Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de

Task 1: CX Decomposition

Description of column of C and rows of X:

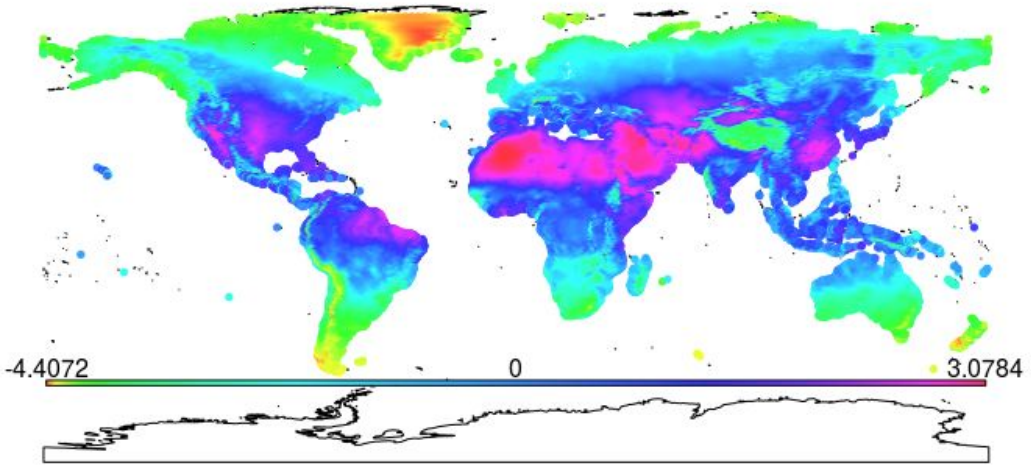
In CX decomposition, some columns of original data are selected via two phase CX algorithm into matrix C. So **C has climate data as rows and locations as columns (a**s worldclim.csv data is transposed). We calculate X by taking pseudo-inverse of C and multiplying it by the original matrix. So comparing with the news data, we can say, **C is a data set which contains climate data versus climatic regions descriptions and X is climatic regions versus position (in terms of latitude and longitude).**

Plots with selected columns from original matrix for matrix C:

For k=5, .i.e. 5 columns are selected after two phase CX algorithm, where column contains location information. A, B, C, D, E is respectively marked in the map below as locations with exact longitude and latitude.
Interpretation: South America, India, China and two locations in Australia are selected as the 5 columns. These interpretations are drawn with respect to the initial choice of 5 locations.

Plot for matrix X [1,]
These are climatic regions for a single factor (min temperature in January) plotted against position (latitude and longitude)
Interpretation: Northern part of South Africa, Gulf countries and India were hottest while the North America, Russia, Australia and Europe were the coldest in January. These interpretations are drawn with respect to the initial choice of 5 locations.

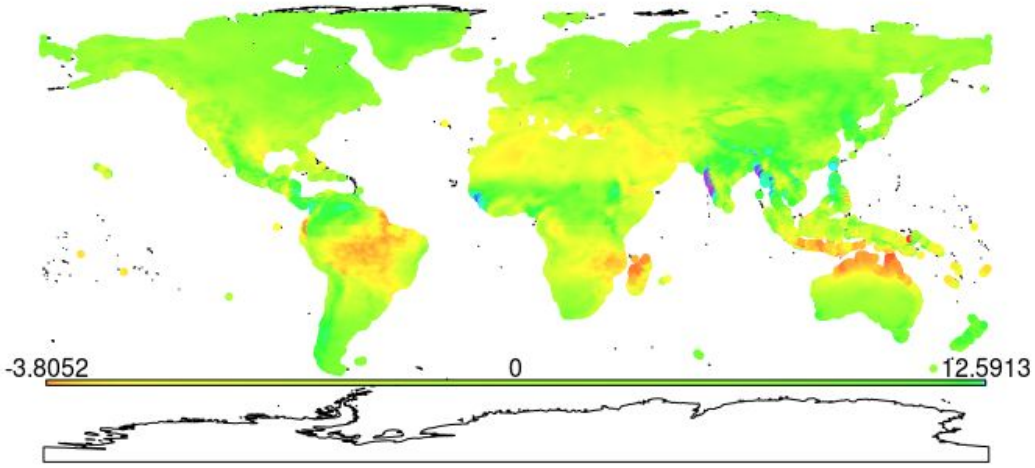
Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



Plot for matrix X [2,]

This are the regions divided according to min temp of February.

Interpretation: While some parts of Africa, North America, Russia and China were hottest, most of the other parts of the world were cold in February. Arctic region was the coldest of all.These interpretations are drawn with respect to the initial choice of 5 locations. This interpretation is entirely different than the previous one.

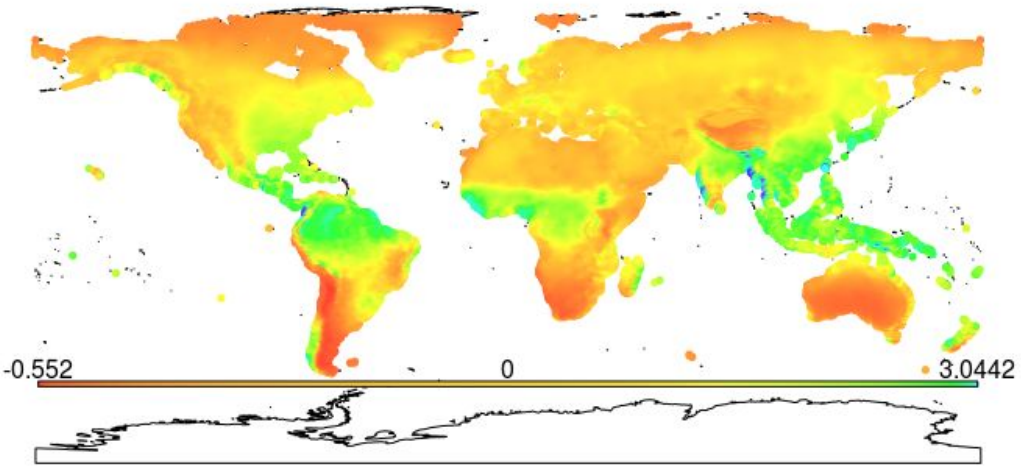


Plot for matrix X [3,]

These are the regions divided according to min temp of March.

Interpretation: This interpretation is again very different from the previous two. These interpretations are drawn with respect to the initial choice of 5 locations. North America. South East Asian countries are the hottest. Arctic region is still chill cold in March while most of South America is colder too. Most of the other countries are facing cold winter in this month.

Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



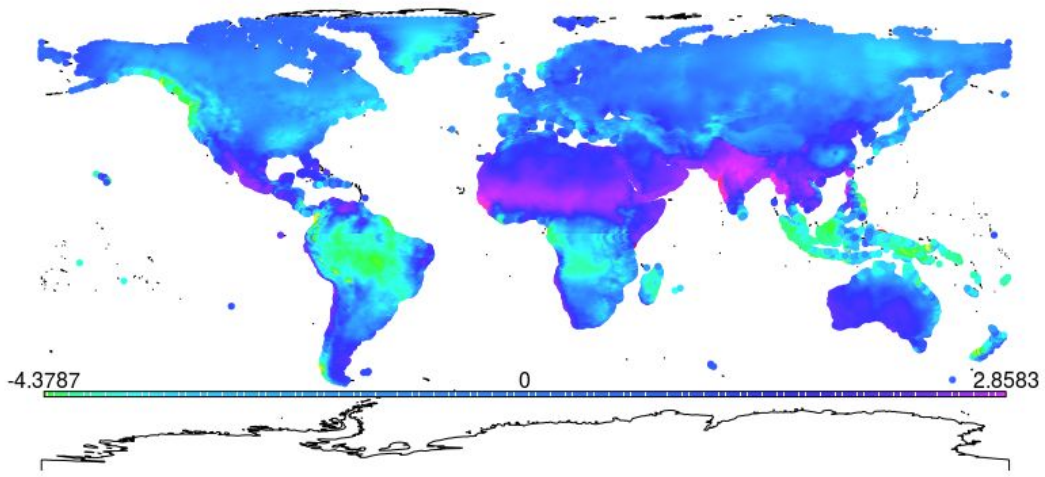
Conclusion: The Arctic region is always extremely cold in the above plots. The Gulf countries and Northern part of South Africa and some parts of Asia are extremely hot. North America and South America have both hot and cold regions like that of Australia throughout the year. These interpretations are drawn with respect to the initial choice of 5 locations. However, for every month, the visualization shows an abrupt change in the climate.

Rank = 4 and Fudge = 8

The locations are different for different locations obtained from C initially as seen in map below.

After two phases, 4 locations are selected and corresponding to these locations regions are mapped according to their minimum temperature in January. From the plot it looks like it was extremely cold in South America as compared to India and some parts of Africa. But the interpretation with k=5 and fudge = 5 was way better as it had more variability in its visualization.

Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de

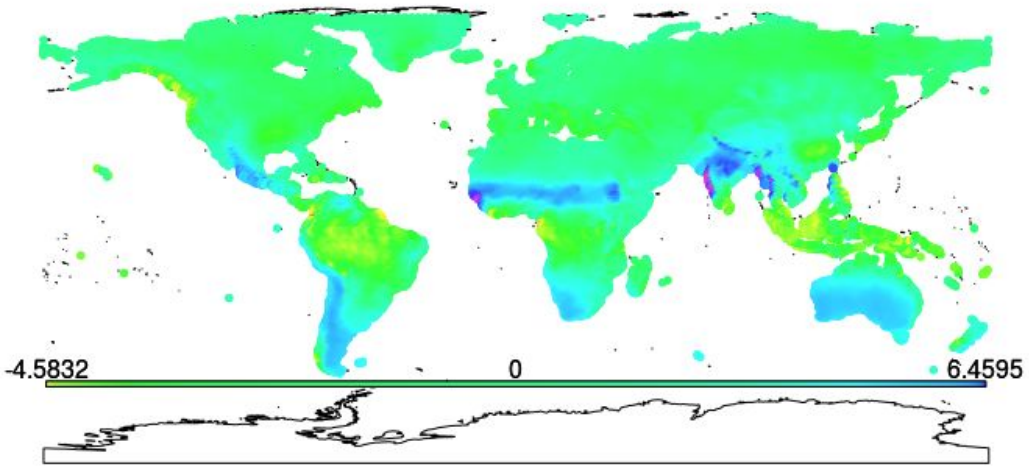


Rank = 5 and Fudge = 1

The locations obtained here are again different from those obtained from fudge = 5.

