Comparing Models using Kolmogorov-Smirnov Test

Dean Gladish June 18, 2018

First test: Kolmogorov-Smirnov.

```
install.packages("ks.test", repos="http://R-Forge.R-project.org")
```

Given the data, I want to use the Kolmogorov-Smirnov test to evaluate the maximum distances between each of the 21 models. To do this, I want to create a matrix that shows the K-S test statistic between all of them.

For every one of the weeks 1 to 52 except for weeks 19 to 42 inclusive, the model makes a prediction for each week up to and including 4 weeks from that week. It also tries to predict the week for which the flu season begins, the peak percentage of flu incidence for this season, and the week during which the flu incidence peaks.

What I want to do is take each pair of the 21 models and run Kolmogorov-Smirnov tests between them for their predictions for weeks 1, 2, 3, and 4. I will do this by setting each of the 52 weeks to be the present week and thus derive 28 (skipping weeks 19 to 42) K-S test statistics for each pair. For each pair I shall generate a scatterplot of the test statistics and present them in a grid.

This simplified data will serve as a starting point for future tests.

```
# Stores the names of all the models in a vector.
modelsVector <- c("CU_EAKFC_SEIRS", "CU_EAKFC_SIRS", "CU_EKF_SEIRS", "CU_EKF_SIRS", "CU_RHF_SEIRS", "CU
```

It looks like we are comparing 22 models. Here I have created a vector to store their names and another vector to store the data.

```
# I create another vector to store the Kolmogorov-Smirnov test statistics for week 1.

KS_Stats_Week_1 <- numeric(choose(22, 2))

sysGlobNames <- numeric(22)

for (i in 1:22) {
    sysGlobNames[i] <- paste("C:/Users/gladi/Documents/GitHub/cdc-flusight-ensemble/model-forecasts/real-}

CU_EAKFC_SEIRS_Files <- lapply(Sys.glob(sysGlobNames[1]), read.csv)

CU_EAKFC_SEIRS_Files_Vector <- vector("list", length(CU_EAKFC_SEIRS_Files)) }

for (i in 1:length(CU_EAKFC_SEIRS_Files)) {
    # Removes all entries where Location is not US National

CU_EAKFC_SEIRS_Files[[i]] <- CU_EAKFC_SEIRS_Files[[i]][!(CU_EAKFC_SEIRS_Files[[i]]$Location != "US Na"

# Removes all entries where Type is not equal to Point

CU_EAKFC_SEIRS_Files[[i]] <- CU_EAKFC_SEIRS_Files[[i]][!(CU_EAKFC_SEIRS_Files[[i]]$Type != "Point"),]

# Now that we're removed some data that I have deemed to be extraneous for this demo test, I

# will now extract the first four rows that represent the predicted flu incidences for the weeks ahea

CU_EAKFC_SEIRS_Files_Vector[[i]] <- CU_EAKFC_SEIRS_Files[[i]][1:4,]$Value
}
```

