Using the K-means Data Cluster Algorithm to Classify Yield Curve Shape

Overview

Exploratory data analysis (EDA) is an approach for data analysis that employs a variety of techniques to help gain insight into a data set and uncover underlying structure. EDA postpones the usual assumptions about what kind of model the data follow, with the more direct approach of allowing the data itself to reveal its underlying structure and model. The idea behind EDA techniques is to help maximize our natural pattern-recognition abilities.

Some typical EDA techniques include histograms and scatter plots, but a little known (or maybe forgotten) technique is that of "data clustering". The purpose of this note is to expose the reader to the application of data clustering as a potentially useful EDA technique for data pattern recognition, which can inform subsequent EDA or model building.

The data to which clustering technique will be applied is the daily US government spot yield curve, and the goal is to classify the curve shape as being either normal, steep, very steep, or flat/inverted. The sections will cover:

- the basics of yield curve shape.
- the data to be analyzed.
- some basic EDA techniques to begin the exploration.
- the clustering technique used.
- how clustering assisted in yield curve shape identification.

Yield Curve Shape Basics

It is well known in the financial community that the yield curve can be qualitatively described by its shape. In particular, students of the yield curve have ascribed the following common descriptions for its shape: normal, steep, inverted, humped, and flat. The yield curve shapes, as per commonly accepted market convention, are briefly described in the following paragraphs.

A **normal** yield curve means that yields moderately rise as maturity lengthens (i.e., the slope of the yield curve is positive). This positive slope reflects investor expectations for the economy to grow in the future and, importantly, for this growth to be associated with a greater expectation that inflation will rise in the future rather than fall. This expectation of higher inflation leads to expectations that the central bank will tighten monetary policy by raising short term interest rates in the future to slow economic growth and dampen inflationary pressure. It also creates a need for a risk premium associated with the uncertainty about the future rate of inflation and the risk this poses to the future value of cash flows. Investors price these risks into the yield curve by demanding higher yields for maturities further into the future.

A **steep** yield curve, where the slope rises more sharply, indicates that the economy is expected to improve quickly in the future. This type of curve can be seen at the beginning of an economic expansion,

or after the end of a recession. Historically, during a steep yield curve environment, the 20-year Treasury bond yield has averaged approximately two percentage points above that of three-month Treasury bills.

An **inverted** yield curve occurs when long-term yields fall below short-term yields. Under these unusual circumstances, long-term investors will settle for lower yields now if they think the economy will slow or even decline in the future.

Lastly, a **flat** yield curve is observed when all maturities have similar yields. A flat curve sends signals of uncertainty in the economy.

Note that there is an additional shape classification of **humped**, which results when short-term and long-term yields are approximately equal and medium-term yields are higher than those of the short-term and long-term. However, for purposes of this note this classification mainly due to its rarity in the U.S. interest rate market.

However, the main issues with the above shape classifications are:

- they are qualitative and based on heuristics.
- there is no market-accepted set of crisp definitions to describe yield curve shape (shape is often in the eye of the beholder), thus there is no market-accepted quantitative metric to measure shape.
- In general there exists no established quantitative techniques for assigning a shape classification to the daily yield curve (otherwise this note would not be written).

Furthermore, it was interesting to discover that not even the Federal Reserve Bank maintains formal definitions (either qualitative or quantitative) of yield curve shape[1] . Given that the Fed Funds rate is very close to zero right now, previous economic implications about steep and very steep yield curves may have to be tempered by the fact that the Fed has nailed short term rates to the floor.

Data and Assumptions

The following data and assumptions were used in the analysis:

- Eighteen and one half years of historical U.S. government par yield curve data, covering the period from April 16th, 1991 to October 16th, 2009, was used.
- The tenors chosen, fifteen in all, were: three month (3M), six month (6M), one year (1Y), two year (2Y), three year (3Y), four year (4Y), five year (5Y), seven year (7Y), eight year (8Y), nine year (9Y), ten year (10Y), fifteen year (15Y), twenty year (20Y), twenty five year (25Y), and thirty year (30Y).
- The data was reviewed for bad/missing data points, and where necessary standard cleansing procedures such as cross-checking against alternative data sources and linear interpolation/extrapolation were used.

[1] As per an email exchange with a senior economist at the St. Louis Federal Reserve Bank.

- The par yield data was subsequently transformed into its corresponding spot (zero coupon bond) yields. Similar to what was done with the original par yield data, the spot yields were also reviewed for suspect data points.
- The use of spot rates was motivated by their ease of use in financial computations.
- All computations and plotting were performed using Matlab.

Level, Slope, and Curvature

Statistical decomposition of the yield curve's daily returns demonstrates that there are three principal factors explaining a majority of the variation: **level**, **slope**, and **curvature**. Roughly speaking the **level** of the curve is typically defined as the longest tenor yield or the long-run yield, the **slope** is the difference between long term and short term yields, and **curvature** as the mid-curve yield relative to an average of long term and short term yields.

While changes in the level account for the major part of the variation in the yield curve, level changes may not be a significant source for changes in value for positions that involve spreads across maturities (i.e. barbells, steepeners, and flatteners).

Firstly, clients often specify a target or benchmark duration for their portfolio which limits the trader's ability to take directional bets on the level of interest rates. Secondly, changes in the slope and curvature are often considered to be more predictable than the level of the yield curve. The reason is that the slope and curvature characteristics tend to exhibit predictable mean reversion[2]. The slope characteristic, which tends to be a reflection of the state of the economy, has a cycle which reflects the economy's business cycle. The curvature characteristic, which reflects the volatility of interest rates, tends to exhibit greater volatility and even faster mean reversion[3]. This is why experienced fixed income traders tend to make yield curve bets based on expectations for slope and curvature, and not level.

Fig. A depicts a highly stylized yield curve and its corresponding first three components.

[2] Litterman, R., Scheinkman, J. and Weiss, L. (1991), 'Volatility and the Yield Curve', Journal of Fixed Income (June), 49-53.

[3] Litterman, R. and Scheinkman, J. (1991), "Common Factors Affecting the Bond Returns", Journal of Fixed Income (June), 54-61.