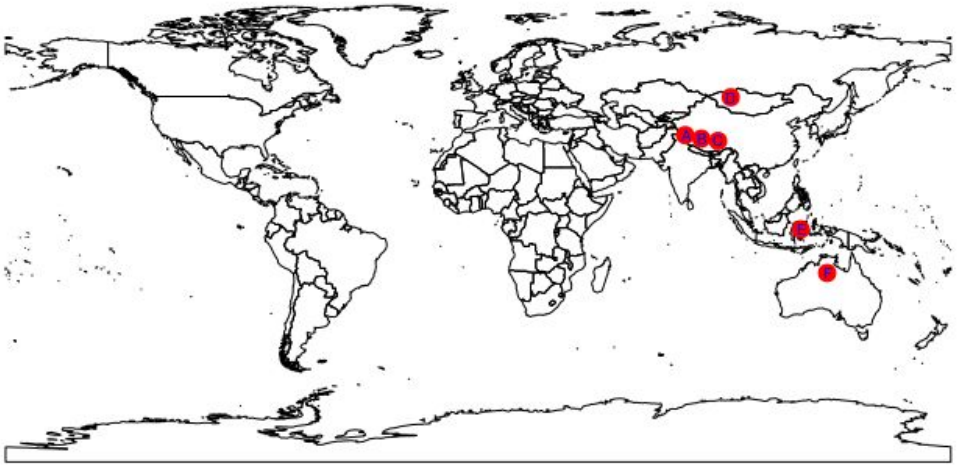
This visualization is showing as if majority of the countries were cold in January. For fudge= 1, except some parts of India, Australia and south of Southern America, all are marked as almost colder regions. In conclusion, the variability of the visualization is lost here which was way better in previous plots. Thus, we conclude that sampling ensures that the variability of the data is captured in the reduced representation and thus, I will give my vote to sampling

for this kind of dataset.



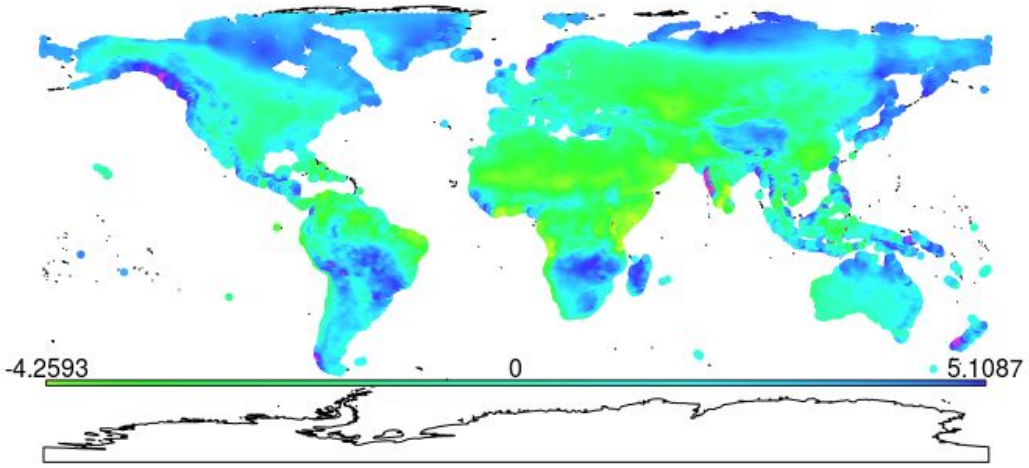
Rank = 6 and Fudge = 6

The locations are different for different locations obtained from C initially as in the plot below.



After two phases, 4 locations are selected and corresponding to these locations regions are mapped according to their minimum temperature in January. This plot marks South Africa, Russia, Gulf Countries and most of Asian countries as colder countries as compared to

North America which is again strangely different from previous interpretations.



Interpretation from all above different results:

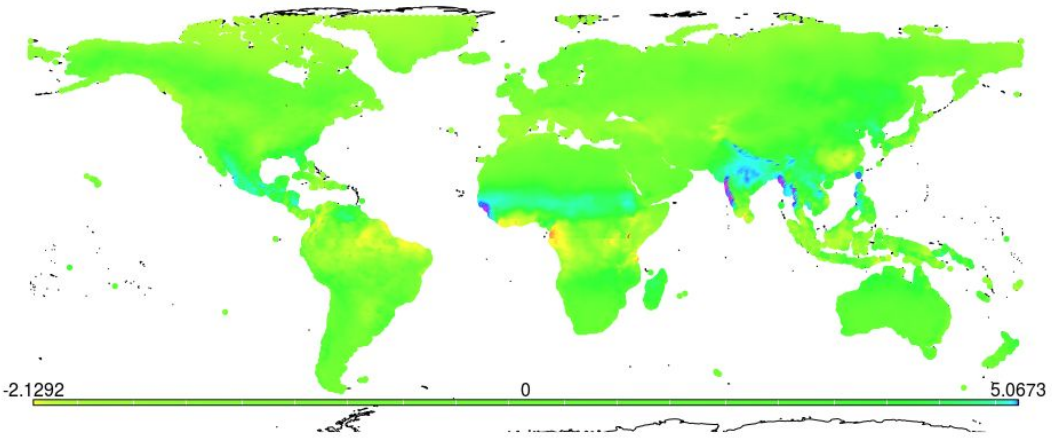
When we change the parameters i.e. the fudge value and the rank, the visualization plots are different for different value of parameters and show abrupt changes. The interpretations were not same from the different plots above. However, to study individual components in greater detail, CX decomposition is a good choice. Thus, if we consider the climate map data, then, CX representation will be good match if we have to study individual components like minimum temperature in a particular month or precipitation in a particular. But for overall interpretation, one has to be sure about the parameter values in order to get the best of the interesting result sets. Also, sampling should be done for this kind of dataset otherwise the variability in the data is not captured finely as can be seen in the above plots.

Task 2: Nonnegative CX Decomposition

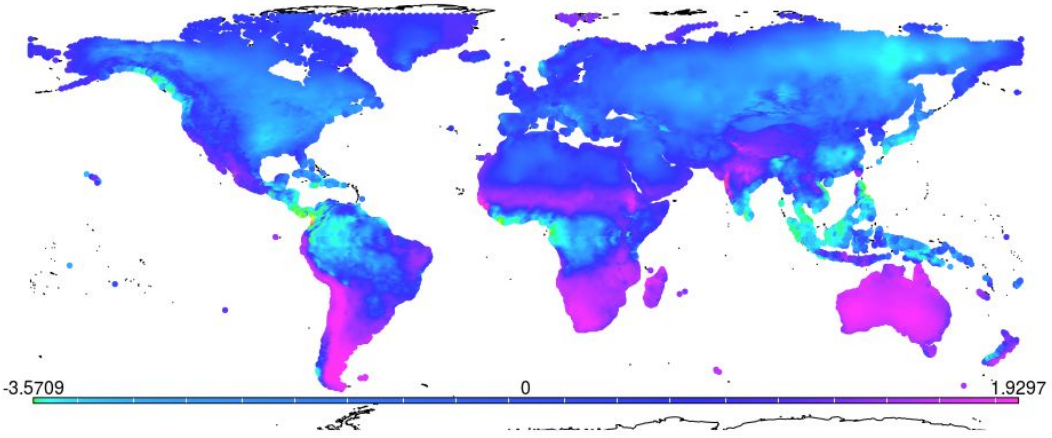
Interpretation using Non-negative CX Decomposition gives the following results.

For k=5, the locations obtained are as in the plot below. These are different than before:

And the corresponding visualization for the minimum temperature in January with respect to these locations is:

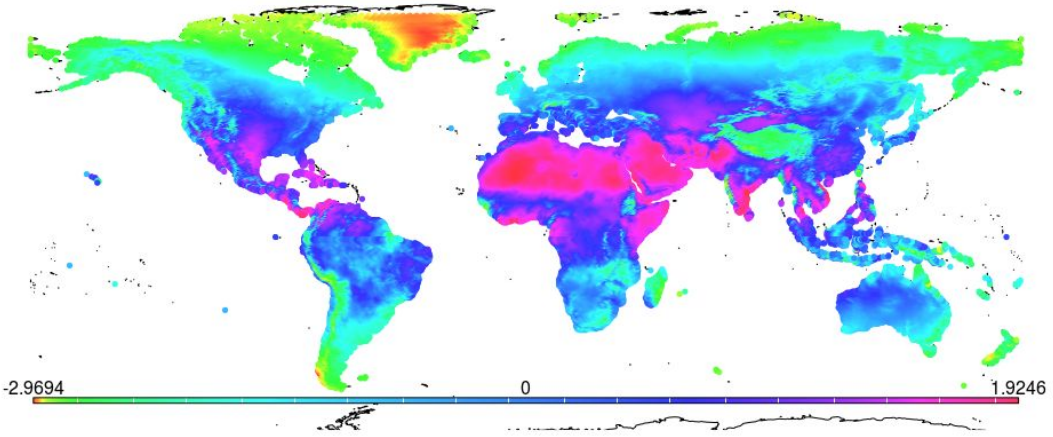


The visualization here is not as clear as that of the Task 1. It is difficult to find the correct interpretation. However, India and some regions in South Africa are still the hottest and the Arctic region is the coldest which is a similar interpretation we got from the Task 1. However, the interpretation in Task 1 was more detailed.
The visualization for minimum temperature in February is:



The visualization for minimum temperature in March is:

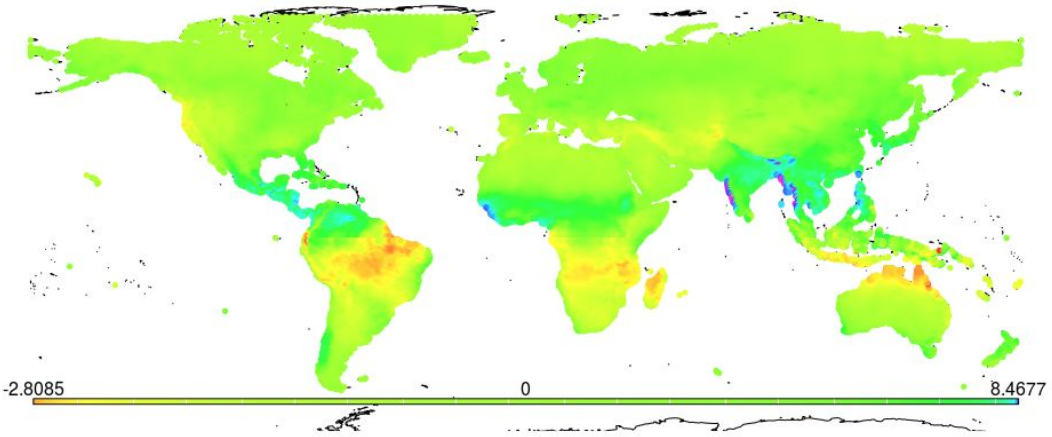
Here the visualization has variability in it and makes it a better representation than that of the Task 1. However, for other settings, CX gives better visualization.



