KJW update_3.10

2022-03-10

Overview

Sample analysis of Ofenbach's HSKT stream count over multiple countries. DV is Stream Count because Chart Ranking is conflated with other factors. Descriptive analysis revealed a trend in this song (since it is a dance song, then it peaks in East European countries first).

Step 1: Data Overivew

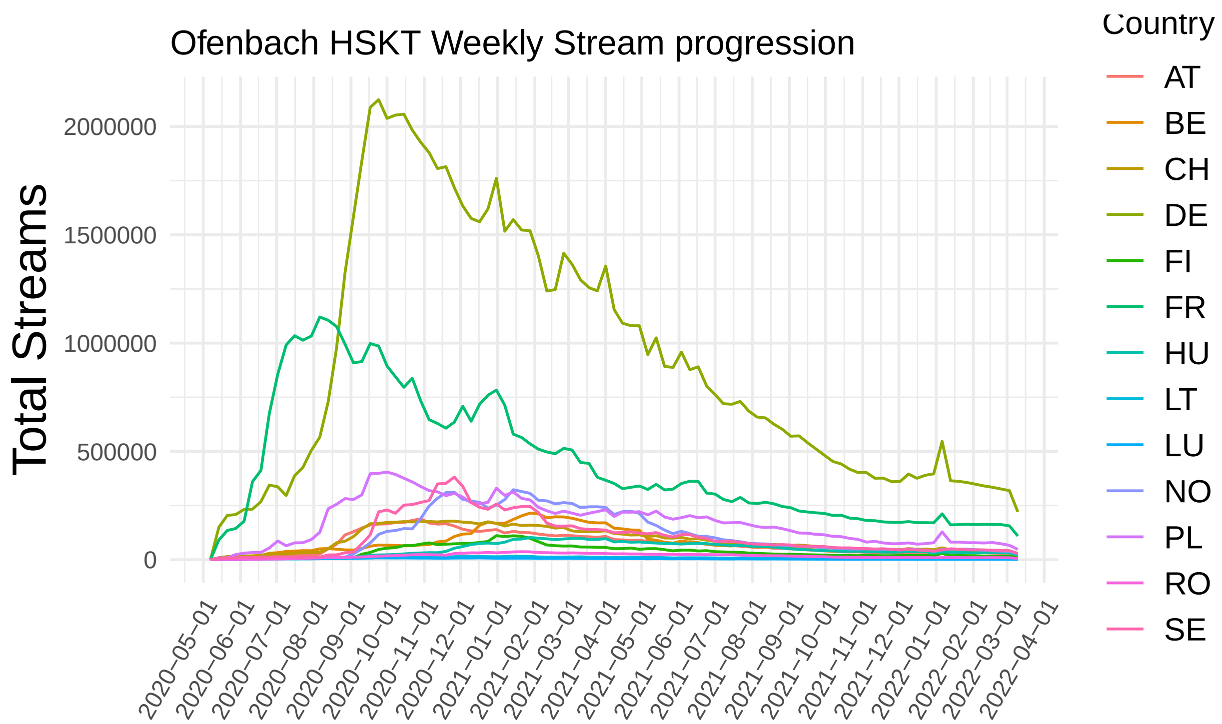
Re-shape data of weekly Ofenbach HSKT Spotify streams, so each row is a date. Each column is the weekly streams, by country. Sample of the data frame:

```
library(tidyverse)
charts <- read_tsv('/cloud/project/raw/weekly_offennbach.tsv')
charts_total <- charts %>%
  filter(COUNTRY_CODE %in% c("FR", "LU", "LT", "DE", "PL", "BE",
                             "CH",  "AT", "WW", "RO", "NO", "HU", "SE", "FI")) %>%
  filter(PRODUCT_TITLE == "Head Shoulders Knees & Toes (feat. Norma Jean Martine)") %>%
  select(COUNTRY_CODE, TOTAL_STREAMS, DATE_KEY)
## Step 1A: reshape
test <- charts_total %>%
  select(TOTAL_STREAMS, COUNTRY_CODE, DATE_KEY) %>%
  group_by_at(vars(-TOTAL_STREAMS)) %>%
  dplyr::mutate(row_id = 1:n()) %>%
  ungroup() %>%
  spread(key = COUNTRY_CODE, value = TOTAL_STREAMS)
test[is.na(test)] = 0
head(test)
```

```
## # A tibble: 6 x 15
##   DATE_KEY   row_id    AT    BE    CH     DE    FI     FR    HU    LT    LU
##   <date>      <int> <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 2020-05-07      1    27    14    39    524    33    317     4     0     2
## 2 2020-05-14      1  7989  3987  9240 150045  6011  91169  2487   555   369
## 3 2020-05-21      1 11096  6902 14553 204561  2856 134973  3089   824   841
## 4 2020-05-28      1 11778  7098 15692 208395  2719 144075  4409  1251   817
## 5 2020-06-04      1 12666  9360 17327 232271  2898 177827  4781  1857   921
## 6 2020-06-11      1 13498 15502 17845 233380  4605 360321  5286  2619  1671
## # ... with 4 more variables: NO <dbl>, PL <dbl>, RO <dbl>, SE <dbl>
```