CU_EAKFC_SIRS_Files <- lapply(Sys.glob(sysGlobNames[2]), read.csv)</pre>

CU_EAKFC_SIRS_Files_Vector <- vector("list", length(CU_EAKFC_SIRS_Files))</pre>

```
for (i in 1:length(CU_EAKFC_SIRS_Files)) {
   CU_EAKFC_SIRS_Files[[i]] <- CU_EAKFC_SIRS_Files[[i]][!(CU_EAKFC_SIRS_Files[[i]]$Location != "US Nation")
   CU_EAKFC_SIRS_Files[[i]] <- CU_EAKFC_SIRS_Files[[i]][!(CU_EAKFC_SIRS_Files[[i]]$Type != "Point"),]
   CU_EAKFC_SIRS_Files_Vector[[i]] <- CU_EAKFC_SIRS_Files[[i]][1:4,]$Value
}
CU_EKF_SEIRS_Files <- lapply(Sys.glob(sysGlobNames[3]), read.csv)</pre>
CU EKF SEIRS Files Vector <- vector("list", length(CU EKF SEIRS Files))
for (i in 1:length(CU EKF SEIRS Files)) {
   CU_EKF_SEIRS_Files[[i]] <- CU_EKF_SEIRS_Files[[i]][!(CU_EKF_SEIRS_Files[[i]]$Location != "US National
   CU_EKF_SEIRS_Files[[i]] <- CU_EKF_SEIRS_Files[[i]][!(CU_EKF_SEIRS_Files[[i]]$Type != "Point"),]
   CU_EKF_SEIRS_Files_Vector[[i]] <- CU_EKF_SEIRS_Files[[i]][1:4,]$Value
}
CU_EKF_SIRS_Files <- lapply(Sys.glob(sysGlobNames[4]), read.csv)
CU_EKF_SIRS_Files_Vector <- vector("list", length(CU_EKF_SIRS_Files))</pre>
for (i in 1:length(CU_EKF_SIRS_Files)) {
   CU_EKF_SIRS_Files[[i]] <- CU_EKF_SIRS_Files[[i]][!(CU_EKF_SIRS_Files[[i]]$Location != "US National"),
   CU_EKF_SIRS_Files[[i]] <- CU_EKF_SIRS_Files[[i]][!(CU_EKF_SIRS_Files[[i])$Type != "Point"),]</pre>
   CU_EKF_SIRS_Files_Vector[[i]] <- CU_EKF_SIRS_Files[[i]][1:4,]

Value
}
CU_RHF_SEIRS_Files <- lapply(Sys.glob(sysGlobNames[5]), read.csv)</pre>
CU_RHF_SEIRS_Files_Vector <- vector("list", length(CU_RHF_SEIRS_Files))</pre>
for (i in 1:length(CU_RHF_SEIRS_Files)) {
   CU_RHF_SEIRS_Files[[i]] <- CU_RHF_SEIRS_Files[[i]][!(CU_RHF_SEIRS_Files[[i]]$Location != "US National
   CU_RHF_SEIRS_Files[[i]] <- CU_RHF_SEIRS_Files[[i]][!(CU_RHF_SEIRS_Files[[i]]$Type != "Point"),]
   CU_RHF_SEIRS_Files_Vector[[i]] <- CU_RHF_SEIRS_Files[[i]][1:4,]$Value
CU_RHF_SIRS_Files <- lapply(Sys.glob(sysGlobNames[6]), read.csv)</pre>
CU_RHF_SIRS_Files_Vector <- vector("list", length(CU_RHF_SIRS_Files))</pre>
for (i in 1:length(CU_RHF_SIRS_Files)) {
   CU_RHF_SIRS_Files[[i]] <- CU_RHF_SIRS_Files[[i]][!(CU_RHF_SIRS_Files[[i]]$Location != "US National"),
   CU_RHF_SIRS_Files[[i]] <- CU_RHF_SIRS_Files[[i]][!(CU_RHF_SIRS_Files[[i]]$Type != "Point"),]
   CU_RHF_SIRS_Files_Vector[[i]] <- CU_RHF_SIRS_Files[[i]][1:4,]$Value
CUBMA_Files <- lapply(Sys.glob(sysGlobNames[7]), read.csv)
CUBMA_Files_Vector <- vector("list", length(CUBMA_Files))</pre>
for (i in 1:length(CUBMA_Files)) {
   CUBMA_Files[[i]] <- CUBMA_Files[[i]][!(CUBMA_Files[[i]]$Location != "US National"),]
   CUBMA_Files[[i]] <- CUBMA_Files[[i]][!(CUBMA_Files[[i]]$Type != "Point"),]</pre>
   CUBMA_Files_Vector[[i]] <- CUBMA_Files[[i]][1:4,]$Value</pre>
}
Delphi_BasisRegression_PackageDefaults_Files <- lapply(Sys.glob(sysGlobNames[8]), read.csv)
Delphi_BasisRegression_PackageDefaults_Files_Vector <- vector("list", length(Delphi_BasisRegression_PackageDefaults_Files_Vector <- vector("list", length(Delphi_BasisRegression_PackageDefaults_Files
for (i in 1:length(Delphi_BasisRegression_PackageDefaults_Files)) {
   Delphi_BasisRegression_PackageDefaults_Files[[i]] <- Delphi_BasisRegression_PackageDefaults_Files[[i]]