For k=4, the locations obtained are:

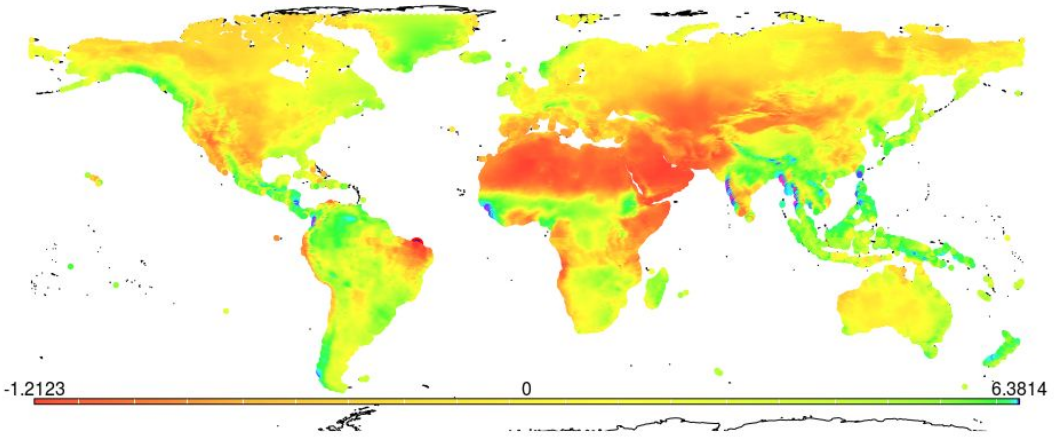
And the corresponding distribution plot obtained with respect to these locations for minimum temp in January is:

Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



For k=6, the locations obtained are:

And the corresponding distribution plot obtained with respect to these locations for minimum temp in January is:



In the above settings, the interpretation is similar to that in Task 1. However, the reconstruction error is worse here in convex cone as compared to cx decomposition. This possibly seems right as in convex_cone, we are in attempt to maximize the volume of the cone formed as a result of the extremal cone vectors while in the case of cx decomposition, the focus is on doing sampling again and again so as to avoide any kind of redundancy. Interpretation is better in CX decomposition.
In conclusion, I will choose CX decomposition for the World Climate data because it involves random sampling in its representation. These samples are a good representation of the data on which we are assuming that there is some redundancy present. Thus, with repetitive sampling, it ensures to hold more variability in its data representation and thus, avoids redundancy to its core.

Task 3: ICA for housing prices

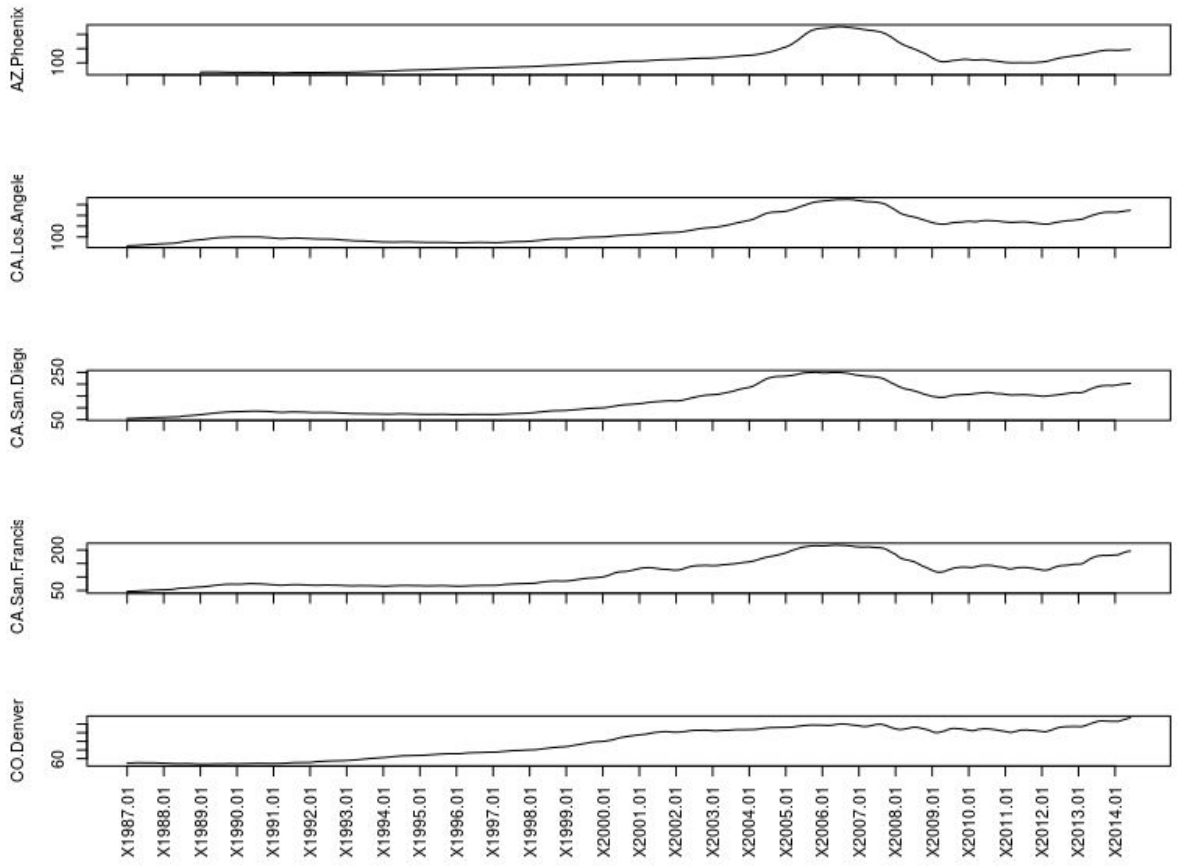
Is normalization a good idea?
The index values vary for different locations in the data. Thus, we normalized the data to Z-scores. Yes, it is a good idea to normalize the data as if data will come under a certain range, then, interpretation becomes easier and better convergence is achieved. So, the different indexes are now brought to the same scale centered around mean to have a better comparison against each other. One can argue that with normalisation, we wouldn't be getting the exact value of the cost at that place, but for the purpose of analysis, it is a good choice.

Different components and what do they represent
After computing the fastICA of the transposed normalized data, we get 5 components: W, K, X, A, S where, K is the pre-whitening factor .i.e. the number of independent components which we need to extract out of the given data. Here we have 20 locations whose

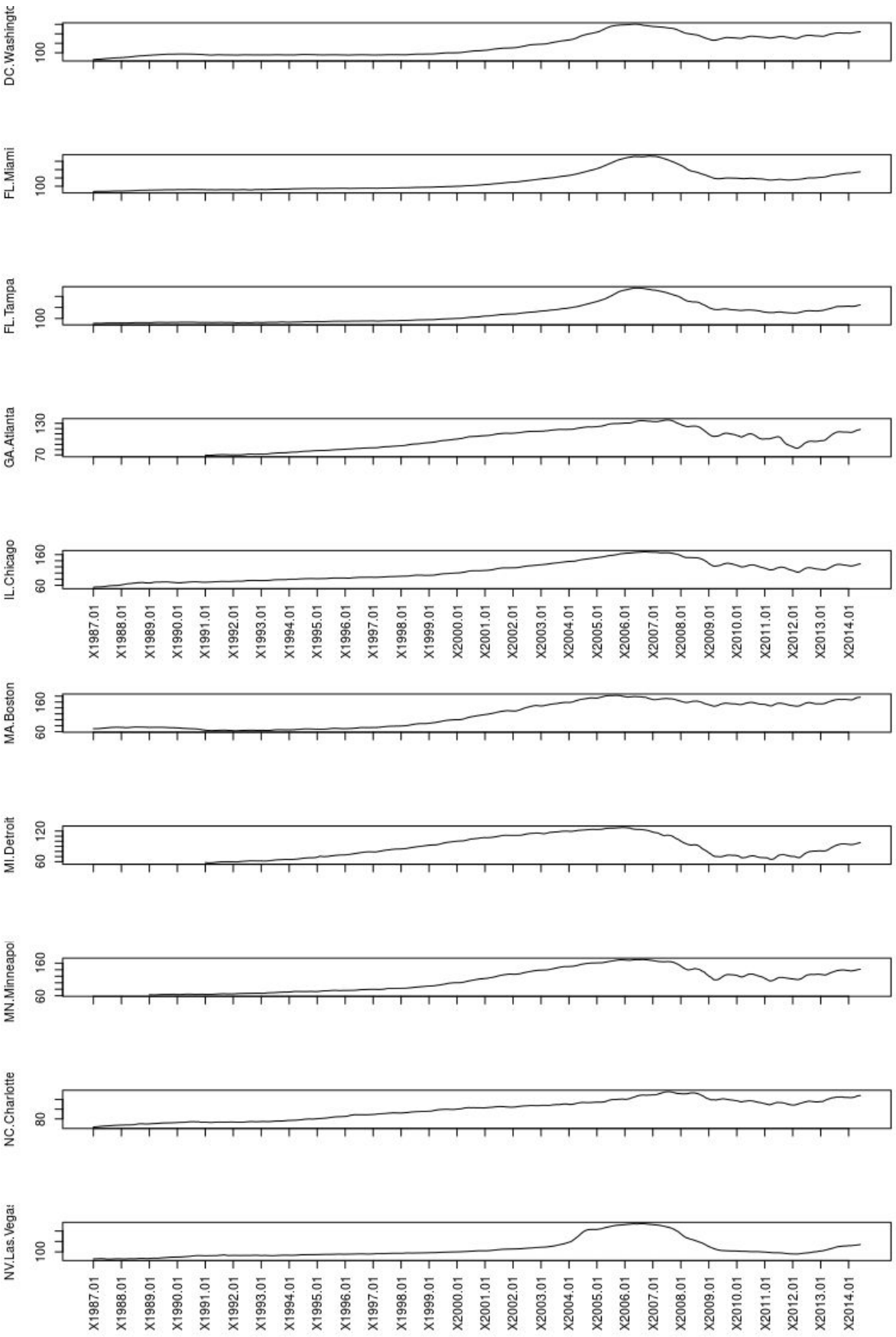
independent components are to be computed. So in our case S are the independent components, alongside A is the mixing matrix as per definition (‘ica_intro’ slide). Finally X is our data matrix with the help of which A and S is computed.