Step 2: Pairwise Country Visualizations

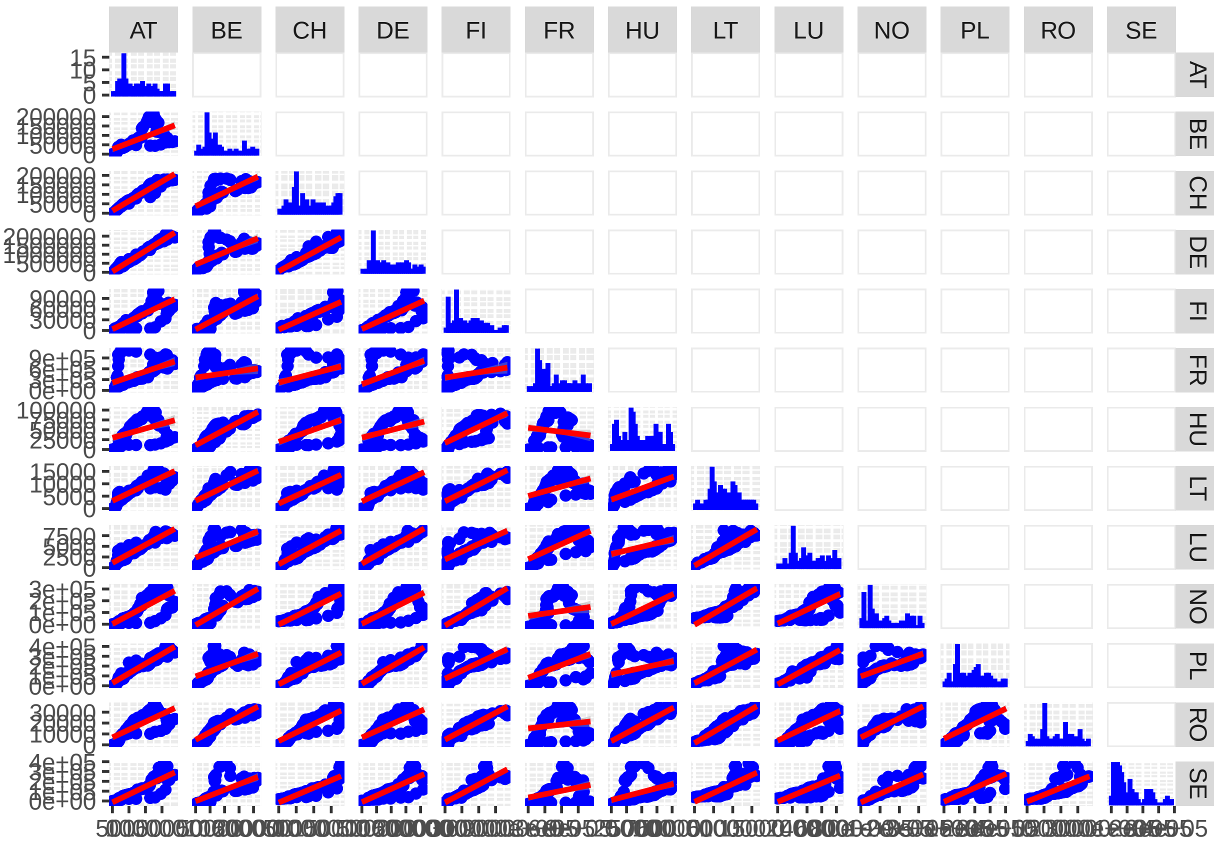
For all countries, visualize the pattern of stream count. FR peaks before the rest, as the artist is from France, then Luxembourg, Lithuania, Germany, Poland, Belgium, Switzerland, Austria. Pattern of development across Western Europe and into Eastern Europe, then Scandinavia, before global chart.



