Sup. Analysis for the Reviewers

Sulyok et al.

2019 július 7

```r
library( lattice )
library(readr)



masterall <- read_delim("fsmeundzecken.csv", ";", escape_double = FALSE, col_types = cols(time = col_da
View(masterall)

masterall[is.na(masterall)] <- 0
summary(masterall)
```

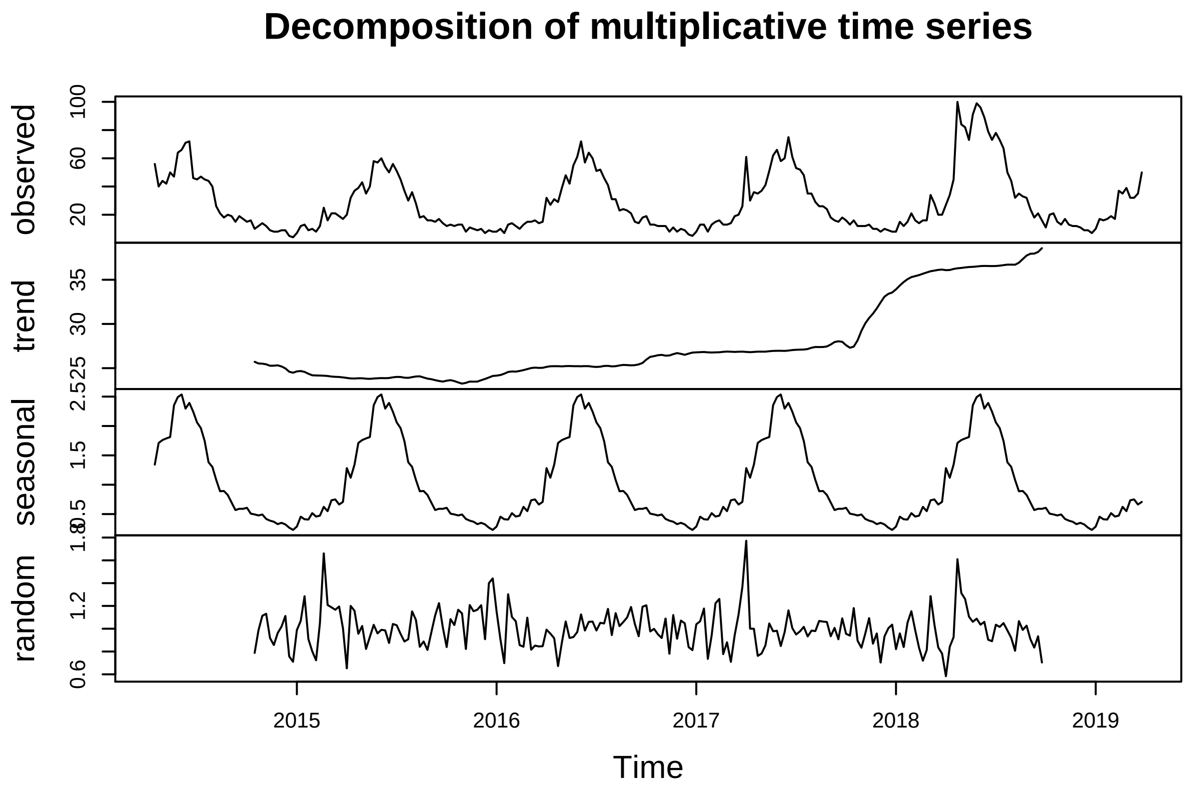
```
##       time                FSMEgt          zeckengt           rki
##  Min.   :2014-04-20   Min.   :  4.00   Min.   :  4.00   Min.   : 0.000
##  1st Qu.:2015-07-13   1st Qu.: 13.00   1st Qu.:  7.00   1st Qu.: 1.000
##  Median :2016-10-05   Median : 19.00   Median : 14.00   Median : 4.000
##  Mean   :2016-10-05   Mean   : 28.39   Mean   : 24.64   Mean   : 7.717
##  3rd Qu.:2017-12-29   3rd Qu.: 40.00   3rd Qu.: 39.00   3rd Qu.:11.750
##  Max.   :2019-03-24   Max.   :100.00   Max.   :100.00   Max.   :54.000
```

```r
cor.test(masterall$FSMEgt, masterall$rki, method="kendall")
```

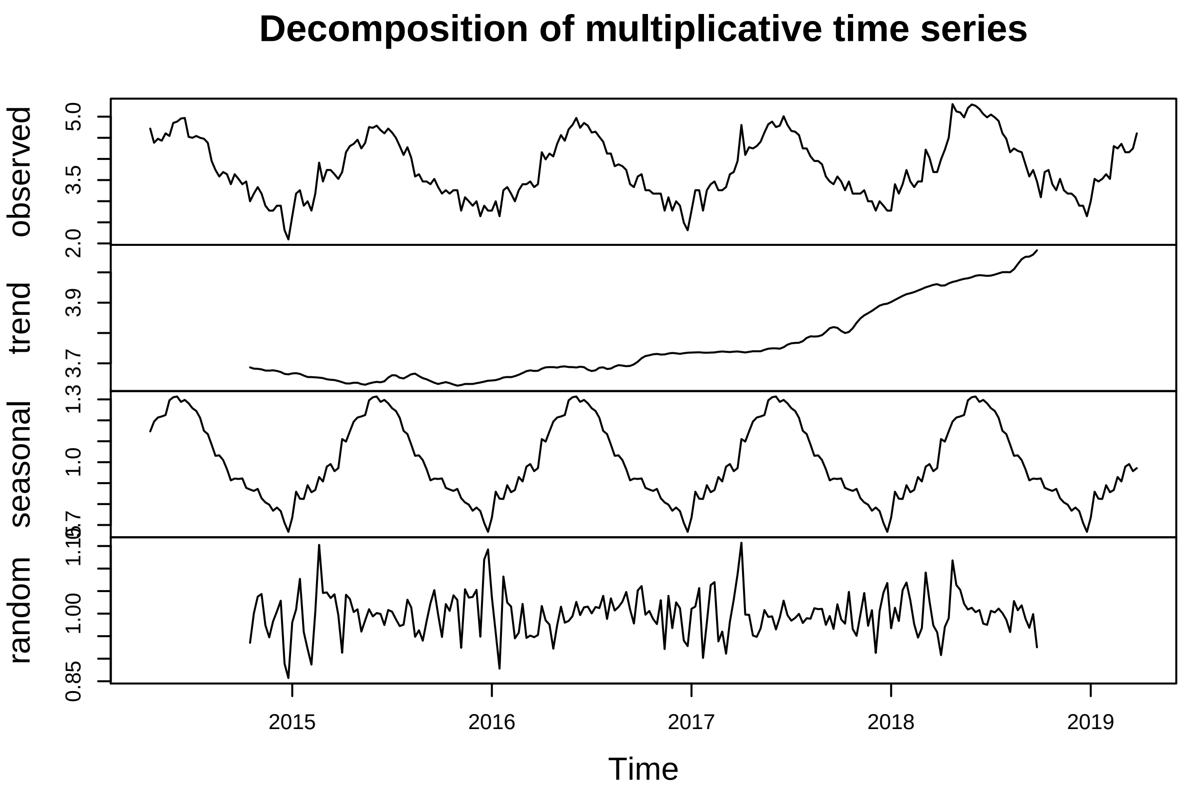
```
##
##  Kendall's rank correlation tau
##
## data:  masterall$FSMEgt and masterall$rki
## z = 7.064, p-value = 1.618e-12
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.3084665
```

```r
library(forecast)

ihs <- function(x) {
  y <- log(x + sqrt(x ^ 2 + 1))
  return(y)
}


hs <- function(x) {
  y <- 0.5*exp(-x)*(exp(2*x)-1)
  return(y)
}


gts <- ts( ihs(masterall$FSMEgt), start=c(2014, 16), end=c(2019, 13), frequency=52)

rkts <- ts( ihs(masterall$rki) , start=c(2014, 16), end=c(2019, 13), frequency=52)
```

```r
gt<-ts( masterall$FSMEgt, start=c(2014, 16), end=c(2019, 13), frequency=52)
decompose_gt <- decompose(gt, "multiplicative")
plot(decompose_gt)
```




```r
gtseasonal<-as.ts(decompose_gt$seasonal)
gttrend<-as.ts(decompose_gt$trend)
gtrandom<-as.ts(decompose_gt$random)

gts <- ts( ihs(masterall$FSMEgt), start=c(2014, 16), end=c(2019, 13), frequency=52)
decompose_gts <- decompose(gts, "multiplicative")
plot(decompose_gts)
```




```
gtseasonal<-as.ts(decompose_gts$seasonal)
gtstrend<-as.ts(decompose_gts$trend)
gtsrandom<-as.ts(decompose_gts$random)
#lets remove the random component
gts<-gts-gtsrandom

summary(gts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.238   2.198   2.602   2.758   3.355   4.275      52
```

