Foundations of Data Science – Assignment 2

Time Series Analysis

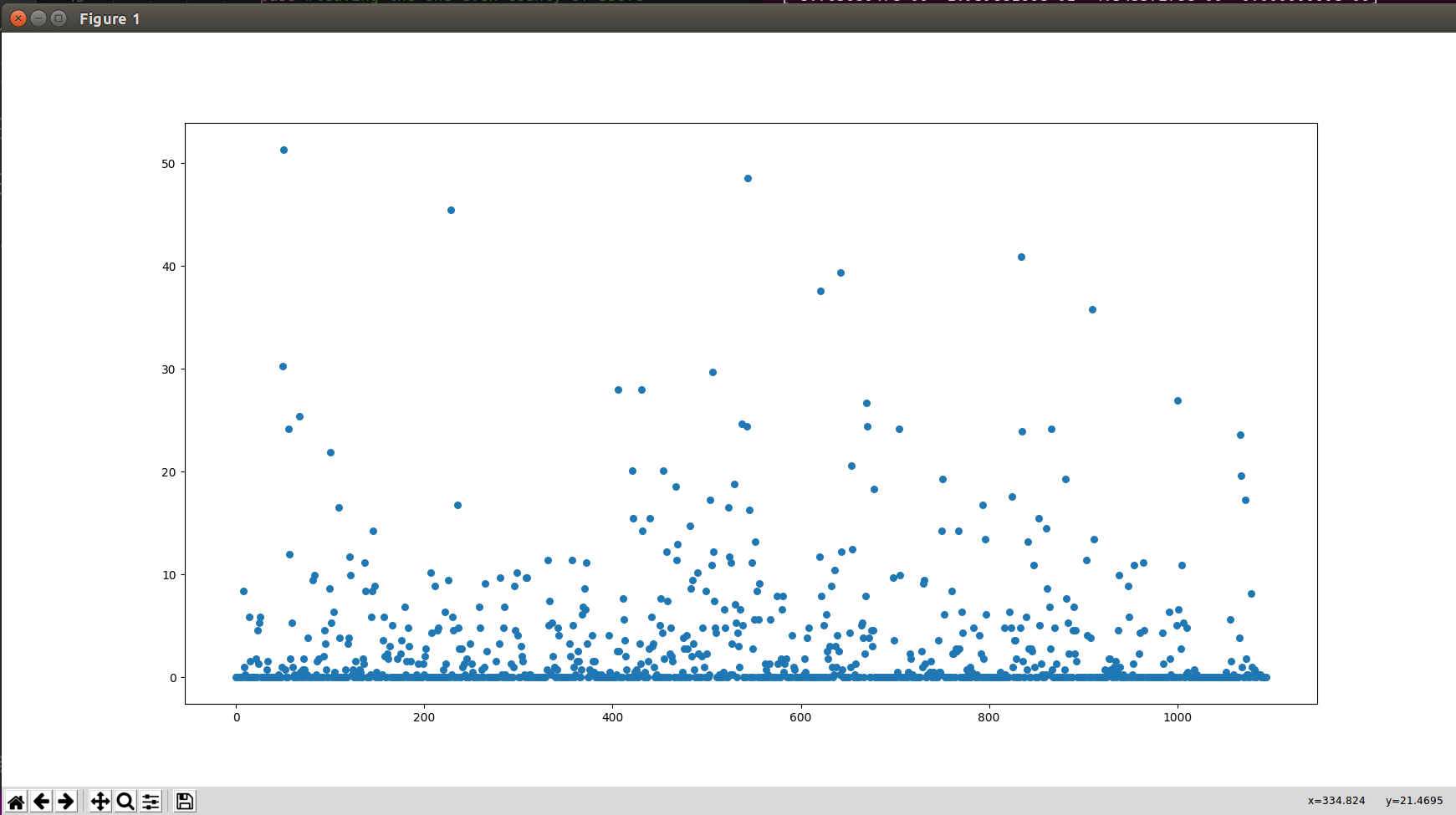
Yash Vijay – 2017A7PS0072P

Jithin Kallukalam Sojan – 2017A7PS0163P

Venkat Nalla – 2016A7PS0030P

**The Dataset**

We used the *crop.txt* and *climate.txt* files from the NC-94 dataset, an extensive dataset with records of special and temporal variables for crop, climate, and soil in parts of the United States. There are 23 parameters, out of which we used only the first four (MaxTemp, MinTemp, Radiant, and Precipitation) for clustering. The data contains values from years 1971-2001, but we only used data for the three years from 1997-1999. The classification was done using regression on production values throughout the year for the crop type *wheat-all.*



A plot of rainfall vs number of days from 1997 is shown above for county 17001.

