



Tutorial: Using TempEst for data exploration

This tutorial describes the use of TempEst to examine the temporal signal of a data set and to look for problematic or erroneous sequences.

In this tutorial, we will explore the use of the interactive graphical program TempEst (tempest) (formerly known as Path-O-Gen) to examine virus sequence data that has been sampled through time to look for problematic sequences and to explore the degree and pattern of temporal signal. This can be a useful way of examining the data for potential issues before committing significant time to running BEAST (beast).

Step by step:

1. Building a non-molecular clock tree (IQ-Tree)
2. Running TempEst and loading the tree

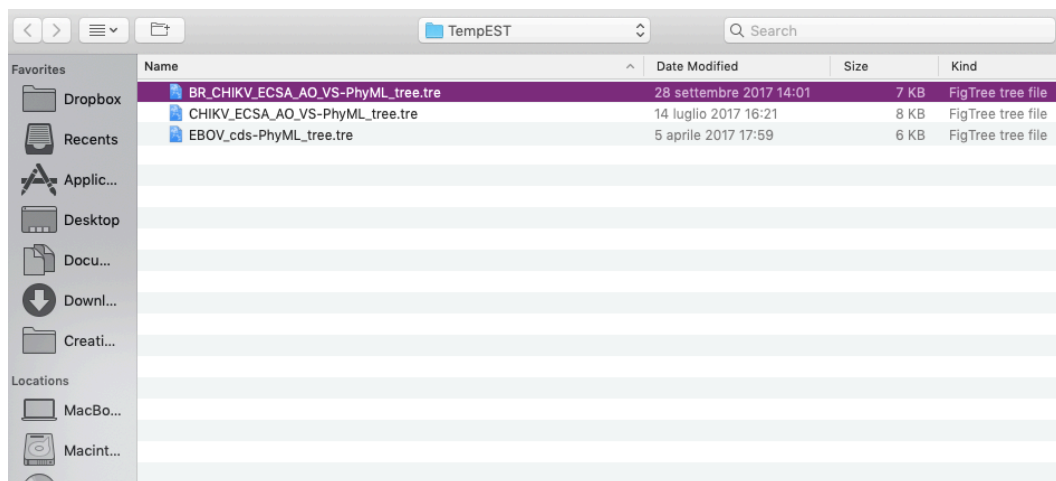
2. Running TempEst and loading the tree

Run TempEst (tempest) by double clicking on its icon.

TempEst is an interactive graphical application for examining the temporal signal in a tree of time-stamped sequences by plotting the divergence of each tip from the root against the date of sampling (a root-to-tip plot).

Once running, TempEst will look similar irrespective of which computer system it is running on. For this tutorial, the Mac OS X version will be shown but the Linux & Windows versions will have exactly the same layout and functionality.

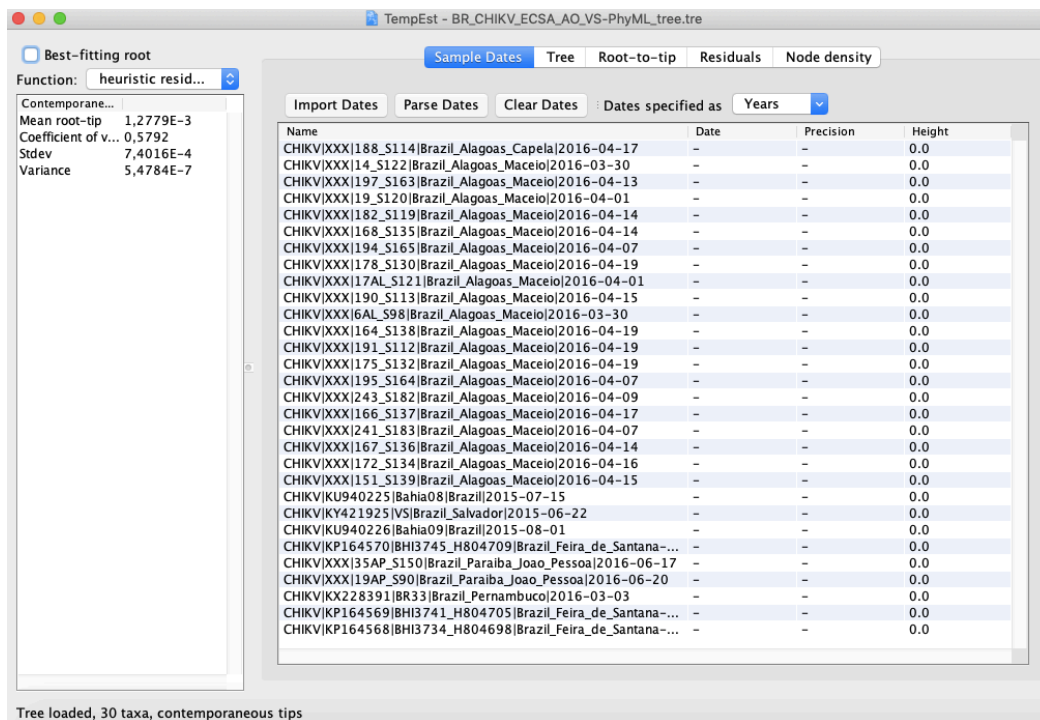
When started, TempEst will immediately display a file selection dialog box in which you can select the tree that you made in the previous section.