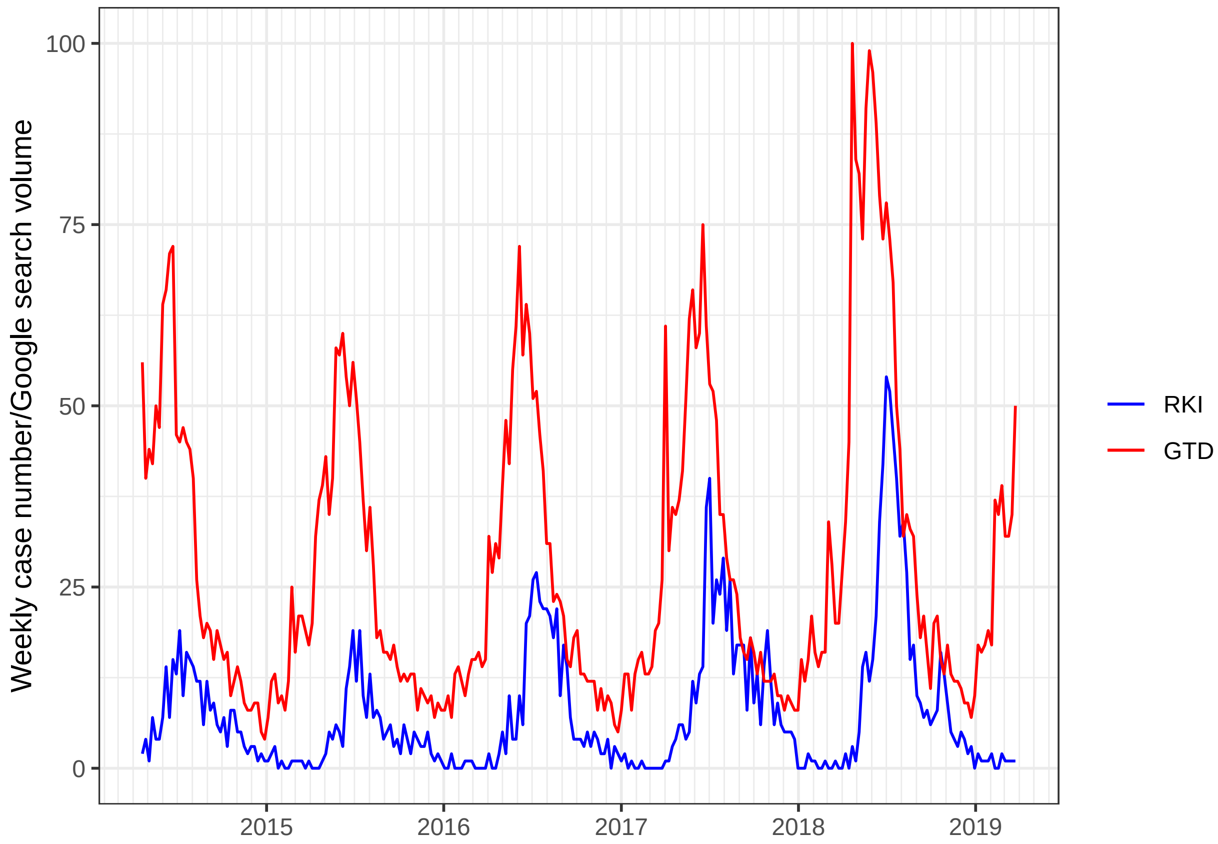
```
gts17<-ts(gts[208:258], start=c(2018, 13), end=c(2019, 13), frequency=52)
gts<-ts(gts[1:207], start=c(2014, 16), end=c(2018, 12), frequency=52)

rkts17<-ts(rkts[208:258], start=c(2018, 13), end=c(2019, 13), frequency=52)
rkts<-ts(rkts[1:207], start=c(2014, 16), end=c(2018, 12), frequency=52)

summary(rkts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.8814  2.0947  1.9695  2.9982  4.3822
```

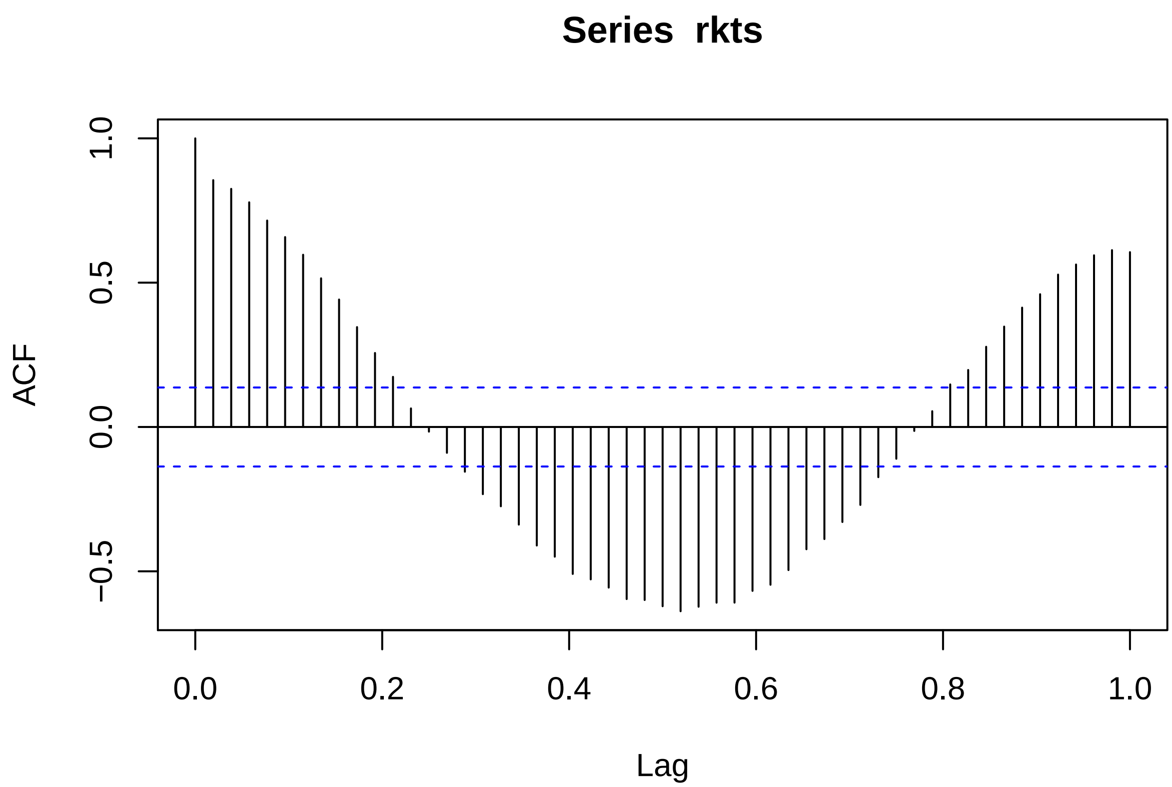
```
summary(rkts17)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.444   2.312   2.348   3.402   4.682
```

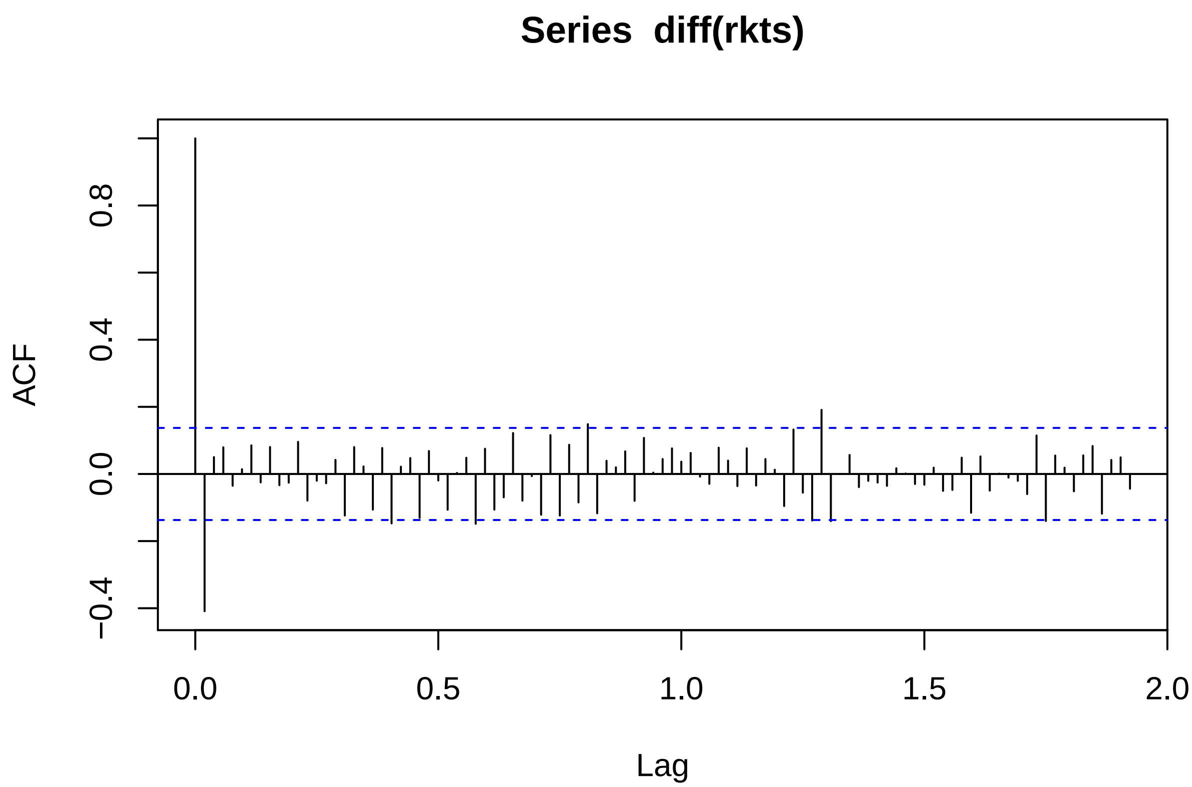
```r
library(ggplot2)

ggplot(data = masterall, aes(x = time)) +
  geom_line(aes(y = rki, colour = "RKI")) +
  geom_line(aes(y = FSMEgt, colour = "GTD")) +
  scale_colour_manual("",
                      breaks = c("RKI", "GTD"),
                      values = c("red", "blue")) +
  scale_x_date(date_minor_breaks = "1 month") +
  xlab(NULL) +
  ylab("Weekly case number/Google search volume") +
  theme_bw()
```