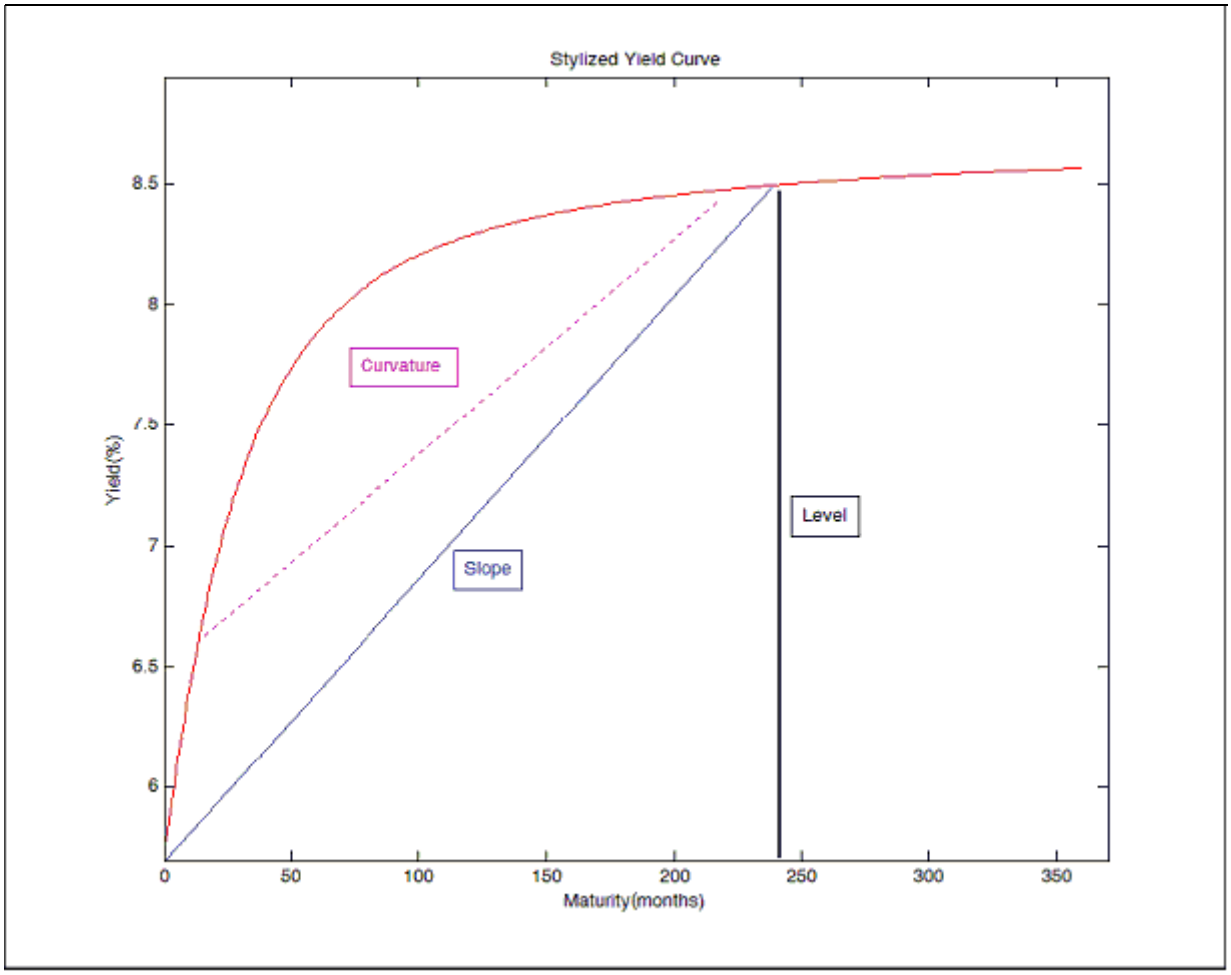

Fig. A - First three components of the yield curve (Level, Slope, Curvature).

EDA begins by first heuristically defining the empirical proxy for the level (L), slope (S), and curvature (C) factors, and seeking to find some interesting relationships among them. The empirical proxies, defined in terms of spot rates, are as follows:

- Empirical Proxy for level (L): 10Y[4]
- Empirical Proxy for slope (S): -(10Y - 3M)
- Empirical Proxy for curvature (C): 2*2Y -3M - 10Y

[4] The ten year yield was chosen given its liquidity and benchmarking importance for long term fixed mortgage rates and corporate bond yield spreads.

Using the historical yield curve data, Fig. 1 plots the empirical proxies for level, slope, and curvature.