Go to the `/home/manager/Phylogenetics/3.Temporal_Signal/TempEST` folder, select the `BR_CHIKV_ECSA_AO_VS-PhyML_tree.tre` file and click Open .

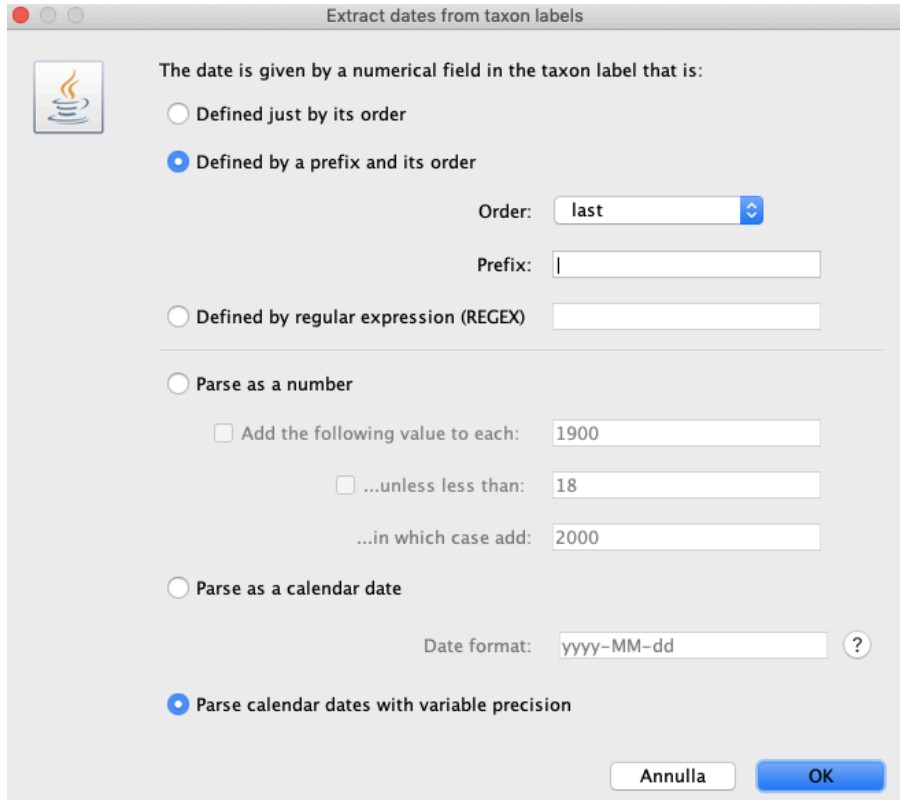
Parsing dates of sampling

Once the tree is loaded the main window will appear and look like this:



Ignore the panel on the left for the moment. The first thing that needs doing is to give the date of sampling to each of the sequences.

The actual year of sampling is given at the end of the name of each taxon. To specify the dates of the sequences in BEAUti we will use the Parse Dates button at the top of the panel. Clicking this will make a dialog box appear:



Extract dates from taxon labels

The date is given by a numerical field in the taxon label that is:

☐ Defined just by its order

☒ Defined by a prefix and its order

Order: last

Prefix: |

☐ Defined by regular expression (REGEX)

☐ Parse as a number

☐ Add the following value to each: 1900

☐ ...unless less than: 18

☐ ...in which case add: 2000

☐ Parse as a calendar date

Date format: yyyy-MM-dd ?