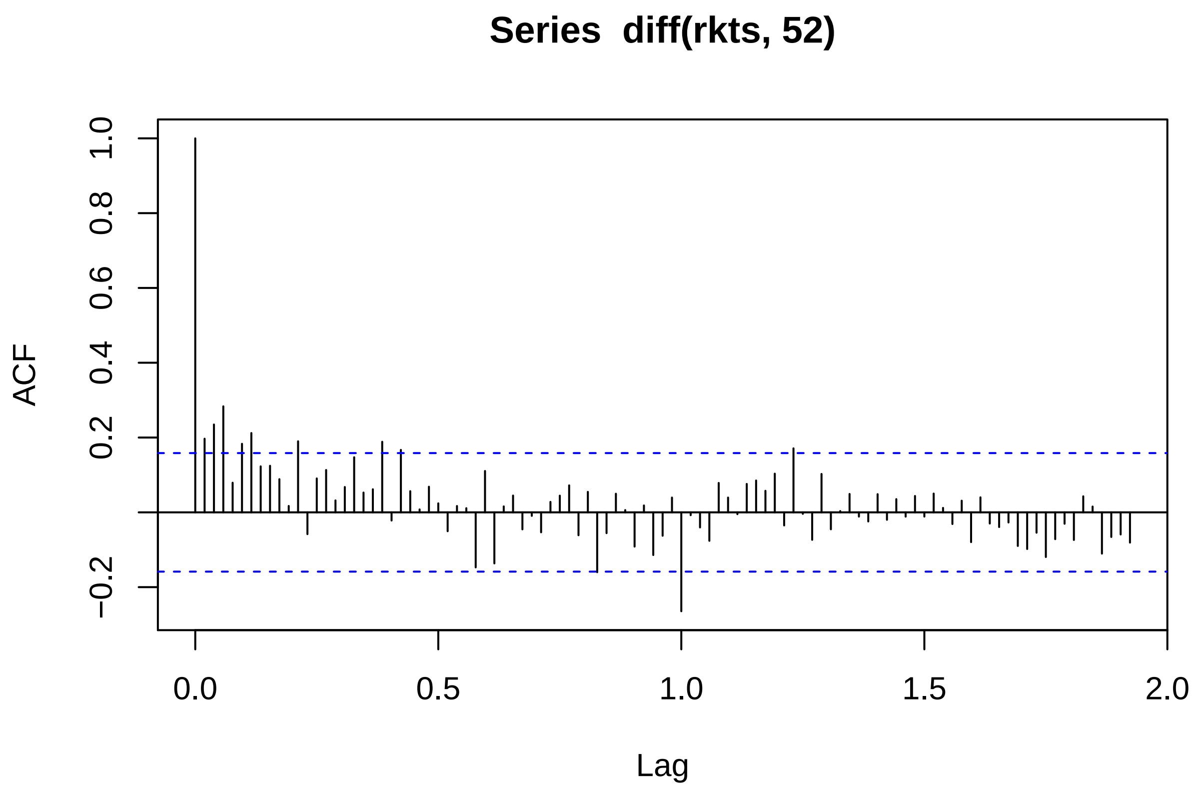

```r
acf( rkts, lag.max = 52 )
```



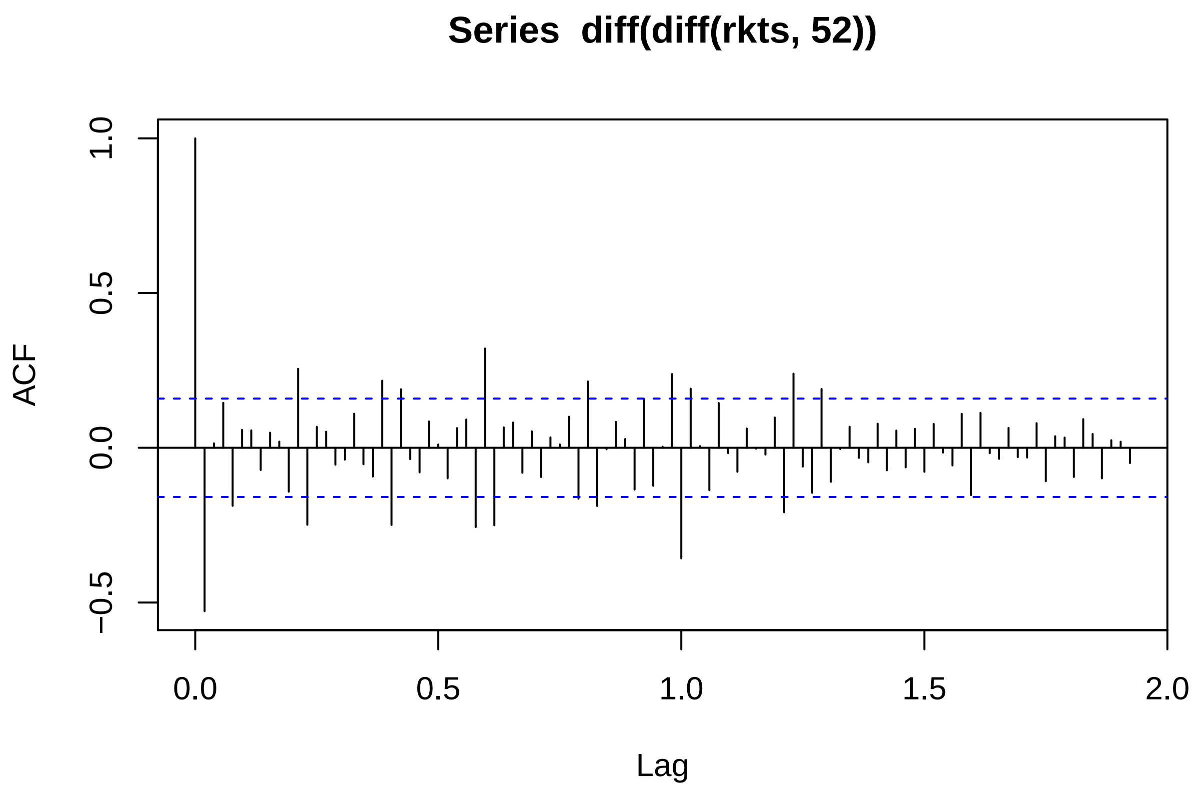

```r
acf( diff( rkts ), lag.max = 100 )
```




```
acf( diff( rkts, 52 ), lag.max = 100 )
```




```
acf( diff( diff( rkts, 52 ) ), lag.max = 100 )
```




```r
fit <- auto.arima( rkts, trace = TRUE, approximation = FALSE )
```

```
##
##  ARIMA(2,1,2)(1,1,1)[52]                    : 314.3923
##  ARIMA(0,1,0)(0,1,0)[52]                    : 412.4703
##  ARIMA(1,1,0)(1,1,0)[52]                    : 338.1085
##  ARIMA(0,1,1)(0,1,1)[52]                    : Inf
##  ARIMA(2,1,2)(0,1,1)[52]                    : Inf
##  ARIMA(2,1,2)(1,1,0)[52]                    : 314.3731
##  ARIMA(1,1,2)(1,1,0)[52]                    : 312.6514
##  ARIMA(1,1,1)(1,1,0)[52]                    : Inf
##  ARIMA(1,1,3)(1,1,0)[52]                    : Inf
##  ARIMA(0,1,1)(1,1,0)[52]                    : 308.5394
##  ARIMA(0,1,1)(0,1,0)[52]                    : 330.3721
##  ARIMA(0,1,1)(1,1,1)[52]                    : 308.3871
##  ARIMA(1,1,1)(1,1,1)[52]                    : Inf
##  ARIMA(0,1,0)(1,1,1)[52]                    : 382.3484
##  ARIMA(0,1,2)(1,1,1)[52]                    : 310.5246
##  ARIMA(1,1,2)(1,1,1)[52]                    : Inf
##
##  Best model: ARIMA(0,1,1)(1,1,1)[52]
```

```r
fit
```

```
## Series: rkts
## ARIMA(0,1,1)(1,1,1)[52]
##
```

```
## Coefficients:
##           ma1     sar1     sma1
##       -0.8262  -0.1086  -0.5085
## s.e.   0.0549   0.2846   0.3923
##
## sigma^2 estimated as 0.3709:  log likelihood=-150.06
## AIC=308.12   AICc=308.39   BIC=320.21
```

```
fc <- forecast( fit, h = length( gts17 ) )
accuracy( fc )
```

```
##                      ME      RMSE       MAE MPE MAPE      MASE        ACF1
## Training set 0.03280024 0.5192268 0.3539646 NaN  Inf 0.5976872 0.007367345
```

```
accuracy( fc, rkts17 )
```

```
##                      ME      RMSE       MAE  MPE MAPE      MASE
## Training set 0.03280024 0.5192268 0.3539646  NaN  Inf 0.5976872
## Test set     0.17811228 0.7126147 0.5901539 -Inf  Inf 0.9965047
##                     ACF1 Theil's U
## Training set 0.007367345        NA
## Test set     0.306890505         0
```

```
fitGT <- auto.arima( rkts, trace = TRUE, approximation = FALSE, xreg = gts )
```

```
##
##  ARIMA(2,1,2)(1,1,1)[52]                    : Inf
##  ARIMA(0,1,0)(0,1,0)[52]                    : 351.0963
##  ARIMA(1,1,0)(1,1,0)[52]                    : 291.3424
##  ARIMA(0,1,1)(0,1,1)[52]                    : 272.553
##  ARIMA(0,1,1)(1,1,1)[52]                    : Inf
##  ARIMA(0,1,1)(0,1,0)[52]                    : 287.93
##  ARIMA(1,1,1)(0,1,1)[52]                    : 274.4624
##  ARIMA(0,1,0)(0,1,1)[52]                    : Inf
##  ARIMA(0,1,2)(0,1,1)[52]                    : 274.5185
##  ARIMA(1,1,2)(0,1,1)[52]                    : 276.7895
##
##  Best model: Regression with ARIMA(0,1,1)(0,1,1)[52] errors
```

```
fitGT
```

```
## Series: rkts
## Regression with ARIMA(0,1,1)(0,1,1)[52] errors
##
## Coefficients:
##           ma1     sma1    xreg
##       -0.8061  -0.5060  0.1007
## s.e.   0.0657   0.1514  0.3432
##
## sigma^2 estimated as 0.3345:  log likelihood=-132.14
## AIC=272.28   AICc=272.55   BIC=284.38
```