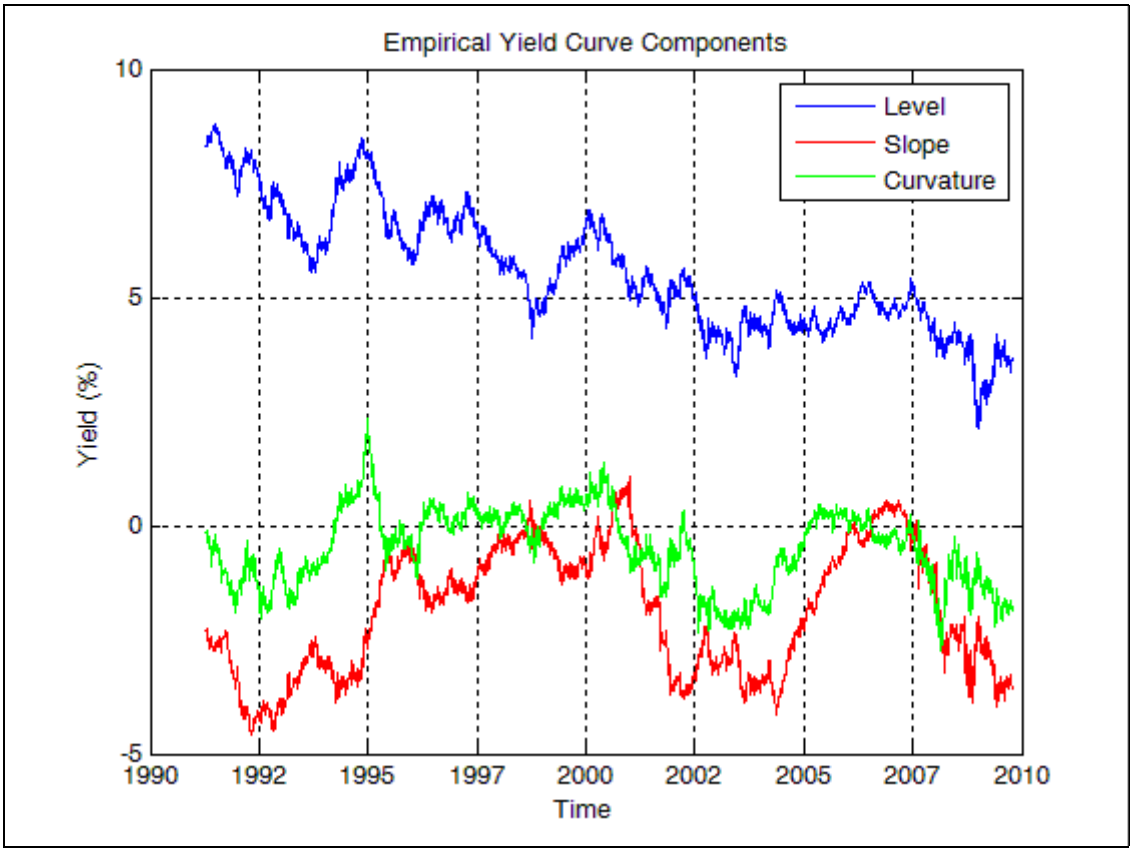

Fig. 1 - Plot of Empirical Proxies for Level, Slope, and Curvature (L, S, C as previously defined).

In order to account for the overall level of interest rates, S and C are normalized by L, such that the new variables, NS and NC, are defined as follows:

- Normalized Slope (NS): -(10Y - 3M)/L
- Normalized Curvature (NC): (2*2Y -3M - 10Y)/L

This normalization accounts for the level of interest rates while still preserving the shape of the yield curve which is predominantly determined by slope and curvature. The (NS, NC) time series is the data set to which EDA and data clustering will be applied. Fig. 2 is a scatter plot of NC versus NS.

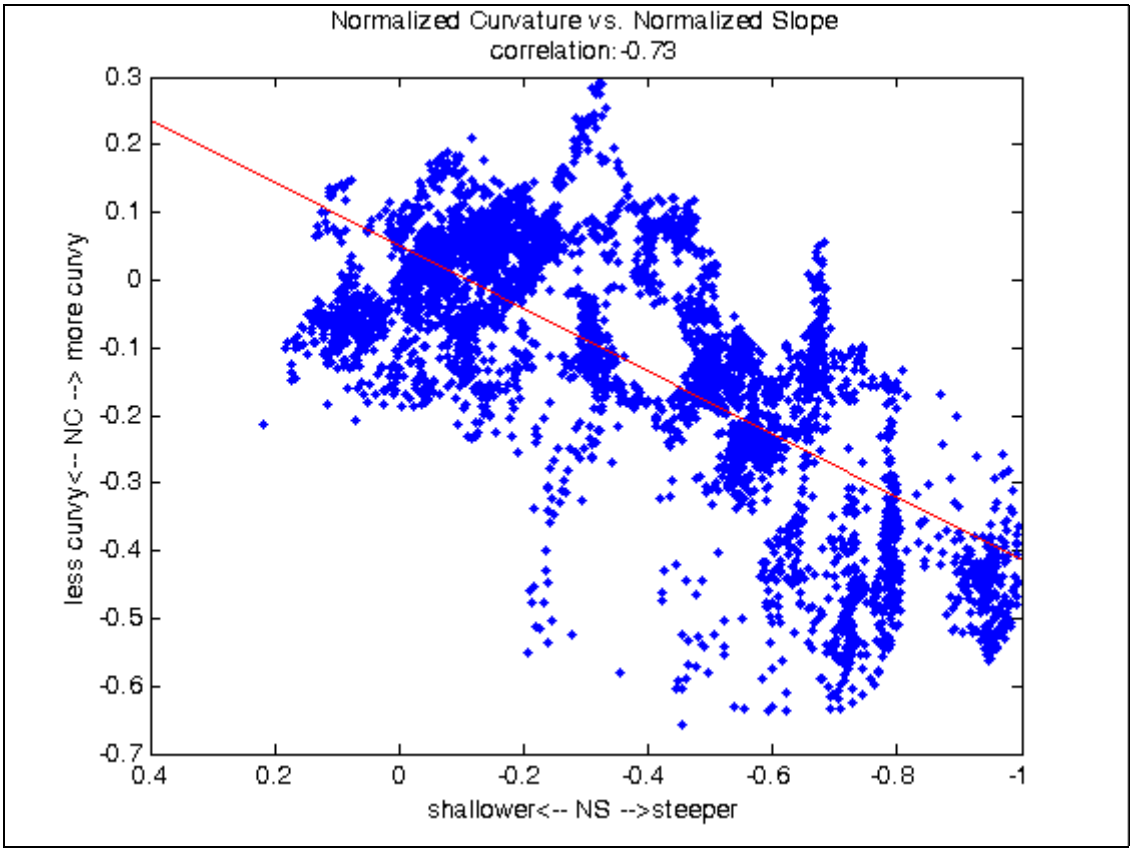

Fig. 2 - Scatter Plot of Normalized Empirical Proxies for Slope and Curvature (NS, NC as previously defined).

While the scatter plot does indicate an overall negative relationship between NC and NS, it only partly assists in the goal of classifying yield curve shape. What one wants to observe is if there might be any naturally occurring concentrations of data. To do this the histogram is applied.

Visualization of Curve Shapes

EDA continues by taking the same data used to produce Fig.2 and plotting a histogram, with the hope that concentrations can be observed which might visually indicate groups of data sharing the same curve shape. From Fig. 3 and Fig. 4, four concentrations, or groups, can be observed. Each of the four groups can now initially be assigned to one of the aforementioned yield curve shape classifications.