How to reconstruct the housing price index of California, LA?
If we multiply S with A, then we are actually getting back the original data. So basically the second column of matrix S is the first independent component of the overall matrix. So it represents Los Angeles, CA data as an independent component.

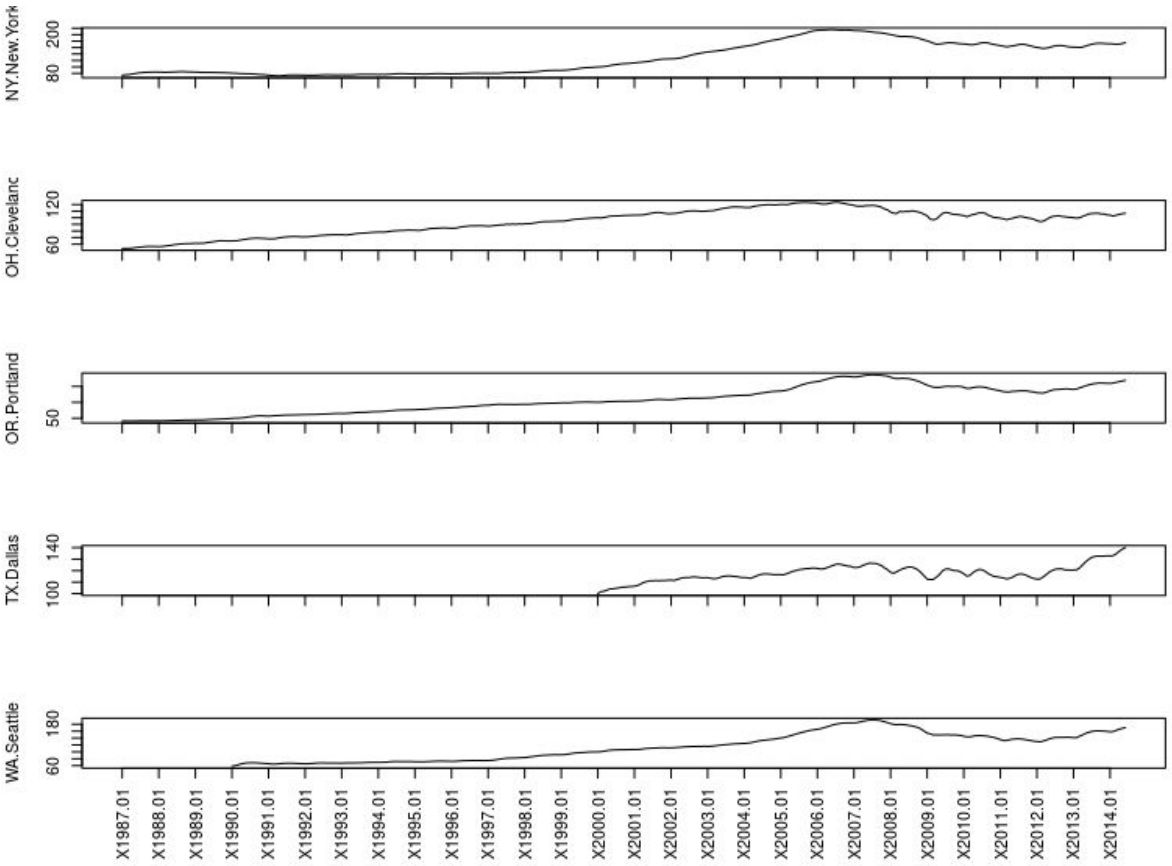
Plots:
The plots for the 20 different countries is as below:



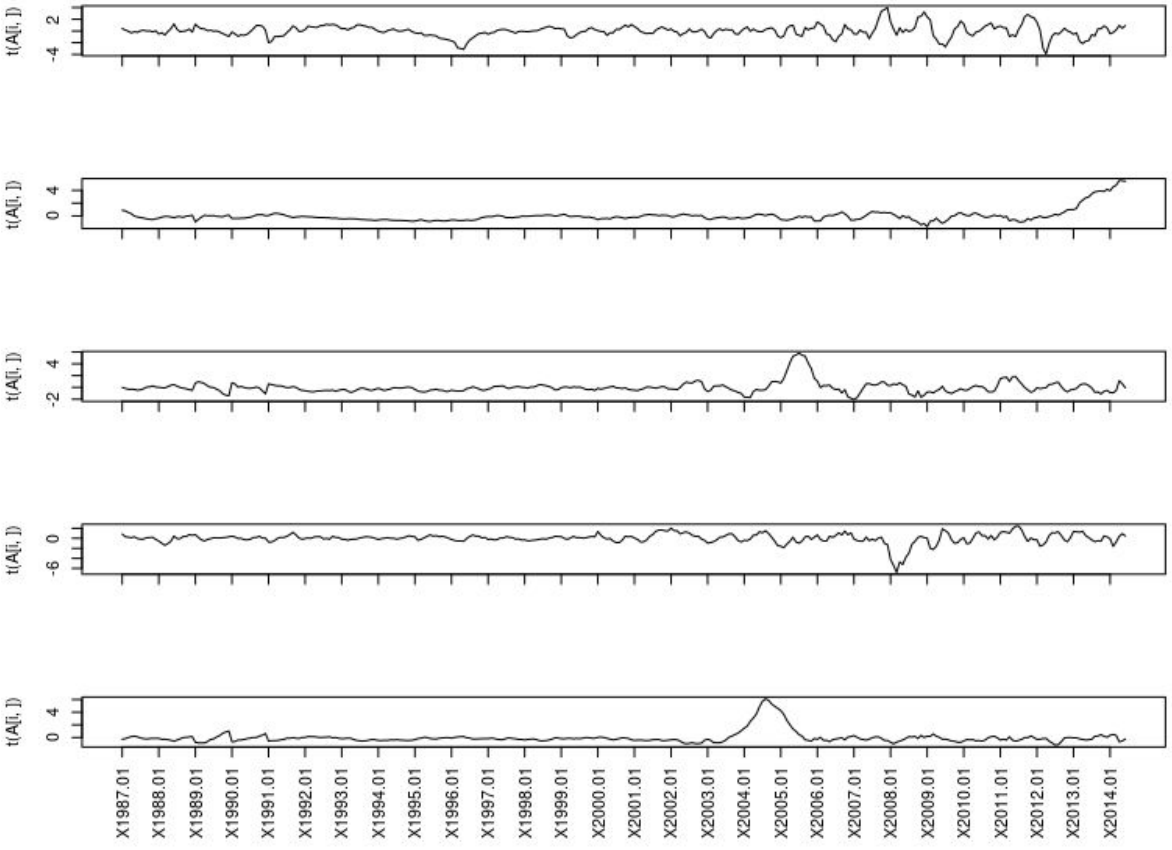
Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



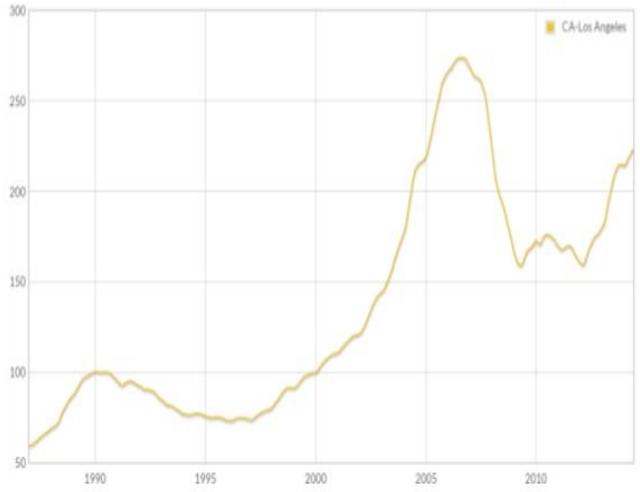
The plot for the independent components of the first 5 countries is

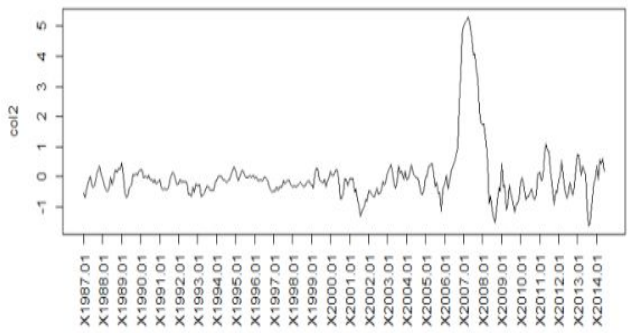


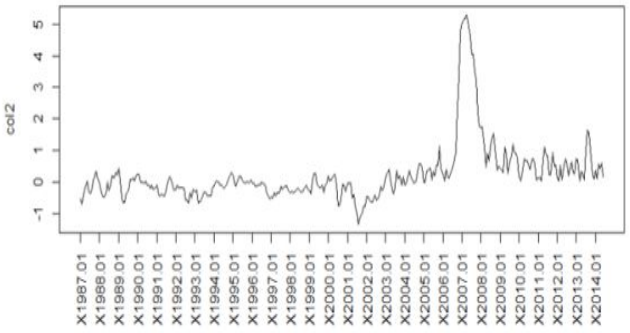
Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de

From the context above first we visualize the independent LA-CA housing price index. By studying the data from the provided link, we get,