Stream count on Spotify

HSNT hits WW charts (global) in week 16

Next, visualize the pairwise comparisons of each country. Is there a relationship between pairs of countries and their vectors of stream counts over time?



Step 3: Pairwise Country Covariance and Autocorrelation Charts

Covariance/Correlation of the Stream

For one song, we have the vector of streams for country A and country B. Covariance and correlation is the measure of dependence between the respective country variances, given by:

$$Cov[X, y] = \frac{\sum(X_i - \bar{X})(Y_j - \bar{Y})}{n-1}$$

and Correlation is a standardized measure of that Covariance, given by:

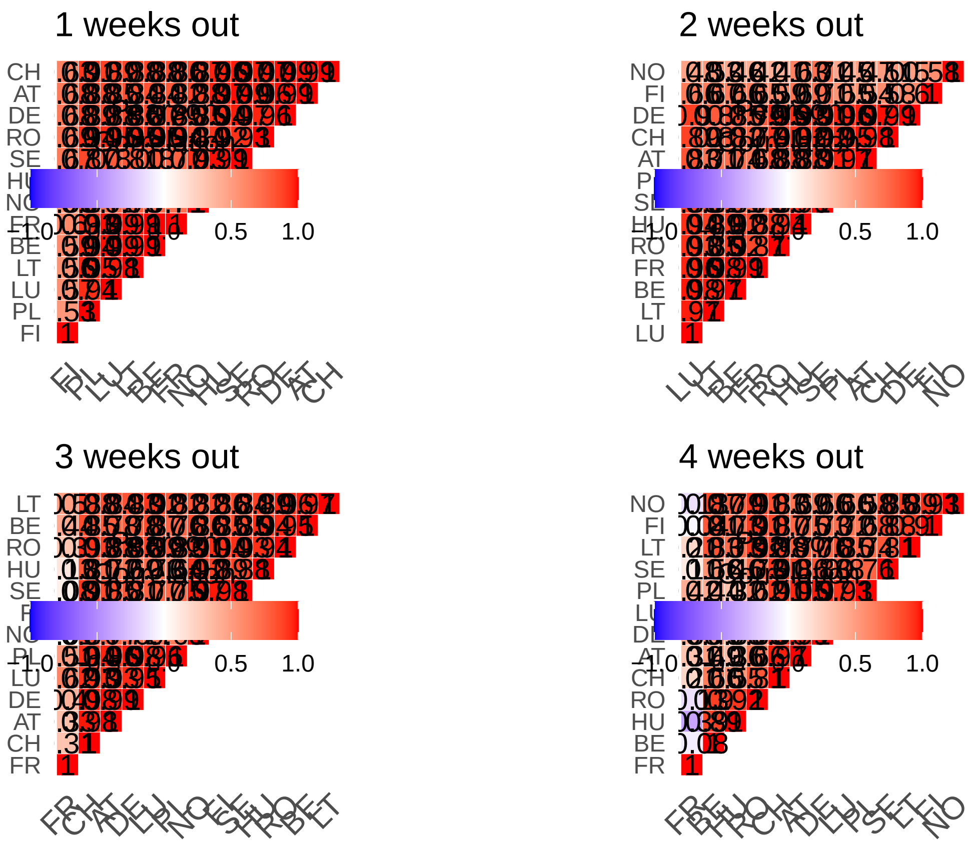
$$Corr[X, Y] = Cov[X, Y]/\sqrt{Var[X]Var[Y]}$$

