.surface) # length = 98
```

```
## [1] 98  5
```

```r
Paul.N.surface <- filter(NTL_PeterPaul_TotalN, lakename == "Paul Lake")
dim(Paul.N.surface) # length = 99
```

```
## [1] 99  5
```

```r
# PETER LAKE
# Run a Mann-Kendall test
mk.test(Peter.N.surface$concentration)
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.N.surface$concentration
## z = 7.2927, n = 98, p-value = 3.039e-13
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##            S         varS          tau
## 2.377000e+03 1.061503e+05 5.001052e-01
```

```r
# Test Peter Lake for changepoints
pettitt.test(Peter.N.surface$concentration) # change point at time = 36
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.N.surface$concentration
## U* = 1884, p-value = 3.744e-10
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              36
```

```r
Peter.N.surface$sampledate[35] # 1993-05-26
```

```
## [1] "1993-05-26"
```

```r
Peter.N.surface$sampledate[36] # 1993-06-02
```

```
## [1] "1993-06-02"
```

```r
# change point at ~ 1993-05-29

# Run separate Mann-Kendall for each change point
mk.test(Peter.N.surface$concentration[1:35])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.N.surface$concentration[1:35]
## z = -0.22722, n = 35, p-value = 0.8203
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
##  -17.00000000 4958.33333333   -0.02857143
```

```r
mk.test(Peter.N.surface$concentration[36:98])
```

```
##
##  Mann-Kendall trend test
##
## data:  Peter.N.surface$concentration[36:98]
## z = 3.1909, n = 63, p-value = 0.001418
## alternative hypothesis: true S is not equal to 0
```

```
## sample estimates:
##            S         varS          tau
## 5.390000e+02 2.842700e+04 2.759857e-01
```

```
# check for second change point
pettitt.test(Peter.N.surface$concentration[36:98]) # change point at time = 36+21 = 57
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Peter.N.surface$concentration[36:98]
## U* = 560, p-value = 0.001213
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              21
```

```
Peter.N.surface$sampledate[56] # 1994-06-22
```

```
## [1] "1994-06-22"
```

```
Peter.N.surface$sampledate[57] # 1994-06-29
```

```
## [1] "1994-06-29"
```

```
# change point at ~ 1994-06-25

# PAUL LAKE
# Run a Mann-Kendall test
mk.test(Paul.N.surface$concentration)
```

```
##
##  Mann-Kendall trend test
##
## data:  Paul.N.surface$concentration
## z = -0.35068, n = 99, p-value = 0.7258
## alternative hypothesis: true S is not equal to 0
## sample estimates:
##             S          varS           tau
## -1.170000e+02  1.094170e+05 -2.411874e-02
```

```
# Test Paul Lake for changepoints
pettitt.test(Paul.N.surface$concentration) # possible change point at time = 16,
```

```
##
##  Pettitt's test for single change-point detection
##
## data:  Paul.N.surface$concentration
## U* = 704, p-value = 0.09624
## alternative hypothesis: two.sided
## sample estimates:
## probable change point at time K
##                              16
```

```
#but p value is not significant
```

What are the results of this test?

ANSWER: Peter Lake shows a significant change in total N surface concentrations over time. It

is seen to have 2 change points: the first ~1993-05-29 ($p < 0.01$), and the second ~1994-06-25 ($p < 0.01$). Paul Lake does not show any significant change points.

5. Generate a graph that illustrates the TN concentrations over time, coloring by lake and adding vertical line(s) representing changepoint(s).

```
# graph with changepoints
Total_N_plot_changepoints <- ggplot(NTL_PeterPaul_TotalN, aes(x = sampledate,
                                                              y = concentration, color = lakename)) +
  geom_point() +
  scale_color_manual(values = c("#7fcdbb", "#253494")) +
  ggtitle("Total N in Peter and Paul Lakes") + # add main title
  xlab("Date") +
  ylab("Total N (\U003BCg/L)") +
  geom_vline(xintercept = as.Date("1993-05-29"), linetype=2,
             color = "#253494", size=1.5) +
  geom_vline(xintercept = as.Date("1994-06-25"), linetype=2,
             color = "#253494", size=1.5)
print(Total_N_plot_changepoints)
```