☒ Parse calendar dates with variable precision

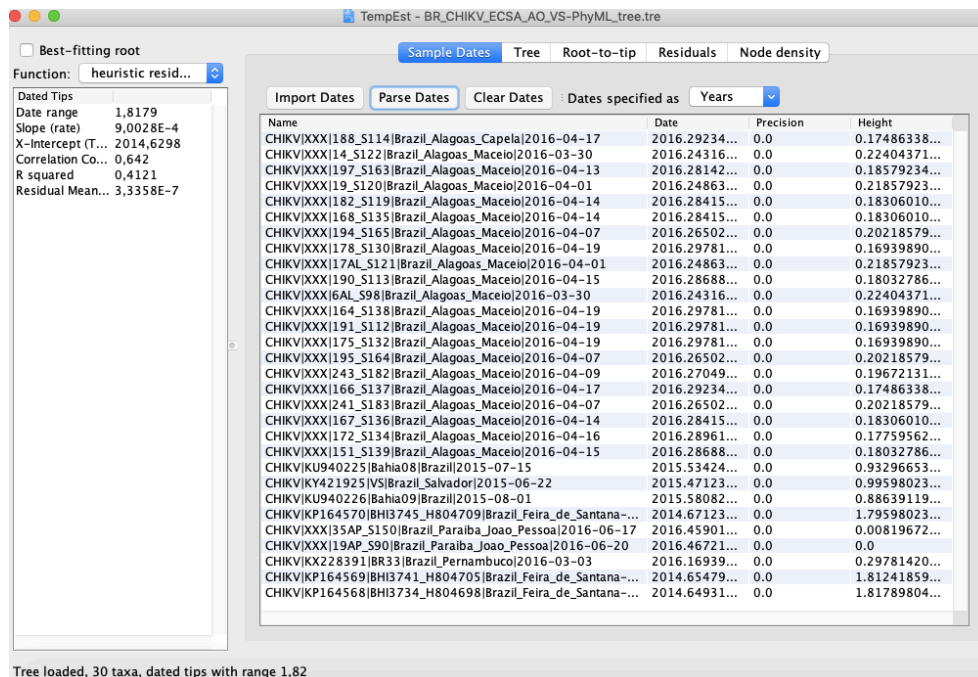
Annulla OK

This operation attempts to extract the dates from the taxon names. It works by trying to find a numerical field within each name. This dialog box is the same as that in BEAUti (beauti) and there are a wide range of options for doing this.

For these sequences you can set the options to look like the figure above: Defined just by its order
Order: last and Parse calendar dates with variable precision.

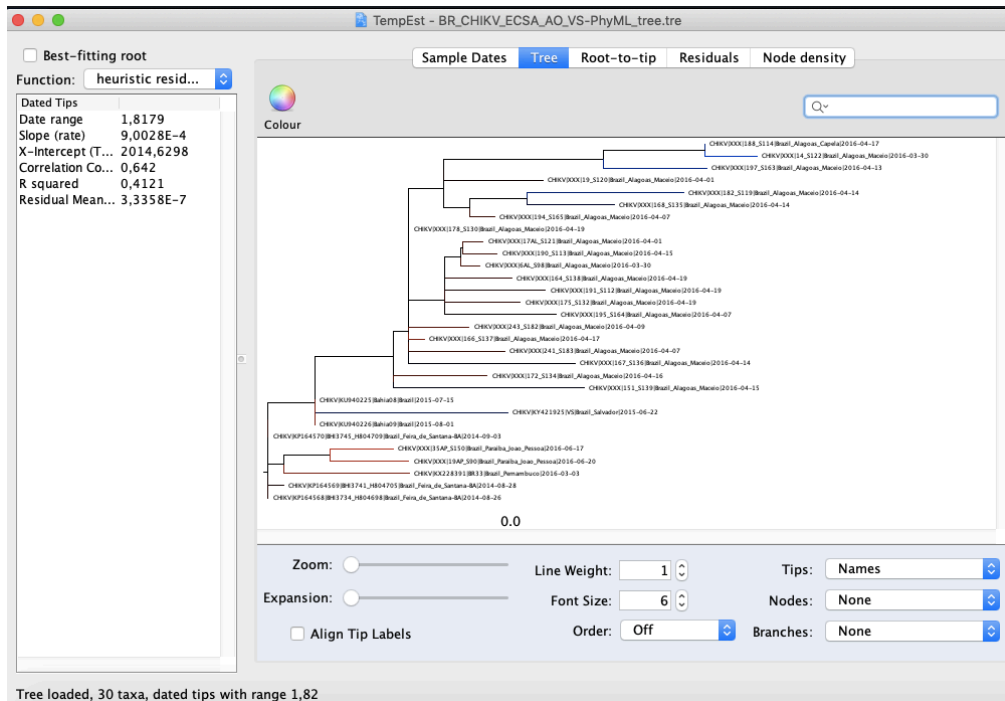
So, we can press OK .