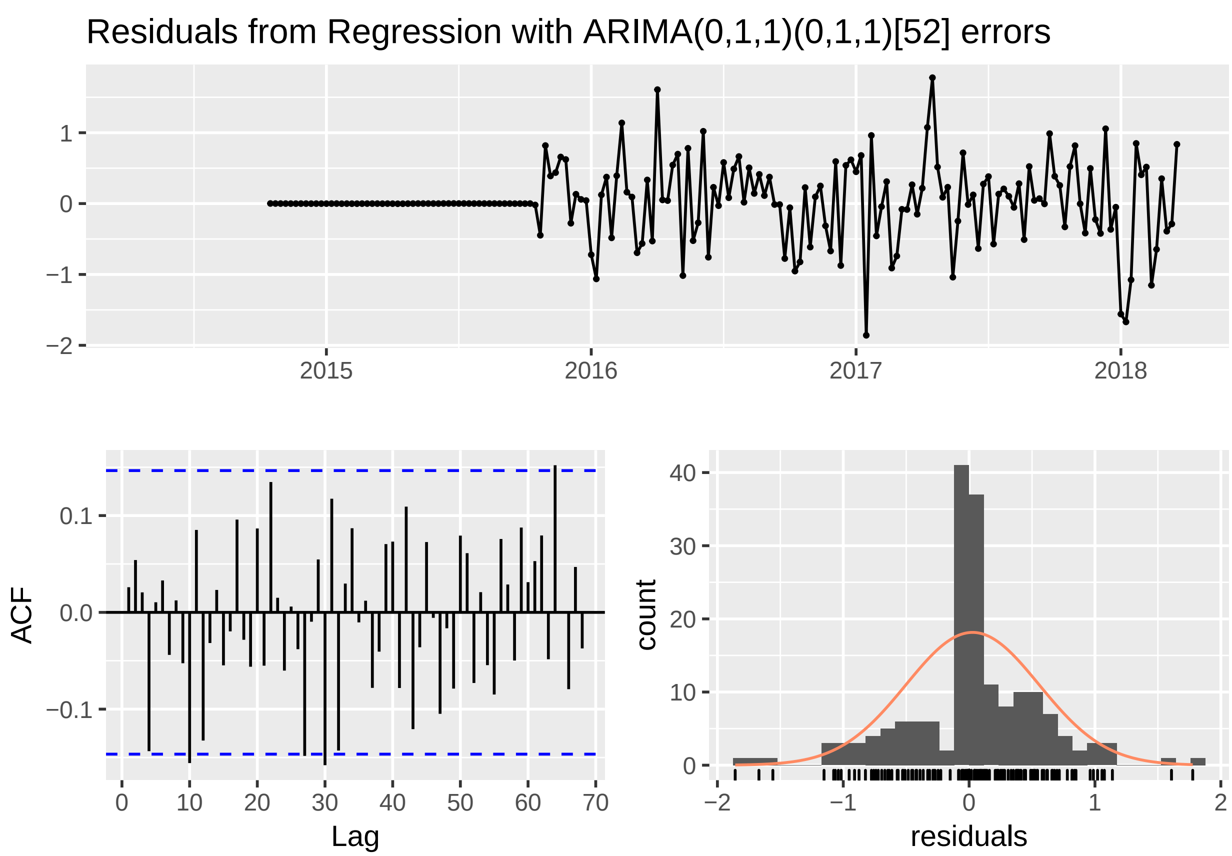
```
fcGT<- forecast( fitGT, xreg = gts17 )
```

```
## Warning in forecast.Arima(fitGT, xreg = gts17): Upper prediction intervals
## are not finite.
```

```r
accuracy( fcGT, rkts17 )
```

```
##                      ME      RMSE       MAE  MPE MAPE     MASE       ACF1
## Training set 0.02717721 0.5277001 0.3458416  NaN  Inf 0.583971 0.02598171
## Test set     0.28974285 0.7378966 0.6552872 -Inf  Inf 1.106486 0.31099206
##              Theil's U
## Training set        NA
## Test set             0
```

```r
checkresiduals(fitGT)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,1)(0,1,1)[52] errors
## Q* = 51.577, df = 38, p-value = 0.06977
##
## Model df: 3.   Total lags used: 41
```

```r
#AIC improved , but forecasting is somewhat worse
#now lets see without Trend decomp. data:
gts <- ts( ihs(masterall$FSMEgt), start=c(2014, 16), end=c(2019, 13), frequency=52)
gt<-ts( masterall$FSMEgt, start=c(2014, 16), end=c(2019, 13), frequency=52)

#lets remove the random component
gts<-gts-(gtsrandom+gtstrend)
gts17<-ts(gts[208:258], start=c(2018, 13), end=c(2019, 13), frequency=52)
```

```r
gts<-ts(gts[1:207], start=c(2014, 16), end=c(2018, 12), frequency=52)

fitGT <- auto.arima( rkts, trace = TRUE, approximation = FALSE, xreg = gts )
```

```
##
##  ARIMA(2,1,2)(1,1,1)[52]                    : Inf
##  ARIMA(0,1,0)(0,1,0)[52]                    : 351.1109
##  ARIMA(1,1,0)(1,1,0)[52]                    : 291.3428
##  ARIMA(0,1,1)(0,1,1)[52]                    : 272.4788
##  ARIMA(0,1,1)(1,1,1)[52]                    : Inf
##  ARIMA(0,1,1)(0,1,0)[52]                    : 287.9054
##  ARIMA(1,1,1)(0,1,1)[52]                    : 274.3932
##  ARIMA(0,1,0)(0,1,1)[52]                    : Inf
##  ARIMA(0,1,2)(0,1,1)[52]                    : 274.4468
##  ARIMA(1,1,2)(0,1,1)[52]                    : 276.7121
##
##  Best model: Regression with ARIMA(0,1,1)(0,1,1)[52] errors
```

```r
fitGT
```

```
## Series: rkts
## Regression with ARIMA(0,1,1)(0,1,1)[52] errors
##
## Coefficients:
##           ma1     sma1    xreg
##       -0.8073  -0.5067  0.1376
## s.e.   0.0655   0.1516  0.3438
##
## sigma^2 estimated as 0.3342:  log likelihood=-132.1
## AIC=272.21   AICc=272.48   BIC=284.3
```