**Tools used**

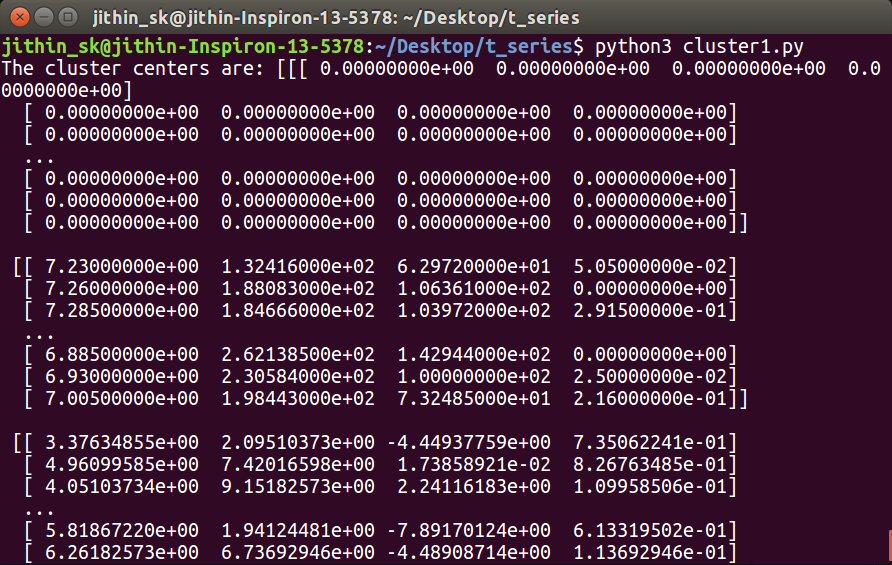
We used python for the assignment. We used functions from several libraries like *numpy*, *tslearn*, *sklearn* and *pandas. matplotlib* was used for visualization of data. *tslearn* is the python package that provides machine learning tools for the analysis of time series data.

**Clustering**

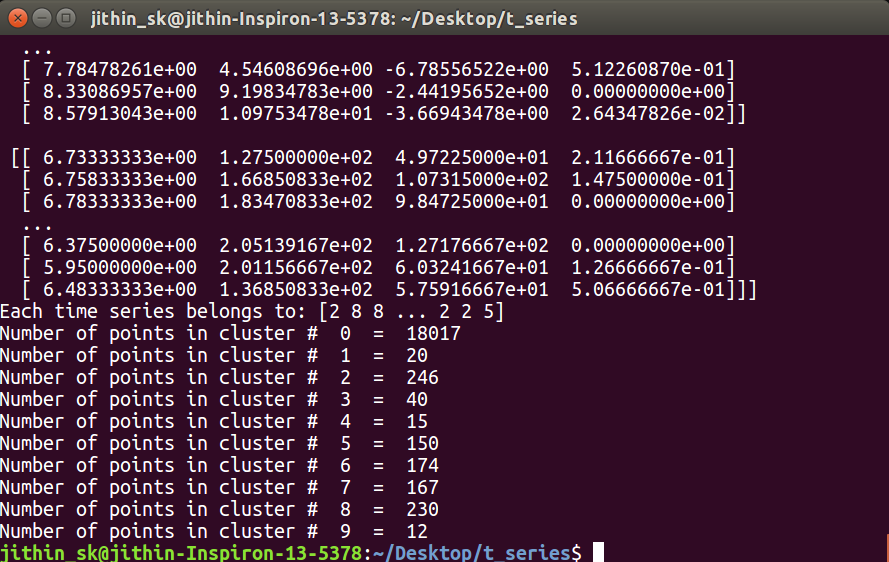
*tslearn.clustering.timeseriesKmeans*  is used for K-means clustering for time series data.

class tslearn.clustering.TimeSeriesKMeans(n\_clusters=3, max\_iter=50, tol=1e-06, n\_init=1, metric='euclidean', max\_iter\_barycenter=100, metric\_params=None, n\_jobs=None, dtw\_inertia=False, verbose=False, random\_state=None, init='k-means++')