The table will now have the year of sampling for each virus in the Dates column. Click on the Dates column header to sort the dates and check that they are all correct.

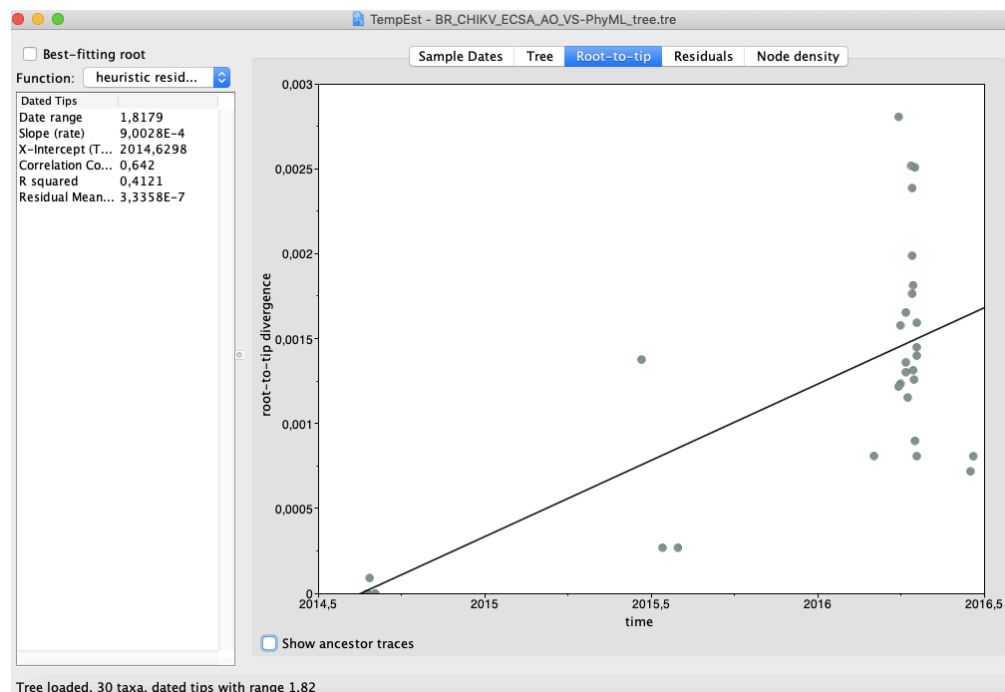


The temporal signal and rooting

We can now explore the data using the tabs at the top of the window - Tree , Root-to-tip & Residuals. If you click on the Tree tab you will see the tree as loaded from the tree file. Because we constructed this tree using a non-molecular-clock model, it will be arbitrarily rooted. If you look at the date of each virus in the tree you will see that there is no correlation with the horizontal position:



Now switch to the Root-to-tip panel. This shows a plot of the divergence from the root of the tree against time of sampling (a so-called 'Root to tip plot'):