Covariance matrix below. Since covariance is not standardized, this is difficult to interpret.

```
##             AT          BE          CH           DE          FI           FR
## AT  2583955150  1741985112  2667378423  29479084655  1155033172   8069252063
## BE  1741985112  3163387170  2239453689  20334671945  1348533382   3735872127
## CH  2667378423  2239453689  2899799599  30428475168  1361756715   7335949064
## DE 29479084655 20334671945 30428475168 342394503516 12992639295 105119115745
## FI  1155033172  1348533382  1361756715  12992639295   801815691   2000829563
## FR  8069252063  3735872127  7335949064 105119115745  2000829563  88820397780
## HU   610649136  1440707345   902886371   6565096678   616521311  -1549291897
## LT   165222259   189077906   189797785   1903965417   100142374    559189950
## LU   109240120    85999315   115209265   1285612179    50077156    489018269
## NO  3869305773  5044056814  4660314519  43266625785  2672385385   5700135606
```

```
## PL  5059968449  3335768891  5197367838  58799030623  2190356872  18886609134
## RO   368910818   508306430   451584205   4129382601   258186895    496682726
## SE  4172678959  3374827877  4609049331  46453533750  2361948689   9452044191
##             HU         LT         LU          NO          PL         RO
## AT   610649136  165222259  109240120  3869305773  5059968449  368910818
## BE  1440707345  189077906   85999315  5044056814  3335768891  508306430
## CH   902886371  189797785  115209265  4660314519  5197367838  451584205
## DE  6565096678 1903965417 1285612179 43266625785 58799030623 4129382601
## FI   616521311  100142374   50077156  2672385385  2190356872  258186895
## FR -1549291897  559189950  489018269  5700135606 18886609134  496682726
## HU   893173271   81070744   27042784  2303247549  1175088724  267392979
## LT    81070744   14987720    7925533   350116277   329138745   35465356
## LU    27042784    7925533    5205732   169337759   223697840   16600081
## NO  2303247549  350116277  169337759  9957118091  7026097537  911369021
## PL  1175088724  329138745  223697840  7026097537 10646069666  725817961
## RO   267392979   35465356   16600081   911369021   725817961   98936411
## SE  1342128187  304019807  175444189  8156231861  7741787447  713380528
##             SE
## AT  4172678959
## BE  3374827877
## CH  4609049331
## DE 46453533750
## FI  2361948689
## FR  9452044191
## HU  1342128187
## LT   304019807
## LU   175444189
## NO  8156231861
## PL  7741787447
## RO   713380528
## SE  8610408239
```

Correlation matrix is easier to interpret:



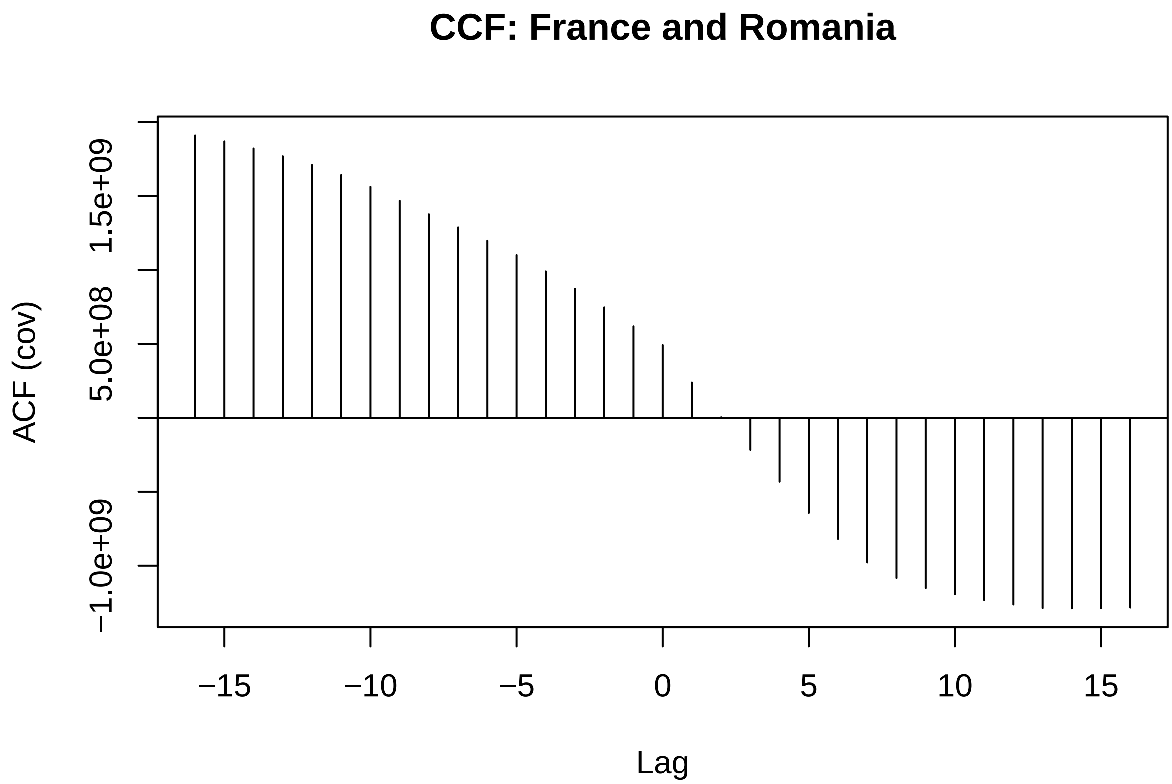
Step 4: Cross-Covariance Function

Cross-Covariance Function

The CCF identifies lags of the x-variable (a predictor country at time t) that might be useful predictors of y_t (the predicted country at time t). The sample CCF is the set of sample covariances between x_{t+h} and y_t for lags (or h's) =0, +-1, +-2, etc. Negative value for h is a covariance between the x variable at a time before t and the y variable at time t. When h=-2, then the CCF gives the covariance between X_{t-2}, the streams of the predicted country at 2 lags behind time t, and y_t, the streams of the predicted country at time t.