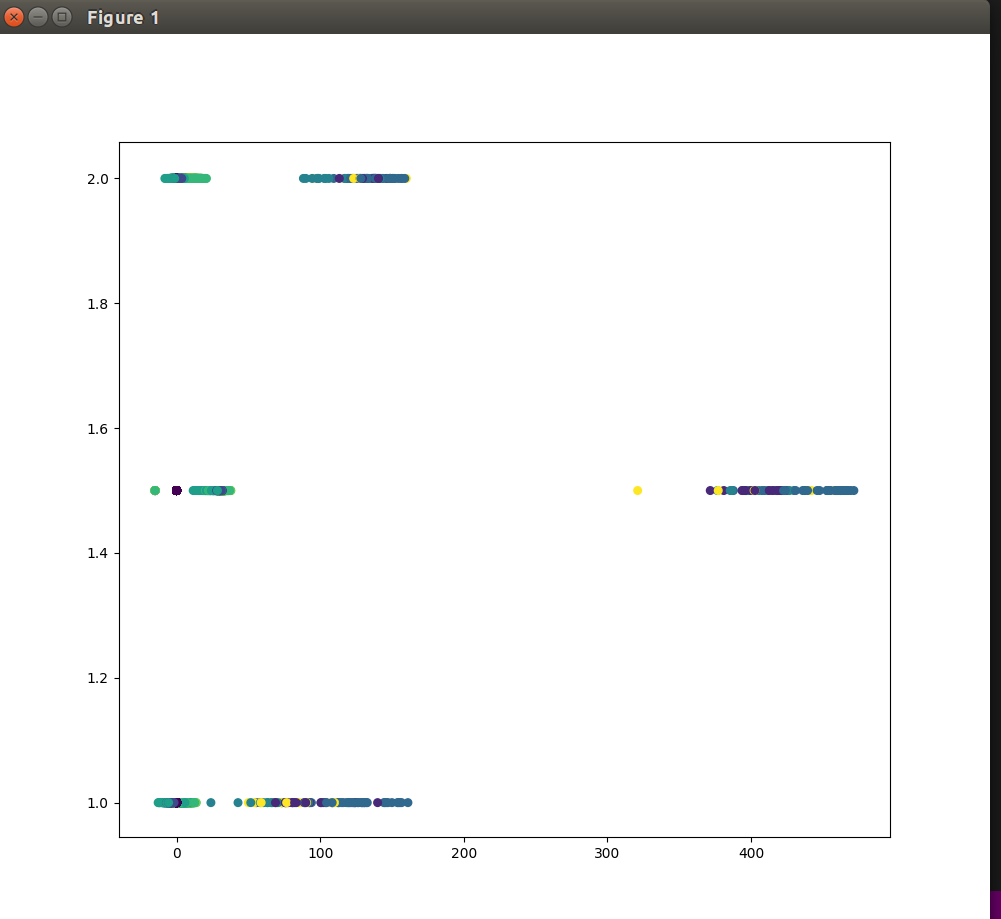
We used the *euclidean* metric as all the time series were of the same sizes.



We have clustered the data into 10 clusters. The labels array ( [2 8 8 . . . 2 2 5] ) holds the cluster numbers for each of the 19000 time series. The cluster centers are of 1095x4 dimensions.



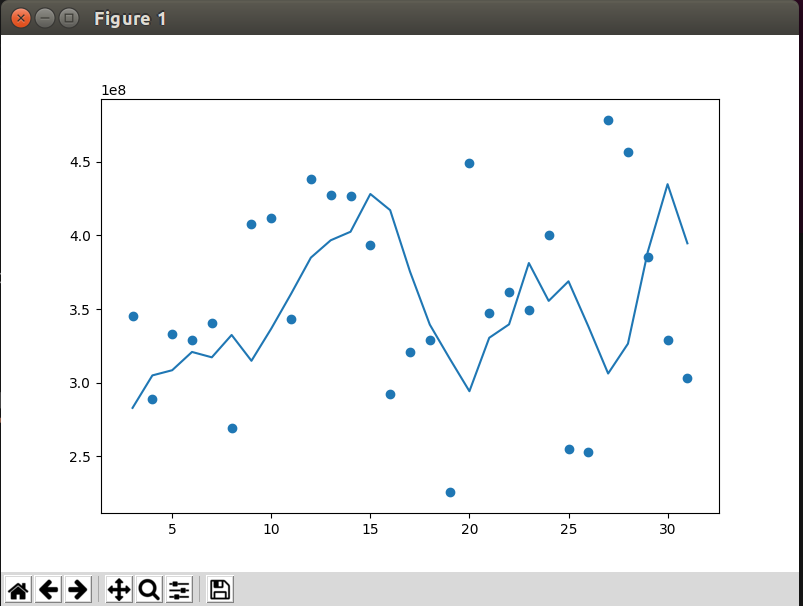
Cluster one has 18000-plus points, which is disproportionately large.