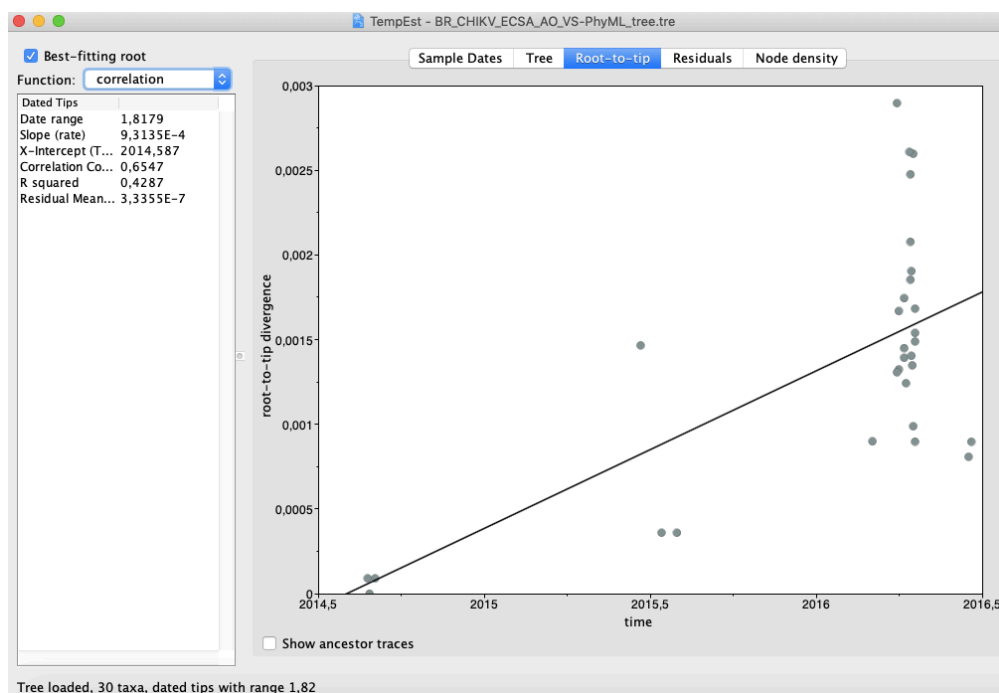


You can see that there is a quite nice correlation in this plot (the line is the best-fit regression). In the table on the left you can see the Correlation Coefficient is 0.64. Despite this, the correlation we are

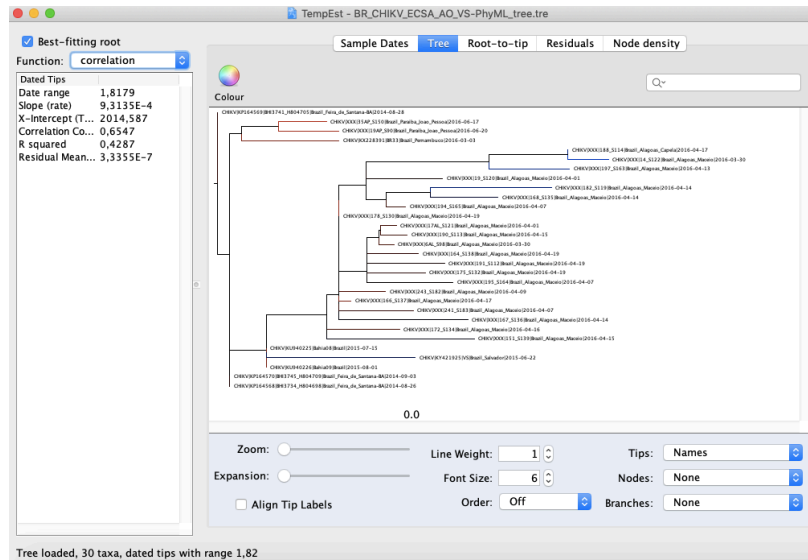
observed is not the correct one as the root is arbitrarily set by the phylogeny reconstruction software and thus divergence from root is meaningless.

TempEst can try to find the root of the tree that optimizes the temporal signal. It does this by trying all possible roots and picks the one that produces the optimal value of a range of statistics. The function it uses is selected in the menu at the top left. The options are to minimize the mean of the squares of the residuals (residual-meansquared), or to maximize the correlation coefficient (correlation) or R (R squared). These are all ad hoc procedures and no particular one is best but residual-mean-squared may be most consistent with the investigations here.

Click Best-fitting root to root the tree at the place that minimizes the mean of the squares of the residuals.