1. Independent component for Los Angeles, CA:

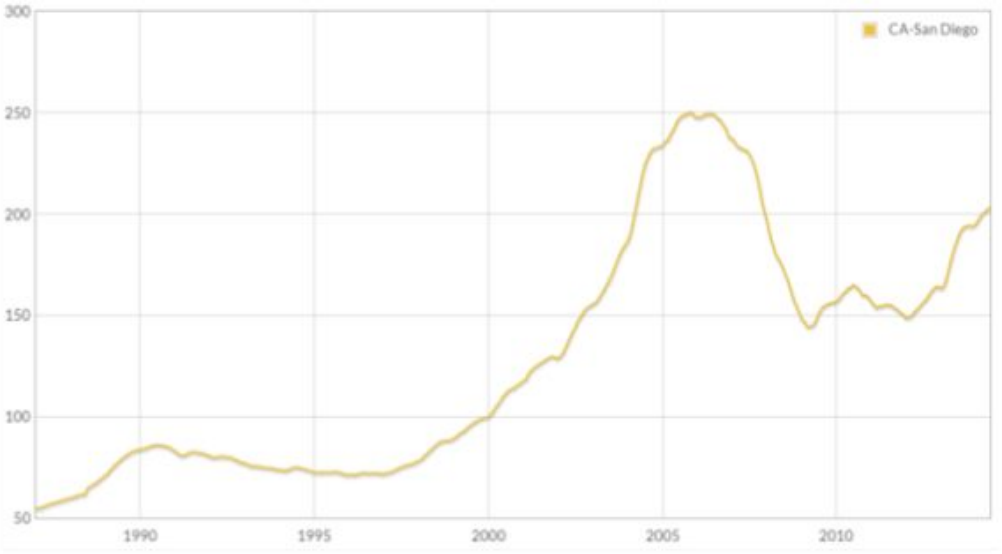




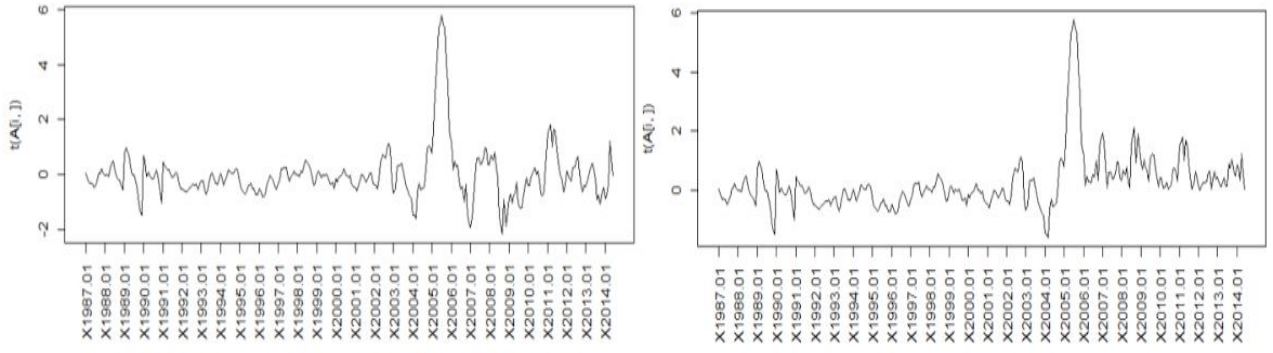


To analyze the data, I adjusted few columns which were needed to be multiplied by -1 in order to get a perfect analysis, the before and after effects of multiplication are shown. Now if compared with the original data, things do match in a certain way. Like,
1. The highest peak is at 2007 only.
2. The range from 1990 to 2007 is more or less same (the positive slope)

Independent component for San Diego, CA

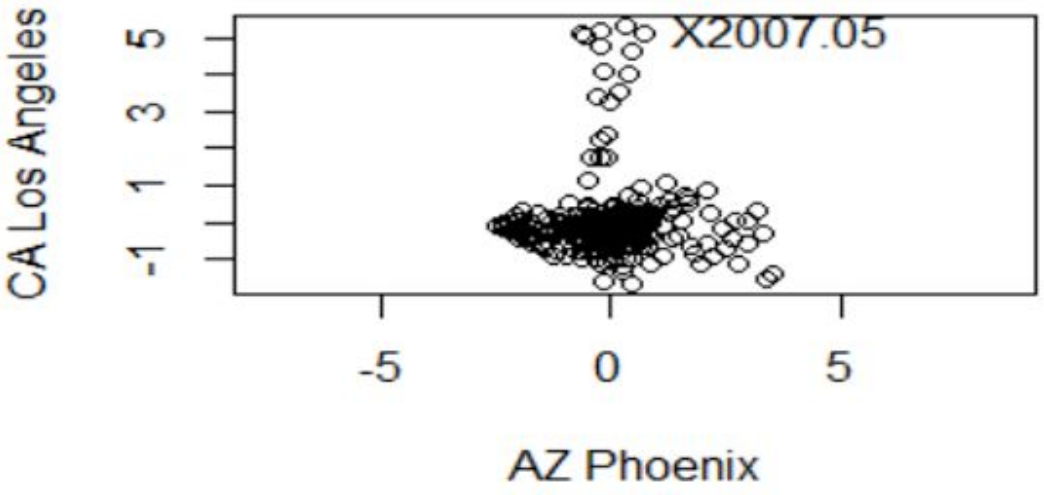


Similarly for the San Diego, CA data, after multiplying by -1, the original projection becomes clear. Like the LA data here also highest peak of the independent component and the observed data matches perfectly. Plots are shown before and after correction respectively.

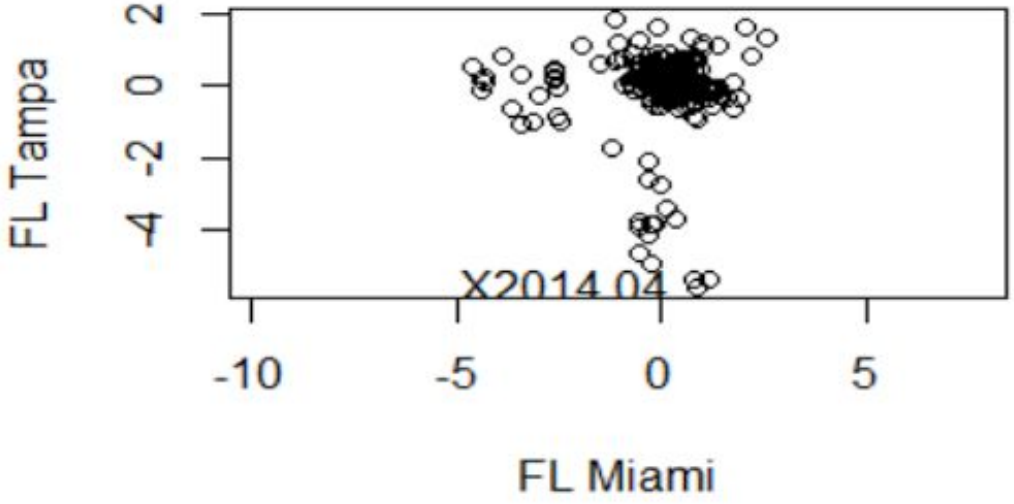


Plot the first two independent components:
The scatter plot shows, independent components representing CA Los Angeles and AZ Phoenix plotted against each other and we get the outliers at 2007 May. In both Phoenix and LA the highest peaks are in the year 2007. So outliers are bound to be found at 2007. This is the value reasonably higher than the others.

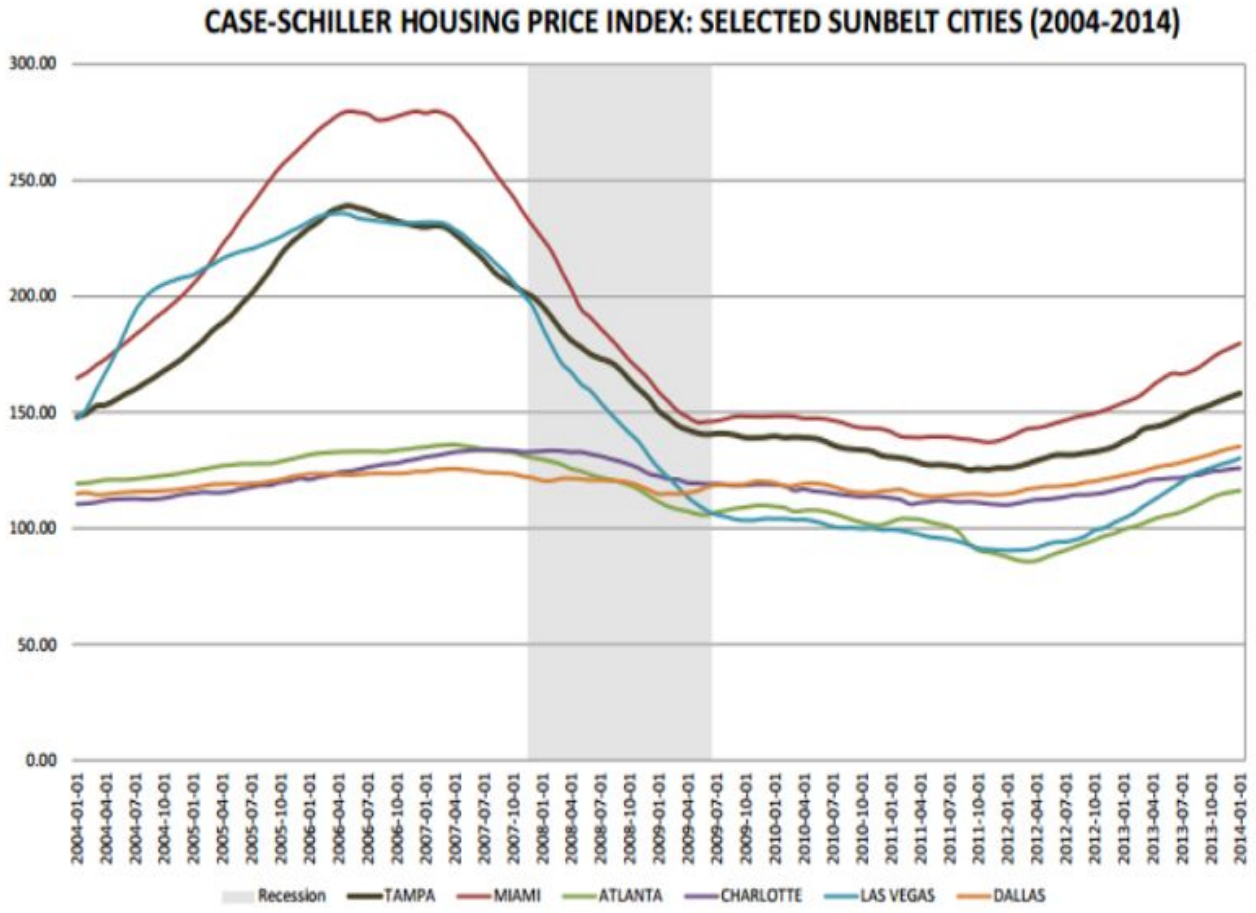
Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