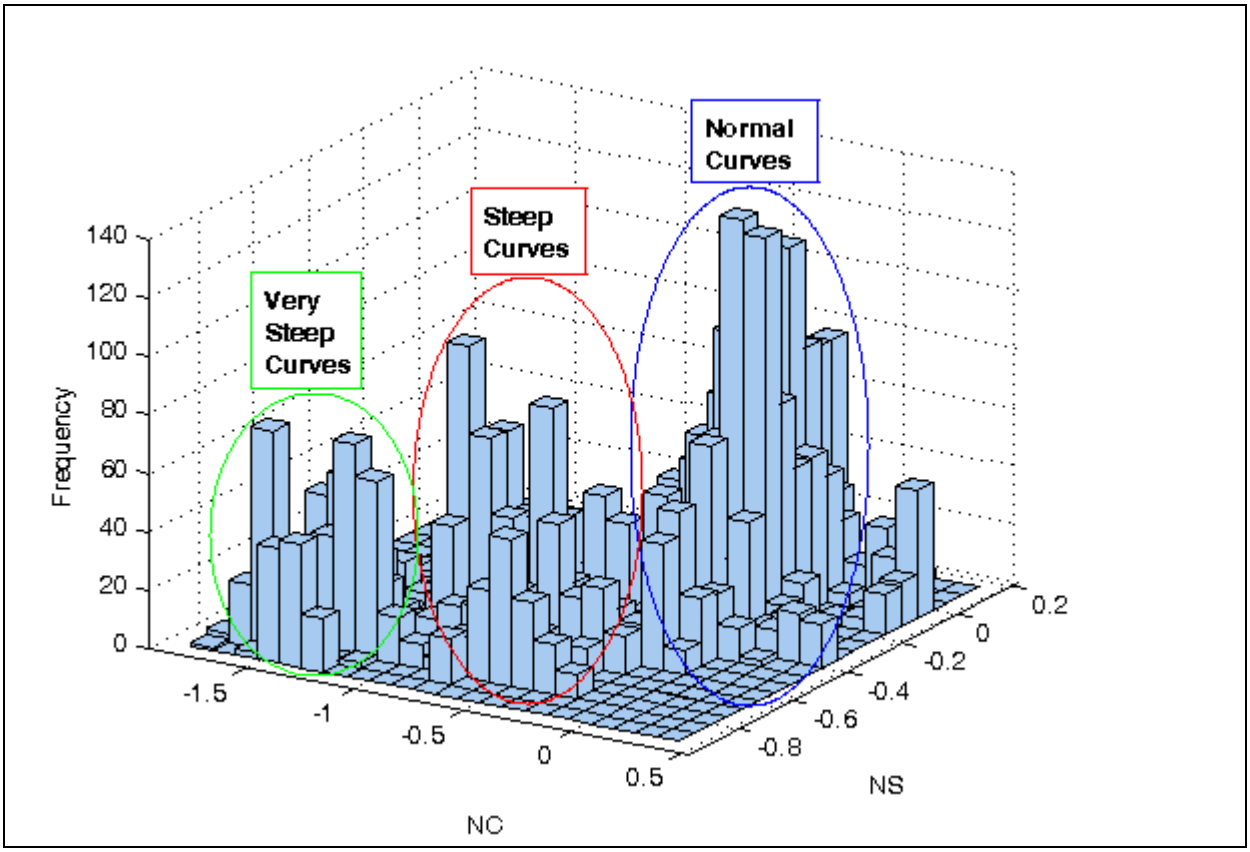

Fig. 3 - Histogram of (NS, NC) data. Note the three main groupings which didn't reveal themselves when viewing the data as a scatter plot in Fig. 2.

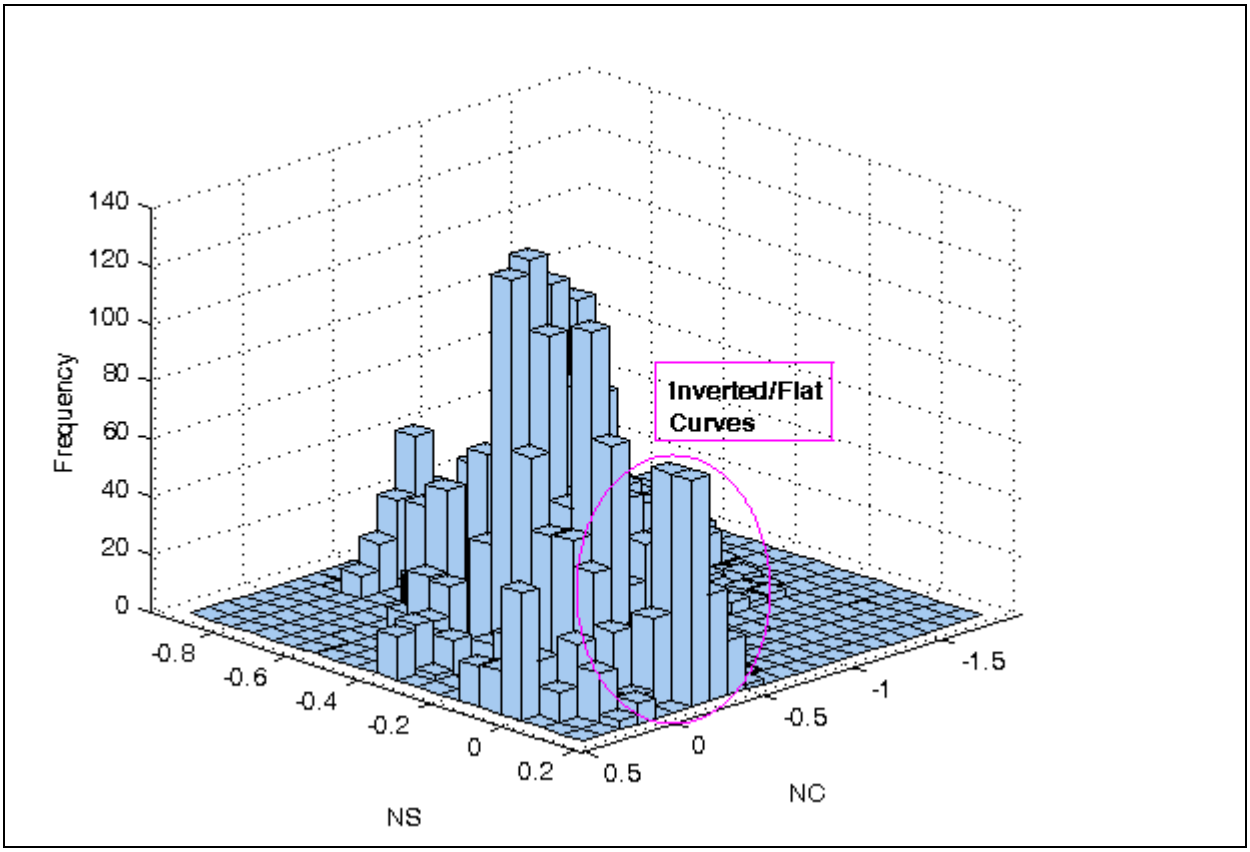

Fig. 4 - This plot is same as Fig. 3, but has been rotated to better display where the Inverted/Flat curves live.