   Delphi_BasisRegression_PackageDefaults_Files[[i]] <- Delphi_BasisRegression_PackageDefaults_Files[[i]]
   Delphi_BasisRegression_PackageDefaults_Files_Vector[[i]] <- Delphi_BasisRegression_PackageDefaults_Fi
```

```
Delphi_DeltaDensity_PackageDefaults_Files <- lapply(Sys.glob(sysGlobNames[9]), read.csv)
Delphi_DeltaDensity_PackageDefaults_Files_Vector <- vector("list", length(Delphi_DeltaDensity_PackageDe
for (i in 1:length(Delphi_DeltaDensity_PackageDefaults_Files)) {
   Delphi_DeltaDensity_PackageDefaults_Files[[i]] <- Delphi_DeltaDensity_PackageDefaults_Files[[i]][!(De
   Delphi_DeltaDensity_PackageDefaults_Files[[i]] <- Delphi_DeltaDensity_PackageDefaults_Files[[i]][!(De
   Delphi_DeltaDensity_PackageDefaults_Files_Vector[[i]] <- Delphi_DeltaDensity_PackageDefaults_Files[[i]
}
Delphi_EmpiricalBayes_Cond4_Files <- lapply(Sys.glob(sysGlobNames[10]), read.csv)
Delphi_EmpiricalBayes_Cond4_Files_Vector <- vector("list", length(Delphi_EmpiricalBayes_Cond4_Files))
for (i in 1:length(Delphi_EmpiricalBayes_Cond4_Files)) {
   Delphi_EmpiricalBayes_Cond4_Files[[i]] <- Delphi_EmpiricalBayes_Cond4_Files[[i]][!(Delphi_EmpiricalBa
   Delphi_EmpiricalBayes_Cond4_Files[[i]] <- Delphi_EmpiricalBayes_Cond4_Files[[i]][!(Delphi_EmpiricalBa
   Delphi_EmpiricalBayes_Cond4_Files_Vector[[i]] <- Delphi_EmpiricalBayes_Cond4_Files[[i]][1:4,]$Value
}
Delphi_EmpiricalBayes_PackageDefaults_Files <- lapply(Sys.glob(sysGlobNames[11]), read.csv)
Delphi_EmpiricalBayes_PackageDefaults_Files_Vector <- vector("list", length(Delphi_EmpiricalBayes_PackageDefaults_Files_Vector <- vector("list", length(Delphi_EmpiricalBayes_PackageDefaults_Files_Files_Vector <- vector("list", length(Delphi_EmpiricalBayes_PackageDefaults_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_Files_
for (i in 1:length(Delphi_EmpiricalBayes_PackageDefaults_Files)) {
   Delphi_EmpiricalBayes_PackageDefaults_Files[[i]] <- Delphi_EmpiricalBayes_PackageDefaults_Files[[i]][
   Delphi_EmpiricalBayes_PackageDefaults_Files[[i]] <- Delphi_EmpiricalBayes_PackageDefaults_Files[[i]][
   Delphi_EmpiricalBayes_PackageDefaults_Files_Vector[[i]] <- Delphi_EmpiricalBayes_PackageDefaults_File
}
Delphi_EmpiricalFutures_PackageDefaults_Files <- lapply(Sys.glob(sysGlobNames[12]), read.csv)
Delphi_EmpiricalFutures_PackageDefaults_Files_Vector <- vector("list", length(Delphi_EmpiricalFutures_P
for (i in 1:length(Delphi_EmpiricalFutures_PackageDefaults_Files)) {
   Delphi_EmpiricalFutures_PackageDefaults_Files[[i]] <- Delphi_EmpiricalFutures_PackageDefaults_Files[[
   Delphi_EmpiricalFutures_PackageDefaults_Files[[i]] <- Delphi_EmpiricalFutures_PackageDefaults_Files[[
   Delphi_EmpiricalFutures_PackageDefaults_Files_Vector[[i]] <- Delphi_EmpiricalFutures_PackageDefaults_
}
Delphi_EmpiricalTrajectories_PackageDefaults_Files <- lapply(Sys.glob(sysGlobNames[13]), read.csv)
Delphi_EmpiricalTrajectories_PackageDefaults_Files_Vector <- vector("list", length(Delphi_EmpiricalTraj
for (i in 1:length(Delphi_EmpiricalTrajectories_PackageDefaults_Files)) {
   Delphi_EmpiricalTrajectories_PackageDefaults_Files[[i]] <- Delphi_EmpiricalTrajectories_PackageDefaul
   Delphi_EmpiricalTrajectories_PackageDefaults_Files[[i]] <- Delphi_EmpiricalTrajectories_PackageDefaul