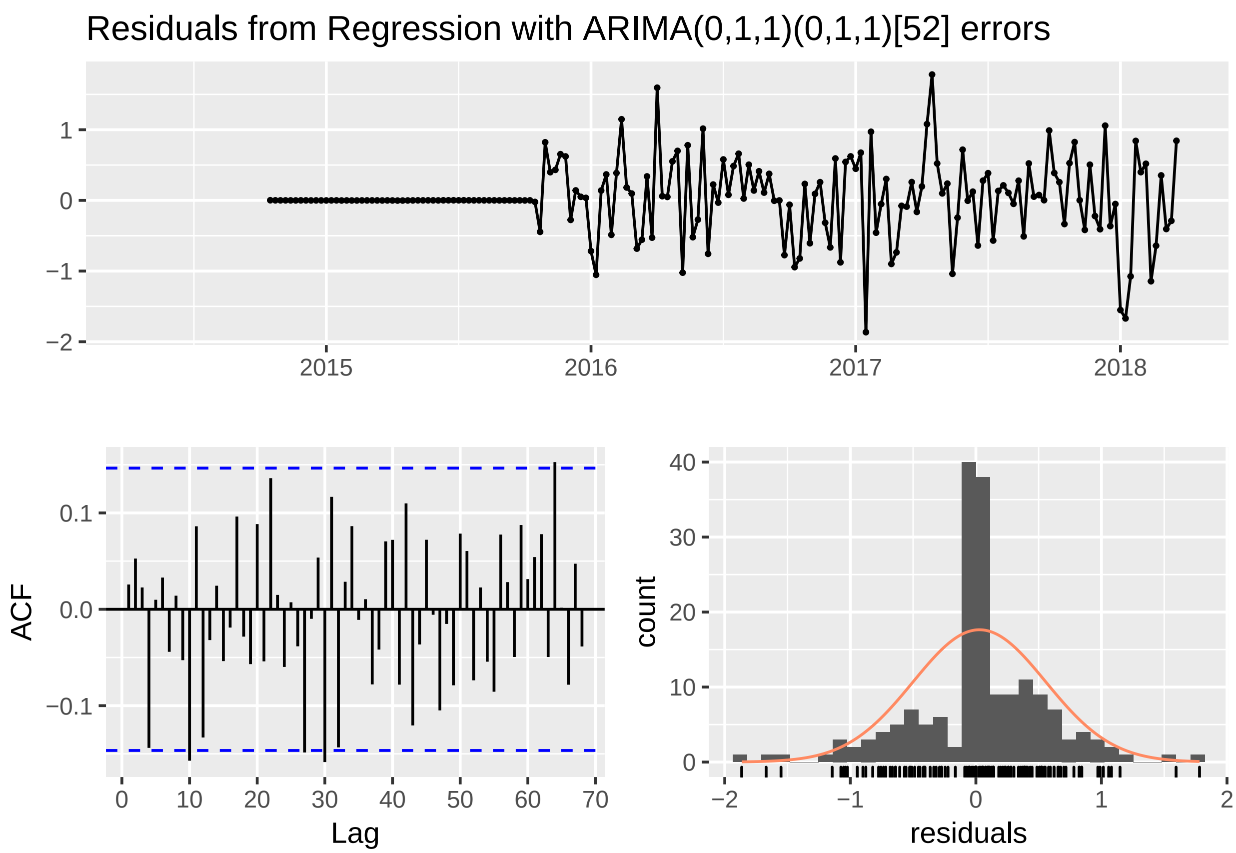
```r
fcGT<- forecast( fitGT, xreg = gts17 )
```

```
## Warning in forecast.Arima(fitGT, xreg = gts17): Upper prediction intervals
## are not finite.
```

```r
accuracy( fcGT, rkts17 )
```

```
##                      ME      RMSE       MAE  MPE MAPE      MASE       ACF1
## Training set 0.02818823 0.5274497 0.3459788  NaN  Inf 0.5842027 0.02572708
## Test set     0.29178412 0.7339938 0.6514540 -Inf  Inf 1.1000131 0.30945779
##              Theil's U
## Training set        NA
## Test set             0
```

```r
checkresiduals(fitGT)
```



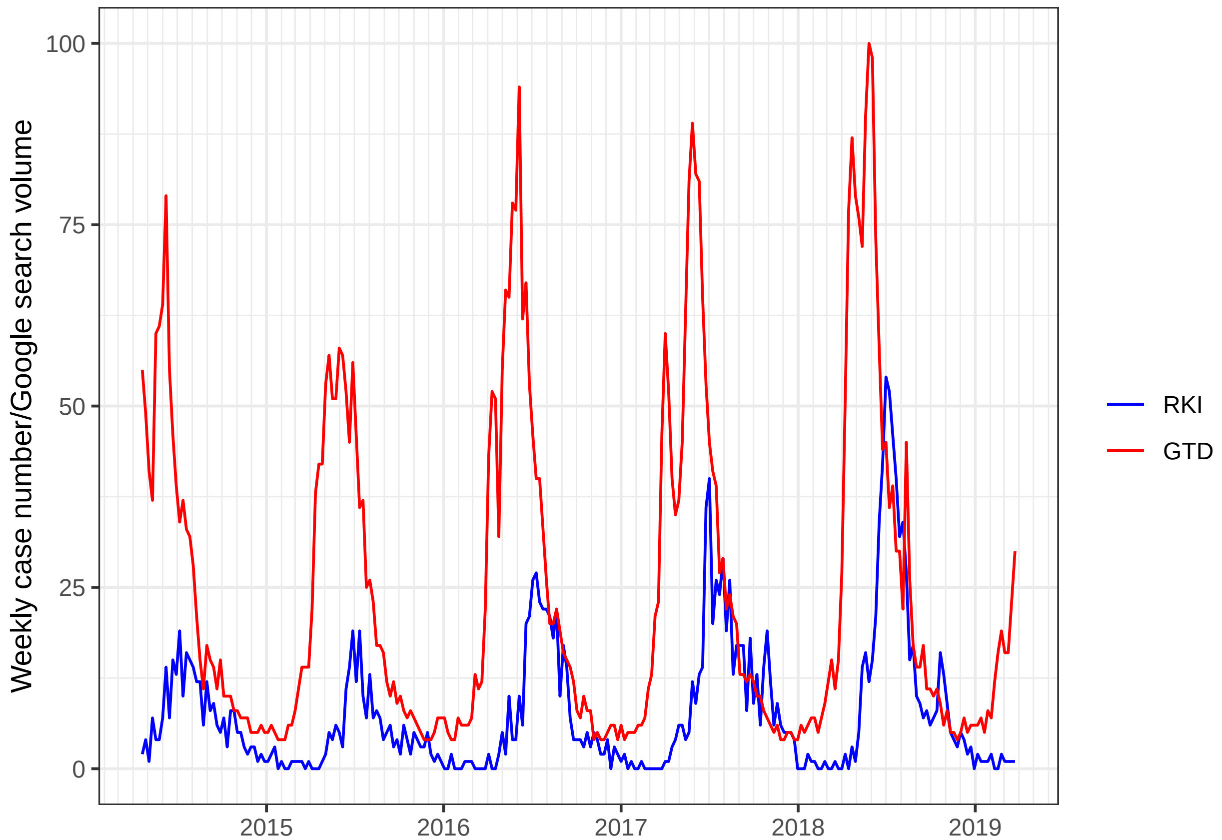

```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,1)(0,1,1)[52] errors
## Q* = 51.853, df = 38, p-value = 0.06638
##
## Model df: 3.   Total lags used: 41
```

```
#so AIC is the same as previously, but prediction is worse than previously- so we probably need the non


# now according to the suggestion lets use something less specifical Google correlate-maybe it will pro
#the problem is that we cannot download any searches to that intervall anymore.

gts <- ts( ihs(masterall$zeckengt), start=c(2014, 16), end=c(2019, 13), frequency=52)


summary(gts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.095   2.644   3.333   3.453   4.357   5.298
```

```
gts17<-ts(gts[208:258], start=c(2018, 13), end=c(2019, 13), frequency=52)
gts<-ts(gts[1:207], start=c(2014, 16), end=c(2018, 12), frequency=52)
```

```r
library(ggplot2)

ggplot(data = masterall, aes(x = time)) +
  geom_line(aes(y = rki, colour = "RKI")) +
  geom_line(aes(y = zeckengt, colour = "GTD")) +
  scale_colour_manual("",
                      breaks = c("RKI", "GTD"),
                      values = c("red", "blue")) +
  scale_x_date(date_minor_breaks = "1 month") +
  xlab(NULL) +
  ylab("Weekly case number/Google search volume") +
  theme_bw()
```




```r
#seems pretty much the same
fitGT <- auto.arima( rkts, trace = TRUE, approximation = FALSE, xreg = gts )
```

```
##
##  ARIMA(2,1,2)(1,1,1)[52]                    : 316.3835
##  ARIMA(0,1,0)(0,1,0)[52]                    : 414.4073
##  ARIMA(1,1,0)(1,1,0)[52]                    : 339.3594
##  ARIMA(0,1,1)(0,1,1)[52]                    : Inf
##  ARIMA(2,1,2)(0,1,1)[52]                    : Inf
##  ARIMA(2,1,2)(1,1,0)[52]                    : 316.4028
##  ARIMA(2,1,2)(0,1,0)[52]                    : 336.4264
##  ARIMA(1,1,2)(1,1,1)[52]                    : 313.1249
##  ARIMA(1,1,1)(1,1,1)[52]                    : Inf
```

```
##  ARIMA(1,1,3)(1,1,1)[52]                    : Inf
##  ARIMA(0,1,1)(1,1,1)[52]                    : 310.2383
##  ARIMA(0,1,1)(1,1,0)[52]                    : 310.4284
##  ARIMA(0,1,1)(0,1,0)[52]                    : 332.3927
##  ARIMA(0,1,0)(1,1,1)[52]                    : 384.24
##  ARIMA(0,1,2)(1,1,1)[52]                    : 312.4067
##
##  Best model: Regression with ARIMA(0,1,1)(1,1,1)[52] errors
```

```r
fitGT
```

```
## Series: rkts
## Regression with ARIMA(0,1,1)(1,1,1)[52] errors
##
## Coefficients:
##           ma1     sar1     sma1     xreg
##       -0.8243  -0.1129  -0.5093  -0.1163
## s.e.   0.0545   0.2808   0.3875   0.2168
##
## sigma^2 estimated as 0.372:  log likelihood=-149.91
## AIC=309.83   AICc=310.24   BIC=324.95
```