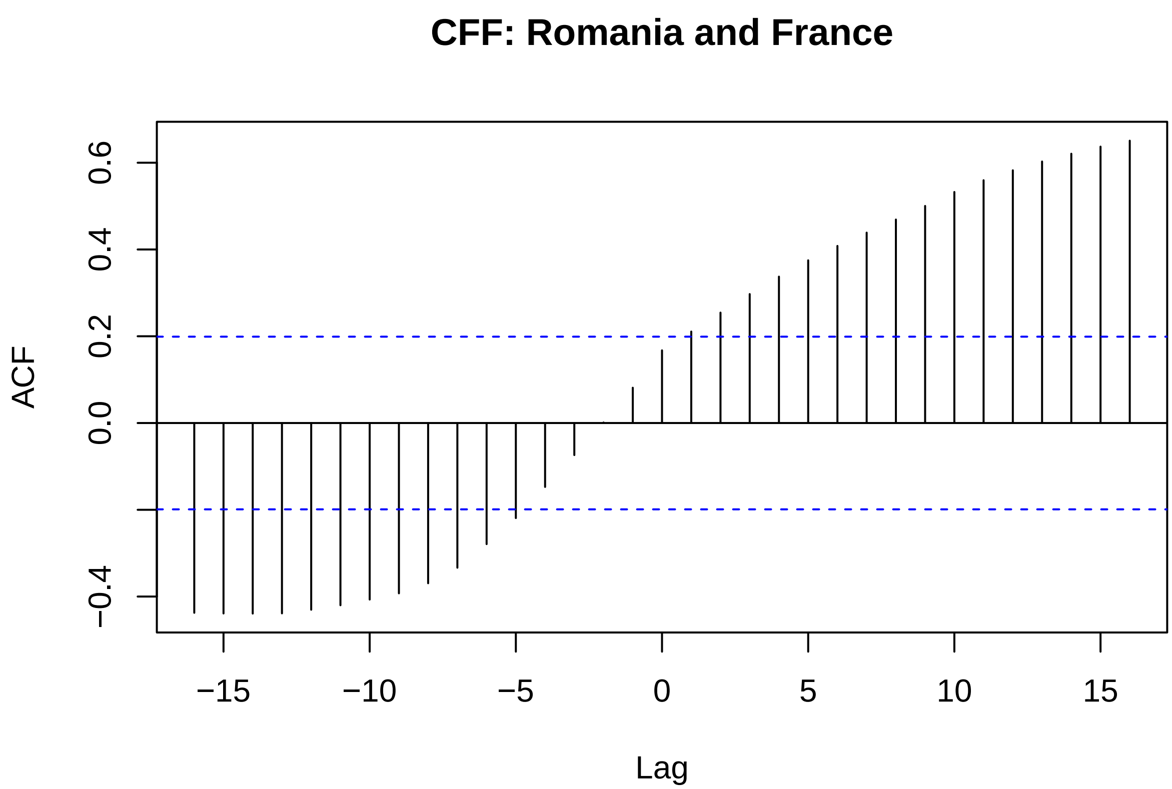
$$CCF(X_t, Y_t)$$

We know from visualizations that Romania is going to lag France. Let's confirm it with these CCF plots. The most dominant cross covariances occur at h=-15 to -10. The maximum correlations in this region are positive, indicating that an above average value of FR streams is likely to lead to an above average value of RO streams, and that this will be realized at lag -15 to lag -10. We see negative covariances at future lags, but these would not make sence to interpret, since the structure is not seasonal. Since many x_{t+h}, with h negative, are predictors of y_t, means that x leads y, or FR leads RO.




```
##
## Autocovariances of series 'X', by lag
##
##       -16       -15       -14       -13       -12       -11       -10        -9
##  1.91e+09  1.87e+09  1.82e+09  1.77e+09  1.71e+09  1.64e+09  1.56e+09  1.47e+09
##        -8        -7        -6        -5        -4        -3        -2        -1
##  1.38e+09  1.29e+09  1.20e+09  1.10e+09  9.90e+08  8.72e+08  7.47e+08  6.19e+08
##         0         1         2         3         4         5         6         7
##  4.92e+08  2.39e+08  3.65e+06 -2.17e+08 -4.32e+08 -6.43e+08 -8.19e+08 -9.79e+08
##         8         9        10        11        12        13        14        15
## -1.08e+09 -1.15e+09 -1.19e+09 -1.23e+09 -1.26e+09 -1.29e+09 -1.29e+09 -1.29e+09
##        16
## -1.28e+09
```

If you switch, then does Romania predict France? When one or more x_{t+h}, with h positive, are predictors of y_t, then x lags y. Did this one with correlation, just to make more interpretable.




```
##
## Autocorrelations of series 'X', by lag
##
##    -16    -15    -14    -13    -12    -11    -10     -9     -8     -7     -6
## -0.437 -0.439 -0.439 -0.439 -0.430 -0.420 -0.407 -0.393 -0.369 -0.334 -0.279
##     -5     -4     -3     -2     -1      0      1      2      3      4      5
## -0.219 -0.147 -0.074  0.001  0.081  0.168  0.211  0.254  0.297  0.337  0.375
##      6      7      8      9     10     11     12     13     14     15     16
##  0.408  0.439  0.469  0.500  0.533  0.560  0.583  0.603  0.621  0.637  0.651
```