   Delphi_EmpiricalTrajectories_PackageDefaults_Files_Vector[[i]] <- Delphi_EmpiricalTrajectories_Packag
}
Delphi_MarkovianDeltaDensity_PackageDefaults_Files <- lapply(Sys.glob(sysGlobNames[14]), read.csv)
Delphi_MarkovianDeltaDensity_PackageDefaults_Files_Vector <- vector("list", length(Delphi_MarkovianDelt
for (i in 1:length(Delphi_MarkovianDeltaDensity_PackageDefaults_Files)) {
   Delphi_MarkovianDeltaDensity_PackageDefaults_Files[[i]] <- Delphi_MarkovianDeltaDensity_PackageDefaul
   Delphi_MarkovianDeltaDensity_PackageDefaults_Files[[i]] <- Delphi_MarkovianDeltaDensity_PackageDefaul
   Delphi_MarkovianDeltaDensity_PackageDefaults_Files_Vector[[i]] <- Delphi_MarkovianDeltaDensity_Packag
}
Delphi_Stat_FewerComponentsNoBackcastNoNowcast_Files <- lapply(Sys.glob(sysGlobNames[15]), read.csv)
Delphi_Stat_FewerComponentsNoBackcastNoNowcast_Files_Vector <- vector("list", length(Delphi_Stat_FewerC
for (i in 1:length(Delphi_Stat_FewerComponentsNoBackcastNoNowcast_Files)) {
   Delphi_Stat_FewerComponentsNoBackcastNoNowcast_Files[[i]] <- Delphi_Stat_FewerComponentsNoBackcastNoN
   Delphi_Stat_FewerComponentsNoBackcastNoNowcast_Files[[i]] <- Delphi_Stat_FewerComponentsNoBackcastNoN
```

```
Delphi_Stat_FewerComponentsNoBackcastNoNowcast_Files_Vector[[i]] <- Delphi_Stat_FewerComponentsNoBack
}
Delphi_Uniform_Files <- lapply(Sys.glob(sysGlobNames[16]), read.csv)</pre>
Delphi_Uniform_Files_Vector <- vector("list", length(Delphi_Uniform_Files))</pre>
for (i in 1:length(Delphi_Uniform_Files)) {
     Delphi_Uniform_Files[[i]] <- Delphi_Uniform_Files[[i]][!(Delphi_Uniform_Files[[i]]$Location != "US Na
     Delphi_Uniform_Files[[i]] <- Delphi_Uniform_Files[[i]][!(Delphi_Uniform_Files[[i]]$Type != "Point"),]
     Delphi_Uniform_Files_Vector[[i]] <- Delphi_Uniform_Files[[i]][1:4,]$Value
}
LANL_DBM_Files <- lapply(Sys.glob(sysGlobNames[17]), read.csv)
LANL_DBM_Files_Vector <- vector("list", length(LANL_DBM_Files))
for (i in 1:length(LANL_DBM_Files)) {
     LANL_DBM_Files[[i]] <- LANL_DBM_Files[[i]][!(LANL_DBM_Files[[i]]$Location != "US National"),]
     LANL_DBM_Files[[i]] <- LANL_DBM_Files[[i]][!(LANL_DBM_Files[[i]]$Type != "Point"),]
     LANL_DBM_Files_Vector[[i]] <- LANL_DBM_Files[[i]][1:4,]$Value
}
ReichLab_kcde_Files <- lapply(Sys.glob(sysGlobNames[18]), read.csv)</pre>
ReichLab_kcde_Files_Vector <- vector("list", length(ReichLab_kcde_Files))</pre>
for (i in 1:length(ReichLab_kcde_Files)) {
     ReichLab_kcde_Files[[i]] <- ReichLab_kcde_Files[[i]][!(ReichLab_kcde_Files[[i]]$Location != "US Nation".
     ReichLab_kcde_Files[[i]] <- ReichLab_kcde_Files[[i]][!(ReichLab_kcde_Files[[i])$Type != "Point"),]</pre>
     ReichLab_kcde_Files_Vector[[i]] <- ReichLab_kcde_Files[[i]][1:4,]$Value
}
ReichLab_kde_Files <- lapply(Sys.glob(sysGlobNames[19]), read.csv)</pre>
ReichLab_kde_Files_Vector <- vector("list", length(ReichLab_kde_Files))</pre>
for (i in 1:length(ReichLab_kde_Files)) {
     ReichLab_kde_Files[[i]] <- ReichLab_kde_Files[[i]][!(ReichLab_kde_Files[[i]]$Location != "US National
     ReichLab_kde_Files[[i]] <- ReichLab_kde_Files[[i]][!(ReichLab_kde_Files[[i]]$Type != "Point"),]</pre>
     ReichLab_kde_Files_Vector[[i]] <- ReichLab_kde_Files[[i]][1:4,]$Value
}
ReichLab_sarima_seasonal_difference_FALSE_Files <- lapply(Sys.glob(sysGlobNames[20]), read.csv)