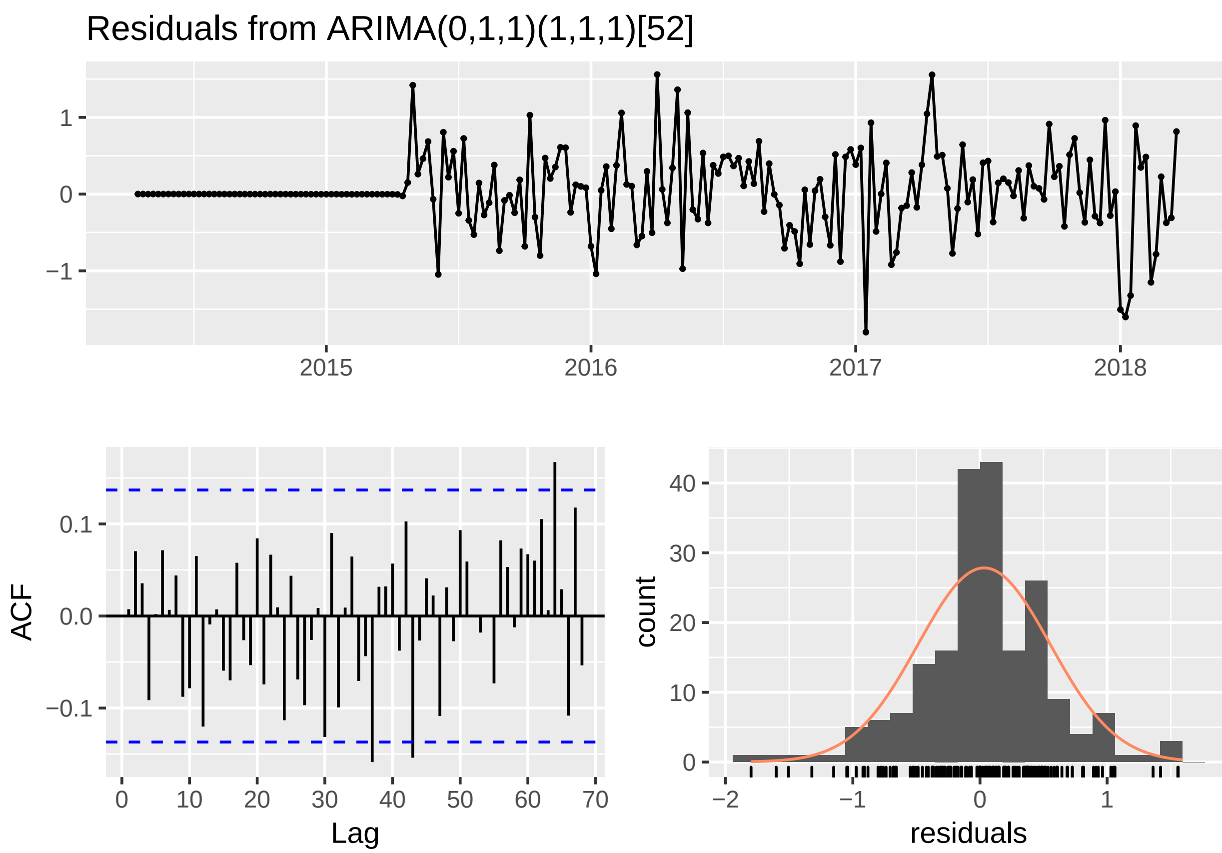
```r
fcGT<- forecast( fitGT, xreg = gts17 )


# result:
accuracy( fc, rkts17 )
```

```
##                      ME      RMSE       MAE  MPE MAPE      MASE
## Training set 0.03280024 0.5192268 0.3539646  NaN  Inf 0.5976872
## Test set     0.17811228 0.7126147 0.5901539 -Inf  Inf 0.9965047
##                     ACF1 Theil's U
## Training set 0.007367345        NA
## Test set     0.306890505         0
```

```r
accuracy( fcGT, rkts17 )
```

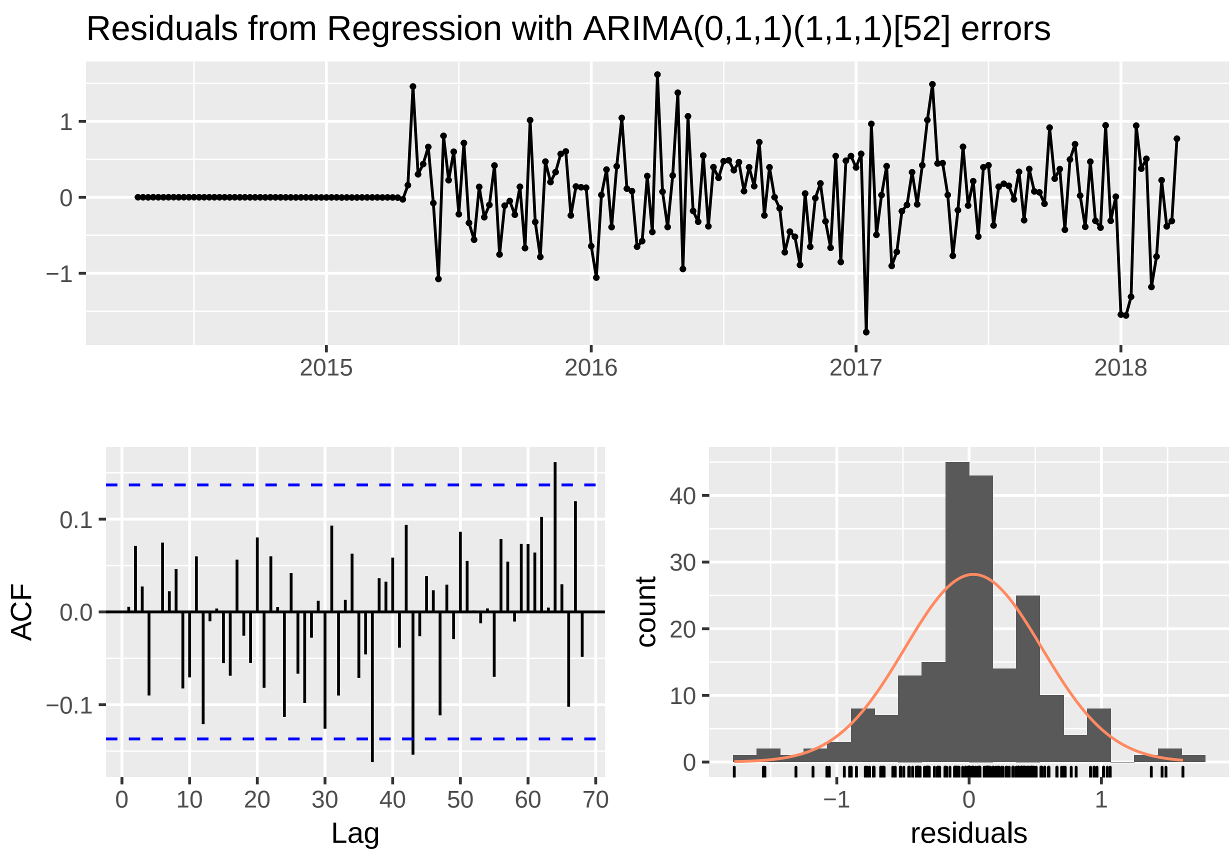
```
##                      ME      RMSE       MAE  MPE MAPE      MASE       ACF1
## Training set 0.03269403 0.5182138 0.3528404  NaN  Inf 0.5957889 0.00549601
## Test set     0.21271942 0.7270932 0.6024214 -Inf  Inf 1.0172190 0.31088510
##              Theil's U
## Training set        NA
## Test set             0
```

```r
checkresiduals(fit)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)(1,1,1)[52]
## Q* = 45.557, df = 38, p-value = 0.1866
##
## Model df: 3.   Total lags used: 41
```

```
checkresiduals(fitGT)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(0,1,1)(1,1,1)[52] errors
## Q* = 44.668, df = 37, p-value = 0.1807
##
## Model df: 4.   Total lags used: 41
```

```
fitGT ## xreg is insignificant
```

```
## Series: rkts
## Regression with ARIMA(0,1,1)(1,1,1)[52] errors
##
## Coefficients:
##           ma1      sar1      sma1      xreg
##       -0.8243   -0.1129   -0.5093   -0.1163
## s.e.   0.0545    0.2808    0.3875    0.2168
##
## sigma^2 estimated as 0.372:  log likelihood=-149.91
## AIC=309.83   AICc=310.24   BIC=324.95
```