Based on the histogram plots and a subsequent review of the actual yield curves relative to the (NS, NC) data, four main groupings were identified to which each was assigned to one of the previously defined yield curve shapes: **normal**, **steep**, **very steep**, or **inverted/flat**. As it turned out the first three groupings were more easily observable in Fig.3. The fourth grouping, however, revealed itself only after rotating Fig. 3 and by a closer inspection of the data. More on this in next two paragraphs.

For the fourth grouping, inverted and flat yield curves were combined into a single classification due to their similarity, in that the difference between a flat yield curve and an inverted yield curve can often be a matter of only a few basis points, thus making their individual detection very "cuspy".

Upon additional review of the (NS,NC) data it was determined that the *inverted/flat* curves could be directly identified where NS >= 0, as curves which are either flat or inverted must, according to how the normalized slope variable NS was defined, have a value greater than or equal to zero. Given that inverted/flat yield curves were directly defined, the (NS, NC) data corresponding to them will not be used during the subsequent data clustering analysis. Fig.4 indicates where the inverted/flat curves occur in the data.

The initial data grouping and corresponding yield curve shape assignments are currently as follows:

- Grouping 1: Normal curves
- Grouping 2: Steep curves
- Grouping 3: Very Steep curves

- Grouping 4: Inverted/Flat curves (where NS >= 0)

To confirm the preceding visual identification of the first three curve shape groupings only, the K-means data clustering technique is applied to the (NS, NC) data.

K-means Clustering

K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set via a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in an intelligent way because different locations will create different results. Therefore, the better choice is to place them as far away from each other as possible. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an initial clustering is achieved. At this point one needs to re-calculate k new centroids as mass centers of the clusters resulting from the previous step. After arriving at the k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. At this point a loop has been generated. If one continues looping he will notice that the k centroids change their location until, at some point, the centroids do not move any more.

The k-means algorithm aims to minimize an objective function, which is typically some sort of distance measure such as Manhattan distance (L_1) or Euclidean distance (L_2).

More generally, the objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2,$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$, and the cluster center c_j, is an indicator of the distance of the *n* data points from their respective cluster centers, or centroids. The general k-means algorithm is composed of the following steps:

1. *Place K points into the space represented by the data that are being clustered. These points represent initial group centroids.*
2. *Assign each data point to the group that has the closest centroid.*
3. *When all data have been assigned, recalculate the positions of the K centroids.*
4. *Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the data into groups from which the metric to be minimized can be calculated.*

Although it can be proved that the procedure will always terminate, the k-means algorithm does not necessarily find the most optimal configuration corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial randomly selected cluster centers (centroids). To overcome this last point, the k-means algorithm can be run multiple times to greatly

reduce the initialization effect. Additionally, any expert knowledge/opinion one might have about the data should be leveraged in order to assist the cluster formation, in particular, in choosing the number of clusters k to use.

K-means Identification of Curve Shapes

Letting the objective function be squared Euclidean distance, excluding data where NS >=0, and setting k equal to three, K-means was applied (invoking the *kmeans* Matlab command) to the (NS, NC) data. Fig. 5 is the graphical output of analysis and is called a "silhouette plot" . After several runs, the best run produced an average silhouette value of 0.7355, which is relatively high, and indicates the choice of three clusters is probably correct. Visually, one can observe three distinct "knife blade" shapes in the plot and very few negative values, which is a more intuitive indication that the clusters are probably correct. Also, the majority of the data can be seen falling into the normal curve shape designation, followed by the steep and very steep shape designations, respectively, and is in line with the frequency of the data we observed in the histogram plots, as well as expert knowledge/opinion.

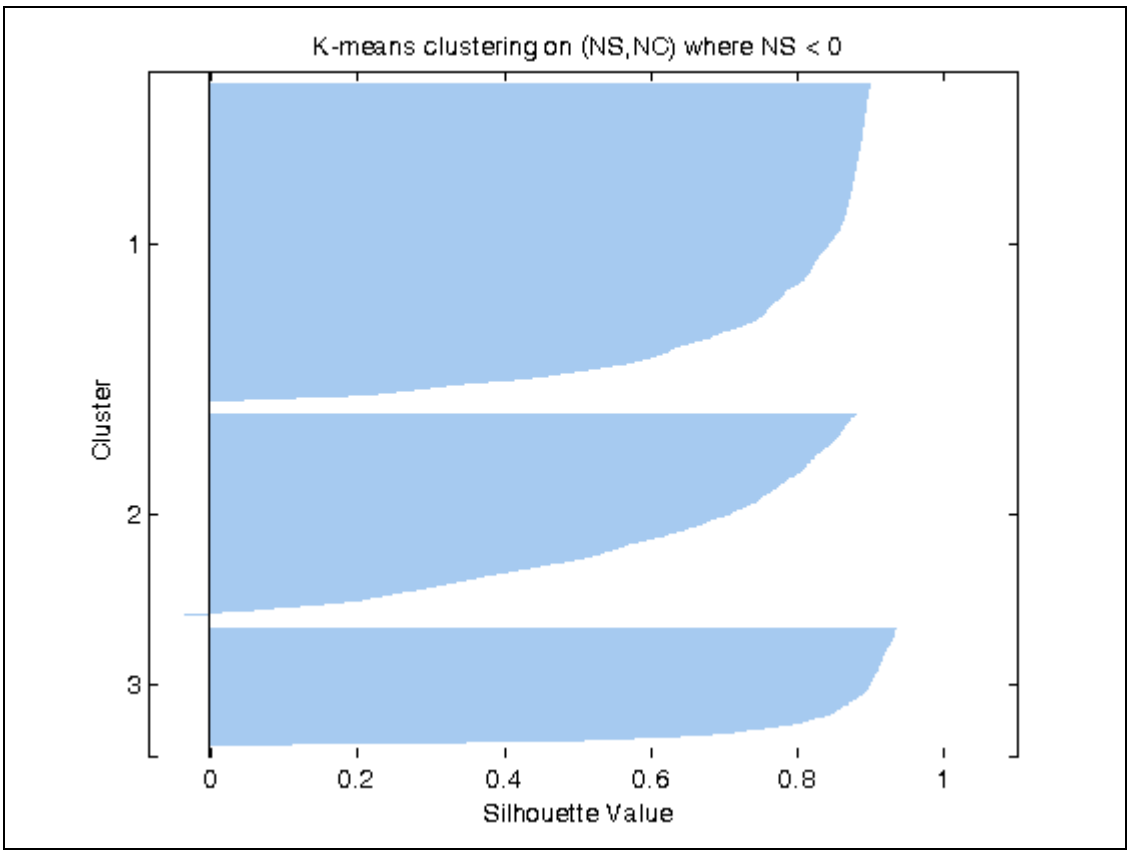

Fig. 5 – Silhouette plot visually indicates that K=3 is probably the correct choice. Three clusters also coincides with what is known about yield curve shapes.