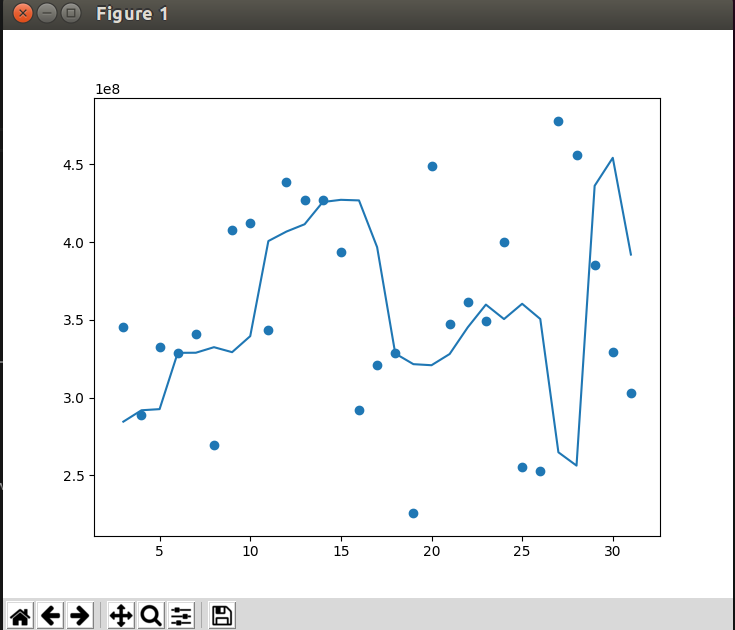
We are showing clusters for 0th, 182nd and 364th day at heights 2, 1.5 and 1. Since it is a time series clustering the affinity of data points is hard to visualize.

**Regression**

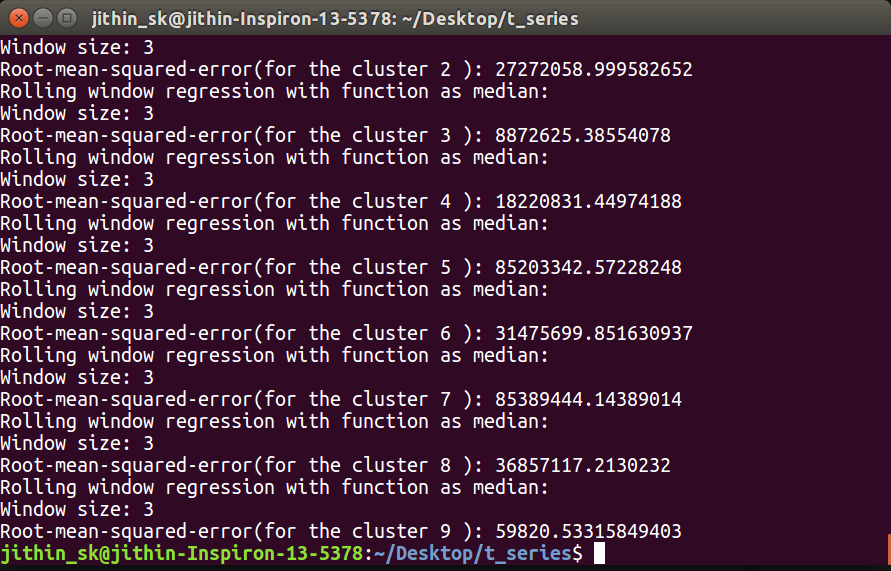
We used the labels array mentioned above to get the sum of values of production of *wheat-all* throughout the 32 years from 1970-2001. We used rolling window model to build a model that predicts crop yield in the (k+1)th time point using a window size of k. The training data points keep changing according to the values that we are predicting as we keep adding the actual value of the previous predictions into account. We compared average and median and found that median gave marginally better results based on root mean squared error of each regression. We did regression in each of the ten cluster and found RMSE for each. We used exponential smoothing with alpha=0.9 (alpha belongs to (0,1)) using trial and error. In the graphs below, the points are the original values, and the curve represents the predicted values using time series regression. The plots below are taken for the 7th cluster.



Regression using moving average with exponential smoothing with window size = 3.



Regression using moving median with exponential smoothing with window size = 3.



RMSE for moving window with median is shown here for each cluster. The large values might seem deceptive as the data itself goes in powers of 9-10.

**Classification: Showing the strength of regression**

For classification we simply used the results obtained from regression. We took the maximum and minimum values attained by the crop yields. Then, we divided this range in 3 partitions (Which could be treated as low, medium and high). Classification was done for each cluster separately, which included getting the minimum and maximum values. We then compared the corresponding classes for the predicted and actual data to get the total accuracy (43%).