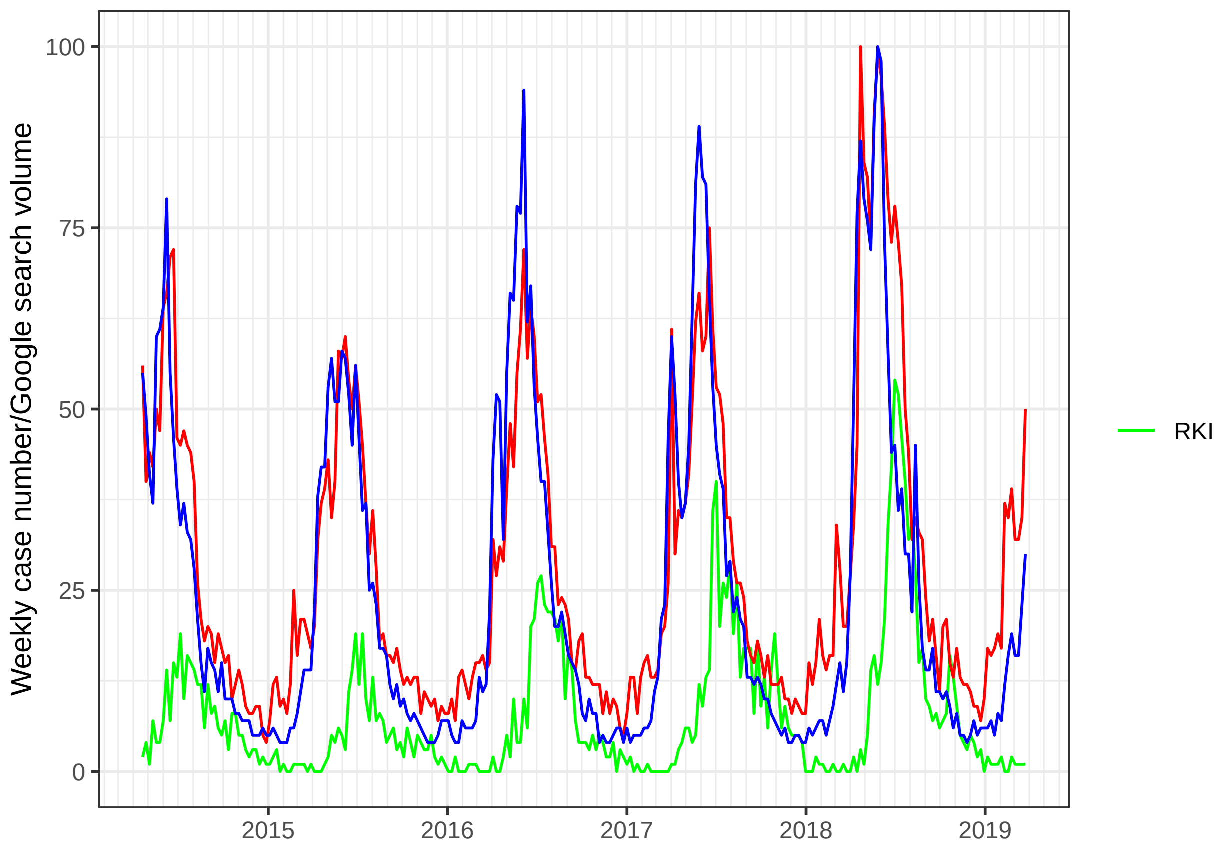
```
a<-AIC( fit, fitGT ) ## fit has a favorable AIC
a
```

```
##       df      AIC
## fit    4 308.1150
## fitGT  5 309.8274
```

```r
dm.test( rkts17-fc$mean, rkts17-fcGT$mean )
```

```
##
##  Diebold-Mariano Test
##
## data:  rkts17 - fc$meanrkts17 - fcGT$mean
## DM = -1.6787, Forecast horizon = 1, Loss function power = 2,
## p-value = 0.09922
## alternative hypothesis: two.sided
```

```r
# so AIC and prediction is worse with the more "robust" searching term

#if we add the FSME -term also (the original)

gtsf <- ts( ihs(masterall$FSMEgt), start=c(2014, 16), end=c(2019, 13), frequency=52)
gtsz <- ts( ihs(masterall$zeckengt), start=c(2014, 16), end=c(2019, 13), frequency=52)

gts17f<-ts(gtsf[208:258], start=c(2018, 13), end=c(2019, 13), frequency=52)
gtsf<-ts(gtsf[1:207], start=c(2014, 16), end=c(2018, 12), frequency=52)
gts17z<-ts(gtsz[208:258], start=c(2018, 13), end=c(2019, 13), frequency=52)
gtsz<-ts(gtsz[1:207], start=c(2014, 16), end=c(2018, 12), frequency=52)





library(ggplot2)

ggplot(data = masterall, aes(x = time)) +
  geom_line(aes(y = rki, colour = "RKI")) +
  geom_line(aes(y = FSMEgt, colour = "GTD TBE")) +
  geom_line(aes(y = zeckengt, colour = "GTD Tick")) +
  scale_colour_manual("",
                      breaks = c("RKI", "GTD"),
                      values = c("red", "blue", "green")) +
  scale_x_date(date_minor_breaks = "1 month") +
  xlab(NULL) +
  ylab("Weekly case number/Google search volume") +
  theme_bw()
```




```
gts<-cbind(gtsf, gtsz)
gts17<-cbind(gts17f, gts17z)
fitGT <- auto.arima( rkts, trace = TRUE, approximation = FALSE, xreg = gts )
```

```
##
##  ARIMA(2,1,2)(1,1,1)[52]                    : 317.4088
##  ARIMA(0,1,0)(0,1,0)[52]                    : 416.3767
##  ARIMA(1,1,0)(1,1,0)[52]                    : 341.4217
##  ARIMA(0,1,1)(0,1,1)[52]                    : Inf
##  ARIMA(2,1,2)(0,1,1)[52]                    : Inf
##  ARIMA(2,1,2)(1,1,0)[52]                    : 317.6819
##  ARIMA(2,1,2)(0,1,0)[52]                    : 337.9827
##  ARIMA(1,1,2)(1,1,1)[52]                    : Inf
##  ARIMA(3,1,2)(1,1,1)[52]                    : 318.1365
##  ARIMA(2,1,1)(1,1,1)[52]                    : Inf
##  ARIMA(2,1,3)(1,1,1)[52]                    : Inf
##  ARIMA(1,1,1)(1,1,1)[52]                    : Inf
##  ARIMA(3,1,3)(1,1,1)[52]                    : Inf
##
##  Best model: Regression with ARIMA(2,1,2)(1,1,1)[52] errors
```

```
fitGT
```

```
## Series: rkts
## Regression with ARIMA(2,1,2)(1,1,1)[52] errors
##
## Coefficients:
```

```
##           ar1      ar2     ma1      ma2     sar1     sma1    gtsf     gtsz
##       -0.9672  -0.0591  0.1296  -0.7549  -0.0753  -0.5607  0.2757  -0.2069
## s.e.   0.1056   0.1153  0.0775   0.0791   0.3987   0.5622  0.2475   0.2334
##
## sigma^2 estimated as 0.3735:  log likelihood=-149.07
## AIC=316.14   AICc=317.41   BIC=343.36
```