To see how well the k-means produced clusters match the initial groupings, in Fig. 6 (NS, NC) were scatter plotted with the data points color-coded according to the cluster number to which they had been assigned. Fig. 6 can then be visually compared to the initial groupings determined from Fig.3 and Fig.4 to check how favorably the corresponding data groupings align.

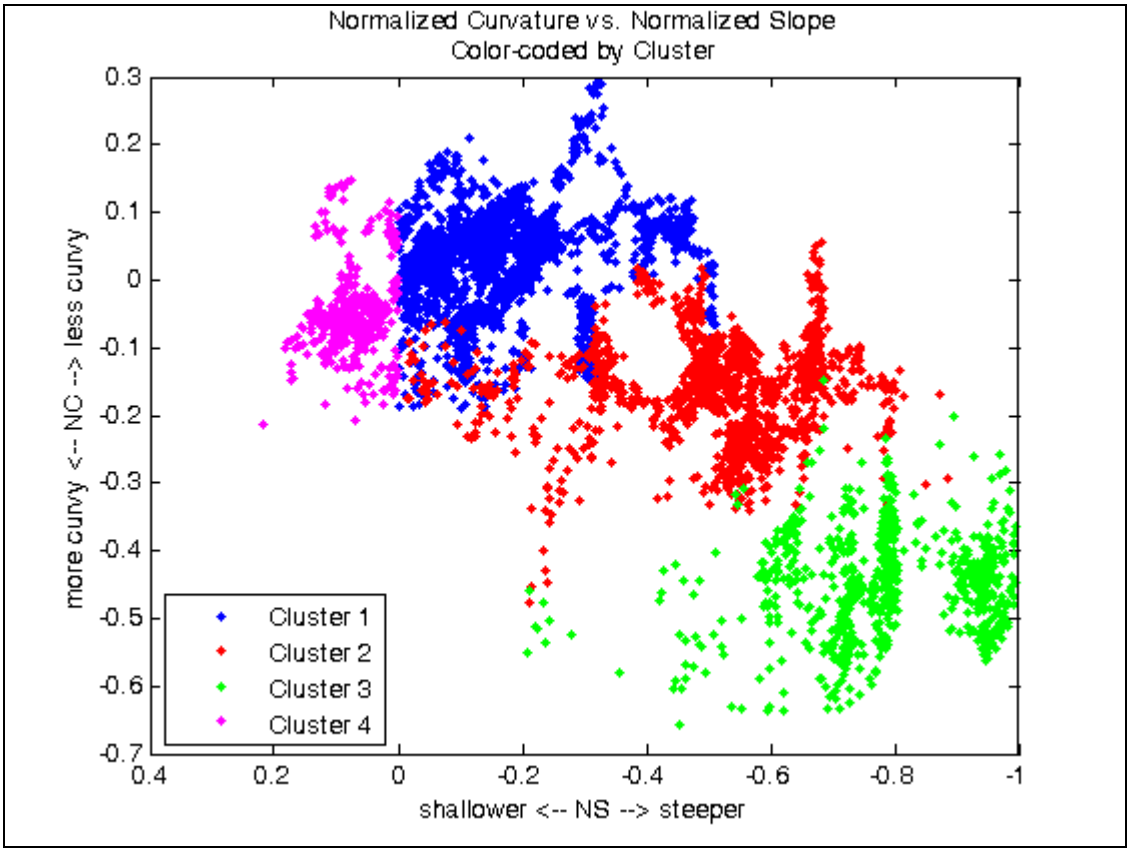

Fig. 6 – This is Fig. 2 with data values color-coded according to the cluster they belong to. Notice how well the colored groups line up with the groupings identified in Fig. 3 and Fig. 4.

Comparing Fig. 6 and Fig. 3, one can observe close agreement between the color-coded scatter plot and the initial visual groupings from the histogram. This indicates that the K-means clustering algorithm performed a satisfactory job of discriminating amongst the normal, steep, and very steep curve shapes embedded in the (NS, NC) data. Again, Cluster 4 was not determined using K-means; it was done directly according to how inverted/flat yield curves were defined.

One can now switch from the informal groupings visually defined from the histogram plots to using the analytically constructed (and mutually exclusive) clusters corresponding to the yield curve shapes previously defined:

- Cluster 1: *Normal curves*
- Cluster 2: *Steep curves*
- Cluster 3: *Very Steep curves*
- Cluster 4: *Inverted/Flat curves*

Finally, one can visually inspect the following plots to see how well the actual yield curves compare to the clusters to which they were assigned, as per Fig. 7, Fig. 8, Fig. 9, and Fig 10 below.