Step 5: Regression Models

DE and FR would be predictive of RO, since RO lags, but RO not predictive of FR, since FR leads.

```
model1 <- lm(RO ~ FR + DE, data = test)
summary(model1)
```

```
##
## Call:
## lm(formula = RO ~ FR + DE, data = test)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14644  -3005    318   4854  10870
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.982e+03  1.262e+03   7.118 2.15e-10 ***
## FR          -1.364e-02  2.695e-03  -5.059 2.08e-06 ***
## DE           1.625e-02  1.373e-03  11.834  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6280 on 94 degrees of freedom
## Multiple R-squared:  0.6096, Adjusted R-squared:  0.6013
## F-statistic:  73.4 on 2 and 94 DF,  p-value: < 2.2e-16
```

```r
model1 <- lm(FR ~ RO , data = test)
summary(model1)
```

```
##
## Call:
## lm(formula = FR ~ RO, data = test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -359067 -216257 -152065  172452  718500
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.593e+05  6.098e+04   5.893 5.75e-08 ***
## RO          5.020e+00  3.031e+00   1.656    0.101
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 295400 on 95 degrees of freedom
## Multiple R-squared:  0.02807,    Adjusted R-squared:  0.01784
## F-statistic: 2.744 on 1 and 95 DF,  p-value: 0.1009
```

Step 6: Autocorrelation Function

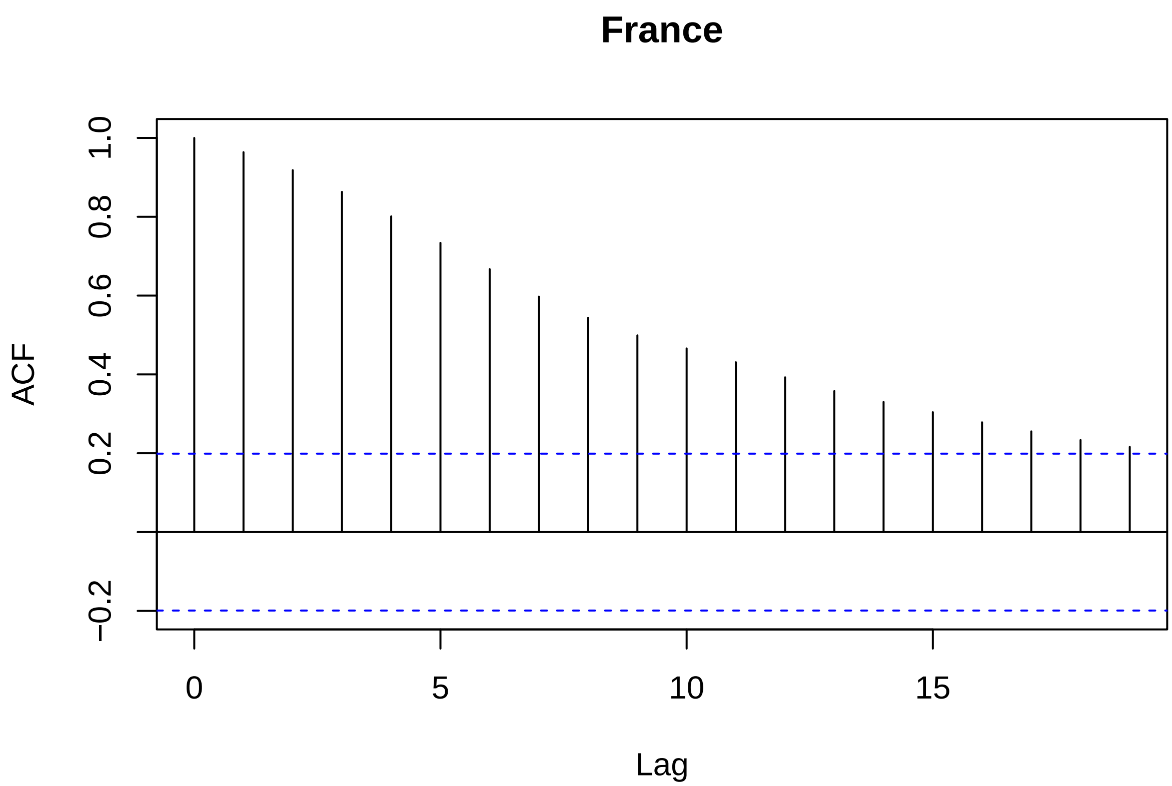
Auto Correlation Function

This models the outcome variable and prior versions of itself

$$ACF(Y_t, T_t y)$$

Most of the countries look like this, since there is no seasonality in stream data. Rather, it has an initial spike from popularity peak.

```r
acf(ts(test[8]), main = "France")
```




```
acf(ts(test[14]), main = "RO")
```