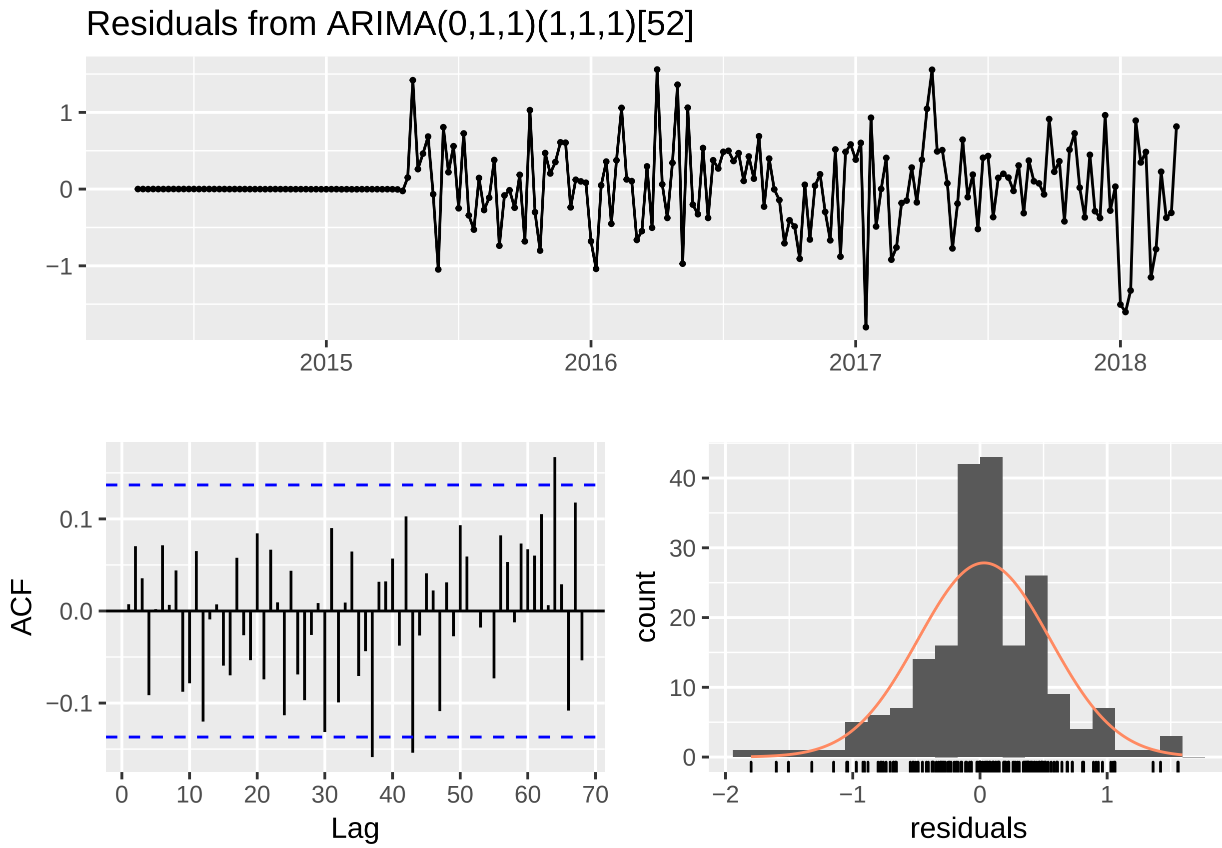
```
fcGT<- forecast( fitGT, xreg = gts17)


# result:
accuracy( fc, rkts17 )
```

```
##                      ME      RMSE       MAE  MPE MAPE      MASE
## Training set 0.03280024 0.5192268 0.3539646  NaN  Inf 0.5976872
## Test set     0.17811228 0.7126147 0.5901539 -Inf  Inf 0.9965047
##                     ACF1 Theil's U
## Training set 0.007367345        NA
## Test set     0.306890505         0
```

```
accuracy( fcGT, rkts17 )
```

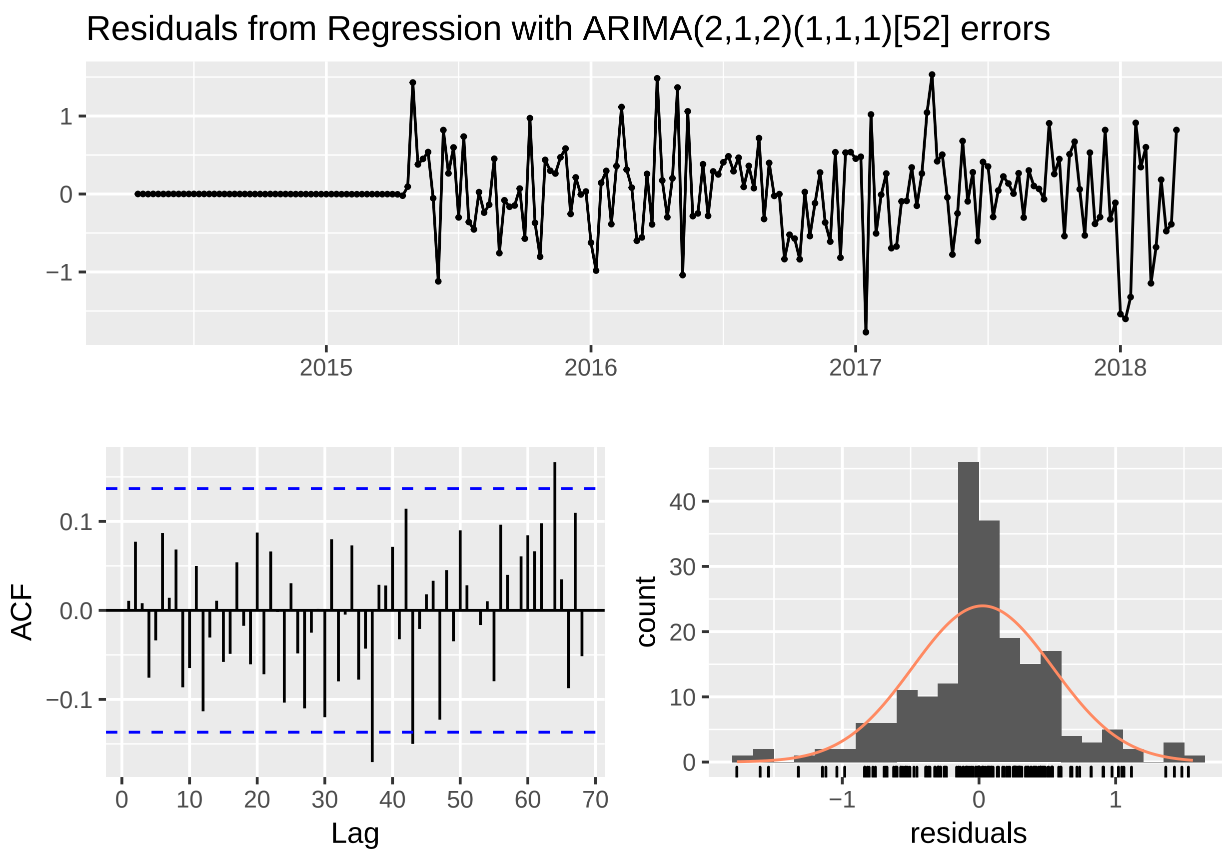
```
##                      ME      RMSE       MAE  MPE MAPE      MASE       ACF1
## Training set 0.02682056 0.5121786 0.3475712  NaN  Inf 0.5868916 0.01074179
## Test set     0.23723261 0.7102719 0.5893764 -Inf  Inf 0.9951919 0.28997753
##              Theil's U
## Training set        NA
## Test set             0
```

```
checkresiduals(fit)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(0,1,1)(1,1,1)[52]
## Q* = 45.557, df = 38, p-value = 0.1866
##
## Model df: 3.   Total lags used: 41
```

```
checkresiduals(fitGT)
```




```
##
##  Ljung-Box test
##
## data:  Residuals from Regression with ARIMA(2,1,2)(1,1,1)[52] errors
## Q* = 43.892, df = 33, p-value = 0.0974
##
## Model df: 8.   Total lags used: 41
```

```
fitGT
```

```
## Series: rkts
## Regression with ARIMA(2,1,2)(1,1,1)[52] errors
##
## Coefficients:
##           ar1      ar2     ma1      ma2      sar1     sma1    gtsf     gtsz
##       -0.9672  -0.0591  0.1296  -0.7549  -0.0753  -0.5607  0.2757  -0.2069
## s.e.   0.1056   0.1153  0.0775   0.0791   0.3987   0.5622  0.2475   0.2334
##
## sigma^2 estimated as 0.3735:  log likelihood=-149.07
## AIC=316.14   AICc=317.41   BIC=343.36
```

```
a<-AIC( fit, fitGT )
a
```

```
##       df      AIC
## fit    4 308.1150
## fitGT  9 316.1412
```

```r
dm.test( rkts17-fc$mean, rkts17-fcGT$mean )
```

```
##
##  Diebold-Mariano Test
##
## data:  rkts17 - fc$meanrkts17 - fcGT$mean
## DM = 0.18236, Forecast horizon = 1, Loss function power = 2,
## p-value = 0.856
## alternative hypothesis: two.sided
```

```r
#•AIC is worse , prediction is also inferior, so it seems adding more "robust" external regressors does
```