Compare 7th and 8th :



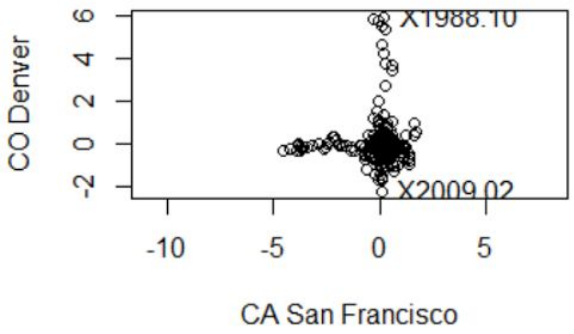
Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de

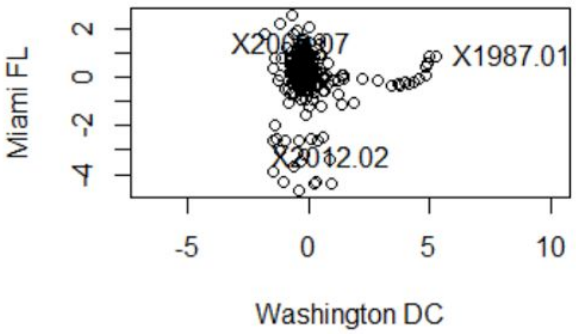


Here unusually the outliers are found at 2014, but there are no such reasons so that we can argue for that. So scatter plot of first two independent components give better interpretations in terms of match with the observed data. But some outliers are during the period of 1988 end. This is complying with the first two independent components. Because we can observe the data that during 1988-end there was an all American crisis so housing price indices were also very less. So this is the general case of all the paired independent Components.

Plot for some components and their interpretation:
For more convenience I have taken the scatter plot between CO- Denver and CA-San Francisco, as well as Washington DC and Miami. This shows the clear economic debt crisis of early 90's and the outliers are perfectly spotted even from the individual graphs it is confirmed that during early 90's these two states were at the worst of their debt crisis. But we can't see the highest peaks as outliers, so interpreting this data is difficult. In both the cases in the year 2009 we can see the slope is linear and neither upwards or downward. So, like the previous case, outliers aren't prominent.

Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de

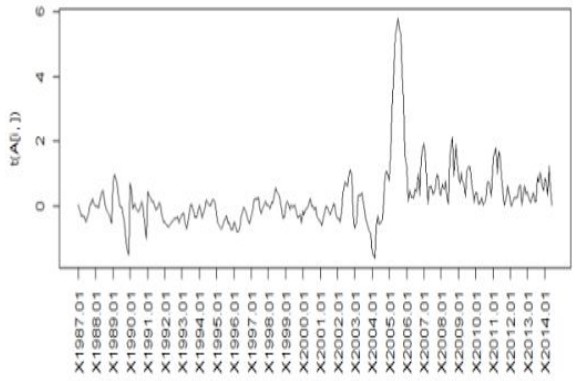


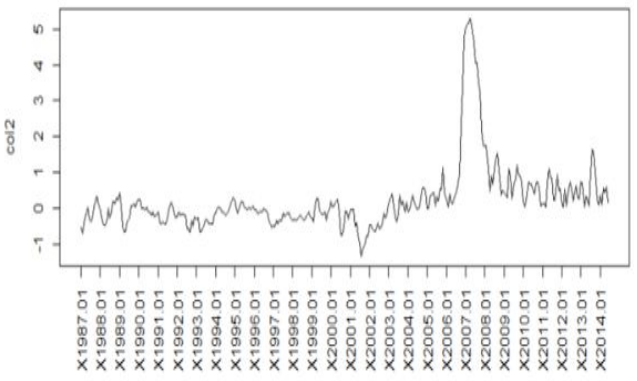


Can you identify the economic transitions from 1988 to 2014 in terms of housing price indices of independent components?

After plotting independent components of Los Angeles, CA and San Diego, CA it is observable that a general pattern is there. The index is slowly rising, from 1988 onwards and then becoming max in the interval 2005-07 and then gradually falling. So it is close to what is actually observed, .i.e. minimum at 1988 then gradually rising, and becoming until 2006 then falling sharp and staying low until 2012 and then rise again. Plots are shown before and after correction respectively.

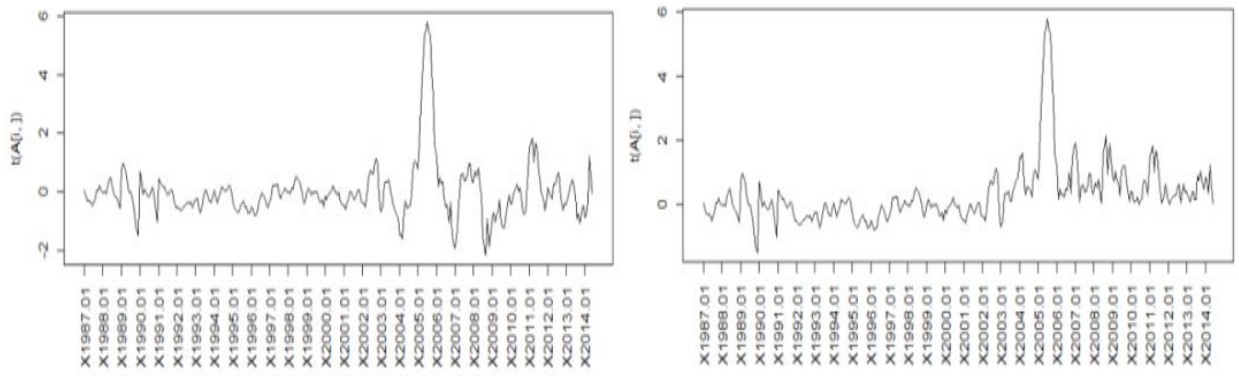
For CA Los Angeles:





For CA San Diego:

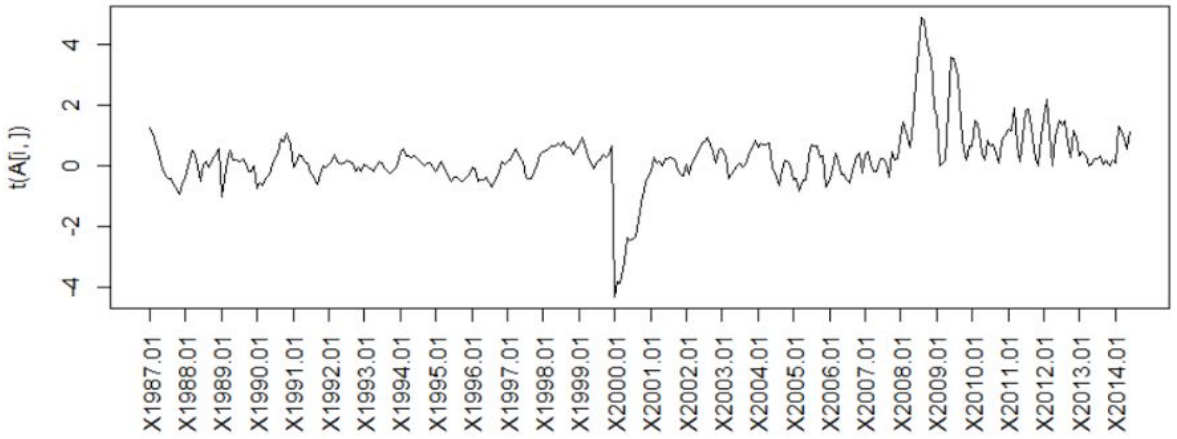
Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



Instead of N.A if we take average or first value in the row:
Taking average of data instead of 0s in place of NAs is in my opinion a very good idea. First of all the missing values are never zeroes, they have their own respective values, so taking average value altogether for all the missing values balances the individual rows. The data looks cleaner:

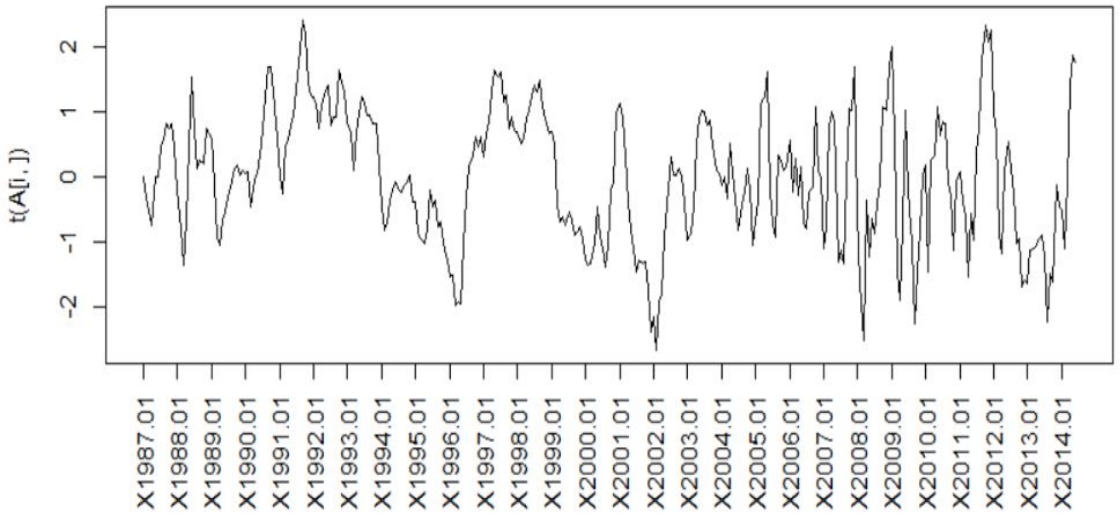
	row.names	X1987.01	X1987.02	X1987.03	X1987.04	X1987.05	X1987.06	X1987.07	X1987.08	X1987.09	X1987.10	X1987.11	X1987.12
1	AZ.Phoenix	111.3257	111.3257	111.3257	111.3257	111.3257	111.3257	111.3257	111.3257	111.3257	111.3257	111.3257	111.3257
2	CA.Los.Angeles	59.3300	59.6500	59.9900	60.8100	61.6700	62.7100	63.6600	64.5600	65.3800	66.2000	66.9400	67.9100
3	CA.San.Diego	54.6700	54.8900	55.1600	55.8500	56.3500	56.8600	57.2600	57.6900	58.1400	58.5300	59.0200	59.4000
4	CA.San.Francisco	46.6100	46.8700	47.3200	47.6900	48.3100	48.8300	49.4900	49.9400	50.6900	51.3300	51.8000	52.0300
5	CO.Denver	50.2000	49.9600	50.1500	50.5500	50.6300	50.5000	50.2800	50.3800	50.1800	50.3800	49.8900	49.8600
6	DC.Washington	64.1100	64.7700	65.7100	66.4000	67.2700	68.7000	69.7900	70.6200	71.7900	72.5700	73.1800	73.4200
7	FL.Miami	68.5000	68.7600	69.2300	69.2000	69.4600	69.3100	69.7000	70.1600	70.9500	71.2800	71.5000	71.4100
8	FL.Tampa	77.3300	77.9300	77.7600	77.5600	77.8500	78.7100	79.1100	79.1400	79.2400	79.0900	79.0500	79.1500
9	GA.Atlanta	101.0365	101.0365	101.0365	101.0365	101.0365	101.0365	101.0365	101.0365	101.0365	101.0365	101.0365	101.0365
10	IL.Chicago	53.5500	54.6400	54.8000	54.8800	55.4300	56.3900	57.5400	58.3700	58.8500	59.3200	59.3400	60.3000
11	MA.Boston	70.0400	70.0800	70.0000	70.7000	71.5100	72.3200	73.0900	73.7900	74.3900	74.6300	74.8300	74.7400
12	MI.Detroit	89.4673	89.4673	89.4673	89.4673	89.4673	89.4673	89.4673	89.4673	89.4673	89.4673	89.4673	89.4673
13	MN.Minneapolis	109.2711	109.2711	109.2711	109.2711	109.2711	109.2711	109.2711	109.2711	109.2711	109.2711	109.2711	109.2711
14	NC.Charlotte	63.3900	63.9400	64.1700	64.8100	65.1800	65.5500	65.7600	66.0800	66.4700	66.7700	67.0900	67.1900

When we are plotting, it in the time series, the graph matches with the actual observation:
The Los Angeles, CA data after correction at specific timelines looks like this:



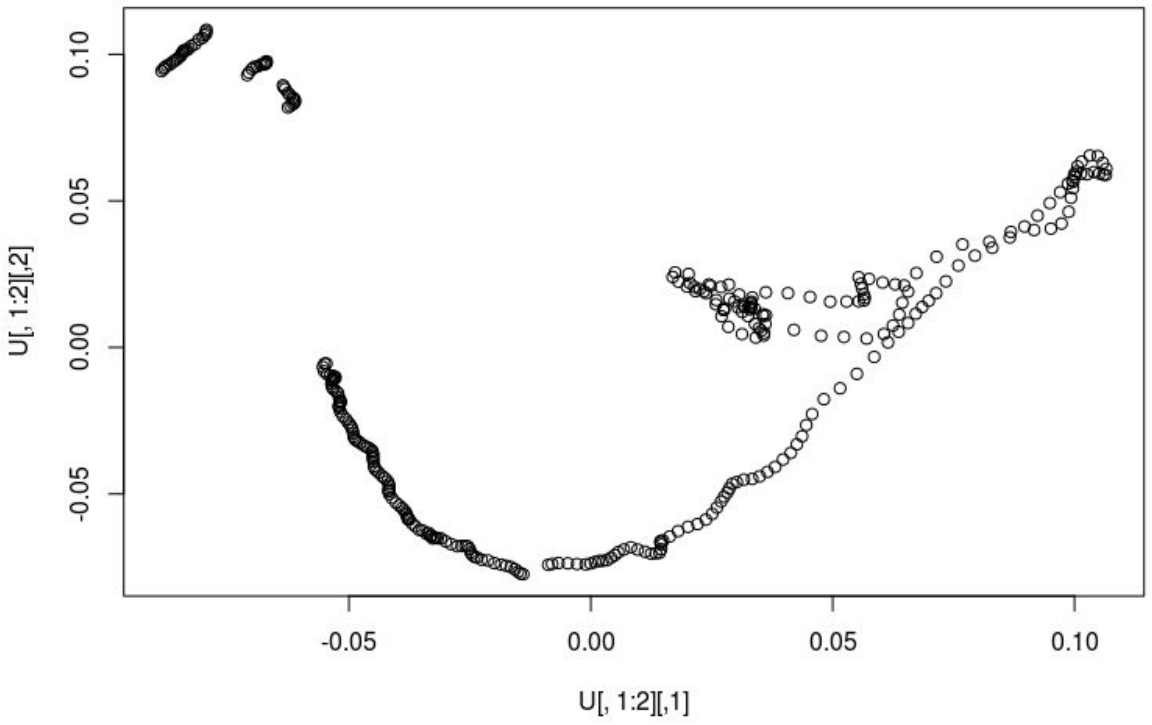
There is indeed a plunge in 2001, for LA, California only. Independent component analysis with replacing NA values with average values of that row is indeed a good idea.
Now replacing the NA values with the first actual value of the row is not a very concrete idea in my opinion, when I plotted the Los Angeles, CA data with this procedure, the time series and the data looked like:

	row.names	X1987.01	X1987.02	X1987.03	X1987.04	X1987.05	X1987.06	X1987.07	X1987.08	X1987.09	X1987.10	X1987.11	X1987.12
1	AZ.Phoenix	67.54	67.54	67.54	67.54	67.54	67.54	67.54	67.54	67.54	67.54	67.54	67.54
2	CA.Los.Angeles	59.33	59.65	59.99	60.81	61.67	62.71	63.66	64.56	65.38	66.20	66.94	67.91
3	CA.San.Diego	54.67	54.89	55.16	55.85	56.35	56.86	57.26	57.69	58.14	58.53	59.02	59.40
4	CA.San.Francisco	46.61	46.87	47.32	47.69	48.31	48.83	49.49	49.94	50.69	51.33	51.80	52.03
5	CO.Denver	50.20	49.96	50.15	50.55	50.63	50.50	50.28	50.38	50.18	50.38	49.89	49.86
6	DC.Washington	64.11	64.77	65.71	66.40	67.27	68.70	69.79	70.62	71.79	72.57	73.18	73.42
7	FL.Miami	68.50	68.76	69.23	69.20	69.46	69.31	69.70	70.16	70.95	71.28	71.50	71.41
8	FL.Tampa	77.33	77.93	77.76	77.56	77.85	78.71	79.11	79.14	79.24	79.09	79.05	79.15
9	GA.Atlanta	69.61	69.61	69.61	69.61	69.61	69.61	69.61	69.61	69.61	69.61	69.61	69.61
10	IL.Chicago	53.55	54.64	54.80	54.88	55.43	56.39	57.54	58.37	58.85	59.32	59.34	60.30
11	MA.Boston	70.04	70.08	70.00	70.70	71.51	72.32	73.09	73.79	74.39	74.63	74.83	74.74
12	MI.Detroit	58.24	58.24	58.24	58.24	58.24	58.24	58.24	58.24	58.24	58.24	58.24	58.24
13	MN.Minneapolis	63.13	63.13	63.13	63.13	63.13	63.13	63.13	63.13	63.13	63.13	63.13	63.13
14	NC.Charlotte	63.39	63.94	64.17	64.81	65.18	65.55	65.76	66.08	66.47	66.77	67.09	67.19



Correction at so many points will actually manipulate the data to a large extent which is not a feasible option.

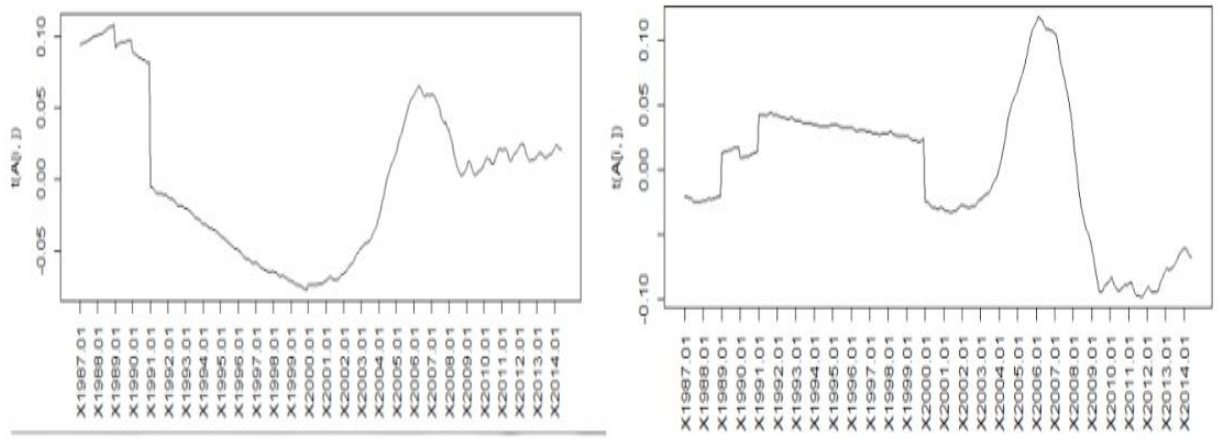
Whitening of the data:
First we plot normal columns of first left singular vector to check if there are any visible patterns. The plot comes out as:



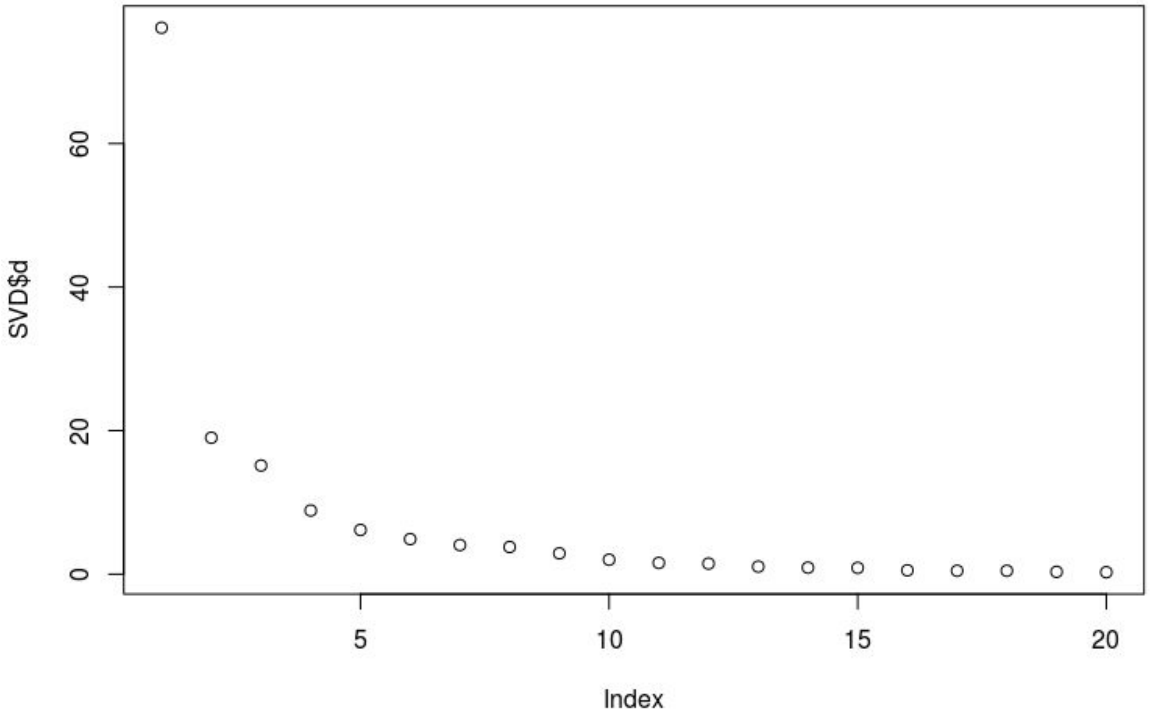
Unfortunately, there are no visible patterns in U plot of first two left singular vectors.

Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de

After calculating SVD of the data, I have plotted the 2nd and 3rd column of the first left singular vector, the first plot shows, Los Angeles, CA data and the 2 nd plot shows San Diego, CA data. The whitened data is perfect but there are few glitches, .e.g. for LA California, 1988 data isn't correct which is same for San Diego also. The peaks more or less comply with the original data in both the cases. For San Diego data the data plots after 2008 are not at all correct.

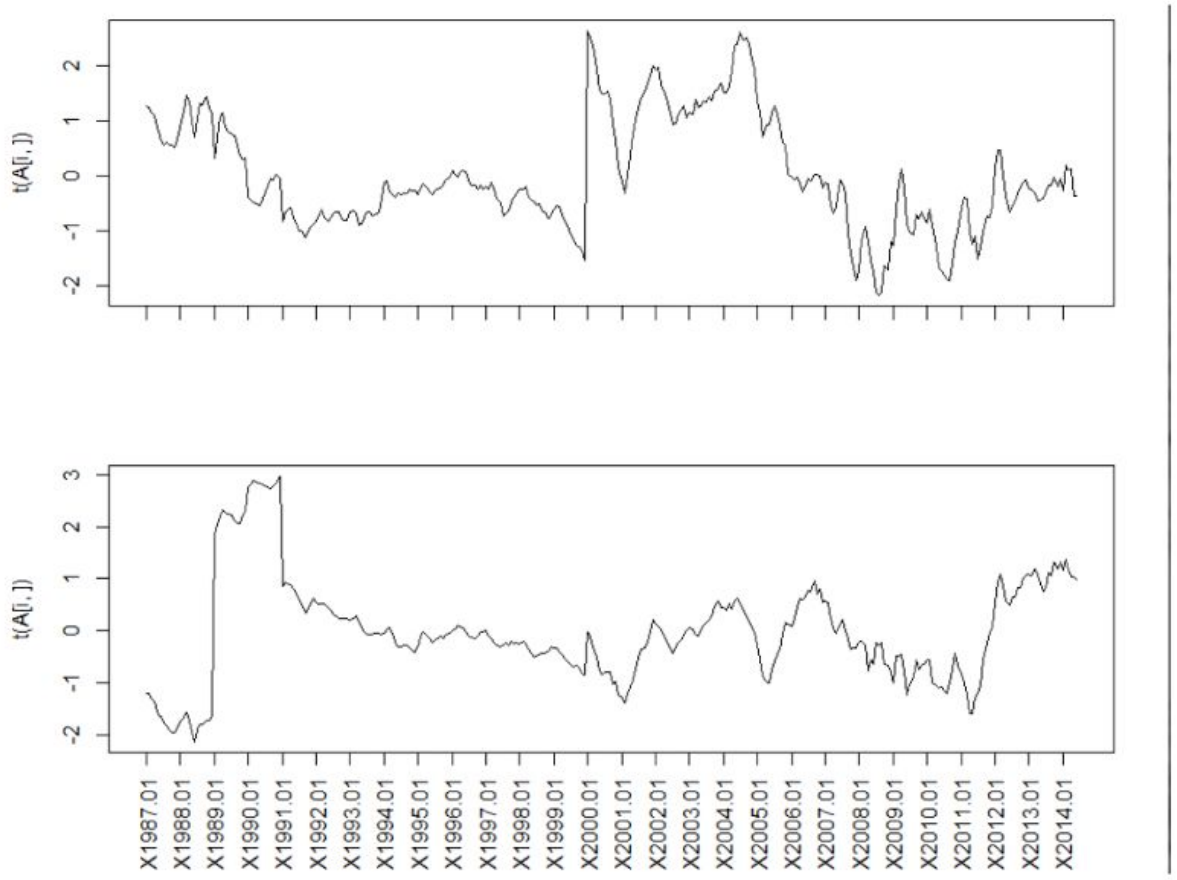


Scree Test Plot:

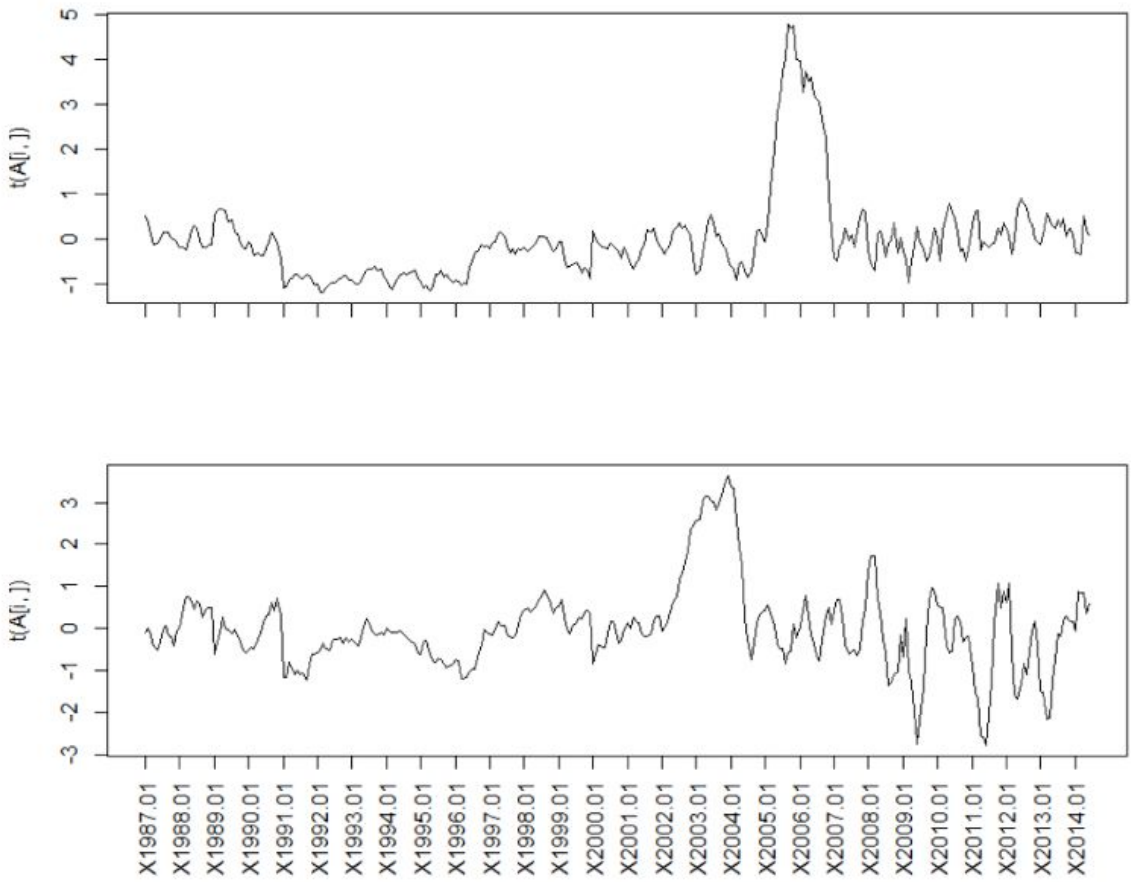


The no of principal components taken here is 2 from the plot.

Report on Hands-on assignment #3: CX & ICA Data Mining and Matrices SS 2017-18	Submitted by: Harshita Jhavar Matriculation Number: 2566267 eMail: s8hajhav@stud.uni-saarland.de



If we take only 2 independent components, the plots are pretty random, no such patterns are found, and above that in both the cases there are peaks greater than or equals to the highest peaks and in the second plot the regions where the peaks are supposed to be are at lows, so interpretation from scree plot are rejected.
Taking 13 as a result from Guttmann Kaiser, the same plot now looks like,



So rank 13 components are better in terms that they are better than the previous but not totally correct in terms of plots, the first component is showing more or less real data, like highest at 2007 and low values are at 1988, though from 2010 onwards, values may not be that low. But the second plot is not at all correct. Highest will never be at 2004.
So in my point of view if whitened data is taken into account, rank choosing method should be Guttmann Kaiser for better understanding of components, but better if all 20 independent components are chosen because 20 is not that greater than 13, in exchange of with we are getting complete analysis of data.