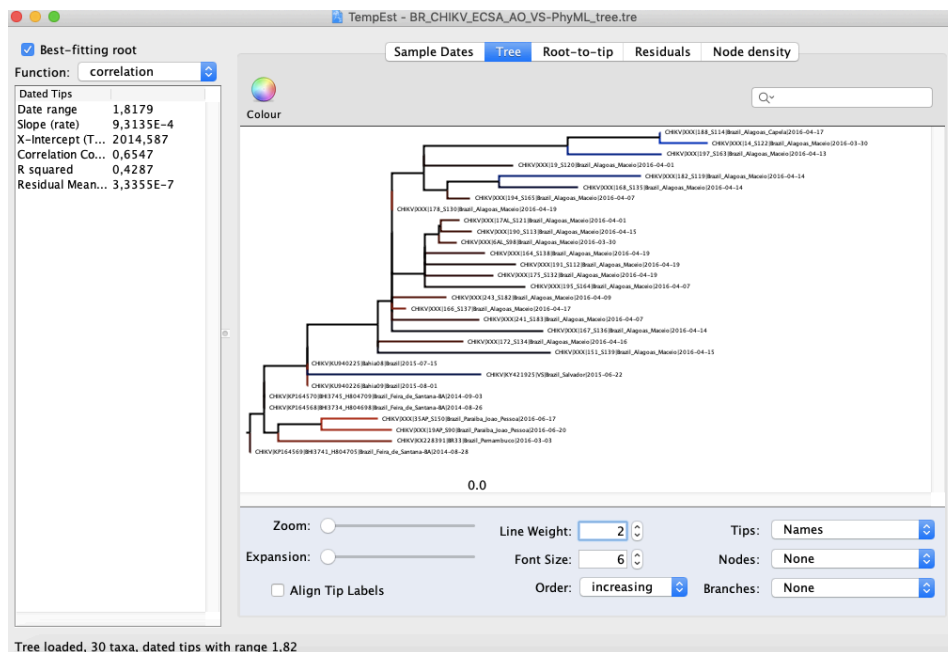


Now there is a better correlation between the dates of the tips and the divergence from the root (the correlation coefficient has nearly doubled). Return to the tree to look where the root was placed:



To make the tree easier to view, switch the Order option in the panel at the bottom to increasing .

The tree branches are coloured to show the residual with blue for tips with positive residuals (above the regression line), red for negative.



On the left hand side of the window there is a table of statistics:

Dated Tips	
Date range	1,8179
Slope (rate)	9,3135E-4
X-Intercept (T...	2014,587
Correlation Co...	0,6547
R squared	0,4287
Residual Mean...	3,3355E-7

As well as the statistical metrics (Correlation Coefficient , R squared and Residual Mean Squared) there are the following:

Date range: The span of dates for the viruses.

Slope (rate): The slope of the regression line. This is an estimate of the rate of evolution in substitutions per site per year.

X-Intercept (TMRCA): The point on the x-axis at which the regression line crosses. This is an estimate of the date of the root of the tree.

You can also extract the roo-to-tip data and plot is using the ggplot package in R Studio.

To do so go to “File” and them click on “Export Data”

The screenshot shows the TempEst software interface. The 'File' menu is open, and the 'Export Data' option is highlighted. The main window displays a table of dated tips with columns for Date, Precision, and Height. The 'Function' dropdown is set to 'correlation'. The status bar at the bottom indicates 'Tree loaded, 30 taxa, dated tips with range 1,82'.

Date	Precision	Height
2016.292349726776	0.0	0.17486338797812095
2016.2431693989072	0.0	0.22404371584684668
2016.2814207650274	0.0	0.18579234972662562
2016.2486338797814	0.0	0.21857923497259435
2016.2841530054645	0.0	0.18306010928949945
2016.2841530054645	0.0	0.18306010928949945
2016.2650273224044	0.0	0.20218579234960998
2016.2978142076502	0.0	0.16939890710386862
2016.2486338797814	0.0	0.21857923497259435
2016.2868852459017	0.0	0.18032786885237329
2016.2431693989072	0.0	0.22404371584684668
2016.2978142076502	0.0	0.16939890710386862
2016.2978142076502	0.0	0.16939890710386862
2016.2978142076502	0.0	0.16939890710386862
2016.2650273224044	0.0	0.20218579234960998
2016.2704918032787	0.0	0.19672131147535765
2016.292349726776	0.0	0.17486338797812095
2016.2650273224044	0.0	0.20218579234960998
2016.2841530054645	0.0	0.18306010928949945
2016.2896174863388	0.0	0.17759562841524712
2016.2868852459017	0.0	0.18032786885237329
2015.5342465753424	0.0	0.9329665394116091
2015.4712328767123	0.0	0.9959802380417386
2015.5808219178082	0.0	0.8863911969458513
2014.6712328767123	0.0	1.7959802380416932
2016.4590163934427	0.0	0.008196721311378496
2016.467213114754	0.0	0.0
2016.1693989071039	0.0	0.29781420765016264
2014.654794520548	0.0	1.8124185942060649
2014.6493150684933	0.0	1.8178980462607797

Let's now estimate the temporal signal in all the other dataset!

And lets discuss results together!