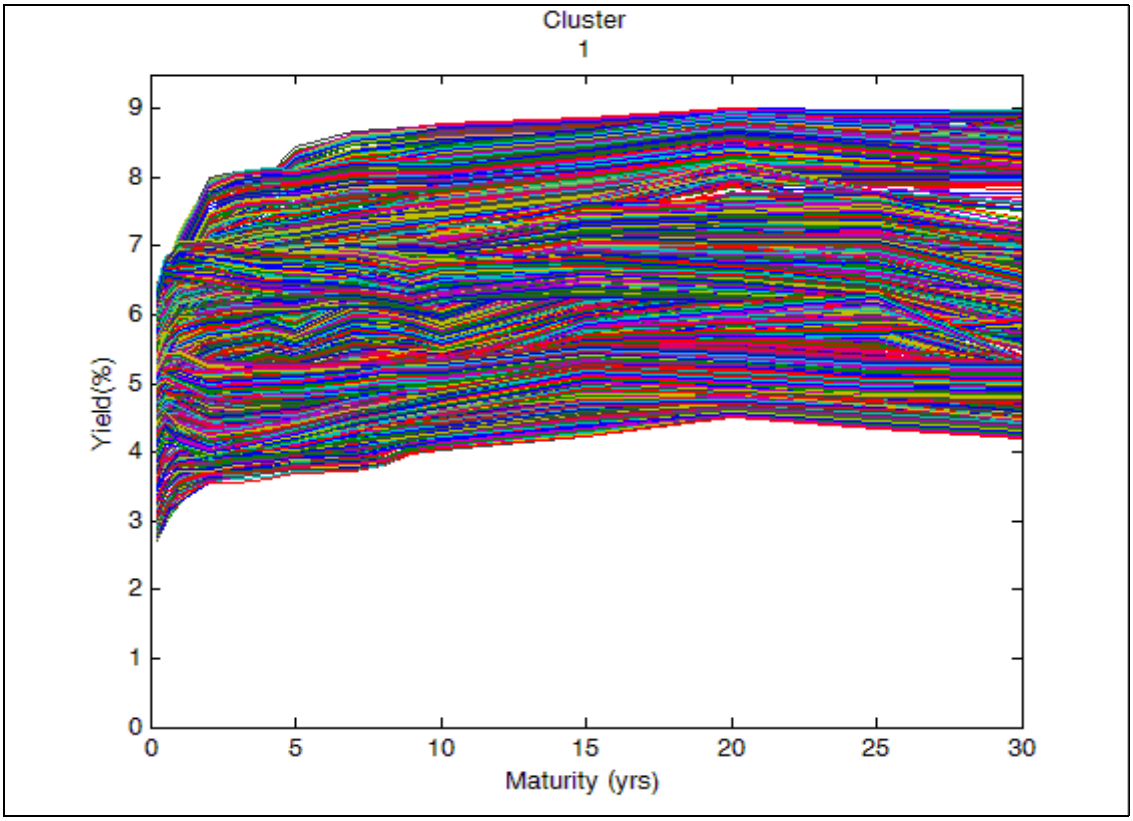

Fig. 7 – Cluster 1: ***Normal curves***

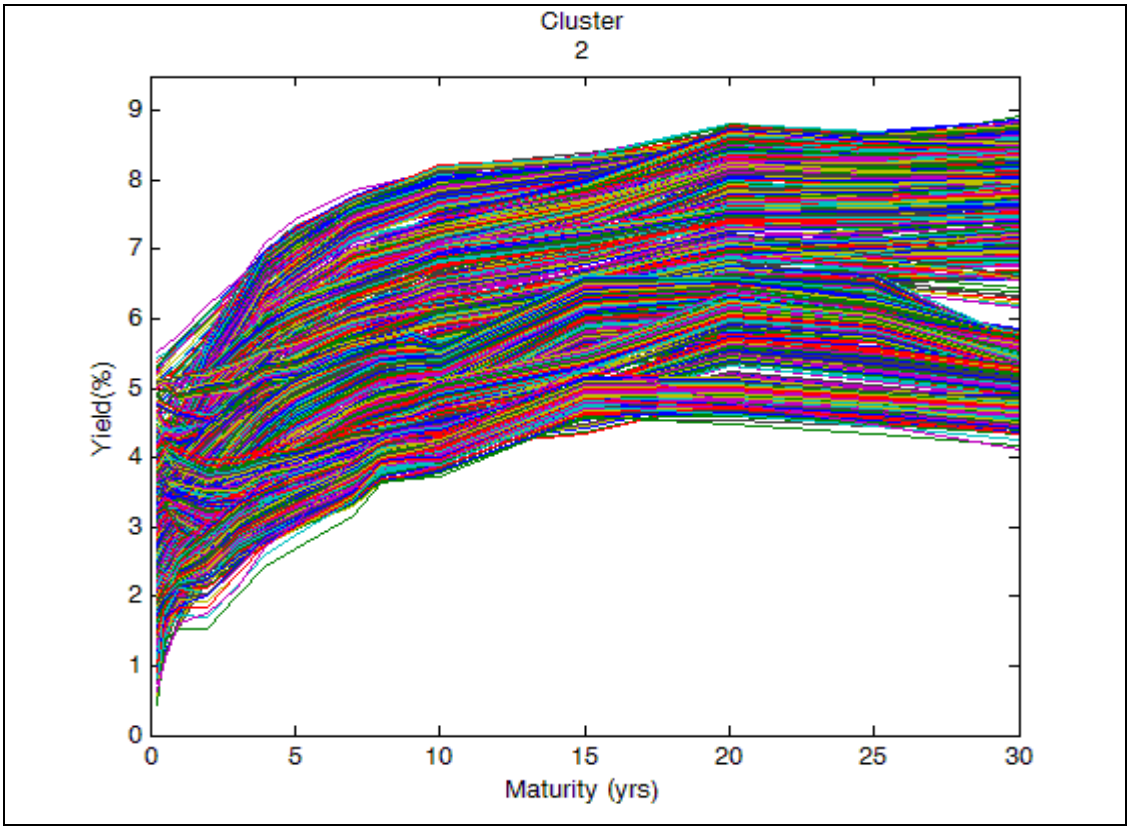

Fig. 8 – Cluster 2: ***Steep curves***

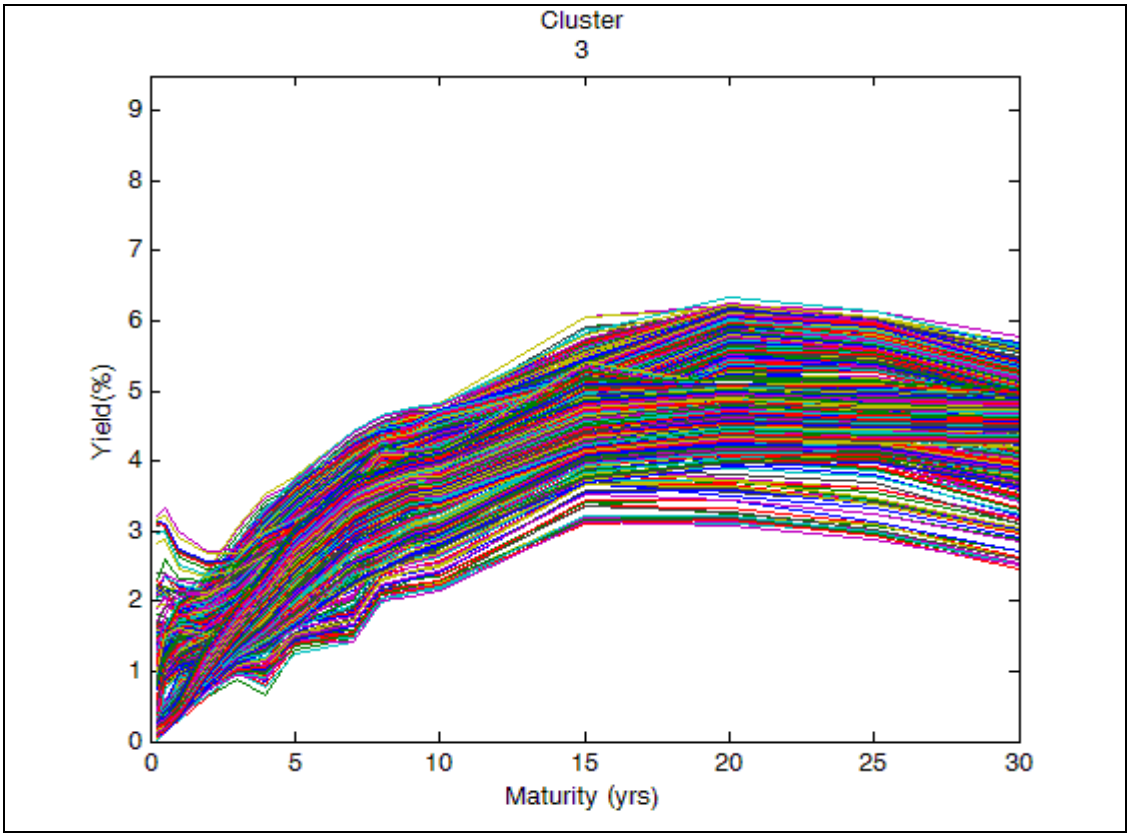

Fig. 9 – Cluster 3: ***Very Steep curves***

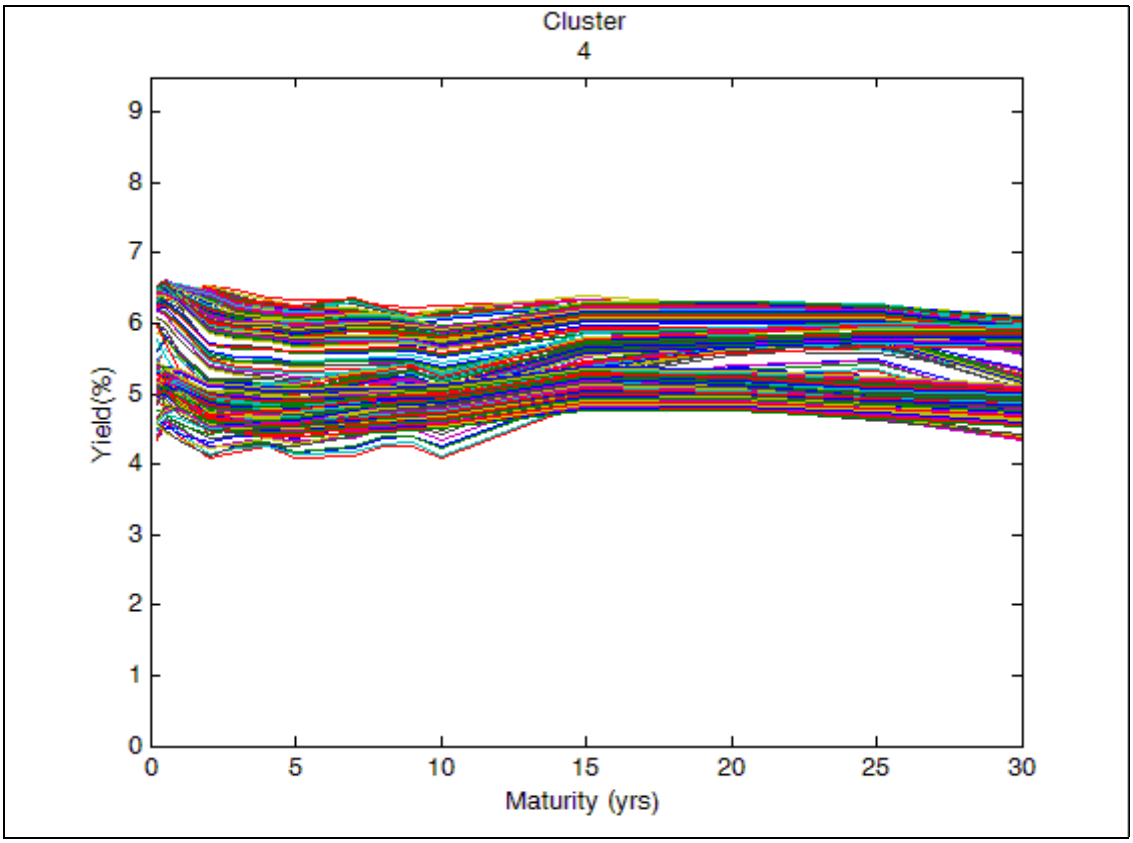

Fig. 10 – Cluster 4: ***Inverted/Flat curves***

What can be observed is that the normal curves look, well, normal, while the distinction between steep and very steep curves seems to have been captured. As already mentioned, the inverted/flat curves were determined directly by definition. Overall, K-means seems to have done a satisfactory job of differentiating the data.

Conclusion

To summarize the preceding, the following high level steps were undertaken in order to assign the daily government spot yield curves according to one of four pre-defined shape classifications:

- Defined the data set to be analyzed (NS, NC) and the classifications to which the data would belong (the yield curve shapes).
- Applied scatter and histogram plots in order to better visualize the data with hope of revealing interesting relationships and intuiting distinct data groupings. From this exercise initial data groupings were determined.
- Applied K-means data clustering algorithm to the data set in order to form data clusters.
- Compared the data cluster results to the initial graphical data groupings determined by the scatter and histogram plots, and visually determined how well the two agreed.
- For further confirmation, graphically compared the K-means cluster assignments derived from the (NS, NC) data to the original untransformed data (e.g. the spot yield curves), and concluded that the clusters found by K-means satisfactorily differentiated the data and allowed one to classify the daily spot yield curve data according to shape.

Having classified the original spot yield curve time series data according to shape, one could conduct further research and use the shape assignments to inform further EDA or subsequent model building.