ReichLab_sarima_seasonal_difference_FALSE_Files_Vector <- vector("list", length(ReichLab_sarima_seasona
for (i in 1:length(ReichLab_sarima_seasonal_difference_FALSE_Files)) {
     ReichLab_sarima_seasonal_difference_FALSE_Files[[i]] <- ReichLab_sarima_seasonal_difference_FALSE_Fil
     \label{lem:condition} ReichLab\_sarima\_seasonal\_difference\_FALSE\_Files[[i]] <- ReichL
     ReichLab_sarima_seasonal_difference_FALSE_Files_Vector[[i]] <- ReichLab_sarima_seasonal_difference_FA
}
ReichLab_sarima_seasonal_difference_TRUE_Files <- lapply(Sys.glob(sysGlobNames[21]), read.csv)
ReichLab_sarima_seasonal_difference_TRUE_Files_Vector <- vector("list", length(ReichLab_sarima_seasonal
for (i in 1:length(ReichLab_sarima_seasonal_difference_TRUE_Files)) {
     \label{lem:condition} ReichLab\_sarima\_seasonal\_difference\_TRUE\_Files[[i]] <- ReichLab\_sarima\_seasonal\_difference\_TRUE\_F
     ReichLab_sarima_seasonal_difference_TRUE_Files[[i]] <- ReichLab_sarima_seasonal_difference_TRUE_Files
     ReichLab_sarima_seasonal_difference_TRUE_Files_Vector[[i]] <- ReichLab_sarima_seasonal_difference_TRU
}
UTAustin_edm_Files <- lapply(Sys.glob(sysGlobNames[22]), read.csv)</pre>
UTAustin_edm_Files_Vector <- vector("list", length(UTAustin_edm_Files))</pre>
```

```
for (i in 1:length(UTAustin_edm_Files)) {
  UTAustin_edm_Files[[i]] <- UTAustin_edm_Files[[i]][!(UTAustin_edm_Files[[i]]$Location != "US National
  UTAustin_edm_Files[[i]] <- UTAustin_edm_Files[[i]][!(UTAustin_edm_Files[[i])$Type != "Point"),]</pre>
  UTAustin_edm_Files_Vector[[i]] <- UTAustin_edm_Files[[i]][1:4,]$Value
}
# Since I first want to look at the K-S test statistic instead of the p-values,
# I will not use the ks.test() function for now.
choose(21, 2)
## [1] 210
# There are 210 possible combinations excluding the last model, UTAustin (which has fewer sets of predi
vectorNames <- c("CU_EAKFC_SEIRS_Files_Vector", "CU_EAKFC_SIRS_Files_Vector", "CU_EKF_SEIRS_Files_Vector")
listofVectors <- vector("list", 210)</pre>
for (i in 1:length(listofVectors)) {
  listofVectors[[i]] <- numeric(28)</pre>
  # We could append more things into each vector (the titles for the plots) later
# Following this process of extracting the necessary data from the models' prediction files
# I now want to graph some K-S test statistics between each pair of models:
# For Now, I will exclude the last one (UTAustin).
# I will also exclude three of the observations on one of the Reich lab kde model's predictions.
j = 1 # Refers to a vector in the list
x = 1 # the first model being compared
while (j <= 210) {
  while (x \le 21) {
    k = x + 1 # the second model being compared
    while (k <= 21) {
      # This for-loop places the 28 K-S comparisons for the pair of models into the jth place
      # in the list of Vectors.
      for (i in 1:28) {
        listofVectors[[j]][i] <- max(abs(eval(as.name(vectorNames[x]))[[i]] - eval(as.name(vectorNames[x]))
      listofVectors[[j]][29] <- paste(vectorNames[x], vectorNames[k], sep = "&")
      k = k + 1
      j = j + 1
    }
    x = x + 1
 }
}
```

































































































































